МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Механико-математический факультет

С. В. Гололобов, А. М. Мацокин

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ АНАЛИЗА И ЛИНЕЙНОЙ АЛГЕБРЫ

ЧАСТЬ 1

Учебно-методическое пособие

Новосибирск
2019

УДК 519.612(075.8)+519.614(075.8)
ББК 22.192.31я73-1
Г614

Рецензент
д-р физ.-мат. наук, проф. *Ю. М. Лаевский*

Гололобов, С. В.

Г614 Вычислительные методы анализа и линейной алгебры : учеб.-метод. пособие : в 2 ч. / С. В. Гололобов, А. М. Мацокин ; Новосиб. гос. ун-т. — Новосибирск : ИПЦ НГУ, 2019. Ч. 1. — 160 с.

ISBN 978-5-4437-0959-8
ISBN 978-5-4437-0960-4 (часть 1)

В пособии излагаются классические разделы вычислительной математики — решение систем линейных алгебраических уравнений и задач о поиске собственных чисел и собственных векторов матрицы. Оно написано на основе лекционной и практической работы авторов со студентами ММФ НГУ. Пособие разделено на лекции и освещает вопросы построения и анализа свойств прямых и итерационных методов.

Предназначено для студентов математических, физических, естественнонаучных факультетов университетов, инженерно-физических и других специальностей вузов, требующих углубленного знания вычислительной математики, а также для всех тех, кто самостоятельно изучает вычислительную линейную алгебру.

УДК 519.612(075.8)+519.614(075.8)
ББК 22.192.31я73-1

Рекомендовано к печати кафедрой «Вычислительная математика» ММФ НГУ (выписка из протокола №139 от 11.07.2019 г.).

ISBN 978-5-4437-0959-8
ISBN 978-5-4437-0960-4 (часть 1)

Оглавление

Введение

Предлагаемое учебно-методическое пособие содержит материал, разработанный для учебного курса с одноименным названием, который читается авторами в третьем семестре на механико-математическом факультете (ММФ) Новосибирского национального исследовательского государственного университета (НГУ) с 2010 г. В основу этого сборника положен личный опыт авторов при обучении студентов.

Данное учебно-методическое пособие предназначено для студентов, заинтересованных в получении базовых навыков в области вычислительной математики. Читатели познакомятся не только с классическими методами вычислительной математики, но и с основными подходами при обосновании применимости этих методов. Следуя поставленной цели, мы стремились построить первую часть сборника таким образом, чтобы раскрыть основные особенности классических подходов при решении систем линейных алгебраических уравнений, а также поиске собственных чисел и векторов. Вторая часть пособия будет посвящена основным методам при поиске корней нелинейных уравнений, численного интегрирования и дифференцирования, и она будет подготовлена и опубликована позже.

Пособие разбито на лекции, чтобы упростить процесс подготовки для начинающих лекторов. Однако данное разбиение достаточно условное, а потому на практике мы рекомендуем адаптировать скорость преподавания материала к уровню студентов и доступному количеству лекционных часов. При обучении более подготовленных студентов можно пропускать хорошо известные факты из линейной алгебры, математического анализа и аналитической геометрии. Для менее подготовленных студентов имеет смысл уменьшить число изучаемых методов, но более детально разобрать оставшиеся методы с повторением или даже изучением материала из курсов высшей математики, который необходим

для понимания данного курса.

Для успешного ознакомления с курсом рекомендуется получить базовые знания в области линейной алгебры, математического анализа и геометрии. Но мы старались приводить основные факты из указанных разделов высшей математики, которые нужны для понимания материала данного пособия. При необходимости доказательство этих фактов можно отыскать в соответствующей литературе и сети Интернет.

Мы надеемся, что данное пособие будет полезно студентам для более полного усвоения курса и применения методов вычислительной математики в их дальнейшей учебной, научно-преподавательской и практической деятельности.

Лекция №1

Введение в вычислительную математику

Для начала вспомним из линейной алгебры то, что читателям должно быть уже знакомо (в противном случае лучше прослушать курс или почитать учебник по линейной алгебре). Матрицей A размера $m \times n$ с элементами a_{ij} будем называть таблицу чисел

$$A \equiv [a_{ij}]_{i,j=1}^{m,n} \equiv [a_{i,j}]_{i,j=1}^{m,n} \equiv$$

$$\equiv \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \equiv \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix},$$

а вектором $\vec{\mathbf{x}}$ столбец из n чисел x_i

$$\vec{\mathbf{x}} \equiv [x_i]_{i=1}^{n} \equiv \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_n} \end{bmatrix}.$$

Мы будем стараться опускать запятую, разделяющую индексы в матрице, там, где это возможно и не запутает читателя.

В этом курсе мы будем предполагать, что матрицы у нас квадратные (т. е. $m = n$), невырожденные и вещественные (устоявшийся жаргон, правильнее их называть вещественнозначными), а вектора — исключительно вектора-столбцы и также вещественные, для определённости. Хотя все утверждения легко переносятся на комплексный случай, материал проще разобрать для вещественного случая и только потом пробовать перенести рассуждения на комплексный. Поэтому вместо того, чтобы говорить «квадратная матрица размера $n \times n$», мы будем говорить просто

«матрица размера n». Заглавные буквы будут использованы для обозначения матриц.

Мы будем называть матрицу *верхней треугольной*, или просто «верхнетреугольной» (жаргонизм!), если все элементы матрицы под главной диагональю равны 0. Мы будем называть матрицу *строго верхней треугольной*, если у неё также равны нулю элементы на главной диагонали. Аналогичным образом будет использоваться понятие *(строго) нижней треугольной*, или просто (строго) «нижнетреугольной» (также жаргонизм) матрицы. Мы будем считать матрицу *диагональной*, если у этой матрицы только элементы на главной диагонали могут быть (но не обязаны!) ненулевыми. Наконец, мы будем полагать матрицу *трёхдиагональной*, если у этой матрицы только элементы на главной диагонали, а также непосредственно под и над ней могут быть (но не обязаны!) ненулевыми.

Линейная алгебра говорит нам, что квадратная матрица A вырождена, если её детерминант (определитель) равен 0. Чтобы посчитать определитель матрицы A, можно воспользоваться формулой

$$|A| = \sum_{(i_1,\ldots,i_n)} (-1)^{r(i_1,\ldots,i_n)} a_{1i_1} \ldots a_{ni_n},$$

т. е. посчитать сумму произведений элементов матрицы по всем перестановкам, каковых насчитывается ровно $n!$. Таким образом, для подсчёта определителя матрицы размера n нам понадобится выполнить $(n-1)n!$ умножений. Если предположить, что компьютер может выполнять один миллиард умножений в секунду, что примерно соответствует частоте процессора в 1 ГГц, то это позволяет утверждать, что мы говорим о вполне современном компьютере. В результате мы имеем:

Размер матрицы n	Время вычисления
10	$\sim 10^{-4}$ секунды
20	~ 17 минут
30	$\sim 400\ 000$ лет

Если вспомнить о том, что мы знаем из алгебры, то метод Крамера для решения системы линейных алгебраических уравнений (далее — система уравнений) $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ с известной *матрицей системы* A и известной *правой частью* $\vec{\mathbf{b}}$ даёт одну компоненту *вектора решения* с помощью двух определителей

$$x_i = \frac{|A_i|}{|A|},$$

где A_i — матрица, в которой столбец с номером i заменён на вектор правой части $\vec{\mathbf{b}}$.

Обратная матрица может быть вычислена с помощью решения n систем уравнений вида $A\vec{\mathbf{x}}^j = \vec{\mathbf{e}}^j$ с правыми частями равными j-му единичному орту для $j = 1, \ldots, n$. Тогда вектора $\vec{\mathbf{x}}^j$ будут столбцами обратной к A матрицы A^{-1}.

Собственные числа и вектора матрицы A удовлетворяют системе уравнений $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$. Для того чтобы вычислить собственные числа матрицы A, нужно найти корни полинома $|A - \lambda E|$, который также вычисляется через определитель. Здесь и далее E обозначает единичную матрицу, т. е. матрицу с единицами на главной диагонали и нулевыми внедиагональными элементами. Собственные вектора являются решениями системы уравнений $(A - \lambda E)\vec{\mathbf{x}} = \vec{\mathbf{0}}$.

Таким образом, при попытке найти решения к указанным выше задачам мы немедленно сталкиваемся с проблемой вычисления детерминанта матрицы, формула для которого, при всей её компактности, требует несоизмеримой работы современного компьютера даже при небольших размерах систем уравнений.

Помимо проблемы времени при решении систем уравнений мы ещё столкнёмся с проблемой точности полученного результата или, другими словами, *погрешности* вычислений. Дело в том, что компьютер умеет работать только с конечным набором рациональных чисел. Соответственно, числа вроде π и $\sqrt{2}$ незамедлительно подменяются их рациональными приближениями. Более того, если результат арифметической операции не попадает в тот конечный набор рациональных чисел, его также подменяют на число из доступного набора. Это приводит к появлению так называемых *ошибок округления*, когда вместо точного результата a мы имеем некоторое приближение к нему $\tilde{a} = a + \varepsilon|a|$ с некоторым, вообще говоря, неизвестным ε, про который можно сказать разве, что $|\varepsilon| \leq 10^{-k}$ для некоторого k.

Вероятна ситуация, когда $k = 6$, матрица A имеет размер $n = 6$, детерминант $|A| = 1$, а элементы $|a_{ij}| \geq 10$. В этом случае каждое произведение из формулы для детерминанта $a_{1i_1} \cdot \ldots \cdot a_{6i_6}$ даст погрешность порядка $O(1)$ за счёт ошибок округления, следовательно, и вычисление всего детерминанта даст погрешность $6! \cdot O(1) = O(1)$, т. е. ошибка при вычислении детерминанта будет порядка самой величины детерминанта. Всё, что будет посчитано таким способом, будет содержать ошибки, которые сделают проведённые вычисления практически непригодными к использованию.

Указанные проблемы проявятся гораздо интенсивнее, если мы будем решать системы линейных алгебраических уравнений порядка (размера) $n \sim 1\,000 \ldots 1\,000\,000$. Именно поэтому требуется изучение вопросов, связанных с поиском алгоритмов, которые бы давали близкий к истинному ответ для классических задач линейной алгебры (и других наук) при проведении вычислений на компьютерах, которые обладают существенно более высокой скоростью счёта, чем человек. Подобными вопросами занимается наука вычислительная математика, с которой мы познакомимся в предлагаемом курсе.

Прежде чем приступить к знакомству с вычислительной математикой, вспомним некоторые сведения из математики, которые пригодятся нам в дальнейшем.

Спектральным радиусом матрицы A размера n называется число, равное максимальному по модулю собственному значению матрицы $\rho(A) = \max\limits_{1 \le i \le n} |\lambda_i|$. Здесь предполагается, что каждое собственное значение встречается столько раз, сколько раз оно появилось на диагонали в нормальной жордановой форме. И поскольку спектральный радиус вычисляется с помощью модуля собственного значения, его можно посчитать даже для комплексных собственных значений.

Скалярным произведением двух элементов u и v называется билинейный (линейный по каждому аргументу) функционал (u, v), который этим элементам некоторого пространства V сопоставляет число (оно может быть комплексным!) и удовлетворяет четырём аксиомам:

1) $\forall v \in V \ (v, v) > 0$, если $v \neq 0$ и $(v, v) = 0$, если $v = 0$;

2) $\forall u, v \in V \ (u, v) = (\bar{v, u})$ (убедитесь, что вы понимаете, что произойдёт с этой аксиомой в вещественном случае);

3) $\forall \alpha \in \mathbb{R}(\mathbb{C}), \forall u, v \in V \ (\alpha u, v) = \alpha(u, v)$ (обратите внимание, что скаляр выносится из первого аргумента; если скаляр стоит перед вторым аргументом, то он вынесется с сопряжением по предыдущей аксиоме);

4) $\forall u, v, w \in V \ (u + v, w) = (u, w) + (v, w)$.

В соответствии с этим определением можно дать определение скалярного произведения вещественных (комплексных) векторов

размера n: $\vec{\mathbf{x}}, \vec{\mathbf{y}} \in \mathbb{R}^n(\mathbb{C}^n)$. Вспомните хотя бы один пример вещественного и комплексного скалярных произведений векторов.

Матрица A^*, *сопряжённая* матрице A относительно скалярного произведения $(\cdot, \cdot)$, определяется тождеством Лагранжа:

$$(A\vec{\mathbf{x}}, \vec{\mathbf{y}}) = (\vec{\mathbf{x}}, A^*\vec{\mathbf{y}}) \; \forall \vec{\mathbf{x}}, \vec{\mathbf{y}} \in \mathbb{R}^n(\mathbb{C}^n).$$

Напомним, что матрица $A^* = [\bar{a}_{ji}]_{i,j=1}^{n,n}$ сопряжена матрице A относительно евклидова скалярного произведения. Здесь $\bar{a}$ обозначает *комплексное сопряжение* числа a. Из алгебры известно, что если матрица A *самосопряжена (эрмитова)*, то все её собственные значения вещественны, а из её собственных векторов можно образовать ортонормированный базис в $\mathbb{R}^n(\mathbb{C}^n)$. Возможно, что эта информация известна вам в другом виде, а именно, что самосопряжённую (эрмитову) матрицу можно привести к вещественному диагональному виду с помощью ортогонального (унитарного) преобразования подобия. В этом случае ортогональный базис образуют столбцы ортогональной (унитарной) матрицы.

Матрица A называется *положительно определённой* в скалярном произведении $(\cdot, \cdot)$ (обозначается $A > 0$), если

$$\forall \vec{\mathbf{x}} \in \mathbb{R}^n(\mathbb{C}^n) \mid \vec{\mathbf{x}} \neq \vec{\mathbf{0}} \; (A\vec{\mathbf{x}}, \vec{\mathbf{x}}) > 0.$$

Это ОСНОВНОЕ определение в вычислительной математике, с которым студенты регулярно сталкиваются за время учёбы, в том числе и на старших курсах. Запомните его! Обратите внимание на обозначение $A > 0$ — это обозначение положительной определённости В ПРИМЕНЕНИИ К МАТРИЦЕ, а вовсе НЕ положительность элементов матрицы! Оно также означает, что для чисел запись $a > 0$ по-прежнему указывает на число больше нуля и то, что оно ВЕЩЕСТВЕННО, так как для комплексных чисел операция сравнения неопределена. Обратите внимание на то, что для положительной определённости требуется, чтобы скалярное произведение давало вещественный результат, даже если вектора и матрица содержат комплексные числа (в противном случае операция > 0 неопределена). Матрица называется *положительно*

полуопределённой в скалярном произведении $(\cdot,\cdot)$ (обозначается $A \geq 0$), если

$$\forall \vec{\mathbf{x}} \in \mathbb{R}^n(\mathbb{C}^n)\ (A\vec{\mathbf{x}}, \vec{\mathbf{x}}) \geq 0.$$

Отличие от положительной определённости состоит только в том, что могут быть ненулевые вектора, которые дадут нуль в скалярном произведении.

Нормой элемента v называется функционал $\|v\|$, который сопоставляет ВЕЩЕСТВЕННОЕ И НЕОТРИЦАТЕЛЬНОЕ число каждому элементу некоторого пространства V и при этом удовлетворяет трём аксиомам:

1) $\forall v \in V\ \|v\| > 0$, если $v \neq 0$ и $\|v\| = 0$, если $v = 0$;
2) $\forall \alpha \in \mathbb{R}, \forall v \in V\ \|\alpha v\| = |\alpha| \cdot \|v\|$ (обратите внимание на модуль скаляра!);
3) $\forall u, v \in V\ \|u + v\| \leq \|u\| + \|v\|$ (*неравенство треугольника*).

Если в нашем пространстве V есть скалярное произведение, то мы всегда можем определить в нём норму, связанную с этим скалярным произведением $\forall u \in V, \|u\| = \sqrt{(u,u)}$, т. е. доказать с помощью аксиом скалярного произведения тот факт, что полученный функционал будет удовлетворять аксиомам нормы. При этом можно доказать неравенство $|(u,v)| \leq \|u\| \cdot \|v\|\ \forall u, v \in V$.

В соответствии с этим определением можно дать определение нормы вещественного (комплексного) вектора $\vec{\mathbf{x}} \in \mathbb{R}^n(\mathbb{C}^n)$ и вещественной (комплексной) матрицы $A \in \mathbb{R}^{n\times n}$ $(\mathbb{C}^{n\times n})$ размера n. При этом мы будем называть такую матричную норму *аддитивной*. Если матричная норма также удовлетворяет условию $\|AB\| \leq \|A\| \cdot \|B\|$, то такую матричную норму будем считать *мультипликативной*. Мультипликативные матричные нормы особо интересны для вычислительной математики, поскольку они позволяют получать всевозможные оценки для алгоритмов,

как будет продемонстрировано в дальнейшем. Ввиду этого, по умолчанию, мы в дальнейшем всегда работаем с мультипликативными матричными нормами.

Матричная норма $\|\cdot\|_M$ называется *согласованной* с векторной нормой $\|\cdot\|_v$, если для любой матрицы A и для любого вектора $\vec{\mathbf{x}}$ справедливо неравенство

$$\|A\vec{\mathbf{x}}\|_v \leq \|A\|_M \cdot \|\vec{\mathbf{x}}\|_v.$$

Матричная норма $\|\cdot\|_v$ называется *подчинённой* векторной норме $\|\cdot\|_v$, если

$$\|A\|_v \equiv \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_v}{\|\vec{\mathbf{x}}\|_v}.$$

В последнем случае матричная норма однозначно определяется через векторную норму, поэтому мы будем использовать у матричной нормы тот же индекс, что и у векторной. Как нетрудно видеть из определения супремума, подчинённая матричная норма является частным случаем согласованной с соответствующей векторной нормой, т. е. выполнено условие согласованности

$$\|A\vec{\mathbf{x}}\|_v \leq \|A\|_v \cdot \|\vec{\mathbf{x}}\|_v.$$

Заметим также, что основное отличие между *sup* и *max* состоит в том, что максимум всегда находится в рассматриваемом множестве, поэтому писать супремум безопаснее, если мы ничего не знаем о свойствах множества, которое содержит бесконечное количество элементов. Для конечного множества всегда можно писать максимум, поскольку он там точно есть, если на множестве определена операция сравнения.

Приведём несколько примеров векторных норм для векторов размера n (докажите, что это действительно НОРМЫ векторов):

1) $\|\vec{\mathbf{x}}\|_1 \equiv |x_1| + \ldots + |x_n|$ — *октаэдрическая*, или просто «*первая*»(жаргонизм!) норма;

2) $\|\vec{\mathbf{x}}\|_2 \equiv \sqrt{|x_1|^2 + \ldots + |x_n|^2}$ — *сферическая*, или *евклидова*, или «*вторая*»(жаргонизм!) норма;

3) $\|\vec{\mathbf{x}}\|_\infty \equiv \max\limits_{1\le i\le n} |x_i|$ — *кубическая*, или *равномерная*, или «*бесконечная*»(жаргонизм!) норма

и для матриц размера n (докажите, что это действительно НОРМЫ матриц):

1) $\|A\|_1 \equiv \sup\limits_{\forall \vec{\mathbf{x}}\neq\vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_1}{\|\vec{\mathbf{x}}\|_1} = \max\limits_{1\le j\le n} \sum_{i=1}^n |a_{ij}|$;

2) $\|A\|_2 \equiv \sup\limits_{\forall \vec{\mathbf{x}}\neq\vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_2}{\|\vec{\mathbf{x}}\|_2} = \sqrt{\rho(A^*A)}$;

3) $\|A\|_\infty \equiv \sup\limits_{\forall \vec{\mathbf{x}}\neq\vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_\infty}{\|\vec{\mathbf{x}}\|_\infty} = \max\limits_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$.

Обратите внимание на тот факт, что октаэдрическая и равномерная нормы матрицы отличаются только порядком индексов i и j под знаками максимума и суммы, что создаёт проблему, если пытаться запомнить порядок следования индексов без понимания того, как эти формулы получены. Однако можно воспользоваться мнемоническим правилом: символ 1 расположен вертикально, значит, сумма идёт по столбцу (по индексу i), а символ ∞ расположен горизонтально, значит, сумма идёт по строке (по индексу j). Помимо этого, из формул следует, что для симметричных (самосопряжённых) матриц эти две нормы дадут одинаковый результат.

Доказательство. Этап 1. Для начала воспользуемся данным ранее определением подчинённой матричной нормы и соответству-

ющей ей векторной нормы в каждом из трёх случаев.

$$\begin{aligned}
\|A\|_1 \equiv \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_1}{\|\vec{\mathbf{x}}\|_1} &= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sum_{i=1}^n |\sum_{j=1}^n a_{ij} x_j|}{\sum_{i=1}^n |x_i|} \leq \\
&\leq \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sum_{i=1}^n \sum_{j=1}^n |a_{ij} x_j|}{\sum_{i=1}^n |x_i|} = \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sum_{j=1}^n \sum_{i=1}^n |a_{ij} x_j|}{\sum_{i=1}^n |x_i|} = \\
&= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sum_{j=1}^n (|x_j| \sum_{i=1}^n |a_{ij}|)}{\sum_{i=1}^n |x_i|} \leq \\
&\leq \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{(\max\limits_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|) \sum_{j=1}^n |x_j|}{\sum_{i=1}^n |x_i|} = \\
&= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} (\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|) = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|.
\end{aligned}$$

Этап 2. Теперь заметим, что «первая» норма матрицы не может превосходить максимума из сумм модулей элементов матрицы по строкам. Пусть он достигается при некотором j_0. Возьмём вектор $\vec{\mathbf{x}} = \vec{\mathbf{e}}^{j_0}$, где $\vec{\mathbf{e}}^{j_0}$ — единичный орт, у которого на позиции j_0 стоит 1, а на всех остальных стоят 0:

$$\|A\|_1 \geq \frac{\|A\vec{\mathbf{e}}^{j_0}\|_1}{\|\vec{\mathbf{e}}^{j_0}\|_1} = \frac{\sum_{i=1}^n |a_{ij_0}|}{1} = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|.$$

Т. е. для любой матрицы мы можем указать вектор, на котором неравенство превращается в равенство, что завершает доказательство формулы для «первой» нормы.

Аналогичным образом получаем результат для «бесконечной»

нормы. Этап 1.

$$\|A\|_\infty \equiv \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_\infty}{\|\vec{\mathbf{x}}\|_\infty} = \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\max\limits_{1 \le i \le n} |\sum_{j=1}^{n} a_{ij} x_j|}{\max\limits_{1 \le i \le n} |x_i|} \le$$

$$\le \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\max\limits_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij} x_j|}{\max\limits_{1 \le i \le n} |x_i|} \le \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\max\limits_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij} \max\limits_{1 \le j \le n} |x_j||}{\max\limits_{1 \le i \le n} |x_i|}$$

$$= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\max\limits_{1 \le j \le n} |x_j| * \max\limits_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|}{\max\limits_{1 \le i \le n} |x_i|} =$$

$$= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} (\max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|) = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

Этап 2. Теперь заметим, что «бесконечная» норма матрицы не может превосходить максимума из сумм модулей элементов матрицы по столбцам. Пусть он достигается при некотором i_0. Возьмём вектор $\vec{\mathbf{x}} = [\ldots, sign(a_{i_0 j}), \ldots]^T$, где $sign()$ — функция знака числа, т. е. она равна -1, если число отрицательное, и 1 в противном случае:

$$\|A\|_\infty \ge \frac{\|A\vec{\mathbf{x}}\|_\infty}{\|\vec{\mathbf{x}}\|_\infty} = \frac{\sum_{j=1}^{n} |a_{i_0 j}|}{1} = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

В этой формуле мы воспользовались тем, что $sign(a) * a = |a|$ для любого вещественного числа a, в том числе и нуля (для проверки понимания происходящего, подумайте о том, что нужно взять вместо функции $sign()$, чтобы получить доказательство в случае комплексной матрицы). Обратите также внимание на то, что другие компоненты вектора $A\vec{\mathbf{x}}$ не могут стать больше, чем $\sum_{j=1}^{n} |a_{ij}|$ при соответствующем номере i и любой комбинации значений -1 и 1 в векторе $\vec{\mathbf{x}}$, как было доказано на первом этапе, т. е. для любой матрицы возможно указать вектор, на котором неравенство превращается в равенство, что завершает доказательство формулы для «бесконечной» нормы.

Доказательство для «второй» нормы также состоит из двух этапов.

Этап 1. Заметим, что матрица A^*A самосопряжена, поэтому у неё есть ортонормированный базис из собственных векторов $\vec{\mathbf{v}}^{\,i}, i = 1, \ldots, n$, соответствующих вещественным собственным числам λ_i, $i = 1, \ldots, n$. Любой вектор представим в виде разложения по базису из собственных векторов $\vec{\mathbf{x}} = \sum_{i=1}^{n} x_i \vec{\mathbf{v}}^{\,i}$. Для этой матрицы все собственные числа будут неотрицательные (докажите!), что также важно для нашего доказательства. Обратите внимание, что хотя у самой матрицы A может не быть базиса из собственных векторов и вещественных собственных чисел, у матриц A^*A и AA^* всегда есть и ортонормированный базис из собственных векторов, и полный набор вещественных собственных чисел. Заметим также, что указанные матрицы могут не совпадать (приведите пример!), а потому местоположение знака сопряжения матрицы важно.

$$\|A\|_2 \equiv \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\|A\vec{\mathbf{x}}\|_2}{\|\vec{\mathbf{x}}\|_2} = \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sqrt{(A\vec{\mathbf{x}}, A\vec{\mathbf{x}})}}{\sqrt{(\vec{\mathbf{x}}, \vec{\mathbf{x}})}} = \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sqrt{(A^*A\vec{\mathbf{x}}, \vec{\mathbf{x}})}}{\sqrt{(\vec{\mathbf{x}}, \vec{\mathbf{x}})}} =$$

$$= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sqrt{\sum_{i=1}^{n} \lambda_i x_i^2}}{\sqrt{\sum_{i=1}^{n} x_i^2}} \leq \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sqrt{\max\limits_{1 \leq i \leq n} \lambda_i \sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n} x_i^2}} =$$

$$= \sup_{\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}} \frac{\sqrt{\max\limits_{1 \leq i \leq n} \lambda_i}}{1} = \sqrt{\max_{1 \leq i \leq n} \lambda_i} = \sqrt{\rho(A^*A)}.$$

Этап 2. Теперь возьмём вектор $\vec{\mathbf{x}} = \vec{\mathbf{v}}^{\,n}$, где собственный вектор $\vec{\mathbf{v}}_n$ соответствует максимальному по МОДУЛЮ собственному числу матрицы A^*A (этого можно достичь соответствующим упорядочиванием собственных чисел и векторов). В нашем случае максимальное по модулю собственное число совпадает с максимальным собственным числом, поэтому $A^*A\vec{\mathbf{v}}^{\,n} = \lambda_n \vec{\mathbf{v}}^{\,n} = \rho(A^*A)\vec{\mathbf{v}}^{\,n}$, и мы получаем:

$$\|A\|_2 \geq \frac{\|A\vec{\mathbf{v}}^{\,n}\|_2}{\|\vec{\mathbf{v}}^{\,n}\|_2} = \frac{\sqrt{(A\vec{\mathbf{v}}^{\,n}, A\vec{\mathbf{v}}^{\,n})}}{\|\vec{\mathbf{v}}^{\,n}\|_2} = \frac{\sqrt{(A^*A\vec{\mathbf{v}}^{\,n}, \vec{\mathbf{v}}^{\,n})}}{\|\vec{\mathbf{v}}^{\,n}\|_2} =$$

$$= \frac{\sqrt{\lambda_n}\|\vec{\mathbf{v}}^n\|_2}{\|\vec{\mathbf{v}}^n\|_2} = \sqrt{\rho(A^*A)}.$$

Для проверки понимания происходящего, подумайте о том, что нужно для получения доказательства в случае комплексной матрицы. Таким образом, для любой матрицы мы можем указать вектор, на котором неравенство превращается в равенство, что завершает доказательство формулы для «второй» нормы.

Обратите внимание, что этап 1 в доказательстве указывает величину, больше которой подчинённая матричная норма не может быть, а этап 2 показывает, что эта величина достигается на определённом векторе для каждой матрицы. И поскольку *sup* достигается, то, вообще говоря, он может быть заменён на *max*, но только по завершении доказательства. До завершения доказательства можно заменить *sup* на *max*, только если читателю известна теорема Вейерштрасса о свойстве конечномерных пространств из математического анализа. □

Будем называть нормы $\|\cdot\|_a$ и $\|\cdot\|_b$ *эквивалентными*, если существуют положительные константы C_1 и C_2 — такие, что $\forall v \in V,\ C_1\|u\|_a \le \|u\|_b \le C_2\|u\|_a$.

Теорема 1. *В конечномерном пространстве любые две нормы эквивалентны.*

Поскольку мы рассмотрели три примера векторных норм, то можем привести примеры констант эквивалентности для них:

$$\|\vec{\mathbf{x}}\|_\infty \le \|\vec{\mathbf{x}}\|_1 \le n\|\vec{\mathbf{x}}\|_\infty,$$

$$\|\vec{\mathbf{x}}\|_\infty \le \|\vec{\mathbf{x}}\|_2 \le \sqrt{n}\|\vec{\mathbf{x}}\|_\infty,$$

$$\|\vec{\mathbf{x}}\|_2 \le \|\vec{\mathbf{x}}\|_1 \le \sqrt{n}\|\vec{\mathbf{x}}\|_2,$$

$$\forall \vec{\mathbf{x}} \in \mathbb{R}^n(\mathbb{C}^n).$$

Докажите, что константы действительно такие и что они неулучшаемые. Обратите внимание на то, что константы эквивалентности зависят от размерности пространства. Именно этот

факт приведёт к тому, что в бесконечномерном пространстве отнюдь не все нормы будут эквивалентными.

Оценка погрешности вычислений

При решении системы линейных уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ могут быть неточно заданы как правая часть $\vec{\mathbf{b}}' = \vec{\mathbf{b}} + \delta\vec{\mathbf{b}}$, так и матрица $A' = A + \delta A$, где компоненты вектора $\delta\vec{\mathbf{b}}$ и элементы матрицы δA «малы» по сравнению с соответствующими элементами исходных вектора и матрицы.

В нашем курсе мы будем понимать малость в смысле норм, т. е. полагать, что некоторая норма вектора $\delta\vec{\mathbf{b}}$ и матрицы δA на несколько порядков меньше, чем соответствующие нормы вектора $\vec{\mathbf{b}}$ и матрицы A. Вообще понятие малости приходит в вычислительную математику из исходных естественнонаучных областей, поскольку только там можно ответить на вопрос о том, километр — это много или мало. По сравнению с размером электрона, километр — это очень много, а по сравнению с парсеком — очень мало. Следует также обратить внимание на то, что в пространстве векторов размера $n = 1\,000\,000$ «первая» норма может быть в миллион раз (6 порядков) больше, чем «бесконечная», как следует из констант эквивалентности. И снова только лишь исходная задача может дать ответ на вопрос, в каком смысле, т. е. в какой норме, измерять малость.

В силу линейности рассматриваемой задачи вместо точного решения $\vec{\mathbf{x}}$ точной задачи мы получим его приближение $\vec{\mathbf{x}}' = \vec{\mathbf{x}} + \delta\vec{\mathbf{x}}$, причем *погрешность*, т. е. компоненты вектора–ошибки $\delta\vec{\mathbf{x}}$, могут быть «большими». Чтобы понять, насколько именно большими, следует оценить норму ошибки посредством норм возмущений правой части и матрицы системы, считая при этом, что матричная норма подчинена векторной норме, чтобы воспользоваться условием согласованности матричной и векторной норм.

Числом обусловленности матрицы A мы будем называть величину

$$\nu(A) \equiv \|A\|\|A^{-1}\|.$$

Она определена только для невырожденных матриц, и конкретное значение зависит от выбора нормы! Мы можем доказать следующие свойства:

- $\nu(A) \geq 1$, поскольку из условия согласованности матричной и векторной норм следует, что $\forall \vec{\mathbf{x}} \neq \vec{\mathbf{0}}\ \|\vec{\mathbf{x}}\| = = \|AA^{-1}\vec{\mathbf{x}}\| \leq \|A\|\|A^{-1}\vec{\mathbf{x}}\| \leq \|A\|\|A^{-1}\|\|\vec{\mathbf{x}}\|$;
- $\nu(AB) \leq \nu(A)\nu(B)$, поскольку из условия мультипликативности матричных норм следует $\nu(AB) = \|AB\|\|(AB)^{-1}\| \leq \|A\|\|B\|\|B^{-1}\|\|A^{-1}\| = \nu(A)\nu(B)$.

А теперь получим две важные оценки с использованием числа обусловленности.

Теорема 2. *Пусть при решении системы* $A\vec{\mathbf{x}} = \vec{\mathbf{b}}, |A| \neq 0, \vec{\mathbf{b}} \neq \vec{\mathbf{0}}$ *была возмущена только правая часть, т. е. правая часть стала* $\vec{\mathbf{b}} + \delta\vec{\mathbf{b}}$. *Тогда решение возмущённой системы* $\vec{\mathbf{x}} + \delta\vec{\mathbf{x}}$ *удовлетворяет оценке*

$$\frac{\|\delta\vec{\mathbf{x}}\|}{\|\vec{\mathbf{x}}\|} \leq \nu(A)\frac{\|\delta\vec{\mathbf{b}}\|}{\|\vec{\mathbf{b}}\|}.$$

Доказательство. Для начала обратим внимание, что здесь $\delta\vec{\mathbf{b}}$ — это некоторый вектор, а НЕ вектор $\vec{\mathbf{b}}$, умноженный на скаляр δ!

Во-первых, в силу линейности системы уравнений $A(\vec{\mathbf{x}}+\delta\vec{\mathbf{x}}) = \vec{\mathbf{b}}+\delta\vec{\mathbf{b}}$ и $A\delta\vec{\mathbf{x}} = \delta\vec{\mathbf{b}}$. Кроме того, в силу невырожденности матрицы $\vec{\mathbf{x}} \neq \vec{\mathbf{0}}$. Во-вторых, в силу условия согласованности норм

$$\|\delta\vec{\mathbf{x}}\| = \|A^{-1}\delta\vec{\mathbf{b}}\| \leq \|A^{-1}\|\|\delta\vec{\mathbf{b}}\|.$$

И в-третьих,

$$\|\vec{\mathbf{b}}\| = \|A\vec{\mathbf{x}}\| \leq \|A\|\|\vec{\mathbf{x}}\|.$$

Разделив первое неравенство на второе, после приведения к требуемому виду мы получаем утверждение теоремы с использованием определения числа обусловленности. □

Теорема 3. *Пусть при решении системы* $A\vec{\mathbf{x}} = \vec{\mathbf{b}}, |A| \neq 0, \vec{\mathbf{b}} \neq \vec{\mathbf{0}}$ *была возмущена и матрица системы, и правая часть, т. е. матрица стала* $A + \delta A$*, а правая часть стала* $\vec{\mathbf{b}} + \delta\vec{\mathbf{b}}$*. Тогда решение возмущённой системы* $\vec{\mathbf{x}} + \delta\vec{\mathbf{x}}$ *удовлетворяет оценке*

$$\frac{\|\delta\vec{\mathbf{x}}\|}{\|\vec{\mathbf{x}}\|} \leq \frac{\nu(A)}{1 - \nu(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta\vec{\mathbf{b}}\|}{\|\vec{\mathbf{b}}\|} + \frac{\|\delta A\|}{\|A\|} \right),$$

если $\|\delta A\|\|A\| < 1$.

Доказательство. Для начала обратим внимание, что здесь δA — это некоторая матрица, а НЕ матрица A, умноженная на скаляр δ!

Во-первых, в силу линейности системы уравнений $(A+\delta A)(\vec{\mathbf{x}}+\delta\vec{\mathbf{x}}) = \vec{\mathbf{b}} + \delta\vec{\mathbf{b}}$. Кроме того, в силу невырожденности матрицы $A \neq 0$ и $\vec{\mathbf{x}} \neq \vec{\mathbf{0}}$.

Во-вторых, матрица $(A + \delta A)^{-1}$ существует. Для доказательства этого промежуточного утверждения заметим, что $A + \delta A = A(E + A^{-1}\delta A)$, где E — единичная матрица. Далее, $\forall\vec{\mathbf{y}} \neq \vec{\mathbf{0}}$ имеет место неравенство

$$0 < (1 - \|A^{-1}\|\|\delta A\|)\|\vec{\mathbf{y}}\| \leq \|\vec{\mathbf{y}}\| - \|A^{-1}\delta A\vec{\mathbf{y}}\| \leq \|(E + A^{-1}\delta A)\vec{\mathbf{y}})\|.$$

Здесь мы использовали условие теоремы для первого неравенств, а далее применили условие согласованности матричной и векторной норм два раза и получили второе неравенство. Наконец, мы использовали неравенство треугольника необычного вида $\|\vec{\mathbf{u}} +$

$\vec{\mathbf{v}} - \vec{\mathbf{v}}\| \leq \|\vec{\mathbf{u}} + \vec{\mathbf{v}}\| + \|\vec{\mathbf{v}}\|$, но тем не менее абсолютно верного и дающего $\|\vec{\mathbf{u}}\| - \|\vec{\mathbf{v}}\| \leq \|\vec{\mathbf{u}} + \vec{\mathbf{v}}\|$ и последнее неравенство. Вспоминая алгебру, мы можем заметить, что указанное неравенство говорит о том, что ни один ненулевой вектор не обращается в ноль при воздействии на него матрицы $E + A^{-1}\delta A$, следовательно, ядро матрицы состоит только из нулевого вектора, что означает, что матрица невырождена, т. е. обратима. В силу условия теоремы матрица A также обратима, что незамедлительно влечёт за собой существование $(A + \delta A)^{-1}$.

Для тех, кто хорошо знаком с матричной алгеброй, можно привести альтернативное доказательство. Если $\|B\| < 1$, то тогда $(E - B)^{-1} = E + B + B^2 + \ldots$ (матричный аналог разложения в ряд Тэйлора функции $(1 - x)^{-1}$) и при этом $\|(E - B)^{-1}\| \leq (1 - \|B\|)^{-1}$. Взяв $B = -A^{-1}\delta A$, получим нужное нам промежуточное утверждение.

В-третьих, получим оценку для нормы матрицы $(A + \delta A)^{-1}$.

$$\|(E + A^{-1}\delta A)^{-1}\| \equiv \sup_{\forall \vec{\mathbf{z}} \neq \vec{\mathbf{0}}} \frac{\|(E + A^{-1}\delta A)^{-1}\vec{\mathbf{z}}\|}{\|\vec{\mathbf{z}}\|} =$$

$$= \sup_{\forall \vec{\mathbf{y}} \neq \vec{\mathbf{0}}} \frac{\|\vec{\mathbf{y}}\|}{\|(E + A^{-1}\delta A)\vec{\mathbf{y}}\|} \leq \frac{1}{1 - \|A^{-1}\|\|\delta A\|}.$$

Первое равенство — определение подчинённой матричной нормы. Второе равенство следует из невырожденности $E + A^{-1}\delta A$, что, в частности, означает однозначную разрешимость уравнения $(E + A^{-1}\delta A)\vec{\mathbf{y}} = \vec{\mathbf{z}},\ \forall \vec{\mathbf{z}}$. Неравенство получается применением оценки, которую мы недавно доказали.

В-четвёртых, воспользуемся тем, что матрица $A + \delta A$ — невырождена, как мы только что показали. Из этого следует, что

$$\delta\vec{\mathbf{x}} = (A + \delta A)^{-1}(\vec{\mathbf{b}} + \delta\vec{\mathbf{b}} - (A + \delta A)\vec{\mathbf{x}}) = (A + \delta A)^{-1}(\delta\vec{\mathbf{b}} - \delta A\vec{\mathbf{x}}).$$

Отсюда получаем оценку

$$\|\delta\vec{\mathbf{x}}\| \leq \|(A + \delta A)^{-1}\|(\|\delta\vec{\mathbf{b}}\| + \|\delta A\|\|\vec{\mathbf{x}}\|)$$

с использованием условия согласованности норм и неравенства треугольника.

В-пятых, вспомним, что условие согласования норм и исходная система уравнений даёт $\|\vec{\mathbf{b}}\| \leq \|A\|\|\vec{\mathbf{x}}\|$. Разделив предпоследнее неравенство на $\|\vec{\mathbf{x}}\|$ и использовав последнее неравенство, получаем утверждение теоремы с использованием определения числа обусловленности

$$\frac{\|\delta\vec{\mathbf{x}}\|}{\|\vec{\mathbf{x}}\|} \leq \|(A+\delta A)^{-1}\|(\frac{\|\delta\vec{\mathbf{b}}\|}{\|\vec{\mathbf{x}}\|} + \|\delta A\|) \leq$$

$$\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|A\|\frac{\|\delta\vec{\mathbf{b}}\|}{\|\vec{\mathbf{b}}\|} + \|\delta A\|) =$$

$$= \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|}}(\frac{\|\delta\vec{\mathbf{b}}\|}{\|\vec{\mathbf{b}}\|} + \frac{\|\delta A\|}{\|A\|}).$$

Обратите внимание, что при $\delta A = 0$ мы получаем в точности оценку из теоремы 2, однако для доказательства нам потребовалось условие на малость возмущений матрицы, которого, по понятным причинам, не было в теореме 2. □

Стоит заметить, что оценки погрешности вычислений при возмущении систем уравнений на основе числа обусловленности получить относительно легко. Однако эти оценки весьма грубые, поскольку описывают наихудший возможный сценарий развития событий, который возможен, например, при симметричной матрице с полным набором собственных векторов, с положительными собственными числами и правой частью, которая является собственным вектором, соответствующим максимальному собственному числу, т. е. $A\vec{\mathbf{b}} = \lambda_{max}\vec{\mathbf{b}}$. Если при этом возмущение правой части является собственным вектором, соответствующим минимальному собственному числу, т. е. $A\delta\vec{\mathbf{b}} = \lambda_{min}\delta\vec{\mathbf{b}}$, тогда

$\vec{\mathbf{x}} = \frac{1}{\lambda_{max}}\vec{\mathbf{b}}$ и $\delta\vec{\mathbf{x}} = \frac{1}{\lambda_{min}}\delta\vec{\mathbf{b}}$. Следовательно,

$$\frac{\|\delta\vec{\mathbf{x}}\|}{\|\vec{\mathbf{x}}\|} = \frac{\lambda_{max}\|\delta\vec{\mathbf{b}}\|}{\lambda_{min}\|\vec{\mathbf{b}}\|}.$$

Если в теореме 2 взять евклидову норму, то получим

$$\frac{\|\delta\vec{\mathbf{x}}\|_2}{\|\vec{\mathbf{x}}\|_2} \leq \nu_2(A)\frac{\|\delta\vec{\mathbf{b}}\|_2}{\|\vec{\mathbf{b}}\|_2}.$$

В этом случае $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = \sqrt{\lambda_{max}^2} = \lambda_{max}$ и $\|A^{-1}\|_2 = \sqrt{\rho(A^{*-1}A^{-1})} = \sqrt{\rho(A^{-2})} = \sqrt{\lambda_{min}^{-2}} = \frac{1}{\lambda_{min}}$. Очевидно, что таким способом достигается точное равенство в оценке влияния возмущений правой части на возмущение решения. В реальной жизни такое случается очень редко, поскольку и правая часть, и возмущения натянуты на большее количество собственных векторов из базиса пространства, а потому реальное возмущение решения может быть существенно меньше, что, однако, никак не проявляется в полученных нами оценках. В следующем году вы узнаете, что число обусловленности матрицы может быть $\nu(A) = \frac{1}{h^2}$ для некоторого параметра h, стремящегося к 0, что означает быстрый рост числа обусловленности, следовательно, и необходимость уточнять оценки погрешности возмущения решения при возмущении правой части, чтобы гарантировать приемлемую точность полученного решения.

Лекция №2

Обратимся теперь к алгоритмам решения систем линейных алгебраических уравнений вида $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$. Как уже говорилось ранее, основная проблема состоит в том, чтобы отыскать алгоритмы, которые бы давали решение за «разумное» количество операций, т. е. разумное время. Понятно, что минимальное количество операций будет соответствовать случаю, когда мы сделали хотя бы одно действие с каждым элементом матрицы и вектора правой части. Если размер матрицы и вектора правой части n, то в матрице будет n^2 элементов, что при растущем n будет доминировать над размером вектора правой части. А потому можно сказать, что меньше, чем за $O(n^2)$ операций решить систему уравнений невозможно. Таким образом, мы будем считать, что «разумное» количество операций для решения системы уравнений — это $O(n^k)$ для некоторого $k \geq 2$, т. е. количество операций определяется по *полиномиальному* закону. При этом нам бы хотелось получить степень k как можно ближе к 2, однако это возможно только при очень жёстких условиях, поэтому мы ограничимся вариантами при $k \leq 3$.

Методы решения систем уравнений условно делятся на два класса: *прямые* и *итерационные*. Прямые методы дают решение системы только по окончании работы алгоритма, а итерационные на каждом шаге алгоритма дают приближение в смысле некоторой нормы к точному решению. Мы начнём с более простых и понятных прямых методов.

Прямые методы для решения систем линейных алгебраических уравнений

Основой для прямых методов служит метод исключения Гаусса, который изучается на практических занятиях по высшей алгебре.

Итак, пусть у нас имеется система уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ размера n — такая, что главные миноры матрицы A невырождены, т. е. $|A_k| \neq 0,\ k = 1, \ldots, n$, где

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{bmatrix}.$$

При $k = n$ мы имеем полную матрицу системы, т. е. она тоже предполагается невырожденной, следовательно, система уравнений (алгебра!) имеет единственное решение.

Воспользуемся методом математической индукции. Сначала запишем базу индукции. Применим к исходной системе преобразование, задаваемое матрицей

$$L_1 = \begin{bmatrix} \frac{1}{a_{11}} & 0 & 0 & \ldots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \ldots & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 & \ldots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ -\frac{a_{n1}}{a_{11}} & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

Следует обратить внимание, что по условию первый главный минор невырожден $A_1 = \begin{bmatrix} a_{11} \end{bmatrix}$. Это значит, что элемент $a_{11} \neq 0$ и деление на него возможно, т. е. матрица L_1 определена корректно. Более того, поскольку её детерминант равен $\frac{1}{a_{11}}$, то она невырождена. В результате имеем

$$L_1 A\vec{\mathbf{x}} = L_1\vec{\mathbf{b}}.$$

Как нетрудно определить с помощью прямых вычислений, матрица $L_1 A = A^{(1)} = [a_{ij}^{(1)}]_{i,j=1}^{n,n}$ имеет нулевые элементы под главной диагональю в первом столбце.

Пусть после $k-1$ шага получена система $A^{(k-1)}\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(k-1)}$, в которой матрица $A^{(k-1)}$ имеет нулевые элементы под главной диагональю в столбцах с номерами от 1 до $k-1$. Применим к этой системе матрицу

$$L_k = \begin{bmatrix} 1 & 0 & & \dots & & & 0 \\ \vdots & \ddots & & & & & \vdots \\ 0 & \dots & 1 & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & \frac{1}{a_{kk}^{(k-1)}} & 0 & \dots & 0 \\ 0 & \dots & 0 & -\frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}} & 1 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \ddots & \\ 0 & \dots & 0 & -\frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & 0 & \dots & 1 \end{bmatrix}.$$

Заметим, что применение матриц $L_1, \dots, L_{k-1}$ к матрице A есть ни что иное как составление некоторых линейных комбинаций из строк матрицы A. Следовательно, если элемент $a_{kk}^{(k-1)} = 0$, то существует некоторая нетривиальная линейная комбинация строк главного минора A_k, которая равна $\vec{\mathbf{0}}^*$. Как известно из алгебры, подобная линейная комбинация означает, что минор A_k вырожден, и это противоречит условию. Таким образом, $a_{kk}^{(k-1)} \neq 0$ и матрица L_k определена корректно, и поскольку и её детерминант равен $\frac{1}{a_{kk}^{(k-1)}}$, то она невырождена. В результате имеем

$$L_k A^{(k-1)}\vec{\mathbf{x}} = L_k \vec{\mathbf{b}}^{(k-1)}.$$

Как нетрудно заметить с помощью прямых вычислений, матрица $L_k A^{(k-1)} = A^{(k)} = [a_{ij}^{(k)}]_{i,j=1}^{n,n}$ имеет нулевые элементы под главной диагональю в k-м столбце, при этом нули под главной диагональю в столбцах $1, \dots, k-1$ также сохранились.

Таким образом, по методу математической индукции мы имеем после n-го шага матрицу $A^{(n)} = U$ — верхняя треугольная матрица с единицами на главной диагонали, а итоговая система

выглядит как $A^{(n)}\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(n)}$. Решить такую систему не составляет труда, поочерёдно вычисляя компоненты вектора $\vec{\mathbf{x}}$, начиная с последней.

Обратите внимание, что все матрицы $L_k, k = 1, \ldots, n$ невырождены, следовательно, преобразование не изменило решение исходной системы.

Отметим, что для вычисления каждой матрицы L_k требуется вычислить величину $c_k = \frac{1}{a_{kk}^{(k-1)}}$, которую впоследствии можно применять для вычисления остальных нетривиальных элементов матрицы L_k с помощью операции умножения. Отсюда, собственно, и появилось название — схема *единственного* деления. Здесь имеется в виду не то, что мы с помощью одного единственного деления преобразуем матрицу к верхнетреугольному виду, а то, что нам нужно одно деление для преобразования одного столбца матрицы — всего делений будет n штук.

Важность этого алгоритма определяется минимальным необходимым количеством операции деления. Дело в том, что компьютер выполняет операцию деления в десятки раз медленнее, чем операции умножения и сложения, а потому минимизация количества операций деления уменьшает трудозатраты компьютера. Однако уже несколько десятилетий скорость выполнения арифметических операций на компьютере существенно превосходит возможности того же компьютера по извлечению данных из памяти, а потому минимизация операций деления уже не критична для эффективной работы алгоритма. Гораздо важнее минимизировать количество обращений в память, ввиду чего актуальность схемы единственного деления стала гораздо ниже, чем это было на заре компьютерной эры.

LU-разложение

Метод Гаусса подаёт идею о том, что для решения системы уравнений можно использовать метод замены исходной системы на эквивалентную, в которой обращение матрицы осуществляется проще, без использования технологии детерминантов. Тот же метод Гаусса указывает на то, что решение систем уравнений с треугольными матрицами не требует значительных усилий, а потому естественным образом возникла идея об LU-разложении. Давайте представим исходную матрицу системы A в виде произведения двух треугольных матриц $A = LU$. Тогда система запишется как $LU\vec{\mathbf{x}} = \vec{\mathbf{b}}$, а потому для её решения сначала нужно будет решить систему $L\vec{\mathbf{y}} = \vec{\mathbf{b}}$ (*прямой ход*), а затем — систему $U\vec{\mathbf{x}} = \vec{\mathbf{y}}$ (*обратный ход*), и мы получим искомое решение $\vec{\mathbf{x}}$. Здесь в обозначениях используются английские буквы L (от англ. *lower triangular* — нижнетреугольная (матрица)) и U (от англ. *upper triangular* — верхнетреугольная (матрица)). Из этих названий становится понятными названия «прямой» и «обратный» ходы. При решении системы уравнений с нижнетреугольной матрицей мы ищем компоненты вектора решения от первой к последней, а при решении системы уравнений с верхнетреугольной матрицей — компоненты вектора решения от последней к первой.

Теорема 4 (об LU-разложении)**.** *Пусть матрица A размера n такова, что её главные миноры удовлетворяют условию $|A_k| \neq 0,\ k = 1, \ldots, n$. Тогда существует разложение матрицы A в произведение нижней и верхней треугольных матриц вида $A = LU$.*

Доказательство. Для доказательства теоремы используем метод математической индукции. Сначала запишем базу индукции. По условию $|A_1| = a_{11}$, следовательно, можно выбрать элементы l_{11} и u_{11} так, чтобы $a_{11} = l_{11}u_{11}$. Таким образом, мы можем построить разложение для первого главного минора матрицы $A_1 = L_1U_1$, при этом разложение будет не единственным,

так как неизвестных 2, а уравнений для их определения только 1 (бесконечно много решений). Отметим также, что матрицы L_1 и U_1 — невырождены в силу условия $0 \neq |A_1| = a_{11}$.

Предположим, что мы построили разложение вида $A_{k-1} = L_{k-1}U_{k-1}$ для $k-1$-го главного минора. Тогда найдём разложение для k-го минора вида:

$$\begin{bmatrix} & & & a_{1,k} \\ & A_{k-1} & & \vdots \\ & & & a_{k-1,k} \\ a_{k,1} & \dots & a_{k,k-1} & a_{k,k} \end{bmatrix} =$$

$$= \begin{bmatrix} & & & 0 \\ & L_{k-1} & & \vdots \\ & & & 0 \\ l_{k,1} & \dots & l_{k,k-1} & l_{k,k} \end{bmatrix} \begin{bmatrix} & & & u_{1,k} \\ & U_{k-1} & & \vdots \\ & & & u_{k-1,k} \\ 0 & \dots & 0 & u_{k,k} \end{bmatrix}.$$

Вспоминая алгебру и условие теоремы, заметим, что $0 \neq |A_{k-1}| = |L_{k-1}||U_{k-1}|$. Это означает, матрицы L_{k-1} и U_{k-1} невырождены. Следовательно, системы уравнений

$$L_{k-1} \begin{bmatrix} \mathbf{u_{1,k}} \\ \vdots \\ \mathbf{u_{k-1,k}} \end{bmatrix} = \begin{bmatrix} \mathbf{a_{1,k}} \\ \vdots \\ \mathbf{a_{k-1,k}} \end{bmatrix}$$

и

$$\begin{bmatrix} \mathbf{l_{k,1}} & \dots & \mathbf{l_{k,k-1}} \end{bmatrix} U_{k-1} = \begin{bmatrix} \mathbf{a_{k,1}} & \dots & \mathbf{a_{k,k-1}} \end{bmatrix}$$

имеют единственное решение при любых правых частях, т. е. любых значениях a_{ij}. Таким образом, последние два неизвестных должны удовлетворять единственному уравнению

$$l_{k,k}u_{k,k} = a_{k,k} - \begin{bmatrix} \mathbf{l_{k,1}} & \dots & \mathbf{l_{k,k-1}} \end{bmatrix} \begin{bmatrix} \mathbf{u_{1,k}} \\ \vdots \\ \mathbf{u_{k-1,k}} \end{bmatrix}.$$

Нет никаких сомнений в том, что при любых значениях в правой части этого уравнения можно подобрать l_{kk} и u_{kk} так, что

уравнение будет верным. Причём этот выбор будет не единственным. Однако в силу условия $|A_k| \neq 0$ мы имеем $l_{kk}u_{kk} \neq 0$, т. е. матрицы L_k и U_k также будут невырождены.

Таким образом, по методу математической индукции, мы имеем после n-го шага разложение $A \equiv A_n = L_n U_n \equiv LU$. □

С практической точки зрения, неединственность LU-разложения как минимум неудобна, ввиду чего обычно полагают, что диагональ матрицы L ИЛИ U состоит из 1. В этом случае мы избавляемся от одной переменной, а вторая определяется через уравнение в доказательстве теоремы единственным образом. Для проверки понимания происходящего следует ответить на вопрос, почему нельзя поставить 0 на диагонали матрицы L или U? Почему нельзя поставить 1 на диагонали как матрицы L, так и матрицы U? С точки зрения вычислений на компьютере проставление 1 на диагонали матрицы L (для определённости) позволяет делать разложение матрицы A прямо в том месте, где матрица A хранится в памяти, поскольку алгоритм устроен так, что вместо 0 под главной диагональю преобразовываемой матрицы A можно хранить внедиагональные элементы матрицы L. Так как мы положили, что диагональные элементы матрицы L равны 1, то провести вычисления не составит проблем без необходимости хранить эти значения.

Заметим, что использовать LU-разложение имеет смысл только в том случае, если у нас есть много разных правых частей. Для одной правой части эффективнее будет не считать матрицу L, а просто применить метод Гаусса. Для понимания происходящего подумайте, как получить LU-разложение из метода Гаусса (подсказка — в результате применения метода Гаусса мы получаем верхнетреугольную матрицу U, так что осталось только найти, что нам даст искомую матрицу L).

Объём вычислений для LU-разложения

Обратимся теперь к одному из вопросов, который мотивировал нас к изучению вычислительной математики: сколько вычислений потребуется на то, чтобы посчитать LU-разложение? Заметим, что на каждом шаге нам требуется решать системы с треугольными матрицами. Пусть размер матрицы k, тогда для вычисления одного решения нам потребуется $1 + 2 + \ldots + k = k(k+1)/2$ операций умножения и деления. Мы ограничимся подсчётом операций умножения и деления, поскольку на современном компьютере операции сложения и умножения могут быть выполнены одновременно. Для проверки понимания происходящего посчитайте, сколько всего операций потребуется (с учётом сложения и вычитания). Поскольку нам надо решить две системы с треугольными матрицами, то нам потребуется $k(k+1)$ операций. А чтобы найти диагональные элементы матриц l_{kk} и u_{kk}, нам понадобится ещё k умножений. Итого мы получаем на всё разложение для матрицы размера n

$$\sum_{k=1}^{n-1}(k(k+1)+k) = \sum_{k=1}^{n-1} k(k+2) = \frac{(n-1)^3}{3} + \frac{3(n-1)^2}{2} + \\ + \frac{7(n-1)}{6} \approx n^3/3$$

умножений и делений. Как нетрудно видеть, это уже гораздо лучше, чем $n!$. И хотя нам не удалось получить количество вычислений пропорционально количеству неизвестных, т. е. пропорционально размеру матрицы n^2, мы всё равно считаем полученный результат достойным. Заметим также, что добавление в расчёты операций сложения и вычитания скажется лишь на множителе перед n^3, но не на его степени. Таким образом, мы получаем вычислительную сложность алгоритма $O(n^3)$.

Вычисление определителя матрицы с помощью LU-разложения

Если мы построили LU-разложение матрицы за $O(n^3)$ операций, и при этом на диагонали матрицы L стоят 1, то, как известно из курса алгебры, $|A| = |L||U| = |U|$, следовательно, для вычисления определителя матрицы A нам понадобится сделать ещё $(n-1)$ умножений диагональных элементов матрицы U. Таким образом, мы свели задачу вычисления определителя матрицы к более простой задаче о вычислении LU-разложения.

Модификации LU-разложения

Теорема 5 (об LDU-разложении)**.** *Пусть матрица A размера n такова, что её главные миноры удовлетворяют условию $|A_k| \neq 0,\ k = 1, \ldots, n$. Тогда существует ЕДИНСТВЕННОЕ разложение матрицы вида $A = LDU$, где матрица L — нижнетреугольная с единицами на главной диагонали, матрица D — диагональная, а матрица U — верхнетреугольная с единицами на главной диагонали.*

Доказательство. Доказательство существования такого разложения следует напрямую из теоремы 4. Полагая $A = \tilde{L}\tilde{U}$, мы незамедлительно получаем уравнение $\tilde{L}\tilde{U} = LDU$, из которого следует, что диагональные элементы матрицы D удовлетворяют условию $d_{kk} = \tilde{l}_{kk}\tilde{u}_{kk} \neq 0$. Зная, как диагональная матрица умножается на любую другую матрицу, мы получаем что $l_{ij} = \tilde{l}_{ij}/\tilde{l}_{jj},\ u_{ij} = \tilde{u}_{ij}/\tilde{u}_{ii},\ \forall i, j = 1, \ldots, n$. Мы сменили обозначения матриц для LU-разложения, чтобы не смешивать два разных разложения. Следует обратить внимание на индексы в формулах и понять, почему они выглядят именно так (вспомните теорию умножения матриц).

Теперь докажем единственность. Воспользуемся методом «от

противного». Предположим, что у нас есть два разложения, т. е.

$$A = L^+D^+U^+ = L^-D^-U^-.$$

Тогда имеем $(L^-)^{-1}L^+ = D^-U^-(D^+U^+)^{-1}$. Поскольку матрицы у нас имеют определённую структуру, воспользуемся этим для доказательства. Во-первых, обратная к нижнетреугольной матрице есть нижнетреугольная матрица, к диагональной — диагональная, а к верхнетреугольной — верхнетреугольная. Во-вторых, произведение верхнетреугольных матриц есть верхнетреугольная матрица, а нижнетреугольных матриц — нижнетреугольная. Наконец, в-третьих, умножение на диагональную матрицу слева или справа оставляет нижнетреугольную матрицу нижнетреугольной, а верхнетреугольную — верхнетреугольной. Следует проверить сделанные утверждения с помощью определения операции умножения для двух матриц. Итак, с левой стороны равенства стоит нижнетреугольная матрица, а с правой — верхнетреугольная. Это означает, что слева и справа должна стоять диагональная матрица, иначе равенства не достичь. Заметим, что на главной диагонали матриц $(L^-)^{-1}$ и L^+ стоят 1, а это значит, что и на диагонали результата стоят 1, т. е. результат — единичная матрица: $(L^-)^{-1}L^+ = D^-U^-(D^+U^+)^{-1} = E$. Отсюда следует, что $L^- = L^+$. Кроме того, $D^-U^- = D^+U^+$.

Аналогично, пользуясь аргументом о структуре матриц и тем, что на диагонали матриц U^- и U^+, а также обратным к ним, стоят 1, получаем, что верно равенство $(D^-)^{-1}D^+ = U^-(U^+)^{-1} = E$, откуда следует $D^- = D^+$ и $U^- = U^+$, что завершает доказательство единственности и всей теоремы. □

Теорема 6 (о разложении Холесского(Холецкого)). *Пусть матрица A размера n такова, что $A = A^* > 0$. Тогда существует единственное разложение матрицы вида $A = LDL^*$, где матрица L — нижнетреугольная с единицами на главной диагонали, матрица D — диагональная и при этом $d_{kk} > 0,\ k = 1, \ldots, n$.*

Доказательство. Воспользуемся теоремой 5. Рассмотрим вектор

размера k:

$$\vec{\mathbf{x}}^k = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_k} \end{bmatrix} \neq \vec{\mathbf{0}}$$

и дополним его нулями до размера n:

$$\vec{\mathbf{y}}^k = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_k} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \neq \vec{\mathbf{0}}$$

для некоторого $k = 1, \ldots, n$. По определению положительной определённости матрицы для любого вектора такого вида справедливо неравенство $(A\vec{\mathbf{y}^k}, \vec{\mathbf{y}^k}) > 0$. Непосредственная проверка показывает, что $(A\vec{\mathbf{y}^k}, \vec{\mathbf{y}^k}) = (A_k\vec{\mathbf{x}^k}, \vec{\mathbf{x}^k})$, где A_k — k-й главный минор матрицы A. Так как мы доказали, что $(A_k\vec{\mathbf{x}^k}, \vec{\mathbf{x}^k}) > 0$ для любого ненулевого вектора размер k, то это значит, что $A_k\vec{\mathbf{x}^k} \neq \vec{\mathbf{0}}$, что, в свою очередь, указывает на то, что ядро матрицы A_k состоит только из нулевого вектора. Следовательно, матрица невырождена и $|A_k| \neq 0$. Данное утверждение справедливо для любого $k = 1, \ldots, n$, значит, условия теоремы 5 выполнены, ввиду чего можно говорить о существовании единственного разложения вида $A = LDU$.

По условию теоремы матрица A — самосопряжённая, т. е. $LDU = A = A^* = U^*D^*L^*$, что означает, что $D^* = D$ и $L^* = U$. Таким образом, существование и единственность разложения Холесского доказаны. При этом, как следует из теоремы 5, на главной диагонали матрицы L стоят единицы.

Из теоремы 5 следует, что матрица L, а значит и L^* — невырождены. Тогда для любого $k = 1, \ldots, n$ можно найти ненулевой вектор $\vec{\mathbf{v}}^k = (L^*)^{-1}\vec{\mathbf{e}}^k$, где $\vec{\mathbf{e}}^k$ — k-й единичный орт. По определению положительной определённости имеем $0 < (A\vec{\mathbf{v}}^k, \vec{\mathbf{v}}^k) =$

$(LDL^*\vec{\mathbf{v}}^k, \vec{\mathbf{v}}^k) = (DL^*\vec{\mathbf{v}}^k, L^*\vec{\mathbf{v}}^k) = (D\vec{\mathbf{e}}^k, \vec{\mathbf{e}}^k) = d_{kk}$. Теорема полностью доказана. □

Метод квадратного корня

Теорема 7. *Пусть матрица A размера n такова, что $A = A^* > 0$. Тогда существует единственное разложение матрицы вида $A = LL^*$, где матрица L — нижнетреугольная, и при этом $\nu_2(L) = \nu_2(L^*) = \sqrt{\nu_2(A)}$.*

Доказательство. По теореме 6 существует единственное разложение вида $A = \tilde{L}\tilde{D}\tilde{L}^*$. Мы сменили обозначения матриц для разложения Холесского, чтобы не смешивать два разных разложения. Из той же теоремы мы знаем, что на диагонали матрицы $\tilde{D}$ стоят положительные числа, следовательно, мы можем определить квадратный корень $\tilde{D}^{1/2} = [\sqrt{d_{kk}}]_{k=1}^n$ из матрицы $\tilde{D}$. Проверьте, что при умножении на саму себя эта матрица действительно даёт исходную матрицу $\tilde{D}$.

Итак, мы имеем

$$A = \tilde{L}\tilde{D}\tilde{L}^* = \tilde{L}\tilde{D}^{1/2}\tilde{D}^{1/2}\tilde{L}^* = (\tilde{L}\tilde{D}^{1/2}) \times (\tilde{L}\tilde{D}^{1/2})^* \equiv LL^*,$$

где $L = \tilde{L}\tilde{D}^{1/2}$. Таким образом, существование и единственность разложения доказаны.

Так как $A = LL^*$, то $\|L^*\|_2 = \sqrt{\rho(LL^*)} \equiv \sqrt{\rho(A)} = \sqrt{\|A\|_2}$. Обратите внимание на то, как мы определили вторую норму матрицы и как использовали это определение. Мы получили ранее, что $\|A\|_2 = \sqrt{\rho(A^*A)}$. Так как матрица A — самосопряжённая, то алгебра нам говорит, что у неё есть полный набор вещественных собственных чисел и полный набор собственных векторов, соответствующий этим собственным числам. Пусть $\vec{\mathbf{v}}$ — собственный вектор, соответствующий некоторому собственному числу λ,

т. е. $A\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$. По определению положительной определённости имеем $0 < (A\vec{\mathbf{v}}, \vec{\mathbf{v}}) = (\lambda\vec{\mathbf{v}}, \vec{\mathbf{v}}) = \lambda(\vec{\mathbf{v}}, \vec{\mathbf{v}})$. Используя первую аксиому скалярного произведения, получаем $\lambda > 0$, так как собственный вектор не может быть нулевым. Это верно для любого собственного числа матрицы A. Так как $A = A^*$, то и собственные числа у них совпадают, т. е. у матрицы A^*A собственные числа равны λ^2. Соответственно, для спектрального радиуса имеем $\rho(A) = \sqrt{\rho(A^*A)} = \|A\|_2$.

Из теоремы 6 следует, что матрица L — невырождена, потому существуют L^{-1} и $(L^*)^{-1}$. Заметим, что $L^*L \neq LL^*$, однако эти две матрицы подобны: $L^{-1}(LL^*)L = L^*L$. Как известно из алгебры, у подобных матриц спектр одинаковый, что означает, что $\rho(L^*L) = \rho(LL^*) \equiv \rho(A)$. Таким образом, получаем $\|L\|_2 = \sqrt{\rho(L^*L)} = \sqrt{\rho(LL^*)} \equiv \sqrt{\rho(A)} = \sqrt{\|A\|_2}$.

Аналогично доказываем утверждение $\|L^{-1}\|_2 = \|(L^*)^{-1}\|_2 \equiv \sqrt{\rho(A^{-1})} = \sqrt{\|A^{-1}\|_2}$.

По определению числа обусловленности мы имеем $\nu_2(L^*) = \nu_2(L) \equiv \|L\|_2\|L^{-1}\|_2 = \sqrt{\nu_2(A)}$. Теорема полностью доказана. $\square$

Обратите внимание, что теорема даёт равенство для чисел обусловленности только для второй матричной нормы. Для других норм это утверждение уже не будет верным.

Используя матричное представление разложения по методу квадратного корня, непосредственными вычислениями по методу математической индукции можно получить формулы метода для $k = 1, \ldots, n$:

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} |l_{kj}|^2},$$

$$l_{k+i,k} = \frac{a_{k+i,k} - \sum_{j=1}^{k-1} l_{k+i,j}\bar{l}_{kj}}{l_{kk}},\ i = 1, \ldots, n-k.$$

Формулы верны как для вещественного, так и для комплексного случая — обратите внимание на знак сопряжения над элементом l_{kj} в формуле. Отметим также, что даже в комплексном случае на диагонали матрицы L будут стоять вещественные положительные числа и только внедиагональные элементы могут быть комплексными. Для проверки понимания происходящего объясните с помощью теории умножения матриц, почему процесс вычисления элементов матрицы L идёт по столбцам от первого к последнему.

Лекция №3

Метод исключения Гаусса с выбором главного элемента (по столбцу)

Метод исключения Гаусса требует знания детерминантов главных миноров, которые обеспечат успешное завершение метода. На практике это непросто гарантировать заблаговременно. Можно, конечно, начинать считать и, если на диагонали попался ноль, то завершать алгоритм по причине вырожденности минора. Однако из алгебры известно, что независимо от того, вырождены или нет главные миноры, если сама матрица невырождена, то у системы есть единственное решение. Мы модифицируем метод исключения Гаусса так, чтобы он работал для невырожденной матрицы независимо от количества вырожденных главных миноров. Разумеется, это справедливо только для случая проведения точных вычислений, которые либо невозможны, либо весьма трудоёмки на реальном компьютере. Тем не менее подход вполне работоспособен и в случае приближённых вычислений, хотя и не гарантирует страховку от проблем с остановкой по причине появления нулей на диагонали, а также неточных результатов счёта.

Начнём с подготовительной работы. Матрица $P = [p_{ij}]_{i,j=1}^{n,n}$ называется *матрицей перестановок*, если в каждой её строке и столбце один ненулевой элемент равен единице. Соответственно, для элементов матрицы перестановок имеем формулу

$$p_{ij} = \begin{cases} 1, & j = k_i, \\ 0, & j \neq k_i. \end{cases}$$

Здесь $(k_1, \ldots, k_n)$ — это перестановка чисел $(1, \ldots, n)$. Непосредственная проверка показывает, что P — ортогональная матрица, т. е. $PP^* = P^*P = PP^{-1} = P^{-1}P = E$ и при этом $\nu_2(P) = 1$. Простейшей матрицей перестановок является единичная матрица, которая оставляет все элементы на месте. Нам понадобится элементарная матрица перестановок P_{kl} — это матрица перестановок на

основе перестановки элементов k и l в n-ке чисел $(1, \ldots, n)$. Опять же непосредственной проверкой можно заметить, что $P_{kl}A$ — это матрица A с переставленными строками k и l, а AP_{kl} — это та же матрица, но уже с переставленными столбцами номер k и l. Заметим, что для элементарной матрицы перестановок верно равенство $P_{kl}^* = P_{kl}$.

Теперь сформулируем метод и сразу обоснуем его свойства по методу математической индукции. Мы предполагаем, что матрица системы A невырождена. Сформулируем базу индукции. В первом столбце матрицы находим номер i_1 — такой, что

$$\max_{i=1,\ldots,n} |a_{i1}| \leq |a_{i_1 1}| \neq 0.$$

Если этот элемент равен нулю, то матрица автоматически вырождена, поскольку все элементы первого столбца равны нулю. С помощью элементарной матрицы перестановок меняем местами строки 1 и i_1 матрицы $A^{(1/2)} \equiv P_{1,i_1}A$. То же преобразование применяем к вектору правой части, чтобы система имела решение, совпадающее с решением исходной задачи $\vec{\mathbf{b}}^{(1/2)} \equiv P_{1,i_1}\vec{\mathbf{b}}$. После этого применим метод исключения к полученной системе с помощью матрицы

$$L_1 = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ -\frac{a_{21}^{(1/2)}}{a_{11}^{(1/2)}} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}^{(1/2)}}{a_{11}^{(1/2)}} & 0 & \ldots & 1 \end{bmatrix}.$$

Итак, $A^{(1)} = L_1 A^{(1/2)}$ и $\vec{\mathbf{b}}^{(1)} = L_1\vec{\mathbf{b}}^{(1/2)}$. Непосредственная проверка показывает, что если $|A| \neq 0$, то и $|A^{(1)}| \neq 0$, так как $|L_1| = 1$. При этом под диагональю в первом столбце стоят только нули.

Пусть после $k-1$ шага имеем систему $A^{(k-1)}\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(k-1)}$ и

$|A^{(k-1)}| \neq 0$. При этом

$$A^{(k-1)} = \begin{bmatrix} U_{k-1} & & U_{k-1,n-k+1} & \\ & a_{k,k}^{(k-1)} & \dots & a_{k,n}^{(k-1)} \\ 0_{n-k+1,k-1} & \vdots & \ddots & \vdots \\ & a_{n,k}^{(k-1)} & \dots & a_{n,n}^{(k-1)} \end{bmatrix},$$

$$\vec{\mathbf{b}}^{(k-1)} = \begin{bmatrix} \vec{\mathbf{f}}^{k-1} \\ \mathbf{b_k^{(k-1)}} \\ \vdots \\ \mathbf{b_n^{(k-1)}} \end{bmatrix}.$$

Обратите внимание на то, что «подвектор» $\vec{\mathbf{f}}^{(k-1)}$ размера $k-1$ не меняется при последующих преобразованиях метода исключения. Теперь находим i_k — такой, что $0 \neq |a_{i_k k}| \leq \max_{i=k,\dots,n} |a_{ik}^{(k-1)}|$. Если этот элемент равен нулю, то матрица $A^{(k-1)}$ (так же как и A) автоматически вырождена, поскольку нижний правый блок матрицы вырожден (первый столбец блока состоит из нулей). Детерминант главного минора U_{k-1} размера $k-1$ уже не играет роли, поскольку детерминант матрицы определяется как произведение детерминантов главного минора U_{k-1} и детерминанта нижнего правого блока. С помощью элементарной матрицы перестановок вновь меняем местами строки k и i_k матрицы $A^{(k-1)}$ и в результате получаем $A^{(k-1/2)} \equiv P_{k,i_k} A^{(k-1)}$. То же преобразование применяем к вектору правой части, чтобы система имела решение, совпадающее с решением исходной задачи $\vec{\mathbf{b}}^{(k-1/2)} \equiv P_{k,i_k} \vec{\mathbf{b}}^{(k-1)}$. После этого применим метод исключения к полученной системе

с помощью матрицы

$$L_k = \begin{bmatrix} 1 & 0 & & & \dots & & 0 \\ \vdots & \ddots & & & & & \vdots \\ 0 & \dots & 1 & & 0 & \dots & 0 \\ 0 & \dots & -\frac{a_{k+1,k}^{(k-1/2)}}{a_{k,k}^{(k-1/2)}} & & 1 & \dots & 0 \\ 0 & \dots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & -\frac{a_{n,k}^{(k-1/2)}}{a_{k,k}^{(k-1/2)}} & & 0 & \dots & 1 \end{bmatrix}.$$

Итак, $A^{(k)} = L_1 A^{(k-1/2)}$ и $\vec{\mathbf{b}}^{(k)} = L_1 \vec{\mathbf{b}}^{(k-1/2)}$. Непосредственная проверка показывает, что если $|A^{(k-1)}| \neq 0$, то и $|A^{(k)}| \neq 0$, так как $|L_k| = 1$. При этом под диагональю в k-м столбце стоят только нули, и нули под диагональными элементами предыдущих столбцов не изменились.

Таким образом, по методу математической индукции после $n-1$ шага мы получим систему с верхнетреугольной матрицей $A^{(n-1)}\vec{\mathbf{x}} \equiv U\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(n-1)}$.

Теорема 8. *Пусть матрица A размера n такова, что $|A| \neq 0$. Тогда существует разложение матрицы PA в произведение нижней и верхней треугольных матриц вида $PA = LU$, при этом $P = P_{n-1,i_{n-1}} \cdot \ldots \cdot P_{1,i_1}$, $L^{-1} = \tilde{L}_{n-1} \cdot \ldots \cdot \tilde{L}_1$, $\tilde{L}_{k+1} = P_{n-1,i_{n-1}} \cdot \ldots \cdot P_{k,i_k} \cdot L_k \cdot P_{k,i_k} \cdot \ldots \cdot P_{n-1,i_{n-1}}$.*

Доказательство. Заметим, что умножение элементарной матрицы перестановок на себя даёт единичную матрицу, что приводит формулы в теореме к той двухшаговой процедуре (перестановка текущего максимального по модулю элемента на диагональ + зануление поддиагональных элементов текущего столбца), которую мы обосновали перед формулированием этой теоремы.

Непосредственная проверка показывает, что матрицы L_k и $\tilde{L}_k$

имеют одинаковую нижнетреугольную структуру, следовательно, и матрица L действительно нижняя треугольная. $\square$

Проверьте, что вы действительно поняли, почему перестановки строк и столбцов именно в указанной последовательности не меняют структуру матриц. Вспомните, почему обращение треугольной матрицы даёт матрицу той же треугольной структуры.

Метод вращений для решения системы уравнений

Мы рассмотрели несколько подходов, в основе которых лежит идея метода исключения Гаусса. Идея состоит в том, чтобы преобразовать квадратную матрицу, которую неясно как обратить быстро, в произведение двух треугольных матриц, каждая из которых обращается достаточно легко. При этом исходная матрица должна быть невырожденной (как минимум), чтобы метод отработал до конца. Однако на практике изначально невырожденная матрица в процессе вычислений и сопутствующих им округлений может стать в том числе и вырожденной. Указанные наблюдения привели исследователей к мысли о том, чтобы попробовать другие подходы. Мы рассмотрим лишь два из них, и первым будет метод вращений.

Проведём подготовительную работу. Мы будем называть *элементарной матрицей вращений* матрицу вида

$$Q_{kp} = \begin{bmatrix} E_{k-1} & & 0 & & 0 \\ & \bar{c}_{kp} & 0 & -\bar{s}_{kp} & \\ 0 & 0 & E_{p-k-1} & 0 & 0 \\ & s_{kp} & 0 & c_{kp} & \\ 0 & & 0 & & E_{n-p} \end{bmatrix}.$$

Здесь предполагается, что $k < p$ и $\bar{c}_{kp}c_{kp} + \bar{s}_{kp}s_{kp} = 1$. Последнее равенство указывает на то, что величины c_{kp} и s_{kp} можно считать

косинусом и синусом некоторого угла (отсюда и название метода — метод вращений). Обратите внимание на то, что элементарная матрица вращений является единичной матрицей, у которой изменены только две строки и два столбца с номерами k и p. При этом сама единичная матрица является тривиальным примером элементарной матрицы вращений.

Непосредственная проверка показывает, что произведение вида $Q_{kp}A$ изменяет только две строки матрицы A с номерами k и p. Все остальные строки матрицы A остаются в неизменном виде. Также можно посчитать и увидеть, что $Q_{kp}Q_{kp}^* = Q_{kp}^*Q_{kp} = E$, т. е. элементарная матрица вращений является *унитарной*, т. е. легко обратимой с помощью операции сопряжения! Заметим, что $|Q_{kp}| = 1$, что также проверяется непосредственными вычислениями.

Теорема 9. *Рассмотрим произвольную матрицу A размера n. Тогда существует разложение матрицы $A = QR$, где матрица Q — унитарная, а матрица R — верхнетреугольная.*

Доказательство. Заметим в первую очередь, что унитарная матрица может быть заполненной ненулевыми элементами, однако она всегда остаётся легко обратимой с помощью операции сопряжения. Как решать систему с верхнетреугольной матрицей, мы уже разбирали ранее.

Воспользуемся методом математической индукции. Пусть после $k-1$ шага имеем систему $A^{(k-1)}\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(k-1)}$. При этом

$$A^{(k-1)} = \begin{bmatrix} R_{k-1} & & R_{k-1,n-k+1} & \\ & a_{k,k}^{(k-1)} & \dots & a_{k,n}^{(k-1)} \\ 0_{n-k+1,k-1} & \vdots & \ddots & \vdots \\ & a_{n,k}^{(k-1)} & \dots & a_{n,n}^{(k-1)} \end{bmatrix},$$

$$\vec{\mathbf{b}}^{\,(k-1)} = \begin{bmatrix} \vec{\mathbf{f}}^{(k-1)} \\ \mathbf{b_k^{(k-1)}} \\ \vdots \\ \mathbf{b_n^{(k-1)}} \end{bmatrix}.$$

«Подвектор» $\vec{\mathbf{f}}^{(k-1)}$ размера $k-1$ не меняется при последующих преобразованиях метода вращений. Для того чтобы написать базу индукции, нам понадобится всего лишь положить $A^{(0)} \equiv A$ и $\vec{\mathbf{b}}^{(0)} \equiv \vec{\mathbf{b}}$.

Возьмём матрицы $Q_{k,k+i},\ i = 1, \ldots, n-k$ с элементами $c_{k,k+i} \equiv \frac{a_{k,k}^{(k-1,i-1)}}{r_{k,k+i}}$ и $s_{k,k+i} \equiv \frac{a_{k+i,k}^{(k-1,i-1)}}{r_{k,k+i}}$ (косинус и синус некоторого угла), если $r_{k,k+i} \equiv \sqrt{|a_{k,k}^{(k-1,i-1)}|^2 + |a_{k+i,k}^{(k-1,i-1)}|^2} \neq 0$. Здесь мы полагаем $a_{ij}^{(k-1,0)} \equiv a_{ij}^{(k-1)}\ \forall i, j = 1, \ldots, n$. Если так оказалось, что $r_{k,k+i} = 0$, то полагаем $Q_{k,k+i} = E$. Применим полученные элементарные матрицы вращений к нашей системе уравнений: $A^{(k-1,i)} = = Q_{k,k+i} A^{(k-1,i-1)}$ и $\vec{\mathbf{b}}^{(k-1,i)} = Q_{k,k+i} \vec{\mathbf{b}}^{(k-1,i-1)},\ i = 1, \ldots, n-k$. Здесь предполагается $A^{(k-1,0)} \equiv A^{(k-1)}$ и $\vec{\mathbf{b}}^{(k-1,0)} \equiv \vec{\mathbf{b}}^{(k-1)}$. Обратите внимание, что k-я строка матрицы $A^{(k-1,i-1)}$ изменяется после применения очередной элементарной матрицы вращений, что и приводит к необходимости использовать составной индекс $^{(k-1,i-1)}$. Заметим, что после применения очередной элементарной матрицы вращения элементы $a_{k+j,k}^{(k-1,i)} = 0,\ j = 1, \ldots, i$.

Пусть $Q_k \equiv Q_{k,n} \cdot \ldots \cdot Q_{k,k+1}$. Тогда $A^{(k)} = Q_k A^{(k-1)}$ и $\vec{\mathbf{b}}^{(k)} = Q_k \vec{\mathbf{b}}^{(k-1)}$. При этом матрица Q_k — унитарная, а матрица $A^{(k)}$ имеет нули под главной диагональю в столбцах с номерами $1, \ldots, k$. Чтобы оформить базу индукции, нам потребуется с помощью указанного алгоритма обнулить все элементы первого столбца, кроме первого элемента.

Таким образом, по методу математической индукции после выполнения $n-1$ шага получим систему с верхнетреугольной матрицей $A^{(n-1)}\vec{\mathbf{x}} \equiv R\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(n-1)}$. Положив $Q^* \equiv Q_{n-1} \cdot \ldots \cdot Q_1$

приходим к утверждению теоремы. $\square$

Обратите внимание, что если $|A| \neq 0$, то $|R| \neq 0$, а если $|A| = 0$, то $|R| = 0$. В любом случае метод даст разложение, даже если матрица вырождена. Проверьте, что вы поняли, что если $A = 0$, то $Q = E$ и $R = 0$. Для проверки усвоения материала получите равенства $\nu_2(A) = \nu_2(Q)\nu_2(R) = \nu_2(R)$.

QR-разложение, полученное в результате применения метода вращений, можно использовать для решения в том числе и вырожденных систем, а также получать информацию о совместности системы уравнений.

Лекция №4

Метод отражений для решения системы уравнений

Проведём подготовительную работу. Будем называть *матрицей отражений* матрицу H, удовлетворяющую равенству $H\vec{\mathbf{a}} = \vec{\mathbf{a}} - 2(\vec{\mathbf{a}}, \vec{\mathbf{w}})\vec{\mathbf{w}} = (E - 2\vec{\mathbf{w}}\vec{\mathbf{w}}^*)\vec{\mathbf{a}}$. В нашем случае предполагается, что вектора $\vec{\mathbf{a}}$ и $H\vec{\mathbf{a}}$ заданы заранее, тогда мы легко получаем формулу для вычисления вектора $\vec{\mathbf{w}} = \frac{\vec{\mathbf{a}} - H\vec{\mathbf{a}}}{\|\vec{\mathbf{a}} - H\vec{\mathbf{a}}\|_2}$. Заметим, что при умножении вектора-столбца на вектор-строку мы получаем матрицу! Именно поэтому возможно вычесть из единичной матрицы матрицу специального вида $2\vec{\mathbf{w}}\vec{\mathbf{w}}^*$. Если переставить операцию * на первый аргумент, то мы получим удвоенное скалярное произведение вектора на себя $2\vec{\mathbf{w}}^*\vec{\mathbf{w}}$, т. е. число. Вычесть число из матрицы нельзя, поскольку такая операция не определена. Подобная ошибка является весьма распространённой среди студентов.

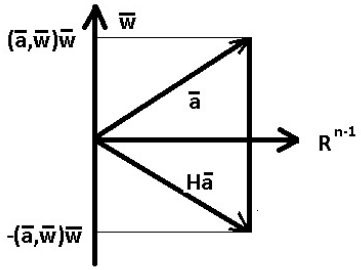

Рис. 1. Геометрическая интерпретация преобразования отражения

Как видно из рис. 1, мы имеем $\vec{\mathbf{w}}$ единичной длины в евклидовой норме, задающим одну координатную ось в пространстве векторов размерности n. Мы отражаем $\vec{\mathbf{a}}$ относительно подпространства размерности $n-1$, ортогонального $\vec{\mathbf{w}}$, и получаем $H\vec{\mathbf{a}}$ (отсюда и название метода — метод отражений). Данная операция возможна для любых $\vec{\mathbf{a}}$ и $\vec{\mathbf{w}}$ и даёт единственное решение. Обратите внимание на то, что единичная матрица является тривиальным примером матрицы отражений в случае, когда $\vec{\mathbf{a}} \perp \vec{\mathbf{w}}$. Если же $\vec{\mathbf{a}} \parallel \vec{\mathbf{w}}$, то матрица отражений будет $(-E)$. Также заметим, что преобразование отражения не меняет

евклидову длину вектора, т. е. $\|\vec{\mathbf{a}}\|_2 = \|H\vec{\mathbf{a}}\|_2$.

Непосредственная проверка показывает, что $H = H^* = H^{-1}$, т. е. матрица отражений является *унитарной*, следовательно, легко обратимой с помощью операции сопряжения! Заметим, что $|H| = -1$, что также проверяется непосредственными вычислениями.

Теорема 10. *Рассмотрим произвольную матрицу A размера n. Тогда существует разложение матрицы $A = HR$, где матрица H — унитарная, а матрица R — верхнетреугольная.*

Доказательство. Заметим в первую очередь, что унитарная матрица может быть заполненной ненулевыми элементами, однако она всегда остаётся легко обратимой с помощью операции сопряжения. Способы решения систем с верхнетреугольной матрицей были разобраны нами ранее.

Воспользуемся методом математической индукции. Пусть после $k-1$-го шага имеем систему $A^{(k-1)}\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(k-1)}$. При этом

$$A^{(k-1)} = \begin{bmatrix} R_{k-1} & & R_{k-1,n-k+1} & \\ & a_{k,k}^{(k-1)} & \dots & a_{k,n}^{(k-1)} \\ 0_{n-k+1,k-1} & \vdots & \ddots & \vdots \\ & a_{n,k}^{(k-1)} & \dots & a_{n,n}^{(k-1)} \end{bmatrix},$$

$$\vec{\mathbf{b}}^{(k-1)} = \begin{bmatrix} \vec{\mathbf{f}}^{(k-1)} \\ \mathbf{b_k^{(k-1)}} \\ \vdots \\ \mathbf{b_n^{(k-1)}} \end{bmatrix}.$$

Следует обратить внимание на то, что «подвектор» $\vec{\mathbf{f}}^{(k-1)}$ размера $k-1$ не меняется при последующих преобразованиях метода отражений. Для того чтобы написать базу индукции, нам понадобится всего лишь положить $A^{(0)} \equiv A$ и $\vec{\mathbf{b}}^{(0)} \equiv \vec{\mathbf{b}}$.

Возьмём матрицу вида $H_k = E - 2\vec{\mathbf{w}}^{(k)}\vec{\mathbf{w}}^{(k)\,*}$, где $\vec{\mathbf{w}}^{(k)}$ определяется таким образом, чтобы элементы под главной диагональю k-го столбца стали нулями. Таким образом, у нас имеется исходный вектор

$$\vec{\mathbf{a}}^{(k)} \equiv \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{a}_{\mathbf{k,k}}^{\mathbf{(k-1)}} \\ \vdots \\ \mathbf{a}_{\mathbf{n,k}}^{\mathbf{(k-1)}} \end{bmatrix}$$

и тот вектор, который мы хотим получить, чтобы матрица приобретала верхнетреугольную структуру

$$H_k\vec{\mathbf{a}}^{(k)} \equiv \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{a}_{\mathbf{k,k}}^{\mathbf{(k)}} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

Мы организовываем преобразование отражения таким образом, чтобы не менять элементы матрицы $A^{(k-1)}$, которые расположены выше k-й строки, т. е. мы сохраняем верхнетреугольную структуру матрицы, полученной к настоящему моменту.

Поскольку преобразование отражения не меняет евклидову длину вектора, то мы утверждаем, что $|a_{k,k}^{(k)}| = \|\vec{\mathbf{a}}\|_2$. Теперь мы можем посчитать $\vec{\mathbf{w}}^{(k)}$, с помощью которого легко построить требуемую нам матрицу отражения на k-м шаге. Очевидно, что если $\vec{\mathbf{a}}^{(k)} = \vec{\mathbf{0}}$, то $\vec{\mathbf{w}}^{(k)} = \vec{\mathbf{0}}$ и $H_k = E$. Таким образом, если в k-м столбце матрицы $A^{(k-1)}$ на диагонали и под ней стоят нули, то ничего делать не надо — у нас уже имеется требуемая структура. Если $\|\vec{\mathbf{a}}^{(k)}\|_2 \neq 0$, то у нас имеется два случая:

1) если $a_{k,k}^{(k-1)} = 0$ (диагональный элемент k-го столбца матрицы $A^{(k-1)}$ равен нулю), то

$$\vec{\mathbf{w}}^{(k)} = \frac{\vec{\mathbf{a}}^{(k)} - \|\vec{\mathbf{a}}^{(k)}\|_2 \vec{\mathbf{e}}^k}{\|\vec{\mathbf{a}}^{(k)} - \|\vec{\mathbf{a}}^{(k)}\|_2 \vec{\mathbf{e}}^k\|_2},$$

где $\vec{\mathbf{e}}^k$ – k-й единичный орт;

2) если $a_{k,k}^{(k-1)} \neq 0$ (диагональный элемент k-го столбца матрицы $A^{(k-1)}$ НЕ равен нулю), то

$$\vec{\mathbf{w}}^{(k)} = \frac{\vec{\mathbf{a}}^{(k)} + sign(a_{k,k}^{(k-1)})\|\vec{\mathbf{a}}^{(k)}\|_2 \vec{\mathbf{e}}^k}{\|\vec{\mathbf{a}}^{(k)} + sign(a_{k,k}^{(k-1)})\|\vec{\mathbf{a}}^{(k)}\|_2 \vec{\mathbf{e}}^k\|_2},$$

где $sign()$ — функция знака числа.

Применим преобразование отражения к системе на $k-1$-м шаге: $A^{(k)} = H_k A^{(k-1)}$ и $\vec{\mathbf{b}}^{(k)} = H_k \vec{\mathbf{b}}^{(k-1)}$. Непосредственные вычисления показывают, что матрица $A^{(k)}$ имеет нули под главной диагональю в столбцах с номерами $1, \ldots, k$. Чтобы оформить базу индукции нам потребуется с помощью указанного алгоритма обнулить все элементы первого столбца, кроме первого элемента.

Таким образом, по методу математической индукции после выполнения $n-1$-го шага получим систему с верхнетреугольной матрицей $A^{(n-1)}\vec{\mathbf{x}} \equiv R\vec{\mathbf{x}} = \vec{\mathbf{b}}^{(n-1)}$. Положив $H^* \equiv H_{n-1} \cdot \ldots \cdot H_1$ приходим к утверждению теоремы. □

Обратите внимание, что если $|A| \neq 0$, то $|R| \neq 0$, а если $|A| = 0$, то $|R| = 0$. И в любом случае метод даст разложение, даже если матрица вырождена. Проверьте, что вы поняли, что если $A = 0$, то $H = E$ и $R = 0$. Для проверки усвоения материала, получите равенства $\nu_2(A) = \nu_2(Q)\nu_2(R) = \nu_2(R)$.

HR-разложение, полученное в результате применения метода вращений, можно использовать для решения в том числе и

вырожденных систем, а также получать информацию о совместности системы уравнений. Об этом мы и поговорим далее.

Решение системы с вырожденной матрицей

HR-разложение с перестановками столбцов матрицы

Воспользуемся тем же приёмом, что мы уже применяли для того, чтобы метод исключения Гаусса работал в более широком диапазоне вариантов — будем переставлять столбцы в процессе выполнения алгоритма HR-разложения. Мы также прибегнем к методу математической индукции, чтобы обосновать алгоритм. Начнём с базы. Выберем номер j_1 — такой, что $\|\vec{\mathbf{a}}^{j_1}\|_2 = \max_{j=1,\ldots,n}\|\vec{\mathbf{a}}^{j}\|_2$, где $\vec{\mathbf{a}}^{j}$ — это j-й столбец матрицы нашей системы A, т. е. выберем столбец матрицы с максимальной евклидовой нормой. С помощью элементарной матрицы перестановок P_{1j_1} сделаем этот столбец первым $A^{(1/2)} \equiv AP_{1j_1}$ и применим к полученной матрице преобразование отражения:

$$A^{(1)} \equiv H_1 A^{(1/2)} = H_1 A P_{1j_1} = \begin{bmatrix} R_1 & & R_{1,n-1} & \\ & a^{(1)}_{2,2} & \ldots & a^{(1)}_{2,n} \\ 0_{n-1,1} & \vdots & \ddots & \vdots \\ & a^{(1)}_{n,2} & \ldots & a^{(1)}_{n,n} \end{bmatrix}.$$

В данном случае матрица R_1 имеет размер 1 и единственный элемент r_{11}. В силу выбора перестановки и особенности метода отражений по сохранению евклидовой нормы преобразуемого вектора получаем $|r_{11}| \geq \|\vec{\mathbf{a}}^{j}\|_2 \geq |a_{ij}|,\ \forall i, j$. Обратите внимание, что r_{11} по модулю не меньше, чем ЛЮБОЙ другой элемент матрицы $A^{(1)}$!

Пусть после $k-1$ шага имеем

$$A^{(k-1)} = \begin{bmatrix} R_{k-1} & & R_{k-1,n-k+1} & \\ & a_{k,k}^{(k-1)} & \dots & a_{k,n}^{(k-1)} \\ 0_{n-k+1,k-1} & \vdots & \ddots & \vdots \\ & a_{n,k}^{(k-1)} & \dots & a_{n,n}^{(k-1)} \end{bmatrix},$$

и при этом

$$|r_{k-1,k-1}| \geq \| \begin{bmatrix} \mathbf{a_{k,j}^{(k-1)}} \\ \vdots \\ \mathbf{a_{n,j}^{(k-1)}} \end{bmatrix} \|_2 \geq |a_{ij}^{(k-1)}|, \ k-1 \leq i,j \leq n.$$

Обратите внимание, что элемент $r_{k-1,k-1}$ по модулю не меньше, чем любой элемент в той строке, где он находится (помимо элементов, которые находятся в строках ниже его)!

Выберем номер j_k — такой, что

$$\| \begin{bmatrix} \mathbf{a_{k,j_k}^{(k-1)}} \\ \vdots \\ \mathbf{a_{n,j_k}^{(k-1)}} \end{bmatrix} \|_2 = \max_{j=k,\dots,n} \| \begin{bmatrix} \mathbf{a_{k,j}^{(k-1)}} \\ \vdots \\ \mathbf{a_{n,j}^{(k-1)}} \end{bmatrix} \|_2,$$

т. е. выберем столбец правого нижнего минора с максимальной евклидовой нормой. С помощью элементарной матрицы перестановок P_{kj_k} сделаем этот столбец первым $A^{(k-1/2)} \equiv A^{(k-1)} P_{kj_k}$ и применим к полученной матрице преобразование отражения:

$$A^{(k)} \equiv H_k A^{(k-1/2)} = H_k A^{(k-1)} P_{kj_k} =$$

$$= \begin{bmatrix} R_k & & R_{k,n-k} & \\ & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ 0_{n-k,k} & \vdots & \ddots & \vdots \\ & a_{n,k+1}^{(k)} & \dots & a_{n,n}^{(k)} \end{bmatrix}.$$

В силу выбора перестановки и особенности метода отражений по сохранению евклидовой нормы преобразуемого вектора получаем

$$|r_{kk}| \geq \| \begin{bmatrix} \mathbf{a}_{\mathbf{k,j}}^{(\mathbf{k})} \\ \vdots \\ \mathbf{a}_{\mathbf{n,j}}^{(\mathbf{k})} \end{bmatrix} \|_2 \geq |a_{ij}^{(k)}|, \ k \leq i, j \leq n.$$

Обратите внимание, что r_{kk} по модулю не меньше, чем ЛЮБОЙ другой элемент правого нижнего минора матрицы $A^{(k)}$ размера $n - k\underline{+\mathbf{1}}$!

Если ядро матрицы A имеет размер m, т. е. $m = \dim(\ker A)$, тогда после $n - m$ шагов получим

$$HAP = R \equiv \begin{bmatrix} R_{n-m} & R_{n-m,m} \\ 0_{n-m,m} & 0_{m,m} \end{bmatrix},$$

где $H \equiv H_{n-m} \cdot \ldots \cdot H_1$ и $P \equiv P_{1,j_1} \cdot \ldots \cdot P_{n-m,j_{n-m}}$ — ортогональные матрицы.

Сведения из высшей алгебры о вырожденных системах уравнений

Вспомним некоторые сведения из алгебры, которые пригодятся нам в дальнейшем.

- Система $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ называется *совместной*, если она имеет хотя бы одно решение.
- Система совместна тогда и только тогда, когда $\vec{\mathbf{b}} \in \mathrm{Im}(A)$.
- Если $\vec{\mathbf{x}}^+$ — некоторое решение системы, то $\{\vec{\mathbf{x}}^+ + \vec{\mathbf{y}}, \ \forall \vec{\mathbf{y}} \in \ker(A)\}$ — общее решение системы.
- Если система $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ совместна, то для любой НЕВЫРОЖДЕННОЙ матрицы B система $BA\vec{\mathbf{x}} = B\vec{\mathbf{b}}$ также совместна и множество решений этих систем совпадают.

- Если $\vec{\mathbf{b}} \notin \mathrm{Im}(A)$, то система несовместна.
- *Обобщённым решением* несовместной системы *относительно некоторой нормы* $\|\cdot\|$ будем называть $\vec{\mathbf{x}}$ — такой, что

$$\|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\| = \min_{\vec{\mathbf{y}}} \|A\vec{\mathbf{y}} - \vec{\mathbf{b}}\|.$$

- Общее решение совместной системы уравнений совпадает с множеством её обобщённых решений.
- Множество обобщённых решений относительно евклидовой нормы $\{\vec{\mathbf{x}} \mid \|A\vec{\mathbf{x}} - \vec{\mathbf{b}}\|_2 = \min_{\vec{\mathbf{y}}} \|A\vec{\mathbf{y}} - \vec{\mathbf{b}}\|_2\}$ совпадает с общим решением системы $A^*A\vec{\mathbf{x}} = A^*\vec{\mathbf{b}}$.

Решение совместной системы с применением HR-разложения с перестановками столбцов

В точной арифметике HR-разложение приводит к результату, который мы уже рассмотрели. Однако при вычислениях на компьютере неизбежно возникают ошибки округления, и потому в результате применения этого разложения получаем систему вида

$$\begin{bmatrix} R_{n-m} & R_{n-m,m} \\ 0_{m,n-m} & \mathcal{E}_m \end{bmatrix} \begin{bmatrix} \vec{\mathbf{y}}^{\,n-m} \\ \vec{\mathbf{y}}^{\,m} \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{g}}^{\,n-m} \\ \vec{\delta}^{\,m} \end{bmatrix}.$$

Если ошибки округления были невелики, то матрица $\mathcal{E}_m$ и вектор $\vec{\delta}_m$ размера m (размерности ядра) будут содержать «маленькие» по модулю элементы. Поскольку в матрице R диагональные элементы мажорируют все элементы, расположенные правее и ниже, то как только диагональный элемент становится заметно меньше (не математический термин, но математические объяснения будут гораздо длиннее и сложнее) по сравнению с предыдущим, мы можем положить все оставшиеся элементы в разложении равными нулю. Операцию обнуления оставшихся

элементов нужно делать в процессе вычисления HR-разложения с перестановками, в противном случае модули элементов могут начать расти за счёт всё тех же ошибок округления и создать неверное впечатление, будто система невырождена, и увести вычисленное решение далеко от истинного. Вспомните, что худшее возможное удаление вычисленного решения от точного определяется числом обусловленности матрицы, а у матрицы $\mathcal{E}_m$ число обусловленности может быть сколь угодно плохим. В результате получим систему

$$\begin{bmatrix} R_{n-m} & R_{n-m,m} \\ 0_{m,n-m} & 0_m \end{bmatrix} \begin{bmatrix} \vec{\mathbf{y}}^{\,n-m} \\ \vec{\mathbf{y}}^{\,m} \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{g}}^{\,n-m} \\ \vec{\mathbf{0}}^{\,m} \end{bmatrix}.$$

У этой системы общее решение находится легко:

$$\begin{bmatrix} \vec{\mathbf{y}}^{\,n-m} \\ \vec{\mathbf{y}}^{\,m} \end{bmatrix} = \begin{bmatrix} R_{n-m}^{-1}(\vec{\mathbf{g}}^{\,n-m} - R_{n-m,m}\vec{\mathbf{y}}^{\,m}) \\ \vec{\mathbf{y}}^{\,m} \end{bmatrix}, \ \forall \vec{\mathbf{y}}^{\,m} \in \mathbb{R}^m,$$

а по нему — и решение исходной системы $\vec{\mathbf{x}} = P\vec{\mathbf{y}}$.

Метод прогонки решения систем с трёхдиагональными матрицами

Рассмотрим один важный частный случай метода исключения Гаусса — метод прогонки. В случае, когда матрица является трёхдиагональной, то система уравнений приобретает вид:

$$\begin{bmatrix} b_1 & c_1 & & 0 \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & a_n & b_n \end{bmatrix} \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \vdots \\ \mathbf{x_n} \end{bmatrix} = \begin{bmatrix} \mathbf{f_1} \\ \vdots \\ \vdots \\ \mathbf{f_n} \end{bmatrix}.$$

Обратите внимание, что компоненты вектора правой части системы уравнений мы обозначили через f_i, поскольку обозначение b_i мы использовали для элементов на главной диагонали матрицы. По теореме 4 в случае, если диагональные миноры матрицы

невырождены, мы можем посчитать LU-разложение. Поскольку матрица у нас имеет трёхдиагональную структуру, то мы получим разложение вида

$$\begin{bmatrix} b_1 & c_1 & & 0 \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & a_n & b_n \end{bmatrix} = \begin{bmatrix} d_1 & 0 & & 0 \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & 0 \\ 0 & & a_n & d_n \end{bmatrix} \begin{bmatrix} 1 & u_1 & & 0 \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & u_{n-1} \\ 0 & & 0 & 1 \end{bmatrix}.$$

Убедитесь, что вы поняли, почему в множителях L и U осталось много нулей, которые не заполнились в процессе вычисления LU-разложения. В данном случае легко рассчитать неизвестные элементы множителей с помощью метода математической индукции:

$$d_1 = b_1,\ u_1 = d_1^{-1}c_1,$$

$$d_2 = b_2 - a_2u_1,\ u_2 = d_2^{-1}c_2,$$

$$\vdots$$

$$d_{n-1} = b_{n-1} - a_{n-1}u_{n-2},\ u_{n-1} = d_{n-1}^{-1}c_{n-1},$$

$$d_n = b_n - a_nu_{n-1}.$$

После того, как посчитаны элементы матриц L и U, мы можем приступить к решению системы. Сначала решаем систему $L\vec{\mathbf{y}} = \vec{\mathbf{f}}$ (это называется *прямой ход* прогонки, поскольку мы считаем неизвестные компоненты $y_1, \ldots, y_n$ вектора $\vec{\mathbf{y}}$ от первой к последней):

$$y_1 = d_1^{-1}f_1 = d_1^{-1}b_1,$$

$$y_2 = d_2^{-1}(f_2 - a_2y_1) = (b_2 - a_2u_2)^{-1}(f_2 - a_2y_1)$$

$$\vdots$$

$$y_n = d_n^{-1}(f_n - a_ny_{n-1}) = (b_n - a_nu_n)^{-1}(f_n - a_ny_{n-1}).$$

Обратите внимание на то, что как для вычисления u_i, так и для вычисления y_i нам нужно вычислить d_i^{-1}. Вычисляя одновременно эти величины, мы можем сэкономить на одной операции (не

надо вычислять d_i^{-1} два раза). Собственно, поэтому метод прогонки является достаточно важным частным случаем.

Далее, решаем систему $U\vec{\mathbf{x}} = \vec{\mathbf{y}}$ (*обратный ход* метода прогонки, поскольку мы считаем неизвестные компоненты $x_1, \ldots, x_n$ вектора $\vec{\mathbf{x}}$ от последней к первой):

$$x_n = y_n,$$

$$x_{n-1} = y_{n-1} - u_{n-1}x_n$$

$$\vdots$$

$$x_1 = y_1 - u_1 x_2.$$

Проверка невырожденности главных миноров не самая простая задача, а потому так же, как и в случае LU-разложения, мы рассмотрим альтернативный вариант проверки этого непростого условия.

Теорема 11. *Пусть трёхдиагональная матрица A размера n такова, что $\forall i = 1, \ldots, n,\ |b_i| > |a_i| + |c_i|$ (здесь мы полагаем $a_1 = c_n = 0$). Тогда существует LU-разложение матрицы и метод прогонки применим.*

Доказательство. Воспользуемся методом от противного. Предположим, что для некоторого k-й главный минор оказался вырожден: $|A_k| = 0$. Тогда существует $\vec{\mathbf{z}} \neq \vec{\mathbf{0}}$ размера k — такой, что $A_k\vec{\mathbf{z}} = \vec{\mathbf{0}}$. У этого вектора найдём максимальную по модулю компоненту $\max_{1 \leq j \leq k} |z_j| = |z_i|$. Не уменьшая степени общности, предположим, что эта компонента стоит на позиции с номером i. Выпишем уравнение с номером i и получим

$$a_i z_{i-1} + b_i z_i + c_i z_{i+1} = 0.$$

Обратите внимание на то, что первое и последнее уравнение содержат на одно слагаемое меньше, но на доказательство это не

влияет. Отсюда следует

$$|b_i| \le |a_i|\frac{|z_{i-1}|}{|z_i|} + |c_i|\frac{|z_{i+1}|}{|z_i|} \le |a_i| + |c_i|,$$

что противоречит условию теоремы, следовательно, наше предположение было неверным, и теорема доказана. □

Приведённую выше теорему можно обосновать с помощью более общей идеи.

Теорема 12 (о кругах Гершгорина). *Пусть дана произвольная матрица A размера n, тогда все её собственные значения лежат в объединении кругов Гершгорина $S_i = \{z \in \mathbb{C} \ : \ |z - a_{ii}| \le r_i \equiv \sum_{j=1, j\neq i}^{n} |a_{ij}|\}, \ i = 1, \ldots, n$.*

Доказательство. Пусть λ — произвольное собственное значение матрицы A. Тогда, по определению, существует некоторый ненулевой вектор $\vec{\mathbf{x}}$ — такой, что $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$. Найдём номер i — такой, что компонента x_i будет максимальной по модулю среди всех компонент вектора. Тогда i-е уравнение для собственного вектора примет вид

$$(\lambda - a_{i,i})x_i = a_{i,1}x_1 + \ldots + a_{i,i-1}x_{i-1} + a_{i,i+1}x_{i+1} + \ldots + a_{i,n}x_n.$$

В силу выбора номера i мы можем утверждать, что $x_i \neq 0$ и

$$|\lambda - a_{i,i}| \le |a_{i,1}|\frac{|x_1|}{|x_i|} + \ldots + |a_{i,i-1}|\frac{|x_{i-1}|}{|x_i|} + |a_{i,i+1}|\frac{|x_{i+1}|}{|x_i|} + \ldots + |a_{i,n}|\frac{|x_n|}{|x_i|}$$

$$\le |a_{i,1}| + \ldots + |a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{i,n}| = r_i.$$

Таким образом, мы доказали, что собственное число λ лежит в некотором круге Гершгорина, следовательно, заведомо лежит в их объединении. В силу произвольности выбора собственного числа теорема доказана. □

Обратите внимание на то, что радиус круга Гершгорина определяется как сумма МОДУЛЕЙ ВНЕДИАГОНАЛЬНЫХ элементов В СТРОКЕ. Аналогично можно доказать теорему, где радиус круга Гершгорина определяется как сумма модулей внедиагональных элементов в столбце. Обратите внимание, что радиусы всех кругов Гершгорина считаются либо по строке, либо по столбцу, но никак не вперемешку.

Следствием теоремы 12 является утверждение о том, что матрица со строгим диагональным преобладанием будет невырожденной. Именно это утверждение можно использовать для доказательства теоремы 11. Так как $|a_{ii}| > r_i$, $\forall i = 1, \ldots, n$ (это и есть *строгое диагональное преобладание*), то ни один круг Гершгорина не проходит через начало координат на комплексной плоскости, следовательно, ноль не является собственным числом матрицы, т. е. матрица невырождена.

Лекция №5

Итерационные методы для решения систем линейных алгебраических уравнений

Мы рассмотрели прямые методы решения систем линейных алгебраических уравнений, а теперь переходим к новой теме — итерационные методы. Непосредственная проверка показывает, что система уравнений с комплексной матрицей сводится к вещественной системе уравнений двойного размера

$$\begin{bmatrix} \operatorname{Re} A & -\operatorname{Im} A \\ \operatorname{Im} A & \operatorname{Re} A \end{bmatrix} \begin{bmatrix} \operatorname{Re} \vec{\mathbf{x}} \\ \operatorname{Im} \vec{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \operatorname{Re} \vec{\mathbf{b}} \\ \operatorname{Im} \vec{\mathbf{b}} \end{bmatrix},$$

что позволяет нам сфокусироваться на вещественных системах уравнений в этом разделе.

Идея итерационных методов состоит в следующем: предположим, что мы посчитали LU-разложение матрицы A из системы уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$. В результате ошибок округления при счёте получим, что $A^{-1} \neq (LU)^{-1} \equiv \tilde{A}^{-1}$. Можно ожидать, что матрица $\tilde{A}^{-1}$ будет «близка» в некотором смысле к матрице A^{-1}. С помощью такого подхода получено решение $\vec{\tilde{\mathbf{x}}} \neq \vec{\mathbf{x}}$, но при этом верно равенство $A(\vec{\tilde{\mathbf{x}}} - \vec{\mathbf{x}}) = A\vec{\tilde{x}} - \vec{\mathbf{b}}$. Приведённую систему уравнений мы можем вновь решить с помощью LU-разложения и вновь получить приближённое решение из-за ошибок округления вида $\underline{\vec{\mathbf{x}}} - \vec{\mathbf{x}} = \tilde{A}^{-1}(A\vec{\tilde{x}} - \vec{\mathbf{b}})$, из которого получаем $\underline{\vec{\mathbf{x}}} = \vec{\mathbf{x}} - \tilde{A}^{-1}(A\vec{\tilde{x}} - \vec{\mathbf{b}})$. Эту процедуру можно повторять вновь и вновь, переписав в индексированном виде:

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - \tilde{A}^{-1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}),\ k = 0, 1, \dots\ .$$

Таким образом мы будем получать *итерационное уточнение* полученного решения. Понятно, что вместо матрицы $\tilde{A}^{-1}$ можно брать другие матрицы, которые приближают нас к искомому решению.

Одношаговым двухслойным итерационным методом решения системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ мы будем называть метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - H_k(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}),\ k = 0, 1, \ldots$$

с заданными начальным приближением $\vec{\mathbf{x}}^{0}$ (его можно взять, например, равным $\vec{\mathbf{0}}$) и матрицами H_k. Мы будем называть вектор $\vec{\mathbf{x}}^{k}$ — k-м *приближением* к решению системы $\vec{\mathbf{x}}$. Обратите внимание, что здесь индекс k — это не номер компоненты вектора, а k-й по порядку вектор, полученный по указанной выше итерационной процедуре. *Ошибкой* на k-й итерации мы будем называть вектор $\vec{\mathbf{z}}^{k} \equiv \vec{\mathbf{x}}^{k} - \vec{\mathbf{x}}$, а *невязкой* — вектор $\vec{\mathbf{r}}^{k} = A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}$. Как нетрудно заметить, $A\vec{\mathbf{z}}^{k} = \vec{\mathbf{r}}^{k}$. Разумеется, нас будет интересовать именно ошибка, поскольку она указывает на то, как близко мы подошли к точному решению системы $\vec{\mathbf{x}}$. Однако если бы было возможно её посчитать, то итерационный метод был бы не нужен, мы смогли бы получить точное решение из простой формулы при известной ошибке и посчитанном итерационном приближении. Поэтому нам остаётся только смотреть на невязку, поскольку невязка может быть посчитана в любой момент при известном приближении к точному решению.

По заданному итерационному методу мы можем определить, каким формулам будут удовлетворять ошибка

$$\vec{\mathbf{z}}^{k+1} = \vec{\mathbf{z}}^{k} - H_k A\vec{\mathbf{z}}^{k} = (E - H_k A)\vec{\mathbf{z}}^{k} \equiv S_k\vec{\mathbf{z}}^{k}$$

и невязка

$$\vec{\mathbf{r}}^{k+1} = \vec{\mathbf{r}}^{k} - AH_k r^{k} = (E - AH_k)\vec{\mathbf{r}}^{k} \equiv T_k\vec{\mathbf{r}}^{k}.$$

Матрица S_k называется *матрицей шага для ошибки*, а матрица T_k — *матрицей шага для невязки*. В случае невырожденной системы уравнений матрицы S_k и T_k подобны, т. е. имеют одинаковый спектр и одновременно либо вырождены, либо нет.

Итерационный метод называется *сходящимся*, если $\forall \vec{\mathbf{x}}^0 \in \mathbb{R}^n$ $\lim_{k\to\infty}\|\vec{\mathbf{z}}^k\| = 0$. Обратите внимание на то, что метод сходится только в том случае, если предел равен нулю для ЛЮБОГО начального вектора $\vec{\mathbf{x}}^0$, а не для какого-либо набора начальных векторов. Заметим также, что определение не зависит от выбора нормы, поскольку в $\mathbb{R}^n$ все нормы эквивалентны.

Стационарный итерационный метод

Метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - H(A\vec{\mathbf{x}}^k - \vec{\mathbf{b}}),\ k = 0, 1, \ldots$$

с заданным начальным приближением $\vec{\mathbf{x}}^0$ и заданной матрицей H мы будем называть *стационарным* одношаговым итерационным методом.

Впредь будем полагать, что $|A| \neq 0$ и $|H| \neq 0$, т. е. искать решения у невырожденных систем с помощью невырожденных матриц. Рассмотрим теперь случаи, при которых стационарный итерационный метод сходится к искомому решению.

Теорема 13. *Если* $\|S\| = \|E - HA\| < 1$, *то стационарный итерационный метод сходится.*

Доказательство. $\|\vec{\mathbf{z}}^k\| = \|S\vec{\mathbf{z}}^{k-1}\| \leq \|S\|\|\vec{\mathbf{z}}^{k-1}\| = \|S\|\|S\vec{\mathbf{z}}^{k-2}\| \leq \ldots \leq \|S\|^k\|\vec{\mathbf{z}}^0\| \to 0\ \forall\vec{\mathbf{z}}^0$. Здесь мы воспользовались свойствами подчинённой нормы и сведениями из математического анализа о сходимости геометрической последовательности. Поскольку вектор $\vec{\mathbf{z}}^0$ произвольный, то и вектор $\vec{\mathbf{x}}^0 = \vec{\mathbf{z}}^0 + \vec{\mathbf{x}}$ будет произвольным. □

Теорема 14. *Если* $\|T\| = \|E - AH\| < 1$, *то стационарный итерационный метод сходится.*

Доказательство.

$$\|\vec{\mathbf{r}}^k\| = \|T\vec{\mathbf{r}}^{k-1}\| \le \|T\|\|\vec{\mathbf{r}}^{k-1}\| =$$

$$= \|T\|\|T\vec{\mathbf{r}}^{k-2}\| \le \ldots \le \|T\|^k\|\vec{\mathbf{r}}^0\| \to 0 \ \forall\vec{\mathbf{r}}^0.$$

Здесь мы также воспользовались свойствами подчинённой нормы и сведениями из математического анализа о сходимости геометрической последовательности. Поскольку вектор $\vec{\mathbf{r}}^0$ произвольный, то вектор $\vec{\mathbf{z}}^0 = A^{-1}\vec{\mathbf{r}}^0$ и, следовательно, $\vec{\mathbf{x}}^0 = \vec{\mathbf{z}}^0 + \vec{\mathbf{x}}$ будут произвольными. Для завершения доказательства заметим, что последовательность $\|\vec{\mathbf{z}}^k\|, k = 1, 2, \ldots$ мажорируется последовательностью $\|\vec{\mathbf{r}}^k\|, k = 1, 2, \ldots$: $\|\vec{\mathbf{z}}^k\| \le \|A^{-1}\vec{\mathbf{r}}^k\| \le \|A\|\|\vec{\mathbf{r}}^k\|$, ввиду чего и последовательность $\|\vec{\mathbf{z}}^k\| \to 0$. □

Теорема 15 (критерий сходимости стационарного итерационного метода)**.** *Стационарный итерационный метод сходится тогда и только тогда, когда* $\rho(S) < 1$, *где* S *— матрица шага для ошибки.*

Доказательство. Необходимость. Будем предполагать, что стационарный итерационный метод сходится, т. е. $\forall\vec{\mathbf{x}}^0 \ \|\vec{\mathbf{z}}^k\| \to 0$ при $k \to \infty$. В силу произвольности выбора вектора $\vec{\mathbf{x}}^0$ мы можем взять его таким, что вектор $\vec{\mathbf{z}}^0 = \vec{\mathbf{x}}^0 - \vec{\mathbf{x}} = \vec{\mathbf{y}} \ne \vec{\mathbf{0}}$ является собственным для матрицы шага для ошибки S, соответствующему некоторому собственному числу λ. По определению имеем $S\vec{\mathbf{y}} = \lambda\vec{\mathbf{y}}$. Тогда $\vec{\mathbf{z}}^k = S\vec{\mathbf{z}}^{k-1} = \ldots = S^k\vec{\mathbf{z}}^0 = \lambda^k\vec{\mathbf{z}}^0$. Так как у вектора $\vec{\mathbf{z}}^k$ от k зависит только величина λ^k, то в случае вещественного собственного вектора $\vec{\mathbf{y}}$ норма $\|\vec{\mathbf{z}}^k\| = |\lambda|^k\|\vec{\mathbf{z}}^0\| \to 0$ при $k \to \infty$ только в случае, если $|\lambda| < 1$. Если собственный вектор $\vec{\mathbf{y}}$ и собственное число λ являются комплексными, нужно рассмотреть вещественные вектора начальной ошибки вида $\vec{\mathbf{z}}^0 = \vec{\mathbf{y}} + \overline{\vec{\mathbf{y}}}$ и $\vec{\mathbf{z}}^0 = i(\vec{\mathbf{y}} - \overline{\vec{\mathbf{y}}})$. В силу произвольности выбора собственного вектора и соответствующего ему собственного числа мы можем утверждать, что все собственные числа матрицы S будут меньше единицы по модулю, что по определению означает, что $\rho(S) < 1$. Необходимое условие доказано.

Достаточность. Как известно из алгебры, матрицу S можно привести к жордановой нормальной форме с помощью, вообще говоря, ортогонального преобразования подобия, т. е. $S = QJQ^{-1}$, где $J = \mathrm{diag}\{J_1, \ldots, J_m\}$, $m \leq n$ и J_i — жорданова клетка вида

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ 0 & & & \lambda_i \end{bmatrix}.$$

Размер жордановой клетки n_i не имеет значения. По условию мы имеем $\rho(S) < 1$, что означает, что $-1 < \lambda_i < 1$. Пусть E_i — единичная матрица размера n_i, а P_i — это матрица с единицами непосредственно над главной диагональю. Тогда по формуле бинома Ньютона для некоторого достаточно большого $k > n_i$ имеем $J_i^k = (\lambda_i E_i + P_i)^k = \sum_{j=0}^{k} C_k^j \lambda_i^{k-j} P_i^j = \sum_{j=0}^{n_i - 1} C_k^j \lambda_i^{k-j} P_i^j$, так как матрица P_i — нильпотентна, т. е. обращается в нулевую матрицу при всех степенях, больших или равных n_i. По определению числа сочетаний $C_k^j \equiv \frac{k!}{j!(k-j)!} = \frac{(k-j+1)\ldots k}{j!} \leq \frac{k^j}{j!} \leq k^n$. Из математического анализа известно, что $\lim_{k\to\infty} k^n \lambda_i^k = 0 \ \forall |\lambda_i| < 1$ (полином растёт медленнее, чем убывает экспонента). Это означает, что $\forall j < n_i \ \lim_{k\to\infty} C_k^j \lambda_i^{k-j} = 0$, следовательно, $J_i \to 0$, следовательно, и матрица $J \to 0$ поэлементно при $k \to \infty$. Так как $S^k = (QJQ^{-1})^k = QJ^kQ^{-1}$, то и матрица $S \to 0$ при $k \to \infty$ поэлементно. Это, в свою очередь, указывает на то, что $\vec{\mathbf{z}}^k = S^k \vec{\mathbf{z}}^0 \to 0$ поэлементно, что влечёт за собой $\|\vec{\mathbf{z}}^k\|_\infty \to 0$. Так как все нормы в $\mathbb{R}^n$ эквивалентны, то $\|\vec{\mathbf{z}}^k\| \to 0$ при $k \to \infty$, что и требовалось доказать. $\square$

После того как мы рассмотрели условия, при которых стационарный метод сходится к искомому решению, возникает естественный вопрос: насколько быстро он сходится? Как известно из математического анализа, последовательность может очень долго находиться сколь угодно далеко от своего предела. А нам (напоминаем!) нужно получить ответ за разумное время.

Рассмотрим вопрос о том, сколько нужно сделать шагов $k = k(\varepsilon)$ стационарного итерационного метода, чтобы начальная ошибка $\|\vec{\mathbf{z}}^0\|$ уменьшилась в ε^{-1} раз (здесь предполагается, что ε является малой величиной).

Теорема 16. *Если* $\|S\| = \|E - HA\| < 1$, *то для уменьшения начальной ошибки в* ε^{-1} *раз потребуется*

$$k(\varepsilon) = \left[\frac{-\ln \varepsilon}{-\ln\|S\|}\right] + 1$$

шагов стационарного итерационного метода.

Доказательство. Квадратные скобки в формуле для $k(\varepsilon)$ обозначают целую часть числа. Так как количество шагов итерационного метода может быть только натуральным, то нам нужно привести вещественное выражение к целому числу. Учитывая, что целая часть числа не делает округлений, нужно добавить единицу, чтобы гарантировать требуемый результат. При любом $k \geq k(\varepsilon)$ имеем $\|\vec{\mathbf{z}}^k\| = \|S\vec{\mathbf{z}}^{k-1}\| \leq \|S\|\|\vec{\mathbf{z}}^{k-1}\| \leq \ldots \leq \|S\|^k\|\vec{\mathbf{z}}^0\| \leq \varepsilon\|\vec{\mathbf{z}}^0\|$. Здесь мы воспользовались свойствами подчинённой нормы. □

Средней скоростью за k *итераций* называется величина $R_k \equiv \ln \sqrt[k]{\|S^k\|}$. Непосредственная проверка с использованием мультипликативности матричной нормы показывает, что $R_k \geq -\ln\|S\|$. Также получаем, что $\|S^k\| = e^{-kR_k} \leq \varepsilon$, если $k \geq \frac{-\ln\varepsilon}{R_k}$. Обратите внимание, что средняя скорость зависит от выбора матричной нормы.

Асимптотической скоростью сходимости называется величина $R_\infty \equiv \lim_{k\to\infty} R_k$.

Теорема 17. *Если* $\rho(S) = \rho(E - HA) < 1$, *то* $R_\infty = -\ln \rho(S)$.

Доказательство. Теорема 15 и определение бесконечной матричной нормы дают нам $\|S^k\|_\infty \le Ck^n\rho^k(S)\ \forall k \ge n$ и некоторой константы C, не зависящей от k. В силу эквивалентности всех норм в конечномерном пространстве существует константа β (не зависит от k) — такая, что $\|S\| \le \beta\|S\|_\infty$. Получаем

$$\bar{\lim}_{k\to\infty}\sqrt[k]{\|S^k\|} \le \bar{\lim}_{k\to\infty}\left[\sqrt[k]{\beta C}(\sqrt[k]{k})^n\rho(S)\right] = \rho(S).$$

Здесь мы воспользовались сведениями из математического анализа о том, что корень k-й степени из числа и полинома в пределе даёт единицу.

Заметим далее, что $\|S^k\| \ge \rho(S^k)$ на основании того факта, что подчинённая матричная норма не может быть меньше спектрального радиуса. По определению подчинённой матричной нормы имеем $\|S\| = \sup_{\vec{\mathbf{x}}\ne 0}\frac{\|S\vec{\mathbf{x}}\|}{\|\vec{\mathbf{x}}\|} \ge |\lambda|$ для любого собственного числа матрицы, так как в качестве произвольного вектора $\vec{\mathbf{x}}$ мы можем взять и собственные вектора матрицы, которые по определению не равны нулю. По аксиомам нормы любой скаляр можно вынести с модулем из-под знака нормы. Следовательно, супремум не может быть меньше, чем модуль любого собственного числа, т. е. не может быть меньше спектрального радиуса.

В итоге получаем

$$\underline{\lim}_{k\to\infty}\sqrt[k]{\|S^k\|} \ge \rho(S).$$

Математический анализ говорит нам, что если верхний и нижний предел последовательности совпадают, то у такой последовательности есть предел, т. е. $\lim_{k\to\infty}\sqrt[k]{\|S^k\|} = \rho(S)$. По определению $R_k = -\ln\sqrt[k]{\|S^k\|} \to -\ln\rho(S)$ при $k\to\infty$, следовательно, $R_\infty = \rho(S,)$ что и требовалось доказать. □

Асимптотическая скорость сходимости стационарного итерационного метода, в отличие от средних скоростей сходимости, не

зависит от выбора матричной нормы, так как корень k-й степени из любой константы (в том числе и константы эквивалентности) в пределе даёт единицу. Обычно полагают, что метод с большей асимптотической скоростью «лучше» метода с меньшей асимптотической скоростью сходимости, но такое мнение не всегда оправдано, поскольку определить количество итераций, необходимое для уменьшения начальной ошибки в ε^{-1} раз в конкретной норме, зная только R_∞ и не зная $R_k\ \forall k$, невозможно. Удобным частным случаем является ситуация, когда $R_\infty = R_k\ \forall k$, однако подобное редко встречается на практике.

11.10.2022

Лекция №6

Итак, мы рассмотрели простой способ построения итерационного метода, а именно стационарный итерационный метод. Далее определим, каким способом можно выбирать матрицу H в стационарном итерационном методе и проанализируем, к каким результатам этот выбор приводит.

Один из общих подходов к построению итерационного метода для решения системы линейных алгебраических уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ состоит в том, чтобы представить матрицу A в виде $A = B - C$, переписать исходную систему в виде $B\vec{\mathbf{x}} = C\vec{\mathbf{x}} + \vec{\mathbf{b}}$ и определить очередное приближение к точному решению по известному приближению на предыдущем шаг из уравнения $B\vec{\mathbf{x}}^{k+1} = C\vec{\mathbf{x}}^{k} + \vec{\mathbf{b}}$. За этим подходом стоит уже хорошо знакомая нам идея — заменить матрицу A, которую мы затрудняемся обратить, на матрицу B, обратную к которой мы можем посчитать достаточно легко. Непосредственная проверка показывает, что указанный выше итерационный метод можно переписать в каноническом виде $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - B^{-1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}})$.

Метод Якоби

Пусть $D \equiv \operatorname{diag} A \equiv \operatorname{diag}\{a_{11}, \ldots, a_{nn}\}$ — диагональная матрица с потенциально ненулевыми элементами только на главной диагонали и равными элементам главной диагонали матрицы A. Тогда стационарный итерационный метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - D^{-1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}})$$

называется *методом Якоби* для решения системы линейных алгебраических уравнений. Поскольку диагональную матрицу обращать легко, то метод реализуется довольно просто. Однако для этого на диагонали матрицы системы не должно быть нулей. Рассмотрим признаки сходимости метода Якоби.

Теорема 18. *Если матрица системы имеет строгое диагональное преобладание по строкам, т. е.*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \geq 0 \ \forall i = 1, \ldots, n,$$

то метод Якоби сходится.

Доказательство. Вычислим i-ю строку матрицы шага для ошибки $S = E - D^{-1}A$:

$$\left[\frac{a_{i,1}}{a_{i,i}}, \ldots, \frac{a_{i,i-1}}{a_{i,i}}, 0, \frac{a_{i,i+1}}{a_{i.i}}, \ldots, \frac{a_{i.n}}{a_{i,i}}\right].$$

Обратим внимание, что $a_{ii} \neq 0, i = 1, \ldots, n$ в силу СТРОГОГО диагонального преобладания по столбцам. По условию теоремы у нас имеется строгое диагональное преобладание по строкам, т. е.

$$|\frac{a_{i,1}}{a_{i,i}}| + \ldots + |\frac{a_{i,i-1}}{a_{i,i}}| + |\frac{a_{i,i+1}}{a_{i.i}}| + \ldots + |\frac{a_{i.n}}{a_{i,i}}| < 1,$$

ввиду чего по определению бесконечной нормы матрицы мы имеем $\|S\|_\infty < 1$. Следовательно, по теореме 13 у нас выполнено достаточное условие сходимости стационарного итерационного метода, и метод Якоби сходится. □

Теорема 19. *Если матрица системы имеет строгое диагональное преобладание по столбцам, т. е.*

$$|a_{jj}| > \sum_{i=1, i\neq j}^{n} |a_{ij}| \geq 0 \ \forall j = 1, \ldots, n,$$

то метод Якоби сходится.

Доказательство. Рассмотрим j-й столбец матрицы шага для невязки $T = E - D^{-1}A$:

$$\left[\frac{a_{1,j}}{a_{j,j}}, \ldots, \frac{a_{j-1,j}}{a_{j,j}}, 0, \frac{a_{j+1,j}}{a_{j,j}}, \ldots, \frac{a_{n,j}}{a_{j,j}}\right]^*.$$

Обратим внимание, что $a_{jj} \neq 0, j = 1, \ldots, n$ в силу СТРОГОГО диагонального преобладания по столбцам. По условию теоремы у нас имеется строгое диагональное преобладание по столбцам, т. е.

$$|\frac{a_{1,j}}{a_{j,j}}| + \ldots + |\frac{a_{j-1,j}}{a_{j,j}}| + |\frac{a_{j+1,j}}{a_{j,j}}| + \ldots + |\frac{a_{n,j}}{a_{j,j}}| < 1,$$

ввиду чего по определению первой нормы матрицы мы имеем $\|T\|_1 < 1$. Отсюда по теореме 14 у нас выполнено достаточное условие сходимости стационарного итерационного метода, и метод Якоби сходится. □

Теорема 20. *Если $A = A^*$ и $D > 0$, то метод Якоби сходится тогда и только тогда, когда $A > 0$ и $2D - A > 0$.*

Доказательство. Поскольку $D > 0$, то на диагонали стоят положительные числа, что проверяется по определению положительной определённости матрицы с помощью единичных ортов $\vec{\mathbf{e}}^i$, $i = 1, \ldots, n$: $0 < (D\vec{\mathbf{e}}^i, \vec{\mathbf{e}}^i) = d_{ii}$. Отсюда следует, что матрица D обратима и мы можем извлечь из неё квадратный корень. Рассмотрим матрицу шага для ошибки метода Якоби $S = E - D^{-1}A$. Заметим, что матрица $D^{1/2}SD^{-1/2} = E - D^{-1/2}AD^{-1/2}$ – самосопряжённая, так как матрица A — самосопряжённая. Как известно из курса алгебры, это означает, что у матрицы все собственные числа вещественные и есть базис из собственных векторов. Поскольку матрица S подобна ей, то у неё также все собственные числа вещественные. При этом мы имеем $\lambda(S) = 1 - \lambda(D^{-1}A)$.

Докажем вспомогательное утверждение о том, что $\rho(S) < 1 \Leftrightarrow \lambda(D^{-1}A) \in (0, 2)$.

Во-первых, так как матрицы $D^{-1}A$ (несамосопряжённая) и $D^{-1/2}AD^{-1/2}$ (самосопряжённая в силу условия теоремы) подобны, они обладают одинаковым набором вещественных собственных чисел, т. е.

$$\lambda(D^{-1}A) = \lambda(D^{-1/2}AD^{-1/2}).$$

Далее, если $A > 0$, т. е.

$$\forall \vec{\mathbf{u}}, \vec{\mathbf{v}} \neq \vec{\mathbf{0}},\ 0 < (A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = (AD^{-1/2}[D^{1/2}\vec{\mathbf{u}}], D^{-1/2}[D^{1/2}\vec{\mathbf{u}}]) \equiv$$
$$\equiv (AD^{-1/2}\vec{\mathbf{v}}, D^{-1/2}\vec{\mathbf{v}}) = (D^{-1/2}AD^{-1/2}\vec{\mathbf{v}}, \vec{\mathbf{v}}),$$

то это неравенство верно и для собственных векторов, которые не бывают нулевыми по определению. Пусть $D^{1/2}\vec{\mathbf{u}} \equiv \vec{\mathbf{v}}$ — собственный вектор, соответствующий некоторому собственному числу $\lambda(D^{-1/2}AD^{-1/2})$, тогда $0 < (A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = \lambda(D^{-1/2}AD^{-1/2})(\vec{\mathbf{v}}, \vec{\mathbf{v}})$. Так как $(\vec{\mathbf{v}}, \vec{\mathbf{v}}) > 0$, то $\lambda(D^{-1/2}AD^{-1/2}) > 0$. Обратите внимание, что даже в комплексном случае собственное число будет вещественным и положительным! Так как указанное утверждение верно для любого собственного числа, то из $A > 0$ следует, что любое собственное число $\lambda(D^{-1}A) > 0$.

Далее заметим, что для любой самосопряжённой матрицы $B = B^*$ имеет место $(B\vec{\mathbf{u}}, \vec{\mathbf{u}}) \geq \lambda_{\min}(B)(\vec{\mathbf{u}}, \vec{\mathbf{u}})$. Данное утверждение основано на том, что у самосопряжённой матрицы есть ортонормированный базис из собственных векторов, по которому мы можем разложить любой вектор $\vec{\mathbf{u}}$. Подставив это разложение в скалярное произведение и воспользовавшись свойством ортонормированности векторов базиса, мы получим требуемое утверждение. Убедитесь, что вы можете провести указанные выкладки самостоятельно. Далее,

$$\forall \vec{\mathbf{v}}, \vec{\mathbf{u}} \neq \vec{\mathbf{0}},\ \lambda_{\min}(D^{-1/2}AD^{-1/2})(\vec{\mathbf{v}}, \vec{\mathbf{v}}) \leq (D^{-1/2}AD^{-1/2}\vec{\mathbf{v}}, \vec{\mathbf{v}}) =$$
$$(AD^{-1/2}\vec{\mathbf{v}}, D^{-1/2}\vec{\mathbf{v}}) = (AD^{-1/2}D^{1/2}\vec{\mathbf{u}}, D^{-1/2}D^{1/2}\vec{\mathbf{u}}) = (A\vec{\mathbf{u}}, \vec{\mathbf{u}}).$$

В силу подобия отсюда следует, что если $\lambda(D^{-1}A) > 0$, то по определению положительной определённости $A > 0$. Таким образом, мы установили, что $\lambda(D^{-1}A) > 0 \Leftrightarrow A > 0$.

Во-вторых, так как матрицы $2E - D^{-1}A$ (несамосопряжённая) и $2E - D^{-1/2}AD^{-1/2}$ (самосопряжённая в силу условия теоремы) подобны, то у них одинаковый набор вещественных собственных чисел, т. е.

$$\lambda(2E - D^{-1}A) = \lambda(2E - D^{-1/2}AD^{-1/2}).$$

Заметим также, что $2E - D^{-1}A = D^{-1}(2D - A)$. А потому

$$\lambda_{\min}(D^{-1}(2D-A))(\vec{\mathbf{u}},\vec{\mathbf{u}}) \le (D^{-1}(2D-A)\vec{\mathbf{u}},\vec{\mathbf{u}}) =$$
$$= (D^{-1/2}D^{-1/2}(2D-A)D^{1/2}D^{-1/2}\vec{\mathbf{u}},\vec{\mathbf{u}}) =$$
$$= (D^{-1/2}(2D-A)D^{1/2}[D^{-1/2}\vec{\mathbf{u}}],[D^{-1/2}\vec{\mathbf{u}}]) \equiv$$
$$\equiv (D^{-1/2}(2D-A)D^{1/2}\vec{\mathbf{v}},\vec{\mathbf{v}}) \ \forall \vec{\mathbf{v}},\vec{\mathbf{u}} \neq \vec{\mathbf{0}}.$$

Аналогично приведённому выше примеру устанавливаем из только что написанной формулы, что $\lambda(2E-D^{-1}A) > 0 \Leftrightarrow D^{-1/2}(2D-A)D^{1/2} > 0$. Поскольку матрицы $2D-A$ и $D^{-1/2}(2D-A)D^{1/2}$ подобны, то

$$((2D-A)\vec{\mathbf{u}},\vec{\mathbf{u}}) \ge \lambda_{\min}(D^{-1/2}(2D-A)D^{1/2})(\vec{\mathbf{u}},\vec{\mathbf{u}}).$$

Отсюда уже опробованным способом мы получаем, что

$$(2D-A) > 0 \Leftrightarrow D^{-1/2}(2D-A)D^{1/2} > 0,$$

следовательно,

$$(2D-A) > 0 \Leftrightarrow \lambda(D^{-1}(2D-A)) > 0.$$

Далее замечаем, что неравенство

$$0 < \lambda(D^{-1}(2D-A)) = \lambda(2E - D^{-1}A) = 2 - \lambda(D^{-1}A)$$

означает, что $\lambda(D^{-1}A) < 2$. Таким образом, мы доказали утверждение $\lambda(D^{-1}A) < 2 \Leftrightarrow 2D - A > 0$.

Так как $\rho(S) < 1 \Leftrightarrow \lambda(D^{-1}A) \in (0,2)$, то, по теореме 15, метод Якоби сходится только при указанных в теореме условиях, что завершает доказательство нашей теоремы. □

Обратите внимание, что мы получили критерий сходимости метода Якоби только в случае самосопряжённой матрицы системы A. Для несамосопряжённых матриц у нас есть только два признака сходимости в виде теорем 18 и 19, но нет критерия сходимости.

Метод Зейделя/Гаусса – Зейделя/Некрасова

Представим матрицу системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ в виде $A = -L + D - R$, где $D \equiv \operatorname{diag} A \equiv \operatorname{diag}\{a_{11}, \ldots, a_{nn}\}$, как и прежде, диагональная матрица с потенциально ненулевыми элементами только на главной диагонали и равными элементам главной диагонали матрицы A, а

$$L = -\begin{bmatrix} 0 & & \ldots & 0 \\ a_{2,1} & 0 & & \\ \vdots & \ddots & \ddots & \vdots \\ a_{n,1} & \ldots & a_{n,n-1} & 0 \end{bmatrix} R = -\begin{bmatrix} 0 & a_{1,2} & \ldots & a_{1,n} \\ \vdots & \ddots & \ddots & \vdots \\ & & 0 & a_{n-1,n} \\ 0 & \ldots & & 0 \end{bmatrix},$$

т. е. строго нижний и строго верхний треугольник матрицы системы, взятые с обратным знаком.

Обратный знак берётся из соображений, что в ряде примеров все ненулевые элементы вне главной диагонали матрицы A имеют отрицательные знаки, следовательно, все элементы матриц D, L, R будут неотрицательными при таком выборе разложения матрицы A в виде суммы.

Мы будем называть *методом Зейделя* для решения системы линейных алгебраических уравнений стационарный итерационный процесс вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - (D - L)^{-1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}).$$

Этот метод определён корректно, если нижнетреугольная матрица $D - L$ невырождена.

Непосредственные вычисления показывают, что матрица шага для ошибки метода Зейделя может быть представлена в виде $S = E - (D - L)^{-1}A = E - (A + R)^{-1}A = (A + R)^{-1}R$.

Теорема 21. *Если $A = A^*$ и $D > 0$, то метод Зейделя сходится тогда и только тогда, когда $A > 0$.*

Доказательство. Заметим, что в силу условия теоремы матрица D невырождена, следовательно, и нижнетреугольная матрица $D - L$ также невырождена, т. е. метод Зейделя определён корректно.

Необходимость. Пусть метод Зейделя сходится, тогда по теореме 15 $\rho(S) \equiv \rho(E - (D - L)^{-1}A) < 1$. Воспользуемся методом «от противного». Предположим, что матрица системы A не положительно определена. Тогда существует ненулевой вектор $\vec{\mathbf{u}}$ такой, что $(A\vec{\mathbf{u}}, \vec{\mathbf{u}}) < 0$. Так как по условию теоремы матрица A — самосопряжённая, то, как известно из алгебры, у неё есть полный набор вещественных чисел и ортонормированных собственных векторов.

В процессе доказательства теоремы 20 мы уже получали вспомогательное утверждение о том, что $(A\vec{\mathbf{v}}, \vec{\mathbf{v}}) \geq \lambda_{\min}(A)(\vec{\mathbf{v}}, \vec{\mathbf{v}})\ \forall \vec{\mathbf{v}}$. Отсюда следует, что минимальное собственное значение матрицы A $\lambda_{\min}(A) < 0$, так как, во-первых, нулевого собственного значения у матрицы A не может быть (мы предположили в начале этого раздела, что мы решаем только невырожденные системы уравнений), а во-вторых, мы предположили, что матрица A не положительно определена. Значит, существует собственный вектор $\vec{\mathbf{u}} \neq \vec{\mathbf{0}}$, соответствующий отрицательному минимальному собственному значению матрицы A, т. е. $A\vec{\mathbf{u}} = \lambda_{\min}(A)\vec{\mathbf{u}}$.

Пусть начальное приближение оказалось таковым, что начальная ошибка итерационного метода $\vec{\mathbf{z}}^0 = \vec{\mathbf{u}}$. Тогда

$$(A\vec{\mathbf{z}}^{k+1}, \vec{\mathbf{z}}^{k+1}) = (A(E - (A + R)^{-1}A)\vec{\mathbf{z}}^k, E - (A + R)^{-1})A\vec{\mathbf{z}}^k) \equiv$$

$$\equiv (A\vec{\mathbf{z}}^k - A\vec{\mathbf{y}}^k, \vec{\mathbf{z}}^k - \vec{\mathbf{y}}^k).$$

Здесь мы положили $\vec{\mathbf{y}}^k \equiv (A + R)^{-1}A\vec{\mathbf{z}}^k$. Заметим, что если $\vec{\mathbf{z}}^k \neq \vec{\mathbf{0}}$, то $\vec{\mathbf{y}}^k \neq \vec{\mathbf{0}}$ также в силу невырожденности матрицы A и, как мы заметили в начале доказательства, невырожденности матрицы $D - L = A + R$. Далее,

$$(A\vec{\mathbf{z}}^k - A\vec{\mathbf{y}}^k, \vec{\mathbf{z}}^k - \vec{\mathbf{y}}^k) =$$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - (A\vec{\mathbf{z}}^k, \vec{\mathbf{y}}^k) - (A\vec{\mathbf{y}}^k, \vec{\mathbf{z}}^k) + (A\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) =$$

воспользуемся свойством самосопряжённости матрицы A и связью между векторами $\vec{\mathbf{z}}^k$ и $\vec{\mathbf{y}}^k$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - ((A+R)\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) - (\vec{\mathbf{y}}^k, (A+R)\vec{\mathbf{y}}^k) + (A\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) =$$

так как для самосопряжённой матрицы верно $(A+R)^* = A + L$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - ((A+R)\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) - ((A+L)\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) + (A\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) =$$

вспоминаем представление матрица A в виде суммы диагонали и двух строго треугольных матриц и то, что вектор $\vec{\mathbf{y}}^k \neq \vec{\mathbf{0}}$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - (D\vec{\mathbf{y}}^k, \vec{\mathbf{y}}^k) < (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) < \ldots < (A\vec{\mathbf{z}}^0, \vec{\mathbf{z}}^0) =$$

$$= \lambda_{\min}(A)(\vec{\mathbf{z}}^0, \vec{\mathbf{z}}^0) \equiv \text{const} < 0.$$

Это означает, что при $k \to \infty$ $\vec{\mathbf{z}}^k$ НЕ стремится к нулевому вектору, что, в свою очередь, ведёт к тому, что никакая норма вектора не стремится к нулю, т. е. $\lim_{k\to\infty}\|\vec{\mathbf{z}}^k\| \geq \text{const} \neq 0$. По определению метод НЕ сходится, что противоречит предположению, и необходимость доказана.

Достаточность. Пусть $A > 0$. Рассмотрим произвольное собственное значение матрицы шага для ошибки S $\lambda(S)$. Заметим, что в силу несамосопряжённости матрицы S собственные числа не обязаны быть вещественными, а также не гарантировано существование базиса из собственных векторов. Но нам этого и не потребуется. Пусть $\vec{\mathbf{u}}$ — собственный вектор, соответствующий этому собственному числу — такой, что $\|\vec{\mathbf{u}}\|_2 = 1$. Тогда $\lambda(S)\vec{\mathbf{u}} = S\vec{\mathbf{u}} = (A+R)^{-1}A\vec{\mathbf{u}}$. Умножим это равенство скалярно на $\vec{\mathbf{u}}$ и преобразуем к виду

$$(R\vec{\mathbf{u}}, \vec{\mathbf{u}}) = \lambda(S)\,[(A\vec{\mathbf{u}}, \vec{\mathbf{u}}) + (R\vec{\mathbf{u}}, \vec{\mathbf{u}})].$$

Пусть $a \equiv (A\vec{\mathbf{u}}, \vec{\mathbf{u}})$. Заметим, что в силу самосопряжённости матрицы A величина $a \in \mathbb{R}$. Так как матрица R не самосопряжена,

то $(R\vec{\mathbf{u}}, \vec{\mathbf{u}}) = \nu + i\mu$ для некоторых $\nu, \mu \in \mathbb{R}$. i здесь обозначает мнимую единицу. Таким образом, наше равенство может быть переписано в виде

$$\nu + i\mu = \lambda(S)\left[a + \nu + i\mu\right].$$

Непосредственные вычисления приводят нас к заключению, что

$$|\lambda(S)|^2 = \frac{\nu^2 + \mu^2}{a(a+2\nu) + \nu^2 + \mu^2}.$$

Отсюда незамедлительно следует, что $|\lambda(S)| < 1$, если $a(a+2\nu) > 0$.

Так как $A > 0$ по предположению, то, во-первых, $(A\vec{\mathbf{u}}, \vec{\mathbf{u}}) \geq \lambda_{\min(A)}(\vec{\mathbf{u}}, \vec{\mathbf{u}}) = \lambda_{\min}(A) > 0$. Здесь мы в очередной раз воспользовались тем, что у самосопряжённой положительно определённой матрицы все собственные числа вещественные и положительные, а также тем, что мы нормировали вектор $\vec{\mathbf{u}}$ в евклидовой норме. Обратите внимание, что вектор $\vec{\mathbf{u}}$ — собственный для S, но он не обязан быть собственным для A.

А во-вторых,

$$a = (A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = (D\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (R\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (L\vec{\mathbf{u}}, \vec{\mathbf{u}}) =$$

в силу самосопряжённости матрицы A для неё верно $R^* = L$

$$= (A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = (D\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (R\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (R^*\vec{\mathbf{u}}, \vec{\mathbf{u}}) =$$

по определению скалярного произведения в комплексном случае (здесь $\bar{\cdot}$ обозначает комплексное сопряжение числа)

$$= (A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = (D\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (R\vec{\mathbf{u}}, \vec{\mathbf{u}}) - (R\bar{\vec{\mathbf{u}}}, \vec{\mathbf{u}}) = d - 2\nu$$

для некоторого вещественного положительного числа d. d вещественное, так как матрица D — самосопряжена в силу самосопряжённости матрицы A и $d > 0$ в силу условия теоремы для матрицы $D > 0$. Таким образом, мы показали, что $a + 2\nu = d > 0$,

т. е. $|\lambda(S)| < 1$. В силу произвольности выбора $\lambda(S)$ это означает, что $\rho(S) < 1$, т. е. по теореме 15 метод Зейделя сходится. Достаточность доказана, а с ней доказана и вся теорема. $\square$

Заметим, что получить критерий сходимости метода Зейделя нам удалось только в случае самосопряжённой матрицы.

Заметим также, что из доказательства достаточности в теореме 21 следует, что

$$\nu^2 + \mu^2 = |(R\vec{\mathbf{u}}, \vec{\mathbf{u}})|^2 \le \|R\|_2^2 \|\vec{\mathbf{u}}\|_2^2 = \|R\|_2^2 \equiv \rho(R^*R).$$

Помимо этого можно заметить, что

$$d \equiv (D\vec{\mathbf{u}}, \vec{\mathbf{u}}) \ge \lambda_{\min}(D)(\vec{\mathbf{u}}, \vec{\mathbf{u}}) = (A\vec{\mathbf{e}}^{\,i}, \vec{\mathbf{e}}^{\,i}) \ge \lambda_{\min}(A).$$

Здесь $\vec{\mathbf{e}}^{\,i}$ — единичный орт, соответствующий позиции минимального диагонального элемента матрицы D. Таким образом, у нас получилось

$$|\lambda(S)|^2 = \frac{\nu^2 + \mu^2}{a(a + 2\nu) + \nu^2 + \mu^2} = \frac{\nu^2 + \mu^2}{a \cdot d + \nu^2 + \mu^2} \le$$

$$\le \frac{\nu^2 + \mu^2}{\lambda_{\min}^2(A) + \nu^2 + \mu^2} \le \max_{0 < \phi \le \rho(R^*R)} \frac{\phi}{\lambda_{\min}^2(A) + \phi} =$$

$$= \frac{\rho(R^*R)}{\lambda_{\min}^2(A) + \rho(R^*R)}.$$

В итоге мы пришли к следующему результату:

$$\rho(S) \le \sqrt{\frac{\rho(R^*R)}{\lambda_{\min}^2(A) + \rho(R^*R)}}.$$

Чем больше минимальное собственное значение матрицы A относительно максимального по модулю собственного значения матрицы R^*R, тем меньше спектральный радиус матрицы шага для ошибки и тем выше асимптотическая скорость сходимости метода Зейделя.

Лекция №7

Итак, мы рассмотрели простой способ построения стационарных итерационных методов на основе разложения матрицы в сумму матриц. Теперь посмотрим на то, как можно расширить диапазон итерационных методов за счёт средств математического и функционального анализа.

Второй из общих подходов к построению итерационного метода для решения системы линейных алгебраических уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ состоит в организации процесса таким образом, чтобы ошибка строго убывала на каждом шаге, т. е. $\|\vec{\mathbf{z}}^{k+1}\| < \|\vec{\mathbf{z}}^{k}\|$ для некоторой нормы. В терминах функционального анализа мы хотим, чтобы *функционал ошибки* $f(\vec{\mathbf{x}}^{k}) \equiv \|\vec{\mathbf{x}}^{k} - \vec{\mathbf{x}}\|$ строго убывал. Сформулируем условия, при которых мы можем получить сходимость итерационного метода.

Теорема 22 (о функционале ошибки)**.** *Если* $f(\vec{\mathbf{x}}^{k+1}) \equiv \|\vec{\mathbf{z}}^{k+1}\| < f(\vec{\mathbf{x}}^{k}) \equiv \|\vec{\mathbf{z}}^{k}\|$ $\forall \vec{\mathbf{z}}^{k} \neq \vec{\mathbf{0}}$, *и оператор шага для ошибки* $S : \mathbb{R}^n \to \mathbb{R}^n$, *действующий по формуле* $\vec{\mathbf{z}}^{k+1} = S(\vec{\mathbf{z}}^{k})$, *непрерывен всюду, кроме, быть может, нуля, то* $\|\vec{\mathbf{z}}^{k}\| \to 0$ *при* $k \to \infty$.

Доказательство. Последовательность норм ошибки ограничена снизу по аксиоме нормы, и по условию теоремы имеем

$$0 \leq \|\vec{\mathbf{z}}^{k+1}\| < \|\vec{\mathbf{z}}^{k}\|.$$

Математический анализ говорит нам о том, что у этой последовательности обязательно есть предел, т. е. $\|\vec{\mathbf{z}}^{k}\| \to \alpha$ при $k \to \infty$. При этом мы можем утверждать, что предел последовательности удовлетворяет условию $\alpha \geq 0$.

Воспользуемся методом «от противного». Предположим, что предел не равен нулю, т. е. $\alpha > 0$. Так как наша последовательность строго монотонно убывающая, то, как следует из матема-

тического анализа, мы можем выбрать сходящуюся подпоследовательность $\vec{\mathbf{z}}^{k_m}$ — такую, что $\vec{\mathbf{z}}^{k_m} \to \vec{\mathbf{z}}$ и при этом $\|\vec{\mathbf{z}}\| = \alpha$.

Столь сложная конструкция нам потребовалось из-за того, что последовательность $\vec{\mathbf{z}}^k$ — не числовая, а векторная, и потому для последовательности вполне возможен вариант, когда вектора в ней просто вращаются (т. е. сохраняют норму), но при этом не сходятся ни к какому вектору. Поэтому последовательность в смысле норм сходится к пределу, а последовательность в смысле векторов не имеет предела. Например, последовательность

$$\vec{\mathbf{z}}^k = \begin{cases} \overrightarrow{(1,0)}, & \text{если } k \text{ чётное} \\ \overrightarrow{(0,1)}, & \text{если } k \text{ нечётное} \end{cases}$$

даёт пример сказанного. Очевидно, что $\|\vec{\mathbf{z}}^k\|_2 = 1$, следовательно, $\lim_{k\to\infty}\|\vec{\mathbf{z}}^k\| = 1$, но при этом $\lim_{k\to\infty} \vec{\mathbf{z}}^k$ не существует. Однако сходящуюся подпоследовательность выбрать можно, так как для каждой компоненты вектора мы имеем ограниченную последовательность. Неограниченной она быть не может, так как тогда бы норма вектора росла, что противоречило бы утверждению о существовании предела для норм векторов из последовательности.

Итак, поскольку $\|\vec{\mathbf{z}}\| = \alpha \neq 0$, то по предположению теоремы $\|S(\vec{\mathbf{z}})\| < \|\vec{\mathbf{z}}\|$, так как $\vec{\mathbf{z}} \neq \vec{\mathbf{0}}$. При этом мы имеем $S(\vec{\mathbf{z}}^{k_m}) \to S(\vec{\mathbf{z}})$ по построению. Выполняя предельный переход в соотношении $\|\vec{\mathbf{z}}^{k_m+1}\| = \|S(\vec{\mathbf{z}}^{k_m})\| < \|\vec{\mathbf{z}}^{k_m}\|$ при $m \to \infty$ с использованием непрерывности оператора S, получаем $\alpha = \|S(\vec{\mathbf{z}})\| < \|\vec{\mathbf{z}}\| = \alpha$, что демонстрирует очевидное противоречие. Таким образом, получаем $\|\vec{\mathbf{z}}^k\| \to 0$ при $k \to \infty$, и теорема доказана. □

Обратите внимание, что мы использовали в теореме более общий термин «оператор S», а не «матрица S». Однако алгебра нам говорит, что линейный оператор можно представить в виде матрицы, если задать базис, чем мы не раз воспользуемся в дальнейшем.

В теореме 22 можно использовать разные нормы, однако в большинстве случаев естественным будет использовать норму, связанную с оператором задачи. Как уже было сказано, оператор представим в виде матрицы в базисе, а потому можно использовать норму вида $\|\vec{\mathbf{u}}\|_C \equiv \sqrt{(C\vec{\mathbf{u}}, \vec{\mathbf{u}})}$. Убедитесь, что вы можете доказать, что если $C = C^* > 0$, то $(\vec{\mathbf{u}}, \vec{\mathbf{v}})_C \equiv (C\vec{\mathbf{u}}, \vec{\mathbf{v}})$ — скалярное произведение и $\|\vec{\mathbf{u}}\|_C \equiv \sqrt{(\vec{\mathbf{u}}, \vec{\mathbf{u}})_C}$ — норма. Самая сложная часть в доказательстве — это доказательство неравенства треугольника. Для этого потребуется прибегнуть к приёму, связанному с использованием вектора $\vec{\mathbf{u}} + \mu\vec{\mathbf{v}}$ с некоторым скаляром μ. По аксиоме скалярного произведения $(\vec{\mathbf{u}} + \mu\vec{\mathbf{v}}, \vec{\mathbf{u}} + \mu\vec{\mathbf{v}})_C \geq 0$ при любых значениях μ, следовательно, у квадратного уравнения относительно μ дискриминант должен быть неположителен, что и приводит к требуемому неравенству $|(\vec{\mathbf{u}}, \vec{\mathbf{u}})_C| \leq \|u_C\|\|v_C\|$, которое вам хорошо известно из курса математического анализа под именем неравенства Коши — Буняковского.

Метод полной релаксации

Применим рассмотренную выше идею в самом простом случае. Пусть для решения системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ с самосопряжённой, положительно определённой матрицей A мы будем строить очередное приближение к точному решению за n шагов, а именно

$$\vec{\mathbf{x}}^{k+i/n} = \vec{\mathbf{x}}^{k+(i-1)/n} - \alpha_{ki}\vec{\mathbf{e}}^i = \begin{bmatrix} \mathbf{x_1^{k+1}} \\ \vdots \\ \mathbf{x_{i-1}^{k+1}} \\ \mathbf{x_i^k - \alpha_{ki}} \\ \mathbf{x_{i+1}^k} \\ \vdots \\ \mathbf{x_n^k} \end{bmatrix},$$

где параметр α_{ki} выбирается из условия минимизации $\|\vec{\mathbf{z}}^{k+i/n}\|_A$. Этот метод мы будем называть *методом полной релаксации.* Об-

ратите внимание, что на каждом шаге мы меняем только одну компоненту вектора приближённого решения, что делает процесс поиска минимума очень простым, однако формула становится сложной. Это приводит к тому, что на i-м шаге не меняются уже насчитанные на предыдущих шагах компоненты с номерами от 1 до $i-1$ вектора приближённого решения, а также компоненты с номерами от $i+1$ до n, которые мы будем пересчитывать на последующих шагах.

Теорема 23. *Для $A = A^* > 0$ метод полной релаксации сходится, и при этом параметр α_{ki} определяется по формуле*

$$\alpha_{ki} = \frac{r_i^{k+(i-1)/n}}{a_{i,i}} \equiv$$

$$\equiv \frac{a_{i,1}x_1^{k+1} + \ldots + a_{i,i-1}x_{i-1}^{k+1} + a_{i,i}x_i^k + \ldots + a_{i,n}x_n^k - b_i}{a_{i,i}}.$$

Доказательство. Так как $A = A^* > 0$, то

$$\|\vec{\mathbf{z}}^{k+i/n}\|_A^2 = (A\vec{\mathbf{z}}^{k+i/n}, \vec{\mathbf{z}}^{k+i/n}) =$$

$$= (A(\vec{\mathbf{z}}^{k+(i-1)/n} - \alpha_{ki}\vec{\mathbf{e}}^{\,i}), \vec{\mathbf{z}}^{k+(i-1)/n} - \alpha_{ki}\vec{\mathbf{e}}^{\,i}) =$$

в силу самосопряжённости матрицы A

$$= \|\vec{\mathbf{z}}^{k+(i-1)/n}\|_A^2 - 2\alpha_{ki}(A\vec{\mathbf{z}}^{k+(i-1)/n}, \vec{\mathbf{e}}^{\,i}) + \alpha_{ki}^2\|\vec{\mathbf{e}}^{\,i}\|_A^2 =$$

вычисляя скалярное определение и используя определение невязки

$$= \|\vec{\mathbf{z}}^{k+(i-1)/n}\|_A^2 - 2\alpha_{ki}r_i^{k+(i-1)/n} + \alpha_{ki}^2 a_{i,i} =$$

с помощью нехитрых алгебраических манипуляций

$$= \|\vec{\mathbf{z}}^{k+(i-1)/n}\|_A^2 - \frac{\left(r_i^{k+(i-1)/n}\right)^2 - \left(\alpha_{ki}a_{i,i} - r_i^{k+(i-1)/n}\right)^2}{a_{i,i}}.$$

Из последнего выражения незамедлительно получаем, что при $\alpha_{ki}a_{i,i} - r_i^{k+(i-1)/n} = 0$ у нас будет максимально возможное в

данной ситуации уменьшение ошибки (т. е. полная релаксация, отсюда и название метода). Тогда

$$\|\vec{\mathbf{z}}^{k+i/n}\|_A^2 = \|\vec{\mathbf{z}}^{k+(i-1)/n}\|_A^2 - \frac{\left(r_i^{k+(i-1)/n}\right)^2}{a_{i,i}}.$$

Таким образом, мы получаем

$$\|\vec{\mathbf{z}}^{k+1}\|_A^2 = \|\vec{\mathbf{z}}^{k}\|_A^2 - \frac{\left(r_1^{k}\right)^2}{a_{1,1}} - \ldots - \frac{\left(r_n^{k+(n-1)/n}\right)^2}{a_{n,n}} < \|\vec{\mathbf{z}}^{k}\|_A^2,$$

если хотя бы одна из компонент вектора невязки не равна нулю. В противном случае, $\vec{\mathbf{r}}^k = \vec{\mathbf{0}}$, что в силу невырожденности матрицы A ведёт к тому, что $\vec{\mathbf{z}}^k = A^{-1}\vec{\mathbf{r}}^k = \vec{\mathbf{0}}$, т. е. мы нашли точное решение системы $\vec{\mathbf{x}}$ на шаге k и дальнейшие итерации не требуются.

Обратите внимание, что мы можем посчитать скалярные произведения вида $(A\vec{\mathbf{e}}^i, \vec{\mathbf{e}}^i) = a_{i,i}$, и, в силу положительной определённости матрицы A, получаем, что $a_{i,i} > 0\ \forall i = 1, \ldots, n$. Если сказать проще, то у положительно определённой матрицы на диагонали стоят вещественные положительные величины, а потому на них можно делить.

Итак, мы получили строго убывающий функционал ошибки. Найдём оператор (матрицу) шага для ошибки, чтобы исследовать его непрерывность. Для i-й компоненты вектора $\vec{\mathbf{x}}^{k+1}$ мы получили представление

$$x_i^{k+1} = x_i^k - \alpha_{ki} = x_i^k - \frac{r_i^{k+(i-1)/n}}{a_{i,i}} = x_i^k - \frac{r_i^{k+(i-1)/n}}{a_{i,i}} \equiv$$

$$\equiv x_i^k - \frac{a_{i,1}x_1^{k+1} + \ldots + a_{i,i-1}x_{i-1}^{k+1} + a_{i,i}x_i^k + \ldots + a_{i,n}x_n^k - b_i}{a_{i,i}}.$$

Заметим теперь, что эту формулу можно переписать в матрично-векторном виде, а именно

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - D^{-1}(-L\vec{\mathbf{x}}^{k+1} + (D-R)\vec{\mathbf{x}}^k - \vec{\mathbf{b}}) =$$

$$= D^{-1}(L\vec{\mathbf{x}}^{k+1} + R\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}).$$

Приведём это выражение к каноническому виду стационарного итерационного метода и получим

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - (D - L)^{-1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}).$$

Как нетрудно видеть, это формула метода Зейделя. По теореме 21 метод Зейделя сходится при $A = A^* > 0$, следовательно, наша теорема доказана.

Мы также можем рассмотреть альтернативный подход к доказательству этой теоремы. Оператор шага для ошибки, представленный в матричном виде как матрица шага для ошибки $S = E - (D - L)^{-1}A$, является всюду непрерывным по определению, следовательно, по теореме 22 метод полной релаксации сходится. □

Мы только что получили двумя разными способами один и тот же метод — метод Зейделя или метод полной релаксации. Попробуем теперь получить другие итерационные методы на основе рассмотренного подхода.

Метод неполной релаксации

Пусть для решения системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ с самосопряжённой, положительно определённой матрицей A мы будем строить очередное приближение к точному решению за n шагов, а именно

$$\vec{\mathbf{x}}^{k+i/n} = \vec{\mathbf{x}}^{k+(i-1)/n} - \omega\alpha_{ki}\vec{\mathbf{e}}^{i},$$

где параметр α_{ki} выбирается как в методе полной релаксации. Этот метод мы будем называть *методом неполной релаксации.* Обратите внимание, что при $\omega = 1$ мы имеем метод полной релаксации. В остальных случаях ошибка будет убывать меньше,

чем в методе полной релаксации независимо от того, будет параметр ω больше или меньше единицы (это мы видели в процессе доказательства теоремы 23).

Теорема 24. *Для $A = A^* > 0$ метод неполной релаксации сходится при $\omega \in (0, 2)$.*

Доказательство. Доказательство теоремы практически полностью совпадает с доказательством теоремы 23, поэтому укажем лишь на основные отличия.

Во-первых,

$$\|\vec{\mathbf{z}}^{k+i/n}\|_A^2 = \|\vec{\mathbf{z}}^{k+(i-1)/n)}\|_A^2 - [1-(\omega-1)^2]\frac{\left(r_i^{k+(i-1)/n}\right)^2}{a_{i,i}} <$$

$$< \|\vec{\mathbf{z}}^{k+(i-1)/n}\|_A^2,$$

если $\omega \in (0, 2)$ и $r_i^{k+(i-1)/n} \neq 0$.

Во-вторых, для i-й компоненты вектора $\vec{\mathbf{x}}^{k+1}$ мы получаем представление

$$x_i^{k+1} = x_i^k - \omega\alpha_{ki} \equiv$$

$$\equiv x_i^k - \omega\frac{a_{i,1}x_1^{k+1} + \ldots + a_{i,i-1}x_{i-1}^{k+1} + a_{i,i}x_i^k + \ldots + a_{i,n}x_n^k - b_i}{a_{i,i}}.$$

И в-третьих, эту формулу можно переписать в матрично-векторном виде, а именно

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \omega D^{-1}(-L\vec{\mathbf{x}}^{k+1} + (D-R)\vec{\mathbf{x}}^k - \vec{\mathbf{b}}).$$

Приведём это выражение к каноническому виду стационарного итерационного метода и получим

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \omega(D-\omega L)^{-1}(A\vec{\mathbf{x}}^k - \vec{\mathbf{b}}).$$

Оператор шага для ошибки, представленный в матричном виде как матрица шага для ошибки $S = E - \omega(D - \omega L)^{-1}A$, является всюду непрерывным по определению, следовательно, по теореме 22 метод неполной релаксации сходится. $\square$

Оценка скорости сходимости методов релаксации в $\mathbb{R}^n$

Рассмотрим теперь, что можно получить для оценки скорости сходимости методов релаксации. Комплексный случай мы не рассматриваем в этом разделе, а потому оператор * означает простое транспонирование до конца этого параграфа. В этом случае для скалярного произведения имеем равенство $(\vec{\mathbf{u}}, \vec{\mathbf{v}}) = (\vec{\mathbf{v}}, \vec{\mathbf{u}})$ (без знака сопряжения, необходимого в комплексном случае).

Мы получили, что ошибка в методах релаксации монотонно убывает в норме $\|\cdot\|_A \equiv \sqrt{(A\cdot, \cdot)}$ по теореме 24. Напоминаем, что для того, чтобы $\|\cdot\|_A$ была нормой, требуется, чтобы $A = A^* > 0$. Далее, из канонического вида стационарного итерационного метода незамедлительно следует, что матрица шага для ошибки $S = E - \omega(D - \omega L)^{-1}A$. По определению подчинённой матричной нормы

$$\|S\|_A^2 = \sup_{\vec{\mathbf{u}} \neq \vec{\mathbf{0}}} \frac{(AS\vec{\mathbf{u}}, \vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})}.$$

Вспоминаем, что для подчинённой (которая также является и согласованной) матричной нормы верно неравенство $\|\vec{\mathbf{z}}^{k+1}\|_A \leq \|S\|_A \|\vec{\mathbf{z}}^k\|_A$. Получим теперь оценку для $\|S\|_A$, чтобы иметь оценку для скорости сходимости методов релаксации.

Теорема 25. *Для $A = A^* > 0$ метод неполной релаксации при $\omega \in (0, 2)$ сходится и при этом*

$$\|S\|_A^2 \leq g(\tau) \equiv \frac{1 - \tau\delta + \tau^2\delta\Delta}{1 + \tau\delta + \tau^2\delta\Delta} < 1 \ \forall \tau > 0,$$

где постоянные $\delta > 0$ и $\Delta > 0$ таковы, что

$$\delta(D\vec{\mathbf{v}}, \vec{\mathbf{v}}) \leq (A\vec{\mathbf{v}}, \vec{\mathbf{v}}), \ (R_1 D^{-1} R_1^* \vec{\mathbf{v}}, \vec{\mathbf{v}}) \leq \Delta(A\vec{\mathbf{v}}, \vec{\mathbf{v}}) \ \forall \vec{\mathbf{v}} \in \mathbb{R}^n (\mathbb{C}^n),$$

$$R_1 = \frac{1}{2}D - L.$$

Доказательство. Заметим, что указанные в теореме условия означают положительную полуопределённость матриц $A - \delta D$ и $\Delta A - R_1 D^{-1} R_1^*$. Непосредственная проверка показывает, что матрицы D и $R_1 D^{-1} R_1^*$ самосопряжены, следовательно, у них вещественный спектр, и вещественный спектр имеют подобные им несамосопряжённые матрицы $D^{-1}A$ и $A^{-1}R_1 D^{-1} R_1^*$. Убедитесь, что вы можете доказать подобие матриц через доказательство существования корня из матриц D (что просто) и A (что сложнее). В результате имеем $\delta \leq \lambda_{\min}(A^{-1}D)$ и $\Delta \geq \lambda_{\max}(A^{-1}R_1 D^{-1} R_1^*)$. Как уже говорилось ранее, получить точные значение собственных чисел — задача более сложная, чем решить систему уравнений, но получить оценку для минимального и максимального собственных чисел иногда бывает возможно без больших усилий.

Проведём вспомогательные преобразования. Так как

$$\omega(D - \omega L)^{-1} = \omega(D + \omega(R_1 - \frac{1}{2}D))^{-1} = \frac{2\omega}{2-\omega}(D + \frac{2\omega}{2-\omega}R_1)^{-1},$$

то

$$S = (E - \tau(D + \tau R_1)^{-1}A \equiv E - \tau B^{-1}A,\ \tau \equiv \frac{2\omega}{2-\omega} \in (0, \infty),$$

$$B \equiv D + \tau R_1.$$

Непосредственные вычисления дают нам

$$\|S\|_A^2 = \sup_{\vec{\mathbf{u}} \neq \vec{\mathbf{0}}} \left[1 - 2\tau \frac{(B^{-1}A\vec{\mathbf{u}}, A\vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})} + \tau^2 \frac{(AB^{-1}A\vec{\mathbf{u}}, A\vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})}\right].$$

Так как $\|\cdot\|_A$ порождена скалярным произведением, то формулы писать проще, если рассматривать эту норму в квадрате.

Далее, учитывая, что

$$\tau(A\vec{\mathbf{u}}, \vec{\mathbf{u}}) = \tau([R_1 + R_1^*]\vec{\mathbf{u}}, \vec{\mathbf{u}}) = 2\tau(R_1\vec{\mathbf{u}}, \vec{\mathbf{u}}) = 2((B - D)\vec{\mathbf{u}}, \vec{\mathbf{u}}),$$

то, во-первых, $R_1 > 0$ (положительно определена), а во-вторых,

$$\tau^2(AB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}}) = 2\tau(BB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}}) -$$

$$-2\tau(DB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}}) = 2\tau(B^{-1}A\vec{\mathbf{u}}, A\vec{\mathbf{u}}) - 2\tau(DB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}}).$$

Отсюда мы получаем

$$\|S\|_A^2 = \sup_{\vec{\mathbf{u}}\neq\vec{\mathbf{0}}} \left[1 - 2\tau\frac{(DB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})}\right] =$$

так как в конечномерном пространстве супремум достигается

$$= \max_{\vec{\mathbf{u}}\neq\vec{\mathbf{0}}} \left[1 - 2\tau\frac{(DB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})}\right] =$$

$$= 1 - 2\tau \min_{\vec{\mathbf{u}}\neq\vec{\mathbf{0}}} \frac{(DB^{-1}A\vec{\mathbf{u}}, B^{-1}A\vec{\mathbf{u}})}{(A\vec{\mathbf{u}}, \vec{\mathbf{u}})} =$$

$$= 1 - 2\tau \min_{\vec{\mathbf{u}}\neq\vec{\mathbf{0}}} \frac{(A^{1/2}(B^{-1})^*DB^{-1}A^{1/2}A^{1/2}\vec{\mathbf{u}}, A^{1/2}\vec{\mathbf{u}})}{(A^{1/2}\vec{\mathbf{u}}, A^{1/2}\vec{\mathbf{u}})} =$$

$A^{1/2}\vec{\mathbf{u}} \equiv \vec{\mathbf{v}}$

$$= 1 - 2\tau \min_{\vec{\mathbf{v}}\neq\vec{\mathbf{0}}} \frac{(A^{1/2}(B^{-1})^*DB^{-1}A^{1/2}\vec{\mathbf{v}}, \vec{\mathbf{v}})}{(\vec{\mathbf{v}}, \vec{\mathbf{v}})} \equiv 1 - 2\tau\gamma,$$

где $\gamma = \lambda_{\min}(A^{1/2}(B^{-1})^*DB^{-1}A^{1/2})$. Заметим, что минимальное собственное число матрицы $A^{1/2}(B^{-1})^*DB^{-1}A^{1/2}$ существует, поскольку матрица самосопряжённая (симметричная в этом параграфе), а потому у неё вещественный спектр. Более того, непосредственная проверка показывает, что эта матрица будет также положительно определена, а потому минимальное собственное число будет больше нуля.

Так как γ — собственное число, то существует вектор $\vec{\mathbf{y}} \neq \vec{\mathbf{0}}$ — такой, что $A^{1/2}(B^{-1})^*DB^{-1}A^{1/2}\vec{\mathbf{y}} = \gamma\vec{\mathbf{y}}$. Пусть $\vec{\mathbf{v}} = A^{-1/2}\vec{\mathbf{y}}$, тогда

$A\vec{\mathbf{v}} = \gamma BD^{-1}B^*\vec{\mathbf{v}}$. Заметим, что $BD^{-1}B^* = (D+\tau R_1)D^{-1}(D+\tau R_1^*) = D + \tau A + \tau^2 R_1 D^{-1} R_1^*$, а потому

$$\gamma = \frac{(A\vec{\mathbf{v}},\vec{\mathbf{v}})}{(D\vec{\mathbf{v}},\vec{\mathbf{v}}) + \tau(A\vec{\mathbf{v}},\vec{\mathbf{v}}) + \tau^2(R_1 D^{-1} R_1^*\vec{\mathbf{v}},\vec{\mathbf{v}})}.$$

Используя условия теоремы, мы легко получаем оценку

$$\gamma \geq \frac{1}{1/\delta + \tau + \tau^2\Delta} = \frac{\delta}{1+\tau\delta+\tau^2\Delta}.$$

Подставив полученную оценку для γ в оценку для $\|S\|_A^2$, получаем утверждение теоремы. □

Непосредственные вычисления дают нам оптимальное значение параметра τ_* в функции $g(\tau)$, а именно

$$\min_{\tau>0} g(\tau) = \frac{1-\sqrt{\delta/(4\Delta)}}{1+\sqrt{\delta/(4\Delta)}}$$

при $\tau_* = \frac{1}{\sqrt{\delta\Delta}}$. При этом оптимальное значение параметра $\omega_* = \frac{2}{1+2\sqrt{\delta\Delta}}$. Убедитесь, что вы можете провести необходимые выкладки и получить указанные величины.

Пример, демонстрирующий различие между методами полной и неполной релаксации

Пусть у нас имеется матрица

$$A = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix},$$

тогда

$$R_1 = \begin{bmatrix} 1 & 0 & & & 0 \\ -1 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 0 \\ 0 & & & -1 & 1 \end{bmatrix}.$$

Для указанной матрицы непосредственная проверка показывает, что $\delta = \lambda_{\min}(D^{-1}A) = 2\sin^2\frac{\pi}{2(n+1)} \approx \frac{\pi^2}{2(n+1)^2} < 1$. Здесь мы воспользовались разложением в ряд Маклорена из математического анализа. Далее, замечая, что справедливо равенство $A = 2R_1D^{-1}R_1^* + \mathrm{diag}\{1,0,\ldots,0\}$, мы получаем, что $0{,}5A \geq R_1D^{-1}R_1^*$ (в смысле положительной полуопределённости!) и $\Delta = 0{,}5$.

По теореме 25 мы получаем для метода полной релаксации $\omega = 1$ и $g(2) \approx 1 - \frac{2\pi^2}{(n+1)^2}$, $\|S\|_A \leq \sqrt{g(2)} \approx \sqrt{1 - \frac{\pi^2}{(n+1)^2}}$, следовательно, для того, чтобы ошибка уменьшилась в $1/\varepsilon$ раз, потребуется сделать $k(\varepsilon) = \frac{(n+1)^2}{\pi^2}\ln\frac{1}{\varepsilon}$ итераций. Здесь мы также активно использовали разложение в ряд Маклорена из математического анализа для получения приближённых значений вблизи нуля для сложных функций.

Аналогично, по теореме 25 мы получаем для метода неполной релаксации оптимальное $\tau_* = \frac{1}{\sqrt{\delta\Delta}}$, тогда $\omega_* = \frac{2}{1+\pi/(n+1)} > 1$ (поэтому метод называют *методом верхней релаксации* в отличие от метода нижней релаксации при $\omega < 1$) и $g(\tau_*) \approx 1 - \frac{\pi}{(n+1)}$, $\|S\|_A \leq \sqrt{g(\tau_*)} \approx \sqrt{1 - \frac{\pi}{(n+1)}}$, следовательно, для того, чтобы ошибка уменьшилась в $1/\varepsilon$ раз, нам потребуется сделать $k(\varepsilon) = \frac{(n+1)}{\pi}\ln\frac{1}{\varepsilon}$ итераций. Здесь мы вновь активно использовали разложение в ряд Маклорена из математического анализа для получения приближённых значений вблизи нуля для сложных функций.

Несмотря на то что метод полной релаксации (он же метод Зейделя) даёт максимально возможное уменьшение ошибки на

каждом шаге, метод неполной релаксации с оптимальным параметром может быть в $(n+1)/(2\pi)$ раз быстрее, что для матриц большого размера с n намного больше 1 будет давать заметный выигрыш по скорости. Это основная причина, по которой на практике используется исключительно метод неполной релаксации с $\omega > 1$ (метод верхней релаксации).

Лекция №8

Градиентные методы

Теперь посмотрим на то, как можно расширить диапазон итерационных методов за счёт нестационарных итерационных методов. Для решения системы линейных алгебраических уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ сформулируем итерационный метод в виде $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k + \alpha\vec{\mathbf{y}}$ и будем искать параметры из условия минимизации ошибки $\|\vec{\mathbf{z}}^{k+1}\|^2 = \min_\alpha \|\vec{\mathbf{z}}^k + \alpha\vec{\mathbf{y}}\|^2$.

Пусть $A = A^* > 0$ и $f(\vec{\mathbf{z}}) = \|\vec{\mathbf{z}}\|_A^2 \equiv (A\vec{\mathbf{z}}, \vec{\mathbf{z}})$. Воспользуемся разложением в ряд Тэйлора:

$$f(\vec{\mathbf{z}}^{k+1}) = f(\vec{\mathbf{z}}^k) + \frac{df(\vec{\mathbf{z}}^k)}{d\alpha}\alpha + O(\alpha^2) =$$

$$= f(\vec{\mathbf{z}}^k) + \left[\frac{\partial f(\vec{\mathbf{z}}^k)}{\partial z_1}y_1 + \ldots + \frac{\partial f(\vec{\mathbf{z}}^k)}{\partial z_n}y_n\right]\alpha + O(\alpha^2) =$$

$$= f(\vec{\mathbf{z}}^k) + (\overrightarrow{\nabla \mathbf{f}}, \vec{\mathbf{y}})\alpha + O(\alpha^2) = f(\vec{\mathbf{z}}^k) + 2\alpha(A\vec{\mathbf{z}}^k, \vec{\mathbf{y}}) + \alpha^2\|\vec{\mathbf{y}}\|_A \approx$$

$$\approx f(\vec{\mathbf{z}}^k) + 2\alpha(A\vec{\mathbf{z}}^k, \vec{\mathbf{y}}) \geq f(\vec{\mathbf{z}}^k) - 2|\alpha|\|A\vec{\mathbf{z}}^k\|_2\|y^k\|_2|cos(A\vec{\mathbf{z}}^k, \vec{\mathbf{y}})|.$$

Отсюда следует, что при $\alpha > 0$ имеет смысл брать $\vec{\mathbf{y}} = -\overrightarrow{\nabla \mathbf{f}} = -2A\vec{\mathbf{z}}^k = -2\vec{\mathbf{r}}^k$, т. е. вектор, направленный вдоль вектора невязки в направлении максимального уменьшения ошибки (отсюда знак «минус»).

Метод наискорейшего спуска

Итак, мы пришли к выводу, что имеет смысл рассматривать итерационные методы вида $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - H_k\vec{\mathbf{r}}^k$, т. е. методы, записанные в каноническом виде. Самый простой вариант этого метода с легко обратимой матрицей — это вариант $H_k = \tau_k E$, т. е. с одним

скаляром, который нужно определить. Идея определения вариантов выбора параметра α возникла благодаря приведённому выше рассуждению. Мы будем называть *методом наискорейшего спуска* метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - \tau_k(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}),$$

где τ_k определяется из условия минимизации $\|\vec{\mathbf{z}}^{k}\|_A^2$.

Теорема 26. *Если $A = A^* > 0$, то метод наискорейшего спуска сходится и при этом*

$$\tau_k = \frac{(\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})}{(A\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})}.$$

Доказательство. Непосредственные вычисления дают нам

$$\|\vec{\mathbf{z}}^{k+1}\|_A^2 = (A(\vec{\mathbf{z}}^{k} - \tau_k\vec{\mathbf{r}}^{k}), \vec{\mathbf{z}}^{k} - \tau_k\vec{\mathbf{r}}^{k}) =$$

$$= \|\vec{\mathbf{z}}^{k}\|_A^2 - 2\tau_k(\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k}) + \tau_k^2(A\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k}).$$

Таким образом, мы имеем квадратное уравнение относительно τ_k. В силу положительной определённости матрицы A мы получаем, что $(A\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k}) > 0 \ \forall \vec{\mathbf{r}}^{k} \neq \vec{\mathbf{0}}$, т. е. парабола направлена ветвями вверх, следовательно, мы можем найти у неё минимум. Школьная проверка даёт нам решение задачи

$$\tau_k = \frac{(\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})}{(A\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})}.$$

При этом выборе параметра τ_k мы получаем, что

$$\|\vec{\mathbf{z}}^{k+1}\|_A^2 = \|\vec{\mathbf{z}}^{k}\|_A^2 - \frac{|(\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})|^2}{(A\vec{\mathbf{r}}^{k}, \vec{\mathbf{r}}^{k})} < \|\vec{\mathbf{z}}^{k}\|_A^2,$$

если $\vec{\mathbf{r}}^{k} \neq \vec{\mathbf{0}}$ (в противном случае итерационный метод уже нашёл точное решение системы уравнений и дальнейшие итерации не требуются).

По определению оператор $S(\vec{\mathbf{z}}^k) = \vec{\mathbf{z}}^k - \tau_k(\vec{\mathbf{z}}^k)A\vec{\mathbf{z}}^k$ непрерывен всюду, кроме, быть может, нуля. Обратите внимание, что параметр τ_k зависит от $\vec{\mathbf{z}}^k$, что делает оператор S, вообще говоря, нелинейным, хотя мы и можем его записать в линейной форме после вычисления параметра τ_k. Таким образом, применение теоремы 22 завершает доказательство сходимости метода наискорейшего спуска и нашей теоремы. □

Метод минимальных невязок

Следующий метод использует ту же идею, что и у метода наискорейшего спуска. Мы будем называть *методом минимальных невязок* метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \tau_k(A\vec{\mathbf{x}}^k - \vec{\mathbf{b}}),$$

где τ_k определяются из условия минимизации $\|\vec{\mathbf{r}}^k\|_2^2$.

Теорема 27. *Если $A > 0$ и вещественна, то метод минимальных невязок сходится и при этом*

$$\tau_k = \frac{(A\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k)}{(A\vec{\mathbf{r}}^k, A\vec{\mathbf{r}}^k)}.$$

Доказательство. Непосредственные вычисления дают нам

$$\|\vec{\mathbf{r}}^{k+1}\|_2^2 = (\vec{\mathbf{r}}^k - \tau_k A\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k - \tau_k A\vec{\mathbf{r}}^k)) =$$

$$= \|\vec{\mathbf{r}}^k\|_2^2 - 2\tau_k(A\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k) + \tau_k^2(A\vec{\mathbf{r}}^k, A\vec{\mathbf{r}}^k).$$

Именно здесь нам понадобилась вещественность матрицы A, чтобы переставить местами вектора в скалярном произведении без необходимости навешивать знак сопряжения. Таким образом, мы имеем квадратное уравнение относительно τ_k. В силу аксиомы нормы мы получаем, что $(A\vec{\mathbf{r}}^k, A\vec{\mathbf{r}}^k) > 0\ \forall \vec{\mathbf{r}}^k \neq \vec{\mathbf{0}}$, т. е. парабола

направлена ветвями вверх, следовательно, мы можем найти у неё минимум. Школьная проверка даёт нам решение задачи

$$\tau_k = \frac{(A\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k)}{(A\vec{\mathbf{r}}^k, A\vec{\mathbf{r}}^k)}.$$

При этом выборе параметра τ_k мы получаем, что

$$\|\vec{\mathbf{r}}^{k+1}\|_2^2 = \|\vec{\mathbf{r}}^k\|_2^2 - \frac{|(A\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k)|^2}{(A\vec{\mathbf{r}}^k, A\vec{\mathbf{r}}^k)} < \|\vec{\mathbf{r}}^k\|_2^2,$$

если $\vec{\mathbf{r}}^k \neq \vec{\mathbf{0}}$ (в противном случае в силу невырожденности матрицы A, итерационный метод уже нашёл точное решение системы уравнений и дальнейшие итерации не требуются).

Перейдём от вектора невязки к вектору ошибки. Так как $\vec{\mathbf{r}}^k = A\vec{\mathbf{z}}^k$, а матрица A^*A симметрична (непосредственная проверка это подтверждает) и положительна определена (следует из невырожденности матрицы A), то мы можем определить норму $\|\cdot\|_{A^*A} = \sqrt{(A^*A\cdot, \cdot)}$. Для этой нормы справедливо $\|\vec{\mathbf{z}}^k\|_{A^*A}^2 = (A^*A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) = (A\vec{\mathbf{z}}^k, A\vec{\mathbf{z}}^k) = (\vec{\mathbf{r}}^k, \vec{\mathbf{r}}^k) = \|\vec{\mathbf{r}}^k\|_2^2$, следовательно,

$$\|\vec{\mathbf{z}}^{k+1}\|_{A^*A} < \|\vec{\mathbf{z}}^k\|_{A^*A}.$$

По определению оператор $S(\vec{\mathbf{z}}^k) = \vec{\mathbf{z}}^k - \tau_k(\vec{\mathbf{z}}^k)A\vec{\mathbf{z}}^k$ непрерывен всюду, кроме, быть может, нуля. Обратите внимание, что параметр τ_k зависит от $\vec{\mathbf{z}}^k$, что делает оператор S, вообще говоря, нелинейным, хотя мы и можем его записать в линейной форме после вычисления параметра τ_k. Таким образом, применение теоремы 22 завершает доказательство сходимости метода минимальных невязок и нашей теоремы. □

Обратите внимание на то, что формулы для вычисления параметра τ_k для методов наискорейшего спуска и минимальных

невязок выглядят очень похоже. Попытки выучить их без понимания того, откуда они получены, неизменно приводят к неразличению их на экзамене. Также нужно обратить внимание на то, что в теореме для метода минимальных невязок не требуется самосопряжённость матрицы системы, но требуется её вещественность.

Метод простой итерации

В методах наискорейшего спуска и минимальных невязок для определения параметра τ_k нужно вычислять два скалярных произведения (с умножением невязки на матрицу системы) на каждом шаге итерационного метода. Использование постоянного параметра τ существенно уменьшает объем вычислений на каждом шаге. Мы будем называть *методом простой итерации* метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - \tau(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}).$$

Теорема 28. *Если $A = A^* > 0$, то метод простой итерации сходится при*

$$\tau \in (0, \frac{2}{\rho(A)}).$$

При этом справедлива оценка

$$\|\vec{\mathbf{z}}^{k}\|_2 \leq \rho_\tau^k \|\vec{\mathbf{z}}^{0}\|_2,$$

где $\rho_\tau = max\{|1 - \tau\lambda_{\min}(A)|, |1 - \tau\lambda_{\max}(A)|\}$. Оптимальные значения параметров:

$$\tau_* = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$$

и

$$\rho_* = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}.$$

Доказательство. Так как метод простой итерации записан в каноническом виде, то матрица шага для ошибки будет $S = E - \tau A$. Отсюда следует, что с помощью свойства подчинённой матричной нормы $\|\vec{\mathbf{z}}^k\|_2 = \|S\vec{\mathbf{z}}^{k-1}\|_2 \leq \|S\|_2\|\vec{\mathbf{z}}^{k-1}\|_2 \leq \ldots \leq \|S\|_2^k\|\vec{\mathbf{z}}^0\|_2$. Если вспомнить выражение для второй матричной нормы, получим

$$\|S\|_2 = \|E - \tau A\|_2 = \rho(E - \tau A) \equiv \rho_\tau.$$

Далее заметим, что в силу самосопряжённости матрицы A у матрицы $E - \tau A$ — вещественный спектр, поэтому мы можем записать

$$\rho_\tau = \max_{\lambda \in \{\lambda_i(A), i=1,\ldots,n\}} |1 - \tau\lambda| = \max\{|1 - \tau\lambda_{\min}(A)|, |1 - \tau\lambda_{\max}(A)|\}$$

в силу того, что функция $|1 - \tau\lambda|$ от аргумента λ выпукла вниз.

По теореме 15 стационарный итерационный метод простой итерации сходится тогда и только тогда, когда $\rho(S) < 1$. Как нетрудно видеть, это означает, что $|1-\tau\lambda| < 1$ или $-1 < 1-\tau\lambda < 1$ или $0 < \tau < 2/\lambda$. Поскольку последнее неравенство должно быть верно для любого собственного числа и они все положительны в силу положительной определённости матрицы A, то метод простой итерации сходится при $0 < \tau < 2/\lambda_{\max}(A)$. Осталось только заметить, что $\lambda_{\max}(A) = \rho(A)$ по определению спектрального радиуса матрицы.

Чтобы выбрать оптимальный параметр τ_*, мы должны минимизировать функцию ρ_τ. По определению модуля эту функцию можно записать в виде

$$\rho_\tau = \begin{cases} 1 - \tau\lambda_{\min}(A), & 0 < \tau \leq \tau_* \\ \tau\lambda_{\max}(A) - 1, & \tau_* \leq \tau < \frac{1}{\lambda_{\max}(A)} \end{cases}.$$

Очевидно, что оптимальное значение достигается в том случае, когда значение функции в левом конце интервала совпадает со значением в правом конце, т. е. $1 - \tau\lambda_{\min}(A) = \tau\lambda_{\max}(A) - 1$ или $\tau_* = \frac{2}{\lambda_{\min}(A)+\lambda_{\max}(A)}$. Подставляя это значение в формулу для ρ_τ, мы незамедлительно приходим к последнему утверждению теоремы. □

Оценка скорости сходимости для методов наискорейшего спуска и минимальных невязок

Воспользуемся оценкой на скорость сходимости метода простой итерации из теоремы 28 для того, чтобы получить оценки для скорости сходимости методов наискорейшего спуска и минимальных невязок.

Теорема 29. *Если $A = A^* > 0$, то для ошибки в методе наискорейшего спуска справедливы оценки*

$$\|\vec{\mathbf{z}}^k\|_2 \le \sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}} \left[\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}\right]^k \|\vec{\mathbf{z}}^0\|_2$$

и

$$\|\vec{\mathbf{z}}^k\|_A \le \left[\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}\right]^k \|\vec{\mathbf{z}}^0\|_A.$$

Доказательство. При доказательстве теоремы 28 мы получили, что при оптимальном выборе параметра τ в методе простой итерации мы имеем

$$\|\vec{\mathbf{z}}^{k+1}\|_A = \min_\tau \|\vec{\mathbf{z}}^k - \tau A\vec{\mathbf{z}}^k\|_A \le \|\vec{\mathbf{z}}^k - \tau_* A\vec{\mathbf{z}}^k\|_A \le \|E - \tau_* A\|_A \|\vec{\mathbf{z}}^k\|_A.$$

Теперь по определению подчинённой матричной нормы с учётом того, что у самосопряжённой положительно определённой матрицы есть невырожденный самосопряжённый квадратный корень, мы получаем

$$\|E - \tau_* A\|_A^2 \equiv \sup_{\vec{\mathbf{y}} \ne \vec{\mathbf{0}}} \frac{(A(E - \tau_* A)\vec{\mathbf{y}}, (E - \tau_* A)\vec{\mathbf{y}})}{(A\vec{\mathbf{y}}, \vec{\mathbf{y}})} =$$

$$= \sup_{\vec{\mathbf{y}} \ne \vec{\mathbf{0}}} \frac{(A^{1/2}A^{1/2}(E - \tau_* A)\vec{\mathbf{y}}, (E - \tau_* A)\vec{\mathbf{y}})}{(A\vec{\mathbf{y}}, \vec{\mathbf{y}})} =$$

$$= \sup_{\vec{\mathbf{y}} \ne \vec{\mathbf{0}}} \frac{(A^{1/2}(E - \tau_* A)\vec{\mathbf{y}}, A^{1/2}(E - \tau_* A)\vec{\mathbf{y}})}{(A^{1/2}A^{1/2}\vec{\mathbf{y}}, \vec{\mathbf{y}})} =$$

за счёт перестановочности матрицы и её корня

$$= \sup_{\vec{\mathbf{y}} \neq \vec{\mathbf{0}}} \frac{((E - \tau_* A)A^{1/2}\vec{\mathbf{y}}, (E - \tau_* A)A^{1/2}\vec{\mathbf{y}})}{(A^{1/2}\vec{\mathbf{y}}, A^{1/2}\vec{\mathbf{y}})} =$$

полагая $\vec{\mathbf{v}} \equiv A^{1/2}\vec{\mathbf{y}}$ и в силу невырожденности корня из матрицы A

$$= \sup_{\vec{\mathbf{v}} \neq \vec{\mathbf{0}}} \frac{((E - \tau_* A)\vec{\mathbf{v}}, (E - \tau_* A)\vec{\mathbf{v}})}{(\vec{\mathbf{v}}, \vec{\mathbf{v}})} \equiv \|E - \tau_* A\|_2^2 \equiv \rho_*^2.$$

Таким образом, мы имеем $\|\vec{\mathbf{z}}^k\|_A \leq \rho_*^k \|\vec{\mathbf{z}}^0\|_A$ — одна из оценок доказана. При доказательстве теоремы 20 мы уже обсуждали, как получить утверждение о том, что для самосопряжённой матрицы верно $\lambda_{\min}(A)(\vec{\mathbf{y}}, \vec{\mathbf{y}}) \leq (A\vec{\mathbf{y}}, \vec{\mathbf{y}}) \leq \lambda_{\max}(A)(\vec{\mathbf{y}}, \vec{\mathbf{y}})\ \forall \vec{\mathbf{y}} \in \mathbb{C}^n$, а потому

$$\sqrt{\lambda_{\min}(A)}\|\vec{\mathbf{y}}\|_2 \leq \|\vec{\mathbf{y}}\|_A \leq \sqrt{\lambda_{\max}(A)}\|\vec{\mathbf{y}}\|_2.$$

Отсюда незамедлительно следует вторая оценка теоремы, и теорема доказана. □

Теорема 30. *Если $A = A^* > 0$, то для ошибки и невязки в методе минимальных невязок справедливы оценки*

$$\|\vec{\mathbf{z}}^k\|_2 \leq \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \left[\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}\right]^k \|\vec{\mathbf{z}}^0\|_2$$

и

$$\|\vec{\mathbf{r}}^k\|_2 \leq \left[\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}\right]^k \|\vec{\mathbf{r}}^0\|_2.$$

Доказательство. Доказательство будем вести аналогично доказательству теоремы 29. Мы получаем, что при оптимальном выборе параметра τ в методе простой итерации

$$\|\vec{\mathbf{r}}^{k+1}\|_2 = \min_{\tau} \|\vec{\mathbf{r}}^k - \tau A\vec{\mathbf{r}}^k\|_2 \leq \|\vec{\mathbf{r}}^k - \tau_* A\vec{\mathbf{r}}^k\|_2 \leq \|E - \tau_* A\|_2 \|\vec{\mathbf{r}}^k\|_2.$$

Так как $\|E - \tau_* A\|_2 \equiv \rho_*$, то получаем $\|\vec{\mathbf{r}}^k\|_2 \leq \rho_*^k \|\vec{\mathbf{r}}^0\|_2$ и оценка для невязки доказана.

Используя те же самые аргументы, что и при доказательстве теоремы 20, мы можем получить утверждение о том, что для самосопряжённой матрицы верно $\lambda^2_{\min}(A)(\vec{\mathbf{y}},\vec{\mathbf{y}}) \leq (A\vec{\mathbf{y}},A\vec{\mathbf{y}}) \leq \lambda^2_{\max}(A)(\vec{\mathbf{y}},\vec{\mathbf{y}})\ \forall\vec{\mathbf{y}} \in \mathbb{C}^n$, а потому с учётом $\vec{\mathbf{r}}^k = A\vec{\mathbf{z}}^k$ получаем

$$\lambda_{\min}(A)\|\vec{\mathbf{z}}^k\|_2 \leq \|\vec{\mathbf{r}}^k\|_2 \leq \lambda_{\max}(A)\|\vec{\mathbf{z}}^k\|_2.$$

Отсюда незамедлительно следует оценка для ошибки, и теорема доказана. □

Обратите внимание, что для получения оценок нам понадобились условия самосопряжённости и положительной определённости матрицы системы в обеих теоремах.

Лекция №9

Метод Ричардсона с чебышёвскими параметрами

Задача оптимизации параметров

В методе простой итерации параметр τ вычисляется один раз, поэтому естественным образом возникает идея о том, чтобы попробовать сделать параметр переменным и за счёт этого добиться ускорения метода, т. е. исследовать нестационарный метод вида $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} - \tau_{k+1}(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}})$. Попробуем взять конечный набор параметров τ_k и выберем их из условия минимизации спектрального радиуса матрицы (оператора) шага для ошибки за m шагов: $\vec{\mathbf{z}}^{m} \equiv S_m(A)\vec{\mathbf{z}}^{0} \equiv (E-\tau_m A)\dots(E-\tau_1 A)\vec{\mathbf{z}}^{0}$. Как мы помним, спектральный радиус матрицы шага для ошибки определяет скорость сходимости итерационного метода. Если все параметры τ_k взять одинаковыми, то мы получим метод простой итерации, который сходится при известных условиях, т. е. предлагаемый способ построения итерационного метода может привести только к лучшему методу.

Мы будем предполагать, что $\lambda(A) \in [\alpha, \beta],\ \alpha > 0$, т. е. ВСЕ собственные значения матрицы A системы линейных уравнений вещественны и положительны.

Так как

$$\min_{\tau_1,\dots,\tau_m} \rho(S_m(A)) =$$

заменим по определению спектральный радиус матрицы на то (вообще говоря, неизвестное) из собственных чисел, которое даёт максимум модуля, и воспользуемся свойствами полинома от матрицы, известными из курса алгебры

$$= \min_{\tau_1,\dots,\tau_m} \left(\max_{\lambda\in\lambda(A)} |S_m(\lambda)| \right) \leq$$

расширим область, в которой мы будем искать максимум полинома

$$\leq \min_{\tau_1,\ldots,\tau_m} \left(\max_{\lambda\in[\alpha,\beta]} |S_m(\lambda)| \right) \equiv \rho_m.$$

Последнюю так называемую *минимаксную задачу* решать проще в силу того, что нам не требуется знать спектр матрицы $S_m(A)$, а достаточно знать лишь границы спектра, что (как уже говорилось ранее) в некоторых случаях представляет из себя более простую задачу, чем поиск всего спектра. Как нетрудно заметить, нам нужно решить задачу о поиске полинома степени m, наименее уклоняющегося от нуля на отрезке $[\alpha, \beta]$. Поскольку любой найденный полином всегда можно умножить на константу (отнормировать) и получить полином, который ещё меньше уклоняется от нуля, то мы сразу зададим условие нормировки, например, $S_m(0) = 1$. Это условие позволяет представить наш полином в естественном виде $S_m(\lambda) = (1 - \tau_m\lambda)\ldots(1 - \tau_1\lambda)$, так как в нуле он очевидно равен единице и выводится из матричного полинома $S_m(A)$.

Перепишем полином в виде

$$S_m(\lambda) = \frac{(-1)^m}{\tau_m \ldots \tau_1}(\lambda - \mu_m)\ldots(\lambda - \mu_1),$$

где $\mu_i = \frac{1}{\tau_i}$.

$\|S_m(A)\|_2 = \rho(S_m(A)) \leq \rho_m$ в случае самосопряжённой матрицы A, т. е. оценка на скорость сходимости метода будет $\|\vec{\mathbf{z}}^k\|_2 \leq \rho_m^k \|\vec{\mathbf{z}}^0\|_2$.

Полином Чебышёва и решение задачи оптимизации параметров

Задача о поиске полинома, наименее отклоняющегося от нуля на заданном отрезке, была решена русским математиком П. Л. Чебышёвым, а потому просто приведём это решение. Для определённости мы будем рассматривать отрезок $[-1, 1]$, тогда решение

задачи задаётся формулой $T_m(x) \equiv \cos(m \arccos(x))$. Несмотря на весьма необычный вид, можно сразу заметить, что $T_0(x) = 1$, а $T_1(x) = x$, если мы не выходим за пределы нашего интервала $[-1, 1]$, т. е. $T_0(x)$ и $T_1(x)$ действительно являются полиномами. Далее по известной из тригонометрии формуле $\cos((k+1)\phi) + \cos((k-1)\phi) = 2cos(\phi)\cos(k\phi)$ при $\phi = \arccos x$ получаем, что $T_{k+1}(x) = 2T_1(x)T_k(x) - T_{k-1}(x)$. По методу математической индукции мы незамедлительно приходим к утверждению о том, что $T_k(x)$ действительно является полиномом степени k при любом $k \geq 0$. Также заметим, что формула с косинусом может использоваться только в пределах интервала $[-1, 1]$, а вот рекуррентная формула уже может использоваться на всей числовой оси.

Вспоминая свойство косинуса, мы легко получаем, что у полинома степени m есть $m + 1$ (ЭТО ОЧЕНЬ ВАЖНО!) точек экстремума $\hat{x}_k = \cos\frac{k\pi}{m}$, $k = 0, \ldots, m$, в которых полином принимает значения ± 1. Непосредственные вычисления показывают, что $-1 = \hat{x}_m < \ldots < \hat{x}_0 = 1$ и $T_m(\hat{x}_k) = (-1)^{k+1}$. Так же просто вычисляется, что корнями полинома $T_m(x)$ являются точки $x_k = \cos\frac{(2k-1)\pi}{2m}, k = 1, \ldots, m$. При этом других корней у полинома нет по основной теореме алгебры (или следствию из неё у некоторых преподавателей). Заметим, что (и это тоже важно) $\hat{x}_k < x_k < \hat{x}_{k-1}$ $\forall k = 1, \ldots, m$.

Теперь с помощью простого линейного преобразования

$$x(\lambda) = \frac{2\lambda - (\beta + \alpha)}{\beta - \alpha}$$

мы можем перевести любой полином из интервала $\lambda \in [\alpha, \beta]$ в интервал $[-1, 1]$, где у нас определён полином Чебышёва. Следовательно, для выполнения условия нормировки мы можем взять

$$S_m(\lambda) = \frac{T_m(x(\lambda))}{T_m(x(0))}.$$

Деление в этой формуле безопасно, так как за пределами интервала $[-1, 1]$ полином $T_m(x)$ не имеет корней.

Корни полинома S_m легко вычисляются по формуле

$$\mu_k = \frac{(\beta-\alpha)x_k + (\beta+\alpha)}{2} \in [\alpha,\beta],\ k = 1,\ldots,n.$$

Теперь осталось лишь проверить, что полином действительно наименее уклоняется от нуля, следовательно, даёт решение нашей задачи оптимизации параметров итерационного метода.

Теорема 31. *Если левая граница интервала для собственных значений матрицы $\alpha > 0$, то*

$$\|S_m(\lambda)\|_{C[\alpha,\beta]} = \inf_{Q_{m-1}(\lambda)} \|1 - \lambda Q_{m-1}(\lambda)\|_{C[\alpha,\beta]}.$$

Доказательство. Для равномерной нормы функции мы будем использовать обозначение

$$\|g(x)\|_{C[a,b]} \equiv \max_{x\in[a,b]} |g(x)|.$$

Проверьте выполнение всех аксиом нормы в этом случае.

Воспользуемся методом «от противного». Предположим, что наш полином $S_m(\lambda)$ на самом деле не является наименее отклоняющимся от нуля, т. е.

$$\|S_m(\lambda)\|_{C[\alpha,\beta]} > \inf_{Q_{m-1}(\lambda)} \|1 - \lambda Q_{m-1}(\lambda)\|_{C[\alpha,\beta]}.$$

Тогда существует вполне конкретный полином $Q_{m-1}(\lambda)$ степени не выше, чем $m-1$, который даёт неравенство $\|S_m(\lambda)\|_{C[\alpha,\beta]} > \|1-\lambda Q_{m-1}(\lambda)\|_{C[\alpha,\beta]}$ в силу конечномерности пространства полиномов.

Теперь заметим, что $S_m(\lambda) = 1-\lambda P_{m-1}(\lambda)$ для некоторого полинома $P_{m-1}(\lambda)$ степени не выше, чем $m-1$. Это представление объясняется тем, что наш полином $S_m(\lambda)$ обращается в единицу

в точке $\lambda = 0$. По построению в точках $\hat{\mu}_k = \frac{(\beta-\alpha)\hat{x}_k+(\beta+\alpha)}{2}$, $k = 0, \ldots, m$ наш полином имеет экстремумы, т. е. выполняется равенство $|S_m(\hat{\mu}_k)| = \|S_m(\lambda)\|_{C[\alpha,\beta]}$. Опять же по построению последовательность $\{S_m(\hat{\mu}_k)\}_k$ знакопеременна, что означает, что полином степени не выше m $R_m(\lambda) \equiv S_m(\lambda) - (1 - \lambda Q_{m-1}(\lambda)) = \lambda(P_{m-1}(\lambda) - Q_{m-1}(\lambda))$ тоже меняет знак в точках $\hat{\mu}_k$, $k = 0, \ldots, m$. Как следует из математического анализа, внутри каждого из интервалов $(\hat{\mu}_{k-1}, \hat{\mu}_k)$, $k = 1, \ldots, m$ имеется хотя бы один корень полинома $R_m(\lambda)$. Это означает, что в интервале $[\alpha, \beta]$ полином $R_m(\lambda)$ имеет не менее, чем m корней! Именно здесь нам пригодилось утверждение о том, что число экстремумов у полинома Чебышёва на один больше, чем число корней.

Поскольку по условию теоремы $\alpha > 0$, то по построению у полинома $R_m(\lambda)$ имеется ещё один корень $\lambda = 0$. Как известно из алгебры, если у полинома степени не выше m имеется не менее $m + 1$-го различного корня, то такой полином обязан быть тождественно нулевым, т. е. $R_m(\lambda) \equiv 0$, а $S_m(\lambda) = 1 - \lambda Q_{m-1}(\lambda)$, что противоречит сделанному предположению. Таким образом, теорема доказана. □

Теперь вычислим константу $\rho_m = \|S_m(\lambda)\|_{C[\alpha,\beta]}$, определяющую скорость сходимости нашего итерационного процесса.

Теорема 32. *Если левая граница интервала для собственных значений матрицы $\alpha > 0$, то*

$$\rho_m = \|S_m(\lambda)\|_{C[\alpha,\beta]} = \frac{2\gamma^m}{1+\gamma^{2m}} < 1, \ \gamma \equiv \frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}.$$

Доказательство. По построению мы незамедлительно получаем, что $\rho_m = \|S_m(\lambda)\|_{C[\alpha,\beta]} = |T_m(x(0))|^{-1}$, при этом $x(0) = -\frac{\beta-\alpha}{\beta+\alpha} < -1$.

Для получения результата нам потребуется ещё одно представление для полинома Чебышёва, которое справедливо только

за пределами интервала $[-1,1]$:

$$T_m(x) = \frac{(x+\sqrt{x^2-1})^m + (x-\sqrt{x^2-1})^m}{2},\ |x|>1.$$

Проверим это представление. База математической индукции будет $T_0(x) = 1$, $T_1(x) = x$. Проверим, что формула $T_{k+1}(x) = 2T_1(x)T_k(x) - T_{k-1}(x)$ верна в предположении, что полиномы степени не выше k могут быть представлены в указанном выше виде. Для упрощения выкладок введём обозначение $y \equiv \sqrt{x^2-1}$. Тогда

$$T_{k+1}(x) = (x+y)^{k+1} + (x-y)^{k+1} = x\left[(x+y)^k + (x-y)^k\right] +$$

$$y\left((x+y)^k - (x-y)^k\right) = x\left[(x+y)^k + (x-y)^k\right] +$$

$$+y\left(x\left[(x+y)^{k-1} - (x-y)^{k-1}\right] + y\left[(x+y)^{k-1} + (x-y)^{k-1}\right]\right) =$$

$$= x\left[(x+y)^k + (x-y)^k\right] + yx\left[(x+y)^{k-1} - (x-y)^{k-1}\right] +$$

так как $y^2 = x^2 - 1$

$$+(x^2-1)\left[(x+y)^{k-1} + (x-y)^{k-1}\right] =$$

$$= 2x\left[(x+y)^k + (x-y)^k\right] - \left[(x+y)^{k-1} + (x-y)^{k-1}\right] +$$

$$+(-x^2 - xy + yx + x^2)(x+y)^{k-1} +$$

$$+(-x^2 + xy - yx + x^2)(x-y)^{k-1} =$$

так как последние два слагаемых равны нулю

$$= 2T_1(x)T_k(x) - T_{k-1}(x),$$

что и требовалось доказать.

Для завершения доказательства теоремы вычислим значение полинома Чебышёва в точке 0:

$$x(0) + \sqrt{x^2(0)-1} = \frac{-\beta + 2\sqrt{\beta}\sqrt{\alpha} - \alpha}{(\sqrt{\beta}+\sqrt{\alpha})(\sqrt{\beta}-\sqrt{\alpha})} = -\frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}} = -\gamma,$$

$$x(0) - \sqrt{x^2(0)-1} = \frac{1}{x(0)+\sqrt{x^2(0)-1}} = -\frac{1}{\gamma}.$$

Таким образом, мы получаем

$$T_m(x(0)) = \frac{(-\gamma)^m + (-\gamma^{-1})^m}{2} = (-1)^m \frac{1+\gamma^{2m}}{2\gamma^m},$$

следовательно,

$$\rho_m = \frac{2\gamma^m}{1+\gamma^{2m}}.$$

Так как $\gamma < 1$, то $\rho_m < 1$, и теорема доказана. □

Циклический метода Ричардсона

Если матрица системы $A = A^* > 0$, и известны оценки границ её спектра $\lambda(A) \in [\alpha, \beta]$, $\alpha > 0$, то *циклическим методом Ричардсона* с длиной цикла m для решения системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ называется нестационарный итерационный метод вида

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \tau_{k+1}(A\vec{\mathbf{x}}^k - \vec{\mathbf{b}}),\ k = 0, 1, \ldots,$$

где остаток от деления j на m, увеличенный на 1, даёт номер k из диапазона от 1 до m, который вычисляется через корни полинома Чебышёва степени m по формуле

$$\tau_k = \frac{2}{(\beta+\alpha) + (\beta-\alpha)\cos\frac{(2k-1)\pi}{2m}},\ k = 1, \ldots, m.$$

Поскольку за каждые m шагов итерационного процесса ошибка уменьшается не менее, чем в ρ_m раз в евклидовой норме, то справедлива оценка

$$\|\vec{\mathbf{z}}^{k\cdot m}\|_2 \le (\rho_m)^k \|\vec{\mathbf{z}}^0\|_2 = \left(\frac{2\gamma^m}{1+\gamma^{2m}}\right)^k \|\vec{\mathbf{z}}^0\|_2,\ \gamma = \frac{\sqrt{\beta}-\sqrt{\alpha}}{\sqrt{\beta}+\sqrt{\alpha}}.$$

Так как $\rho_m < 1$, то в силу непрерывности оператора шага для ошибки за m шагов по теореме 22 циклический метод Ричардсона сходится.

В связи с тем, что норма оператора шага для ошибки $\|E - \tau_k A\|_2$ может быть больше единицы для некоторых шагов, то за счёт ошибок округления некоторые реализации метода Ричардсона могут быть неустойчивы, т. е. НЕ СХОДИТЬСЯ к точному решению задачи. Противоречия здесь нет, так как утверждение о сходимости циклического метода Ричардсона предполагало вычисления в точной арифметике. Заметим также, что мы делали оператор шага для ошибки меньше единицы в евклидовой норме ЗА m ШАГОВ, ничего не говоря о том, какова будет норма оператора шага для ошибки за один шаг.

В данном курсе мы не сможем рассказать теорию устойчивости вычислительных методов, но приведём пример и покажем, как мельчайшие ошибки округления могут влиять на результат вычислений настолько сильно, что эти вычисления могут стать абсолютно непригодными на практике. Подробнее с информацией об устойчивости как математическом понятии можно ознакомиться на последующих курсах по вычислительным методам.

Итак, возьмём метод Ричардсона длины $m = 10000$. В качестве системы уравнений будем использовать систему с матрицей A размера 2×2 :

$$A = \begin{bmatrix} 1 & 0 \\ 0 & m \end{bmatrix}.$$

Также для простоты примера мы будем предполагать, что ошибки округления возникают только на шаге с номером m и они пропорциональны начальному вектору ошибки, т. е. вместо $\vec{\mathbf{z}}^m$ на последнем шаге мы получим некоторый другой вектор ошибки

$$\tilde{\vec{z}}^m = (E - \tau_m A)\left[(E - \tau_{m-1}A)\dots(E - \tau_1 A)\vec{\mathbf{z}}^0 + \varepsilon\vec{\mathbf{z}}^0\right] =$$

$$= \vec{\mathbf{z}}^m + \varepsilon(E - \tau_m A)\vec{\mathbf{z}}^0.$$

В данном примере очевидно, что в качестве границ спектра матрицы можно взять $\alpha = 1$ и $\beta = m = 10000$. По теореме 32 мы получаем, что в этом случае $\gamma = 99/101$, и по формуле замечательного предела из математического анализа $\rho_{10000} \approx 2e^{-200}$. Это означает, что точная ошибка в евклидовой норме $\|\vec{\mathbf{z}}^{10000}\|_2 \leq 2e^{-200}\|\vec{\mathbf{z}}^0\|_2 \approx 0$, т. е. практически нулевая. Однако непосредственные вычисления показывают, что $\tau_{10000} \approx 1$ и

$$E - \tau_{10000}A \approx \begin{bmatrix} 0 & 0 \\ 0 & -9999 \end{bmatrix},$$

а это значит, что возмущённая за счёт округлений ошибка будет $\|\vec{\tilde{z}}^{10000}\|_2 \approx \varepsilon * 10000|z_2^0|$. Если ошибки округления были порядка $\varepsilon = 10^{-4}$, то ошибка после 10 000 шагов метода фактически не уменьшилась и осталась на начальном уровне. Это и есть неустойчивость вычислительных формул метода, поскольку сходящийся метод не сходится из-за ошибок округления.

Однако если мы поменяем порядок параметров в методе Ричардсона на $\tau_2, \ldots, \tau_{10000}, \tau_1$, то

$$\vec{\tilde{z}}^m =$$

$$= (E - \tau_1 A)(E - \tau_{10000}A)\left[(E - \tau_{m-1}A)\ldots(E - \tau_2 A)\vec{\mathbf{z}}^0 + \varepsilon\vec{\mathbf{z}}^0\right] =$$

$$= \vec{\mathbf{z}}^m + \varepsilon(E - \tau_1 A)(E - \tau_{10000}A)\vec{\mathbf{z}}^0$$

— теперь ситуация с ошибкой кардинально поменялась. Поскольку $\tau_1 = 0{,}00001 + O(10000^{-2})$, то

$$(E - \tau_1 A)(E - \tau_{10000}A) \approx \begin{bmatrix} 0 & 0 \\ 0 & O(1) \end{bmatrix},$$

следовательно, $\|\vec{\tilde{z}}^m\|_2 \approx \varepsilon O(|z_2^0|)$, т. е. получен результат с точностью до ошибок округления.

Поскольку переупорядочение параметров метода Ричардсона существенно влияет на устойчивость вычислений, то мы можем сформулировать задачу об оптимальном упорядочении параметров следующим образом. Пусть $p \equiv \{p(1), \ldots, p(m)\}$ является перестановкой m чисел $\{1, \ldots, m\}$. Пусть $\nu_j(p) \equiv \rho((E -$

$\tau_{p(j)}A)\dots(E-\tau_{p(m)}A))$ и $\nu(p)\equiv\sum_{j=1}^{m-1}\tau_{p(j)}\nu_{j+1}(p)+\tau_p(m)$. Нужно найти p_* — такое, что $\nu(p_*)=\inf_p\nu(p)$. При $m=2^k$ для некоторого целого положительного k у этой задачи имеется элегантное решение:

$$\tau_1,\dots,\tau_m\Rightarrow r_1^1\equiv(\tau_m,\tau_1),\dots,r_{m/2}^1\equiv(\tau_{m/2+1},\tau_{m/2})\Rightarrow$$

$$\Rightarrow r_1^2\equiv(r_{m/2}^1,r_1^1),\dots,r_{m/4}^2\equiv(r_{m/4+1}^1,r_{m/4}^1)\Rightarrow\dots\Rightarrow p_*\equiv r_1^k.$$

Трёхчленные формулы реализации метода Ричардсона с чебышёвскими параметрами

Неудобство классического метода Ричардсона состоит в том, что помимо вычисления параметров метода требуется также провести их переупорядочивание для получения устойчивого к ошибкам округления результата. Однако в рамках классической двухслойной схемы итерационного метода решить задачу с автоматическим упорядочиванием параметров не удаётся, для этого потребуется перейти к трёхчленной формуле.

Переформулируем нашу задачу следующим образом. Для решения системы $A\vec{\mathbf{x}}=\vec{\mathbf{b}}$ мы будем строить приближения $\vec{\mathbf{x}}^k$, и при этом для ошибки будет справедлива формула

$$\vec{\mathbf{z}}^k=S_k(A)\vec{\mathbf{z}}^0,\ k\geq 0.$$

Мы знаем, как рассчитать параметры τ_k для фиксированной длины цикла в методе Ричардсона. Очевидно, что $S_0(A)=E$. А для метода простой итерации после одного шага по теореме 28 при оптимальном значении параметра τ : $S_1(A)=E-\frac{2}{\beta+\alpha}A$. Это же значение оптимального параметра нам даст и метод Ричардсона при $m=1$. Таким образом, для первого приближения к точному решению справедлива формула

$$\vec{\mathbf{x}}^1=\vec{\mathbf{x}}^0-\frac{2}{\beta+\alpha}(A\vec{\mathbf{x}}^0-\vec{\mathbf{b}})=\vec{\mathbf{x}}^0-\frac{2}{\beta+\alpha}\vec{\mathbf{r}}^0.$$

Пусть $t_k \equiv T_k(x(0)) = T_k(-\frac{\beta+\alpha}{\beta-\alpha})$. Вспомним теперь, как мы задавали оператор шага для ошибки в методе Ричардсона через полином Чебышёва, тогда с учётом введённого обозначения мы имеем рекуррентную формулу, связывающую между собой операторы шага для ошибки при длине цикла $k-1$, k и $k+1$ в методе:

$$t_{k+1}S_{k+1}(A) = 2T_1(\frac{2}{\beta-\alpha}A - \frac{\beta+\alpha}{\beta-\alpha}E)t_kS_k(A) - t_{k-1}S_{k-1}(A) =$$

так как $T_1(x) = x$ и $t_1 = -\frac{\beta+\alpha}{\beta-\alpha}$

$$= 2t_k\frac{2}{\beta-\alpha}AS_k(A) + 2t_kt_1S_k(A) - t_{k-1}S_{k-1}(A).$$

Полученное равенство мы умножим слева на вектор начальной ошибки $\vec{\mathbf{z}}^0$. По определению невязки и в силу перестановочности матрицы с её любым матричным полиномом получим

$$t_{k+1}\vec{\mathbf{z}}^{k+1} = 2t_k\frac{2}{\beta-\alpha}\vec{\mathbf{r}}^k + 2t_kt_1\vec{\mathbf{z}}^k - t_{k-1}\vec{\mathbf{z}}^{k-1} =$$

$$= 2t_k\frac{2}{\beta-\alpha}\vec{\mathbf{r}}^k + (2t_kt_1 - t_{k-1})\vec{\mathbf{z}}^k + t_{k-1}(\vec{\mathbf{z}}^k - \vec{\mathbf{z}}^{k-1}) =$$

применяем рекуррентную формулу для полинома Чебышёва в точке

$$= 2t_k\frac{2}{\beta-\alpha}\vec{\mathbf{r}}^k + t_{k+1}\vec{\mathbf{z}}^k + t_{k-1}(\vec{\mathbf{z}}^k - \vec{\mathbf{z}}^{k-1}).$$

Вспоминаем определение ошибки и получаем, что

$$t_{k+1}\vec{\mathbf{x}}^{k+1} = 2t_k\frac{2}{\beta-\alpha}\vec{\mathbf{r}}^k + t_{k+1}\vec{\mathbf{x}}^k + t_{k-1}(\vec{\mathbf{x}}^k - \vec{\mathbf{x}}^{k-1}),\ k = 2, 3, \ldots$$

Задавая $\vec{\mathbf{x}}^0$, мы можем вычислить $\vec{\mathbf{x}}^1$ по формуле метода простой итерации с оптимальным параметром τ и после этого рассчитать остальные приближения к точному решению $\vec{\mathbf{x}}^k$ по формуле

$$\vec{\mathbf{x}}^{k+1} = 2\frac{t_k}{t_{k+1}}\frac{2}{\beta-\alpha}\vec{\mathbf{r}}^k + \vec{\mathbf{x}}^k + \frac{t_{k-1}}{t_{k+1}}(\vec{\mathbf{x}}^k - \vec{\mathbf{x}}^{k-1}),\ k = 1, 2, \ldots$$

БЕЗ необходимости заранее задавать количество шагов метода Ричардсона! При этом метод по построению при любом k будет давать такое же решение, как и метод Ричардсона с длиной цикла k.

Преобразуем эту формулу к более простому виду. Пусть $\omega_k \equiv \frac{t_{k-1}}{t_k}$. Тогда $\omega_1 = -\frac{\beta-\alpha}{\beta+\alpha}$ и в силу рекуррентной формулы для полинома Чебышёва $t_{k+1} = 2t_1 t_k - t_{k-1}$

$$\omega_{k+1} = \frac{1}{2t_1 - \omega_k} = \frac{1}{2(\omega_1)^{-1} - \omega_k},\ k = 1, 2, \ldots$$

Заметим также, что

$$2\frac{t_k}{t_{k+1}}\frac{2}{\beta-\alpha} = -\frac{2t_k t_1}{t_{k+1}}\frac{2}{\beta+\alpha} = -\frac{t_{k+1}+t_{k-1}}{t_{k+1}}\frac{2}{\beta+\alpha} =$$

$$= -(1 + \frac{t_{k-1}}{t_k}\frac{t_k}{t_{k+1}})\frac{2}{\beta+\alpha}.$$

В результате мы можем переписать формулы метода в виде

$$\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^{k} + \omega_k\omega_{k-1}(\vec{\mathbf{x}}^{k} - \vec{\mathbf{x}}^{k-1}) -$$

$$-\frac{2}{\beta-\alpha}(1 + \omega_k\omega_{k+1})(A\vec{\mathbf{x}}^{k} - \vec{\mathbf{b}}),\ k = 1, 2, \ldots$$

Это и есть *трёхчленные формулы реализации метода Ричардсона*, записанные в виде двухшагового трёхслойного итерационного процесса.

Лекция №10

Многошаговые методы. Вариационная оптимизация

Для определения параметров метода Ричардсона (метода простой итерации при длине цикла $m = 1$) для решения системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ необходимо предварительное вычисление (точное или приближенное) границ спектра матрицы A, чего не требуется в методах наискорейшего спуска и минимальных невязок. Попытаемся взять простоту вычислительных формул метода Ричардсона и убрать из неё необходимость знать границы спектра матрицы. Для этого мы будем выбирать параметры метода $\vec{\mathbf{x}}^k = \vec{\mathbf{x}}^{k-1} - \tau_k(A\vec{\mathbf{x}}^{k-1} - \vec{\mathbf{b}})$ из условия

$$\|\vec{\mathbf{z}}^{km}\| \equiv \|(E - \tau_m^{(k)}A)\dots(E - \tau_1^{(k)}A)\vec{\mathbf{z}}^{(k-1)m}\| =$$

$$= \min_{\vec{\gamma}}\|(E - \gamma_m A)\dots(E - \gamma_1 A)\vec{\mathbf{z}}^{(k-1)m}\|.$$

Мы можем решать эту задачу при $k = 1$, так как при других значениях k решение будет таким же с точностью до обозначений. Мы также будем предполагать, что $\|\vec{\mathbf{z}}\|^2 \equiv \|\vec{\mathbf{z}}\|_D^2 \equiv (D\vec{\mathbf{z}}, \vec{\mathbf{z}})$ для некоторой матрицы $D = D^* > 0$.

Как нетрудно видеть,

$$\vec{\mathbf{z}}^m = \vec{\mathbf{z}}^0 - q_1(\tau)A\vec{\mathbf{z}}^0 - \dots - q_m(\tau)A^m\vec{\mathbf{z}}^0$$

для некоторых числовых коэффициентов q_i, зависящих от выбранного набора τ_i. Из курса алгебры известно, что вектор ошибки на шаге m есть не что иное как линейная комбинация векторов $\vec{\mathbf{z}}^0, \dots, A^m\vec{\mathbf{z}}^0$. Пусть $L_m \equiv L\{A\vec{\mathbf{z}}^0, \dots, A^m\vec{\mathbf{z}}^0\}$ — линейная оболочка векторов $A\vec{\mathbf{z}}^0, \dots, A^m\vec{\mathbf{z}}^0$. Обратите внимание на то, что в линейное подпространство мы не взяли вектор $\vec{\mathbf{z}}^0$ по причине того, что в линейную комбинацию он всегда входит с известным коэффициентом, равным 1. Опираясь на сведения из курса алгебры, мы в любом линейном подпространстве всегда можем выбрать

какой-нибудь базис, например, $\{\vec{\mathbf{g}}^1, \ldots, \vec{\mathbf{g}}^m\}$. Для того чтобы базис был длины m, нам необходимо, чтобы вектора $A\vec{\mathbf{z}}^0, \ldots, A^m\vec{\mathbf{z}}^0$ были линейно независимы. Предположим, что это так и есть. Позже мы увидим, что появление линейной зависимости нам ничем не мешает в итерационном процессе.

Итак, наша задача выглядит так:

$$\|\vec{\mathbf{z}}\|_D = \min_{\vec{\gamma}} \|(E - \gamma_m A)\ldots(E - \gamma_1 A)\vec{\mathbf{z}}^0\|_D =$$

$$= \min_{\vec{\alpha}} \|\vec{\mathbf{z}}^0 - \alpha_1\vec{\mathbf{g}}^1 - \ldots - \alpha_m\vec{\mathbf{g}}^m\|_D.$$

Как следует из математического анализа, если минимум существует, то компоненты вектора $\vec{\alpha}$ удовлетворяют системе уравнений

$$\frac{\partial(D\vec{\mathbf{z}}^m, \vec{\mathbf{z}}^m)}{\partial\alpha_i} = 0,\ i = 1, \ldots, m.$$

Непосредственные вычисления дают нам

$$\frac{1}{2}\frac{\partial(D\vec{\mathbf{z}}^m, \vec{\mathbf{z}}^m)}{\partial\alpha_i} = (D\frac{\partial\vec{\mathbf{z}}^m}{\partial\alpha_i}, \vec{\mathbf{z}}^m) =$$

$$= (D\vec{\mathbf{g}}^i, \vec{\mathbf{z}}^0 - \alpha_1\vec{\mathbf{g}}^1 - \ldots - \alpha_m\vec{\mathbf{g}}^m),\ i = 1, \ldots, m.$$

Перепишем полученную систему уравнений в матрично-векторном виде

$$\begin{bmatrix} (D\vec{\mathbf{g}}^1, \vec{\mathbf{g}}^1) & \ldots & (D\vec{\mathbf{g}}^1, \vec{\mathbf{g}}^m) \\ \vdots & & \vdots \\ (D\vec{\mathbf{g}}^m, \vec{\mathbf{g}}^1) & \ldots & (D\vec{\mathbf{g}}^m, \vec{\mathbf{g}}^m) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} (D\vec{\mathbf{z}}^0, \vec{\mathbf{g}}^1) \\ \vdots \\ (D\vec{\mathbf{z}}^0, \vec{\mathbf{g}}^m) \end{bmatrix}.$$

Матрица этой системы — это хорошо известная вам из курса алгебры матрица Грамма базиса $\{\vec{\mathbf{g}}^i\}_{i=1}^m$ в подпространстве L_m. Это, в частности, означает, что она невырождена и система однозначно разрешима.

Как нетрудно заметить, вектор правой части системы определяется через $\vec{\mathbf{z}}^0$. Если бы мы знали, чему равен $\vec{\mathbf{z}}^0$, то нам не

нужно было бы использовать никакой итерационный процесс, поскольку мы сразу могли бы посчитать точное решение системы. Однако если взять матрицу $D = A^*HA$ с любой наперёд заданной матрицей $H = H^* > 0$, то в правой части вместо неизвестного вектора ошибки $\vec{\mathbf{z}}^0$ будет стоять легко вычислимый вектор невязки $\vec{\mathbf{r}}^0 = A\vec{\mathbf{z}}^0$.

Как мы уже поняли, решать систему уравнений — задача не из простых, даже если система уравнений стала меньшего размера (в нашем случае размера m). Однако если мы выберем базис $\{\vec{\mathbf{g}}^i\}_{i=1}^m$ *D-ортогональным*, где $(D\vec{\mathbf{g}}^i, \vec{\mathbf{g}}^j) = 0$ при $i \neq j$, то матрица нашей системы станет диагональной, следовательно, легко обратимой. В этом случае

$$\alpha_i = \frac{(D\vec{\mathbf{z}}^0, \vec{\mathbf{g}}^i)}{(D\vec{\mathbf{g}}^i, \vec{\mathbf{g}}^i)}, i = 1, \ldots, m.$$

После получения коэффициентов α_i можно вычислить очередное приближение к точному решению по формуле

$$\vec{\mathbf{x}}^m = \vec{\mathbf{x}}^0 - \alpha_1\vec{\mathbf{g}}^1 - \ldots - \alpha_m\vec{\mathbf{g}}^m.$$

Повторяя указанные вычисления аналогично тому, что мы делали в циклическом методе Ричардсона, возможно получить последующие приближения к точному решению $\vec{\mathbf{x}}^k$ при любом $k > 0$.

Метод сопряжённых градиентов

Пусть матрица системы $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ самосопряжена и положительно определена. Возьмём $D = A$. Тогда мы хотим построить A-ортогональный базис в подпространстве в L_m. Мы будем строить его с помощью метода математической индукции.

Сначала сформулируем базу индукции. Пусть $\vec{\mathbf{g}}^1 = A\vec{\mathbf{z}}^0 = \vec{\mathbf{r}}^0$. Очевидно, что $\vec{\mathbf{g}}^1$ является базисом в подпространстве $L_1 = \{A\vec{\mathbf{z}}^0\}$. Если вдруг так оказалось, что он не является базисом,

т. е. $\vec{\mathbf{g}}^1 = \vec{\mathbf{0}}$, то это означает, что $\vec{\mathbf{r}}^0 = \vec{\mathbf{0}}$. В силу невырожденности положительно определённой матрицы $\vec{\mathbf{z}}^0 = \vec{\mathbf{0}}$, что означает, что $\vec{\mathbf{x}}^0 = \vec{\mathbf{x}}$, т. е. мы угадали точное решение нашей системы уравнений. С точки зрения итерационного процесса дальнейшие вычисления не требуются, и строить базис в L_1 не нужно. Заметим также, что

$$\alpha_1 = \frac{(\vec{\mathbf{r}}^0, \vec{\mathbf{g}}^1)}{(A\vec{\mathbf{g}}^1, \vec{\mathbf{g}}^1)},\ \vec{\mathbf{x}}^1 = \vec{\mathbf{x}}^0 - \alpha_1 \vec{\mathbf{g}}^1,\ \vec{\mathbf{r}}^1 = \vec{\mathbf{r}}^0 - \alpha_1 A\vec{\mathbf{g}}^1.$$

Снова обратим внимание на то, что если $\vec{\mathbf{r}}^1 = \vec{\mathbf{0}}$, то мы нашли точное решение системы и дальнейшие усилия по построению базиса в подпространстве нам не требуются. В противном случае мы замечаем, что

$$\vec{\mathbf{r}}^1 = A\vec{\mathbf{z}}^0 - \alpha_1 A^2 \vec{\mathbf{z}}^0,$$

а это значит, что $\vec{\mathbf{r}}^1 \in L_2$. Кроме того, $(\vec{\mathbf{r}}^1, \vec{\mathbf{g}}^1) = 0$ в силу выбора коэффициента α_1. Обратите внимание на то, что $\alpha_1 = 0$ только если $(\vec{\mathbf{r}}^0, \vec{\mathbf{g}}^1) = 0$, а это возможно только в том случае, если $\vec{\mathbf{r}}^0 \equiv \vec{\mathbf{g}}^1 = \vec{\mathbf{0}}$ — этот вариант мы уже обсудили. Итак, весьма длинная база индукции сформулирована.

Предположим теперь, что после k шагов мы построили A-ортогональный базис в подпространстве $L_k = \{\vec{\mathbf{g}}^1, \ldots, \vec{\mathbf{g}}^k\}$. При этом у нас получилось, что $\vec{\mathbf{r}}^k \in L_{k+1}$ и $\vec{\mathbf{r}}^k$ ортогонален (в классическом евклидовом скалярном произведении!) подпространству L_k, т. е. вектора невязки у нас ортогональны соответствующему подпространству в евклидовом смысле, а базис — A-ортогонален.

Определим теперь $\vec{\mathbf{g}}^{k+1} = \vec{\mathbf{r}}^k - \gamma_1 \vec{\mathbf{g}}^1 - \ldots - \gamma_k \vec{\mathbf{g}}^k$ таким образом, чтобы $(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^i) = 0,\ i = 1, \ldots, k$. Т. е. мы хотим сделать его A-ортогональным векторам из базиса подпространства L_k. В силу предположения индукции мы получаем, что коэффициенты γ_i легко вычисляются по формуле

$$\gamma_i = \frac{(A\vec{\mathbf{r}}^k, \vec{\mathbf{g}}^i)}{(A\vec{\mathbf{g}}^i, \vec{\mathbf{g}}^i)},\ i = 1, \ldots, k$$

ввиду того, что $(A\vec{\mathbf{g}}^i, \vec{\mathbf{g}}^j) = 0,\ i \neq j$ в диапазоне $1, \ldots, k$. Нетрудно видеть, что при таком построении A-ортогональных базисных векторов $A\vec{\mathbf{g}}^i \in L_{i+1}$ и $\vec{\mathbf{r}}^k$ ортогонален в евклидовом смысле всем подпространствам $L_i \in L_k,\ i = 1, \ldots, k$. Это значит, что в силу самосопряжённости матрицы A $(A\vec{\mathbf{r}}^k, \vec{\mathbf{g}}^{i)} = (\vec{\mathbf{r}}^k, A\vec{\mathbf{g}}^i) = 0,\ i = 1, \ldots, k-1$. Следует обратить внимание на диапазон для индекса i. Стандартная ошибка студентов заключается в добавлении индекса k в диапазон, что незамедлительно влечёт $\gamma_k = 0$, и тогда формула для $\vec{\mathbf{g}}^{k+1}$ теряет смысл.

Сделаем шаг индукции. Как мы только что увидели, при

$$\gamma_k = \frac{(A\vec{\mathbf{r}}^k, \vec{\mathbf{g}}^k)}{(A\vec{\mathbf{g}}^k, \vec{\mathbf{g}}^k)}$$

мы имеем $\vec{\mathbf{g}}^{k+1} = \vec{\mathbf{r}}^k - \gamma_k \vec{\mathbf{g}}^k$ и $\{\vec{\mathbf{g}}^1, \ldots, \vec{\mathbf{g}}^{k+1}\}$ — базис в подпространстве L_{k+1}. Если это не так, то $\vec{\mathbf{r}}^k \in L_k$ и $\vec{\mathbf{g}}^{k+1} = \vec{\mathbf{0}}$, т. е. решение системы уравнений найдено на предыдущем шаге, и дальнейшие вычисления не требуются.

По формуле для коэффициента α_{k+1}, с учётом уже полученной A-ортогональности векторов $\vec{\mathbf{g}}^i,\ i = 1, \ldots, k+1$, мы получаем

$$\alpha_{k+1} = \frac{(\vec{\mathbf{r}}^0, \vec{\mathbf{g}}^{k+1})}{(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^{k+1})} = \frac{(\vec{\mathbf{r}}^1 - \alpha_1 A\vec{\mathbf{g}}^1, \vec{\mathbf{g}}^{k+1})}{(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^{k+1})} =$$

$$= \frac{(\vec{\mathbf{r}}^1, \vec{\mathbf{g}}^{k+1})}{(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^{k+1})} = \ldots = \frac{(\vec{\mathbf{r}}^k, \vec{\mathbf{g}}^{k+1})}{(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^{k+1})}.$$

При этом у нас получается, что $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \alpha_{k+1}\vec{\mathbf{g}}^{k+1}$ и $\vec{\mathbf{r}}^{k+1} = \vec{\mathbf{r}}^k - \alpha_{k+1} A\vec{\mathbf{g}}^{k+1}$.

Из формулы для невязки незамедлительно следует, что скалярное произведение $(\vec{\mathbf{r}}^{k+1}, \vec{\mathbf{g}}^{k+1}) = 0$. По предположению индукции $\vec{\mathbf{r}}^k$ ортогонален подпространству L_k. По построению имеем $(A\vec{\mathbf{g}}^{k+1}, \vec{\mathbf{g}}^i) = 0,\ i = 1, \ldots, k$. В результате получаем, что $\vec{\mathbf{r}}^{k+1}$ ортогонален подпространству L_{k+1}.

Из формулы для $\vec{\mathbf{g}}^{k+1}$ незамедлительно следует, что $\vec{\mathbf{r}}^k \in L_{k+1}$. Это означает, что существуют некоторые коэффициенты ν_i — такие, что $\vec{\mathbf{r}}^k = \nu_1 A\vec{\mathbf{z}}^0 + \ldots + \nu_{k+1} A^{k+1}\vec{\mathbf{z}}^0$. Аналогично получаем, что существуют некоторые коэффициенты μ_i — такие, что $\vec{\mathbf{g}}_{k+1} = \mu_1 A\vec{\mathbf{z}}^0 + \ldots + \mu_{k+1} A^{k+1}\vec{\mathbf{z}}^0$. По формуле для невязки мы незамедлительно получаем, что существуют некоторые коэффициенты β_i — такие, что $\vec{\mathbf{r}}^{k+1} = \beta_1 A\vec{\mathbf{z}}^0 + \ldots + \beta_{k+2} A^{k+2}\vec{\mathbf{z}}^0$, т. е. $\vec{\mathbf{r}}^{k+1} \in L_{k+2}$ по определению подпространства L_{k+2}.

Таким образом, по методу математической индукции, формулы верны для любого k. Приведённые формулы для построения $\vec{\mathbf{g}}^i$, $\vec{\mathbf{x}}^i$, $\vec{\mathbf{r}}^i$ через коэффициенты γ_i и α_i называются формулами *метода сопряжённых градиентов.*

Теорема 33. *Если матрица системы уравнений такова, что $A = A^* > 0$, то метод сопряжённых градиентов сходится к точному решению системы уравнений не более чем за n итераций, где n — размерность системы, при этом*

$$\|\vec{\mathbf{z}}^k\|_A \le \frac{2\gamma^m}{1+\gamma^{2m}}\|\vec{\mathbf{z}}^0\|_A,\ \gamma \equiv \frac{\sqrt{\lambda_{\max}(A)} - \sqrt{\lambda_{\min}(A)}}{\sqrt{\lambda_{\max}(A)} + \sqrt{\lambda_{\min}(A)}}.$$

Доказательство. Заметим, что как только $\vec{\mathbf{r}}^k \in L_k$, то единственным вариантом для него быть ортогональным этому же подпространству — это быть нулевым вектором. Поскольку матрица системы невырождена в силу её положительной определённости, это даст нулевой вектор ошибки, что по определению означает нахождение точного решения. Поскольку размер системы равен n, то в худшем случае мы построим подпространство L_n размерности n. После этого какова бы ни была система и начальное приближение к точному решению $\vec{\mathbf{x}}^0$, $\vec{\mathbf{r}}^n \in L_n$, что, как было сказано, будет означать, что мы нашли точное решение. Но так как невязка может оказаться в подпространстве L_k при $k < n$, то метод сопряжённых градиентов может сойтись и раньше.

Далее, поскольку мы строили метод сопряжённых градиентов

на основе идеи о минимизации нормы ошибки, то

$$\|\vec{\mathbf{z}}^k\|_A = \min_{\vec{\tau}} \|S_k(A)\vec{\mathbf{z}}^0\|_A \leq \|S_k(A)\vec{\mathbf{z}}^0\|_A$$

для любого набора τ_i и, в частности, для набора τ_i, построенного для $\lambda(A) \in [\lambda_{\min}(A), \lambda_{\max}(A)]$ по формуле на основе корней полинома Чебышёва степени k. Из алгебры известно, что матрица перестановочна со своим матричным полиномом и что для нашей матрицы существует квадратный корень, поэтому

$$\|S_k(A)\vec{\mathbf{z}}^0\|_A = \|S_k(A)A^{1/2}\vec{\mathbf{z}}^0\|_2.$$

По теореме 32

$$\|S_k(A)A^{1/2}\vec{\mathbf{z}}^0\|_2 \leq \rho_k \|A^{1/2}\vec{\mathbf{z}}^0\|_2 = \rho_k \|\vec{\mathbf{z}}^0\|_A.$$

Отсюда следует утверждение теоремы.

При этом для получения этой оценки нам совершенно не нужно знать минимальное и максимальное собственное значение матрицы A, достаточно лишь знать, что $\lambda_{\min}(A) > 0$, что следует из положительной определённости и самосопряжённости матрицы. Более того, нам не нужно знать количество итераций k заранее, формула для оценки скорости сходимости верна при любом k. $\square$

Переобусловливатель

Если систему уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ преобразовать к системе уравнений $B^{-1}A\vec{\mathbf{x}} = B^{-1}\vec{\mathbf{b}}$, то обусловленность (число обусловленности) матрицы $B^{-1}A$ новой системы может оказаться значительно меньше обусловленности матрицы A исходной системы, и тогда влияние ошибок округления на решение системы уменьшится (как мы помним из оценок с использованием числа обусловленности, которые мы получали ранее).

Матрица B называется *переобусловливателем*, или предобусловливателем, или предобуславливателем (популярные термины, но всё равно жаргонизмы) для матрицы A и для системы уравнений $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$.

Самосопряжённые положительно определенные матрицы A и B называются эквивалентными по спектру с постоянными $C_1 \geq C_0 > 0$, если

$$C_0(B\vec{\mathbf{v}}, \vec{\mathbf{v}}) \leq (A\vec{\mathbf{v}}, \vec{\mathbf{v}}) \leq C_1(B\vec{\mathbf{v}}, \vec{\mathbf{v}}), \ \forall \vec{\mathbf{v}} \in \mathbb{C}^n(\mathbb{R}^n).$$

Теорема 34. *Если матрицы таковы, что $A = A^* > 0$ и $B = B^* > 0$, то все собственные числа матрицы $B^{-1}A$ вещественны, положительны и принадлежат интервалу $[C_0, C_1]$.*

Доказательство. Мы уже неоднократно пользовались тем, что для самосопряжённой положительно определённой матрицы существует квадратный корень с такими же свойствами. Прибегнем к этому и сейчас. Матрицы $B^{-1}A$ (вообще говоря, несамосопряжённая) и $B^{1/2}(B^{-1}A)B^{-1/2} = B^{-1/2}AB^{-1/2}$ (самосопряжённая) подобны, следовательно, у них одинаковый спектр. Так как вторая матрица самосопряжена, то у неё вещественный спектр и вещественные собственные векторы $B^{-1/2}AB^{-1/2}\vec{\mathbf{u}} = \lambda(B^{-1}A)\vec{\mathbf{u}}$. Непосредственная проверка по определению показывает, что матрица $B^{-1/2}AB^{-1/2}$ — положительно определена, так как матрица A положительно определена. Это значит, что все собственные числа положительны.

Заметим, что вектор $\vec{\mathbf{v}} \equiv B^{-1/2}\vec{\mathbf{u}}$ является собственным вектором матрицы $B^{-1}A$ с тем же самым собственным числом $\lambda(B^{-1}A)$. При этом из уравнения для собственного значения следует, что

$$\lambda(B^{-1}A) = \frac{(B^{-1}A\vec{\mathbf{v}}, B\vec{\mathbf{v}})}{(\vec{\mathbf{v}}, B\vec{\mathbf{v}})} = \frac{(A\vec{\mathbf{v}}, \vec{\mathbf{v}})}{(B\vec{\mathbf{v}}, \vec{\mathbf{v}})}$$

в силу самосопряжённости матрицы B. Доказательство теоремы теперь следует из условия эквивалентности по спектру для мат-

риц A и B, а также того факта, что приведённое выше равенство верно для любого собственного значения матрицы $B^{-1}A$. $\square$

Из теорем 34 и 28 незамедлительно следует, что для матрицы шага для ошибки $S_\tau \equiv (E - \tau B^{-1}A)$ справедливо неравенство $\rho(S_\tau) < 1 \; \forall \tau \in (0, 2/C_1)$. Кроме того, в силу самосопряжённости матриц $B^{1/2}S_\tau B^{-1/2}$ и $A^{1/2}S_\tau A^{-1/2}$ мы получаем

$$\|S_\tau\|_B = \|B^{1/2}S_\tau B^{-1/2}\|_2 = \rho(B^{1/2}S_\tau B^{-1/2}) = \rho(S_\tau) < 1$$

и

$$\|S_\tau\|_A = \|A^{1/2}S_\tau A^{-1/2}\|_2 = \rho(A^{1/2}S_\tau A^{-1/2}) = \rho(S_\tau) < 1.$$

Это означает, что *метод простой итерации с переобусловливателем* $\vec{\mathbf{x}}^{k+1} = \vec{\mathbf{x}}^k - \tau B^{-1}(A\vec{\mathbf{x}}^k - \vec{\mathbf{b}})$ сходится при $\tau \in (0, 2/C_1)$ по любой из теорем 13, 14, или 15.

Теорема 35 (часть критерия Самарского)**.** *Если матрицы таковы, что* $A = A^* > 0$ *и* $B = B^* > 0$*, то метод простой итерации с переобусловливателем сходится, если* $B > 0{,}5\tau A$ *при* $\tau > 0$.

Доказательство. Заметим, что оператор S_τ непрерывен всюду по определению, следовательно, для того, чтобы воспользоваться теоремой 22, нам осталось только получить строгое убывание функционала ошибки в какой-нибудь норме, а именно

$$\|\vec{\mathbf{z}}^{k+1}\|_A^2 = (A\vec{\mathbf{z}}^{k+1}, \vec{\mathbf{z}}^{k+1}) = (AS_\tau\vec{\mathbf{z}}^k, S_\tau\vec{\mathbf{z}}^k) =$$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - 2\tau(AB^{-1}A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) + \tau^2(AB^{-1}A\vec{\mathbf{z}}^k, B^{-1}A\vec{\mathbf{z}}^k) =$$

полагаем $\vec{\mathbf{w}}^k \equiv B^{-1}A\vec{\mathbf{z}}^k$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - 2\tau(B\vec{\mathbf{w}}^k, \vec{\mathbf{w}}^k) + \tau^2(A\vec{\mathbf{w}}^k, \vec{\mathbf{w}}^k) =$$

$$= (A\vec{\mathbf{z}}^k, \vec{\mathbf{z}}^k) - 2\tau([B - 0{,}5\tau A]\vec{\mathbf{w}}^k, \vec{\mathbf{w}}^k) < \|\vec{\mathbf{z}}^k\|_A^2$$

если матрица $B - 0{,}5\tau A$ — положительно определена и $\tau > 0$. Теорема доказана. $\square$

Обратите внимание, что выражение $B > 0{,}5\tau A$ означает, что матрица $B - 0{,}5\tau A$ — положительно определена (> 0 в смысле этого обозначения для матриц), а НЕ неравенство для соответствующих элементов матриц.

Положительно определённые матрицы

Мы уже неоднократно сталкивались с положительно определёнными матрицами, что обусловливает их важную роль при доказательстве различных утверждений из вычислительной линейной алгебры. В этом разделе мы рассмотрим основные свойства этих матриц и соберём их все в одном месте. Но для начала следует напомнить определение положительно определённой матрицы. Матрица A называется *положительно определённой* (обозначается как $A > 0$) в $\mathbb{C}^n(\mathbb{R}^n)$, если $(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) > 0 \ \forall \vec{\mathbf{x}} \in \mathbb{C}^n(\mathbb{R}^n), \vec{\mathbf{x}} \neq \vec{\mathbf{0}}$.

Теорема 36. $(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = \mathrm{Re}(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) \ \forall \vec{\mathbf{x}} \in \mathbb{C}^n \iff A = A^*$.

Доказательство. Сразу заметим, что

$$\sum_{j,k=1}^{n} a_{jk} x_k \bar{x}_j = (A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\vec{\mathbf{x}}, A^*\vec{\mathbf{x}}) = \overline{(A^*\vec{\mathbf{x}}, \vec{\mathbf{x}})} = \sum_{j,k=1}^{n} \bar{a}_{kj} \bar{x}_k x_j.$$

Если брать единичные орты в качестве вектора $\vec{\mathbf{x}}$, то мы получим, что на диагонали матрицы должны стоять вещественные числа, чтобы скалярное произведение давало вещественный результат, т. е. $a_{jj} = \bar{a}_{jj} \in \mathbb{R}$. Далее с помощью тех же единичных ортов составляем комбинации вида $\vec{\mathbf{x}} = \vec{\mathbf{e}}^j + i\vec{\mathbf{e}}^k$ и $\vec{\mathbf{x}} = \vec{\mathbf{e}}^j + \vec{\mathbf{e}}^k$ (здесь i обозначает мнимую единицу) и получаем с помощью непосредственных вычислений, что вещественные части элементов a_{jk} и a_{kj} должны совпадать, а мнимые части этих элементов — отличаться только знаком, т. е. $a_{jk} = \bar{a}_{kj}$. Только в этом случае скалярное произведение будет давать вещественный результат. По определению сопряжённой матрицы в евклидовом скалярном произведении мы получаем, что $A = A^*$, и теорема доказана. $\square$

Теорема 37. *Если* $A = A^*$, *то*

$$A > 0 \text{ в } \mathbb{C}^n(\mathbb{R}^n) \iff \forall \lambda(A) > 0,$$

$$\lambda_{\min}(A)(\vec{\mathbf{x}}, \vec{\mathbf{x}}) \leq (A\vec{\mathbf{x}}, \vec{\mathbf{x}}) \leq \lambda_{\max}(A)(\vec{\mathbf{x}}, \vec{\mathbf{x}}) \ \forall \vec{\mathbf{x}} \in \mathbb{C}^n(\mathbb{R}^n).$$

Доказательство. Из курса алгебры известно, что у самосопряжённой матрицы спектр вещественный и есть базис из собственных векторов $A\vec{\mathbf{v}}_j = \lambda_j(A)\vec{\mathbf{v}}_j, j = 1, \ldots, n$. Подставляя в скалярное произведение $(A\vec{\mathbf{x}}, \vec{\mathbf{x}})$ собственные вектора в качестве $\vec{\mathbf{x}}$, мы сразу получаем, что матрица может быть положительно определённой тогда и только тогда, когда все $\lambda_j(A) > 0$. Первое утверждение из теоремы доказано.

Далее раскладываем вектор $\vec{\mathbf{x}}$ по ортонормированному базису из собственных векторов $\vec{\mathbf{x}} = x_1\vec{\mathbf{v}}^1 + \ldots + x_n\vec{\mathbf{v}}^n$ и подставляем в скалярное произведение. В результате получаем $(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = \lambda_1(A)|x_1|^2 + \ldots + \lambda_n(A)|x_n|^2$. Так как $(\vec{\mathbf{x}}, \vec{\mathbf{x}}) = |x_1|^2 + \ldots + |x_n|^2$, то мы сразу переходим ко второму утверждению теоремы с учётом положительности всех собственных чисел. □

Обратите внимание, что указанная теорема верна только для самосопряжённых матриц. Для несамосопряжённых матриц из положительности собственных чисел не следует положительная определённость! Для проверки понимания материала, придумайте треугольную матрицу размера 2×2, у которой все собственные числа положительные, но она не положительно определена.

Теорема 38 (критерий Сильвестра)**.** *Если* $A = A^*$, *то*

$$A > 0 \text{ в } \mathbb{C}^n(\mathbb{R}^n) \iff \forall j = 1, \ldots, n \ |A_j| > 0.$$

Доказательство. По теореме 5 для матрицы существует разложение $A = LDU$, если главные миноры матрицы невырождены.

Так как $A = A^*$, то $L = U^*$. Для любого минора справедливо разложение $A_j = L_j D L_j^*$, и потому $|A_j| = |L_j||D||L_j^*| = 1 \cdot \prod_{k=1}^{j} d_k \cdot 1$, ввиду чего мы незамедлительно получаем, что матрица будет положительно определённой тогда и только тогда, когда все главные миноры положительны. □

Обратите внимание, что и эта теорема верна только для самосопряжённых матриц. Для несамосопряжённых матриц из положительности главных миноров не следует положительная определённость! Для проверки понимания материала, приведите пример треугольной матрицы размера 2×2, у которой все главные миноры положительные, но она не положительно определена.

Теорема 39. *$A > 0$ в $\mathbb{R}^n \iff A + A^* > 0$ в $\mathbb{R}^n$.*

Доказательство.

$$(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\frac{1}{2}A\vec{\mathbf{x}}, \vec{\mathbf{x}}) + (\frac{1}{2}A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\frac{1}{2}A\vec{\mathbf{x}}, \vec{\mathbf{x}}) + (\vec{\mathbf{x}}, \frac{1}{2}A^*\vec{\mathbf{x}}) =$$

поскольку в этой теореме мы работаем с вещественнозначными скалярными произведениями

$$= (\frac{1}{2}A\vec{\mathbf{x}}, \vec{\mathbf{x}}) + (\frac{1}{2}A^*\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\frac{A + A^*}{2}\vec{\mathbf{x}}, \vec{\mathbf{x}}).$$

Отсюда следует утверждение теоремы по определению положительной определённости матрицы. □

Теорема 40. *Если $A = -A^*$ — вещественная кососимметричная матрица, то $A = 0$ в $\mathbb{R}^n$.*

Доказательство. Для любого вектора $\vec{\mathbf{x}}$

$$(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\vec{\mathbf{x}}, A^*\vec{\mathbf{x}}) = (\vec{\mathbf{x}}, -A\vec{\mathbf{x}}) = -(A\vec{\mathbf{x}}, \vec{\mathbf{x}}).$$

Поскольку только одно число равно самому себе со знаком «минус», то $(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = 0$, что и означает $A = 0$ в условии теоремы. □

Данная теорема свидетельствует о том, что скалярное произведение с кососимметричной матрицей и вещественными векторами всегда равно нулю. Обратите внимание, что только в этой теореме курса обозначение $A = 0$ означает не нулевую матрицу, а $(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = 0$.

Теорема 41. *Если* $A > 0$ *в* $\mathbb{R}^n$, *то* $\operatorname{Re}\lambda(A) > 0$.

Доказательство. Нетрудно убедиться, что

$$(A\vec{\mathbf{x}}, \vec{\mathbf{x}}) = (\frac{A + A^*}{2}\vec{\mathbf{x}}, \vec{\mathbf{x}}) + (\frac{A - A^*}{2}\vec{\mathbf{x}}, \vec{\mathbf{x}}) =$$

второе скалярное произведение идёт с кососимметричной матрицей при любом выборе матрицы A, ввиду чего, по теореме 40,

$$= (\frac{A + A^*}{2}\vec{\mathbf{x}}, \vec{\mathbf{x}}).$$

Непосредственная проверка показывает, что

$$\lambda(\frac{A + A^*}{2}) = \operatorname{Re}\lambda(A)$$

и применение теоремы 37 завершает доказательство. □

Для понимания материала постройте пример вещественной несимметричной, но положительно определённой в $\mathbb{R}^n$, матрицы.

Лекция №11

Задача о поиске собственных значений и собственных векторов

Задача формулируется очень просто: для матрицы A нужно найти числа λ и ненулевые векторы $\vec{\mathbf{x}}$ — такие, что $A\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$. Тогда, по определению, λ — собственное значение, а $\vec{\mathbf{x}}$ — собственный вектор. В этом разделе обозначение $\vec{\mathbf{x}}$ будет использоваться уже не для обозначения решения системы уравнений, а для обозначения собственного вектора.

Корректность задачи на собственные значения

Исследуем корректность поставленной задачи, чтобы быть уверенным, что мы решаем задачу, которую можно решить адекватно. Для этого нам потребуется вспомнить некоторые сведения из алгебры, в частности, о том, что все собственные значения матрицы являются корнями характеристического полинома степени не выше, чем размер матрицы

$$P_n(\lambda) = |A - \lambda E| = (-1)^n\lambda^n + p_{n-1}\lambda^{n-1} + \ldots + p_1\lambda + p_0.$$

При этом коэффициенты этого полинома являются непрерывными функциями элементов матрицы A.

Пусть δA — матрица с некоторыми малыми (в смысле абсолютных значений) элементами. Для матрицы $A + \delta A$ мы можем построить характеристический полином $P_{n,\delta}(\lambda)$.

Теорема 42. $\lim_{\delta A \to 0} P_{n,\delta}(\lambda) = P_n(\lambda)\ \forall\lambda \in \mathbb{C}$.

Доказательство. Воспользуемся непрерывной зависимостью коэффициентов характеристического полинома от значений матрицы и незамедлительно получим утверждение теоремы. □

Обратите внимание на то, что в теореме не говорится, каким именно образом элементы матрицы δA стремятся к нулю.

Теорема 43. *В любом круге на комплексной плоскости с центром в любой точке λ_c и радиуса $\sqrt[n]{|P_n(\lambda_c)|}$ лежит хотя бы один корень любого полинома $P_n(\lambda)$.*

Доказательство. Применим разложение функции в ряд Тэйлора, известное из курса математического анализа,

$$P_n(\lambda) = P_n(\lambda_c) + \frac{P_n'(\lambda_c)}{1!}(\lambda - \lambda_c)^1 + \ldots + \frac{P_n^{(n)}(\lambda_c)}{n!}(\lambda - \lambda_c)^n \equiv Q_n(z),$$

где $z \equiv \lambda - \lambda_c$.

Пусть $z_1, \ldots, z_n$ — корни полинома $Q(z)$. Пусть $z_{\min}$ — это корень с минимальным по модулю абсолютным значением. Если их несколько, то можно взять любой из них. Вспомним алгебру и представим полином в виде $Q(z) = (z - z_1)\ldots(z - z_n)$, откуда сразу следует

$$|P_n(\lambda_c)| = |Q(0)| = |z_1 \ldots z_n| \geq |z_{\min}|^n = |\lambda_{\min} - \lambda_c|^n.$$

Обратите внимание, что старший коэффициент полинома $P_n(\lambda)$, следовательно, и полинома $Q(z)$ равен по модулю единице, поэтому произведение корней по модулю идёт со множителем 1, а не с каким-либо другим значением. Из этой формулы мы сразу получаем, что корень $\lambda_{\min}$, соответствующий $z_{\min}$, лежит в круге на комплексной плоскости радиуса $\sqrt[n]{|P_n(\lambda_c)|}$. Теорема доказана. □

Данное утверждение справедливо для любого полинома, не обязательно характеристического, в том числе и с вещественными коэффициентами.

Теорема 44. *Если $\lambda_1, \ldots, \lambda_n$ — все корни полинома $P_n(\lambda)$ без учёта кратности, то существует такая нумерация всех корней*

$\lambda_{1,\delta}, \ldots, \lambda_{n,\delta}$ *полинома* $P_{n,\delta}(\lambda)$, *что* $\lambda_{k,\delta} \to \lambda_k,\ k = 1, \ldots, n$ *при* $\delta A \to 0$.

Доказательство. Для доказательства теоремы применим метод математической индукции. При $n = 1$ имеем $\lambda_{1,\delta} = p_{0,\delta} \to p_0 = \lambda_1$, так как для полинома первой степени со старшим коэффициентом равным (-1) корень вычисляется легко, и он равен в точности младшему коэффициенту. При этом этот младший коэффициент в точности равен единственному коэффициенту матрицы размера 1×1, а потому сходимость очевидна.

Предположим, что мы построили нумерацию корней при $n < k$. Рассмотрим случай $n = k$. По теореме 43 существует корень $\lambda_{1,\delta}$ полинома $P_{k,\delta}(\lambda)$ — такой, что $|\lambda_{1,\delta} - \lambda_1| \leq \sqrt[k]{|P_{k,\delta}(\lambda_1)|}$. По теореме 42 $P_{k,\delta}(\lambda_1) \to 0$ при $\delta A \to 0$, так как λ_1 — корень полинома $P_k(\lambda)$, к которому стремится полином $P_{1,\delta}(\lambda)$.

Как известно из курса алгебры, если мы знаем один корень полинома, то можем разложить полином в произведение вида $P_k(\lambda) = (\lambda - \lambda_1)R_{k-1}(\lambda)$, где $R_{k-1}(\lambda)$ — некоторый полином степени $k-1$. Аналогично получаем $P_{k,\delta}(\lambda) = (\lambda - \lambda_{1,\delta})R_{k-1,\delta}(\lambda)$. По предположению индукции $R_{k-1,\delta}(\lambda) \to R_{k-1}(\lambda)$ и

$$\lambda_{2,\delta} \to \lambda_2, \ldots, \lambda_{k,\delta} \to \lambda_k.$$

Теорема доказана. □

Таким образом, мы доказали, что при малых возмущениях (например, при неизбежном появлении ошибок округления в процессе вычислений), всё равно будем получать собственные числа, близкие к точным. Здесь опять опускаем сложный вопрос о том, можем ли мы вычислить собственные значения близкие к точным с практической точки зрения, поскольку миллиард может быть близким к единице в некоторой норме. Для проверки понимания материала приведите пример нормы, в которой это утверждение

верно. С практической точки зрения эти величины различаются слишком сильно, чтобы считать миллиард приближением к единице.

Степенной метод вычисления максимального собственного значения матрицы

Мы будем рассматривать только самосопряжённые положительно полуопределённые матрицы $A = A^* \geq 0$, чтобы все собственные числа были вещественными и неотрицательными.

В случае несамосопряжённых матриц, помимо проблемы, связанной с наличием комплексных собственных чисел у вещественных матриц, мы имеем проблему неустойчивости вычислений, связанную с клетками Жордана, которые легко распадаются на клетки меньшего размера при сколь угодно малых изменениях элементов матрицы. Такая проблема не имеет адекватного решения, поэтому для несамосопряжённой матрицы B рекомендуется искать сингулярные числа вместо собственных, поскольку в таком случае мы будем работать с самосопряжённой матрицей B^*B.

В том случае, если матрица не является положительно определённой, нам придётся решать проблему с разными знаками у собственных чисел. Чтобы этого не делать, рекомендуется сдвинуть спектр матрицы так, чтобы он стал неотрицательным, например, вместо матрицы $C = C^*$ со знакопеременным спектром рассмотреть матрицу $C + \|C\|_1 E$, которая будет положительно полуопределена. Для проверки понимания материала следует убедиться в возможности обоснования положительной полуопределённости новой матрицы.

Итерационный процесс

$$\vec{\mathbf{x}}^{k+1} = A\frac{\vec{\mathbf{x}}^k}{\|\vec{\mathbf{x}}^k\|},\ k = 0, 1, \ldots,\ \vec{\mathbf{x}}^0 \neq \vec{\mathbf{0}}$$

называется *степенным методом* вычисления максимального собственного числа матрицы $A = A^* \geq 0$.

Теорема 45. *Если проекция начального вектора* $\vec{\mathbf{x}}^0$ *на линейную оболочку собственных векторов, соответствующих максимальному собственному значению матрицы* $A = A^* \geq 0$, *не равна* $\vec{\mathbf{0}}$, *то* $\lim_{k\to\infty} \vec{\mathbf{x}}^k = \vec{\mathbf{x}}$ *и* $\lim_{k\to\infty} \|\vec{\mathbf{x}}^k\| = \rho(A)$, *где* $A\vec{\mathbf{x}} = \rho(A)\vec{\mathbf{x}}$.

Доказательство. Из алгебры известно, что так как матрица A самосопряжена, то у неё вещественный спектр и есть базис из собственных векторов. Так как матрица положительно полуопределена, то все её собственные числа неотрицательны. Для проверки понимания материала следует убедиться, что возможно доказать последнее утверждение. Таким образом, по определению спектрального радиуса $\rho(A) = \lambda_{\max}(A)$.

Пусть $0 \leq \lambda_1 \leq \ldots \leq \lambda_r < \lambda_{r+1} = \ldots = \lambda_n$ — собственные значения матрицы A, а $\vec{\mathbf{v}}^1, \ldots, \vec{\mathbf{v}}^n$ — соответствующие им собственные вектора. Следует обратить внимание, что максимальных собственных чисел может быть несколько штук, например, $n - r$. При этом $\lambda_{r+1} = \ldots = \lambda_n = \rho(A)$. Для проверки понимания материала приведите пример матрицы, у которой все собственные числа максимальные, т. е. $r = 0$.

Поскольку собственные вектора матрицы A образуют базис, то мы можем разложить по нему любой вектор, в том числе и начальный вектор в степенном методе:

$$\vec{\mathbf{x}}^0 = \alpha_1\vec{\mathbf{v}}^1 + \ldots + \alpha_r\vec{\mathbf{v}}^r + \alpha_{r+1}\vec{\mathbf{v}}^{r+1} + \ldots + \alpha_n\vec{\mathbf{v}}^n =$$

$$= \alpha_1\vec{\mathbf{v}}^1 + \ldots + \alpha_r\vec{\mathbf{v}}^r + \vec{\mathbf{y}},$$

где $\vec{\mathbf{y}} \equiv \alpha_{r+1}\vec{\mathbf{v}}^{r+1} + \ldots + \alpha_n\vec{\mathbf{v}}^n$ и $\vec{\mathbf{y}} \neq \vec{\mathbf{0}}$ по предположению теоремы. При этом нетрудно проверить, что $A\vec{\mathbf{y}} = \rho(A)\vec{\mathbf{y}}$, т. е. вектор $\vec{\mathbf{y}}$ — также собственный, соответствующий максимальному собственном значению.

Тогда

$$A^k\vec{\mathbf{x}}^0 = \alpha_1 A^k\vec{\mathbf{v}}^1 + \ldots + \alpha_r A^k\vec{\mathbf{v}}^r + A^k\vec{\mathbf{y}} = \alpha_1\lambda_1^k\vec{\mathbf{v}}^1 + \ldots + \alpha_r\lambda_r^k\vec{\mathbf{v}}^r +$$

$$+\rho^k(A)\vec{\mathbf{y}} = \rho^k(A)\left[\alpha_1\left(\frac{\lambda_1}{\rho(A)}\right)^k\vec{\mathbf{v}}^1 + \ldots + \alpha_r\left(\frac{\lambda_r}{\rho(A)}\right)^k\vec{\mathbf{v}}^r + \vec{\mathbf{y}}\right].$$

Так как

$$0 \le \frac{\lambda_1}{\rho(A)} \le \ldots \le \frac{\lambda_r}{\rho(A)} < 1$$

и

$$\vec{\mathbf{x}}^k = \frac{A\vec{\mathbf{x}}^{k-1}}{\|\vec{\mathbf{x}}^{k-1}\|} = \ldots = \frac{A^k\vec{\mathbf{x}}^0}{\|A^{k-1}\vec{\mathbf{x}}^0\|},$$

то, как следует из курса математического анализа,

$$\|\vec{\mathbf{x}}^k\| =$$

$$= \rho(A)\frac{\|\alpha_1\left(\frac{\lambda_1}{\rho(A)}\right)^k\vec{\mathbf{v}}^1 + \ldots + \alpha_r\left(\frac{\lambda_r}{\rho(A)}\right)^k\vec{\mathbf{v}}^r + \vec{\mathbf{y}}\|}{\|\alpha_1\left(\frac{\lambda_1}{\rho(A)}\right)^{k-1}\vec{\mathbf{v}}^1 + \ldots + \alpha_r\left(\frac{\lambda_r}{\rho(A)}\right)^{k-1}\vec{\mathbf{v}}^r + \vec{\mathbf{y}}\|} \to \rho(A)$$

при $k \to \infty$. Аналогично получаем, что

$$\vec{\mathbf{x}}^k = \rho(A)\frac{\alpha_1\left(\frac{\lambda_1}{\rho(A)}\right)^k\vec{\mathbf{v}}^1 + \ldots + \alpha_r\left(\frac{\lambda_r}{\rho(A)}\right)^k\vec{\mathbf{v}}^r + \vec{\mathbf{y}}}{\|\alpha_1\left(\frac{\lambda_1}{\rho(A)}\right)^{k-1}\vec{\mathbf{v}}^1 + \ldots + \alpha_r\left(\frac{\lambda_r}{\rho(A)}\right)^{k-1}\vec{\mathbf{v}}^r + \vec{\mathbf{y}}\|} \to$$

$$\to \rho(A)\frac{\vec{\mathbf{y}}}{\|\vec{\mathbf{y}}\|}, \text{ при } k \to \infty,$$

и теорема доказана. □

Мы можем выбрать любую норму для формулы степенного метода, на результате это не скажется — мы всё равно будем сходиться к максимальному собственному числу и некоторому собственному вектору, ему соответствующему.

Условие о том, что начальный вектор $\vec{\mathbf{x}}^0$ содержит в разложении по базису из собственных векторов ненулевую компоненту при собственному векторе, соответствующем максимальному собственному числу, проверить весьма затруднительно. Однако на практике, даже если это не так, из-за ошибок округления эта ненулевая компонента практически всегда появляется. Поэтому основная и сложная проблема для степенного метода — понять, когда можно остановить итерации, чтобы не столкнуться с ситуацией нахождения не максимального собственного значения, а некоторого другого собственного числа. Поскольку простого решения не существует, то только понимание исходной естественнонаучной задачи может подсказать, нашли ли вы максимальное собственное значение или ещё нет. Хорошим подспорьем в деле определения момента остановки служит проверка компонент $x_i^k,\ i = 1, \ldots, n$ вектора $\vec{\mathbf{x}}^k$ ВМЕСТЕ С $\|\vec{\mathbf{x}}^k\|$. Если они практически не меняются от итерации к итерации, то можно считать, что метод сошёлся. Если же компоненты вектора $\vec{\mathbf{x}}^k$ сильно меняются от итерации к итерации, например, когла достаточно «большие» компоненты вектора меняют знак или свою величину, то итерации стоит продолжить.

Степенной метод вычисления минимального собственного значения матрицы

Как и в предыдущем параграфе, здесь мы будем рассматривать только самосопряжённые положительно полуопределённые матрицы $A = A^* \geq 0$, чтобы все собственные числа были вещественными и неотрицательными.

Идея метода очень проста. Пусть некоторое неотрицательное число $\beta \geq \rho(A)$, тогда максимальное собственное число матрицы $\beta E - A$ (оно же будет и спектральным радиусом этой матрицы) будет связано с минимальным собственным числом матрицы A по формуле $\rho(\beta E - A) = \beta - \lambda_{\min}(A)$. Это равенство проверяется

через определение собственного числа матрицы.

Чтобы выбрать число β, нужно всего лишь вспомнить уже упоминавшееся в данном курсе утверждение о том, что $\rho(A) \leq \|A\|$. Ранее мы выяснили, что легко вычислить «первую» или «бесконечную» нормы матрицы A. Проверьте понимание материла, ответив на вопрос о том, почему в нашем случае эти две нормы будут давать одинаковый результат.

Итерационный процесс

$$\vec{\mathbf{x}}^{k+1} = (\|A\|_1 E - A)\frac{\vec{\mathbf{x}}^k}{\|\vec{\mathbf{x}}^k\|},\ k = 0, 1, \ldots,\ \vec{\mathbf{x}}^0 \neq \vec{\mathbf{0}}$$

называется *степенным методом* вычисления минимального собственного числа матрицы $A = A^* \geq 0$.

Теорема 46. *Если проекция начального вектора $\vec{\mathbf{x}}^0$ на линейную оболочку собственных векторов, соответствующих минимальному собственному значению матрицы $A = A^* \geq 0$, не равна $\vec{\mathbf{0}}$, то $\lim_{k\to\infty} \vec{\mathbf{x}}^k = \vec{\mathbf{x}}$ и $\lim_{k\to\infty}(\|A\|_1 - \|\vec{\mathbf{x}}^k\|) = \lambda_{\min}(A)$, где $A\vec{\mathbf{x}} = \lambda_{\min}(A)\vec{\mathbf{x}}$.*

Доказательство. Аналогично доказательству теоремы 45. Следует убедиться, что вы можете сделать необходимыми изменения в доказательстве. □

Применение ортогонализации и степенного метода для вычисления очередного собственного значения

Предположим, что собственное значение $\lambda_n(A) = \rho(A)$ и соответствующий ему собственный вектор (какой–то!) $\vec{\mathbf{v}}^n$ матрицы A уже вычислили приближенно, например, степенным методом, т. е. мы имеем $\hat{\lambda}_n \approx \lambda_n$ и $\vec{\hat{v}}_n \approx \vec{\mathbf{v}}^n$. Построим симметричную положительно определенную матрицу $\hat{A}_{n-1} \equiv \hat{P}_n A \hat{P}_n$, где матрица

$\hat{P}_n \equiv E - \vec{\hat{v}}_n \vec{\hat{v}}_n^*$ – ортогональный проектор на подпространство, ортогональное вектору $\vec{\hat{v}}_n$. Следует убедиться, что вы можете проверить требуемые свойства матрицы $\hat{P}_n$.

Непосредственная проверка показывает, что спектр матрицы A_{n-1} ПРИ УСЛОВИИ $\hat{\lambda}_n = \lambda_n$ и $\vec{\hat{v}}_n = \vec{\mathbf{v}}_n$ состоит из собственных значений $\lambda_1, \ldots, \lambda_{n-1}$ матрицы A и нуля, где собственный вектор $\vec{\mathbf{v}}_n$ принадлежит ее ядру. Следовательно, применяя степенной метод для матрицы $\hat{A}_{n-1}$, по теореме 45 мы получим приближение к $\lambda_{n-1}(A)$ и $\vec{\mathbf{v}}_{n-1}$ — очередным собственным значению и вектору матрицы A.

Эту процедуру можно продолжать до тех пор, пока не получим все собственные значения. Если максимальных собственных значений несколько, то в результате применения предложенного подхода первые $n - r$ собственных значений будут одинаковыми, меняться будут только собственные вектора. При этом собственные вектора будут ортогональны (с точностью до ошибок округления) друг другу.

Стоит обратить внимание на то, что указанная процедура весьма чувствительна к ошибкам округления, ввиду чего в зависимости от свойств конкретной матрицы результат может варьироваться в диапазоне от приемлемого до ужасного. Только понимание исходной естественнонаучной задачи, как мы подчёркивали ранее, позволит понять, насколько хорошо вычислены приближения к собственным числам и векторам.

Лекция №12

Метод бисекций (метод деления пополам)

Рассмотрим другой подход к вычислению собственных чисел матрицы. Его идея базируется на хорошо известном из математического анализа методе деления пополам, или методе бисекций, а также законе инерции, известном из курса алгебры. Если самосопряжённую матрицу A конгруэнтным преобразованием привести к диагональному виду $D = T^*AT$, где $|T| \neq 0$, то от матрицы T (другими словами, способа преобразования) не зависят следующие величины:

- $\sigma_-(A)$ – количество отрицательных элементов на диагонали матрицы D;
- $\sigma_0(A)$ – количество нулевых элементов на диагонали матрицы D;
- $\sigma_+(A)$ – количество положительных элементов на диагонали матрицы D.

Теорема 47. *Если самосопряжённая матрица $A = A^*$ такова, что все её главные миноры невырождены, т. е. $|A_k| \neq 0,\ k = 1, \ldots, n$, то количество ее отрицательных собственных значений $\sigma_-(A)$ определяется числом перемен знака в последовательности Штурма $\{1, |A_1|, \ldots, |A_n| \equiv |A|\}$.*

Доказательство. Нам известно из теоремы 5, что если все главные миноры невырождены, т. е. $|A_k| \neq 0,\ k = 1, \ldots, n$, то в силу самосопряжённости матрицы A мы имеем LDU-разложение вида $A = LDL^*$, где $D = \mathrm{diag}\{d_1, \ldots, d_n\}$, $|L| = |L^*| = 1$, $|A_k| = d_1 \ldots d_k$. Следовательно, в этом случае за конечное и разумное число действий мы можем определить

$$\sigma(A) \equiv \{\sigma_-(A), \sigma_0(A), \sigma_+(A)\}.$$

При этом, в силу условия теоремы о невырожденности $A_n \equiv A$, мы имеем $\sigma_0(A) = 0$ до начала каких бы то ни было вычислений.

Опираясь на курс алгебры, мы получаем, что самосопряжённая матрица $A = A^*$ преобразованием подобия ортогональной матрицей Q из собственных векторов (она же задаёт конгруэнтное преобразование в силу ортогональности базиса из собственных векторов) приводится к диагональному виду

$$\Lambda \equiv \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\} = Q\Lambda Q^*.$$

Следовательно,

- $\sigma_-(A)$ – количество отрицательных элементов на диагонали матрицы A;
- $\sigma_0(A)$ – количество нулевых элементов на диагонали матрицы A;
- $\sigma_+(A)$ – количество положительных элементов на диагонали матрицы A.

И, что самое важное, используя LDL^*-разложение, мы можем эти числа определить.

Из равенства $A = LDL^* = Q\Lambda Q^*$ в силу $|L| = |L^*| = |Q| = |Q^*| = 1$ и $|\Lambda| = |D|$ незамедлительно следует, что любая смена знака в последовательности Штурма соответствует одному отрицательному собственному значению. Для проверки понимания материала убедитесь, что вы способны доказать утверждение о том, что $|Q| = |Q^*| = 1$, а также завершить доказательство теоремы, указав, что происходит с главными минорами матрицы A. $\square$

Как мы недавно удостоверились, все собственные числа самосопряжённой матрицы лежат в интервале $[-\|A\|_1, \|A\|_1]$, так как

$\rho(A) \le \|A\|_1 = \|A\|_\infty$. Пусть $a_0 \equiv -\|A\|_1,\ b_0 \equiv \|A\|_1$ и собственные числа матрицы A упорядочены по возрастанию $\lambda_1 \le \ldots \le \lambda_n$.

Для фиксированного номера $j \in 1, \ldots, n$ определим, в какой половине интервала $[a_0, b_0]$ лежит собственное число λ_j. Для этого вычислим $\sigma_-(A - c_0 E)$ – количество собственных значений меньших $c_0 \equiv \frac{a_0+b_0}{2}$. Если $\sigma_-(A - c_0 E) \ge j$, то $\lambda_j \in [a_0, c_0] \equiv [a_1, b_1]$, иначе $\lambda_j \in [c_0, b_0] \equiv [a_1, b_1]$. Через k таких шагов получим $\lambda_j \in [a_k, b_k]$, при этом

$$b_k - a_k = \frac{\|A\|_1}{2^{k-1}} \to 0$$

при $k \to \infty$, т. е. мы можем получить оценку искомого собственного числа с любой точностью. Это и есть *метод бисекций.*

Приведение самосопряжённых матриц к трёхдиагональному виду

Вести расчёт количества отрицательных собственных чисел с помощью LDU-разложения будет весьма затратной процедурой, поскольку это разложение придётся пересчитывать при каждом изменении середины интервала c_j. Поэтому мы рассмотрим альтернативный подход, уменьшающий вычислительные затраты, который основан на приведении самосопряжённой матрицы к трёхдиагональному виду ортогональным преобразованим подобия с помощью матриц вращения. Как и в методе вращений, под элементарной матрицей вращений $Q_{i,j}$ мы будем подразумевать матрицу, отличающуюся от единичной максимум в четырёх элементах:

$$(Q_{i,j})_{i,i} = \bar{c}_{ij},\ (Q_{i,j})_{j,j} = c_{ij},\ (Q_{i,j})_{i,j} = -\bar{s}_{ij},\ (Q_{i,j})_{j,i} = s_{ij},$$

$$|c_{ij}|^2 + |s_{ij}|^2 = 1.$$

Теорема 48. *Любая самосопряжённая матрица $A = A^*$ подобна трёхдиагональной вещественной матрице.*

Доказательство. Для доказательства теоремы применим метод математической индукции. Абсолютно также, как мы делали в методе вращений, с помощью элементарных матриц вращения будем исключать элементы первого столбца, начиная с ТРЕТЬЕГО элемента.

$$A_1 = (Q_{2,n} \dots Q_{2,3}) A (Q_{2,n} \dots Q_{2,3})^* \equiv Q_1 A Q_1^*.$$

В силу самосопряжённости матрицы A за счёт этого унитарного преобразования у нас обнулятся элементы первой строки матрицы, начиная с ТРЕТЬЕГО. При этом очевидно, что матрица A_1 будет самосопряжённой, т. е. $A_1 = A_1^*$.

Предположим, что после $k-1$-го шага у нас есть самосопряжённая матрица A_{k-1}, у которой k-й (обратите внимание на его порядковый номер!) главный минор имеет трёхдиагональную структуру. Тогда на k-м шаге мы исключаем элементы k-го столбца, НАЧИНАЯ С $k+2$-го, с помощью последовательного умножения на унитарные матрицы

$$A_k = (Q_{k+1,n} \dots Q_{k+1,k+2}) A_{k-1} (Q_{k+1,n} \dots Q_{k+1,k+2})^* \equiv$$

$$\equiv Q_k A_{k-1} Q_k^*.$$

В результате мы получим, что минор $k+1$-го порядка будет иметь трёхдиагональную структуру, а матрица A_k будет самосопряжённой.

Таким образом, после $n-2$ шагов, где n — размер матрицы, мы получим самосопряжённую матрицу

$$T \equiv A_{n-2} = (Q_{n-2} \dots Q_1) A (Q_{n-2} \dots Q_1)^* \equiv$$

$$\equiv Q A Q^* = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \bar{\beta}_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \bar{\beta}_{n-1} & \alpha_n \end{bmatrix}.$$

Мы только что доказали, что любая самосопряжённая матрица подобна трёхдиагональной (но не обязательно вещественной) матрице.

Предположим, что $\beta_i \neq 0,\ i = 1, \ldots, n-1$. Определим диагональную матрицу $D = \operatorname{diag}\{d_1, \ldots, d_n\}$ с элементами

$$d_1 = 1,\ d_2 = \frac{\beta_1}{|\beta_1|}, \ldots, d_n = \frac{\beta_{n-1}}{|\beta_{n-1}|}.$$

Непосредственные вычисления показывают, что матрица D будет унитарной, т. е. $D^* = D^{-1}$. Применим её к матрице T и получим, что

$$J \equiv DTD^* = \begin{bmatrix} \alpha_1 & |\beta_1| & & \\ |\bar{\beta}_1| & \ddots & \ddots & \\ & \ddots & \ddots & |\beta_{n-1}| \\ & & |\bar{\beta}_{n-1}| & \alpha_n \end{bmatrix}.$$

Так как модули сопряжённых чисел совпадают, то матрица J — вещественная и симметричная.

Нам осталось только указать, что делать в случае, если какие-то β_i оказались равными нулю. Пусть $\beta_k = 0$ при некотором $k = 1, \ldots, n-1$. Тогда

$$T = \begin{bmatrix} \hat{T}_k & 0 \\ 0 & \hat{T}_{n-k} \end{bmatrix},$$

следовательно, спектр матрицы T является объединением спектров матриц $\hat{T}_k$ и $\hat{T}_{n-k}$. Если в этих матрицах нет нулевых элементов β_i, то мы выполняем приведённое выше преобразование подобия к каждой из подматриц и снова получаем вещественную матрицу J. В противном случае повторяем процедуру разделения матрицы на блоки дальше. Теорема доказана. □

По построению матрицы Q и D из теоремы 48 — унитарные, т. е. $Q^* = Q^{-1}$ и $D^* = D^{-1}$, что означает, что матрицы A и J подобны. Как известно из курса алгебры, у подобных матриц спектры

совпадают, т. е. у матриц A и J одинаковые собственные числа. Таким образом, поиск собственных значений самосопряжённой матрицы можно свести к поиску собственных значений трёхдиагональной матрицы с помощью матриц вращения.

Следует обратить внимание на то, какие элементы столбца (и строки) мы обнуляем в теореме 48. Стандартной ошибкой является попытка обнулить ВСЕ элементы столбца, расположенные ниже главной диагонали. В общем случае с помощью матриц вращения этого добиться невозможно.

Якобиевы матрицы

Вещественная трёхдиагональная матрица

$$J = \begin{bmatrix} a_1 & b_1 & & 0 \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & c_n & a_n \end{bmatrix}$$

— такая, что $b_i c_{i+1} > 0,\ i = 1, \ldots, n-1$, называется *якобиевой*. Следует обратить внимание на условие для внедиагональных элементов, которое говорит о том, что НЕ любая трёхдиагональная матрица является якобиевой. Кроме того, нужно не путать якобиеву матрицу с матрицей Якоби и якобианом.

По теореме 48 мы получаем в результате преобразований вращения симметричную матрицу, а потому мы будем полагать в дальнейшем, что в якобиевой матрице $c_i = b_{i-1},\ i = 2, \ldots, n$. Для начала изучим основные свойства якобиевых матриц.

Теорема 49. *Если J — якобиева матрица, то для детерминантов её главных миноров справедливы следующие утверждения:*

- $|J_0| \equiv 1,\ |J_1| = a_1,\ |J_{i+1}| = a_{i+1}|J_i| - b_i^2|J_{i-1}|,\ i = 1, \ldots, n-1.$

- *Если при* $i < n$ $|J_i| = 0$, *то* $|J_{i-1}||J_{i+1}| < 0$. *Если* $|J_n| \equiv |J| = 0$, *то* $|J_{n-1}| \neq 0$.

Доказательство. Для доказательства первого утверждения воспользуемся методом математической индукции. База индукции очевидна, а переход с шага k на шаг $k+1$ осуществляется с помощью хорошо известного из курса алгебры метода вычисления определителя матрицы с помощью разложения по строке или столбцу, которые и дают рекуррентную формулу из теоремы.

Пользуясь рекуррентной формулой, при условии $|J_i| = 0$, получаем, что $|J_{i+1}| = -b_i^2|J_{i-1}|$, откуда следует, что определители миноров $i-1$-го и $i+1$-го порядков имеют разные знаки при $i < n$. А для $i = n$ по той же формуле получаем, что $0 = a_n|J_{n-1}| - b_i^2|J_{n-2}|$. Отсюда по методу «от противного» следует, что если $|J_{n-1}| = 0$, то и все остальные детерминанты главных миноров тоже равны нулю, что противоречит определению якобиевой матрицы. Обратите внимание, что равенство нулю диагональных элементов якобиевой матрицы вполне возможно, поэтому нам нужно воспользоваться тем, что из $|J_2| = 0$ следует, что $b_1 = 0$, что противоречит определению якобиевой матрицы. Теорема доказана. □

Теорема 50. *Собственные значения якобиевой* J *матрицы попарно различные (простые, в терминах алгебры).*

Доказательство. Из курса алгебры известно, что размерность ядра матрицы $J_\lambda \equiv J - \lambda(J)E$ в точности совпадает с кратностью собственного числа $\lambda(J)$. Кроме того, по определению матрица J_λ является якобиевой, следовательно, по теореме 49 у вырожденной (так как $\lambda(J)$ — собственное значение) матрицы J_λ детерминант $n-1$-го главного минора не может равняться нулю. Это означает, что ранг матрицы J_λ равен $n-1$, а размер ядра — 1. Таким образом, $\lambda(J)$ — простое собственное число якобиевой матрицы

J. Поскольку эти рассуждения верны для любого собственного числа матрицы J, то теорема доказана. $\square$

Теперь докажем теорему, которая лежит в основе метода бисекций.

Теорема 51. *Если J — якобиева матрица, то количество отрицательных собственных чисел в ней $\sigma_-(J)$ равно числу перемен знака в последовательности Штурма $\{1, |J_1|, \ldots, |J_n| \equiv |J|\}$ при условии, что мы приписываем $|J_k| = 0$ знак детерминанта предыдущего главного минора $|J_{k-1}|$.*

Доказательство. Если $|J_k| \neq 0$. $k = 1, \ldots, n$, то утверждение теоремы следует из теоремы 47. Поэтому предположим, что при некотором k у нас $|J_k| = 0$. В соответствии с условием теоремы будем считать, что знак детерминанта этого главного минора совпадает со знаком детерминанта предыдущего главного минора, который не может равняться нулю в силу теоремы 49. Возьмём $\varepsilon_0 = \min_{\lambda(J_i)\neq 0, i=1,\ldots,n} |\lambda|$, т. е. минимальное по модулю из всех НЕНУЛЕВЫХ собственных чисел ВСЕХ главных миноров матрицы J. Рассмотрим якобиевы матрицы вида $J_{\pm\varepsilon} \equiv J \pm \varepsilon E$, $\varepsilon \in (0, \varepsilon_0)$.

Практически очевидно, что $\lambda(J_{\pm\varepsilon,i}) = \lambda(J_i) \pm \varepsilon \neq 0$, $i = 1, \ldots, n$. Заметим далее, что в силу сделанного выбора ε мы имеем $|J_{\pm\varepsilon,i}| \neq 0$, $i = 1, \ldots, n$, поскольку, как известно из курса алгебры, определитель матрицы равен произведению её собственных чисел.

Если $|J_i| \neq 0$, то знаки детерминантов главных миноров i-го порядка исходной и возмущённых якобиевых матриц совпадают, т. е. $\mathrm{sign}(|J_{\varepsilon,i}|) = \mathrm{sign}(|J_{-\varepsilon,i}|) = \mathrm{sign}(|J_i|)$. Если $|J_i| = 0$, то у детерминантов главных миноров i-го порядка возмущённых матриц знаки противоположные, т. е. $\mathrm{sign}(|J_{\varepsilon,i}|) \cdot \mathrm{sign}(|J_{-\varepsilon,i}|) < 0$. Это следует из теоремы 50, поскольку у якобиевых матриц все

собственные числа простые. По построению у возмущённой якобиевой матрицы $J_{-\varepsilon,i}$ отрицательных собственных чисел на одно больше, чем у возмущённой якобиевой матрицы $J_{\varepsilon,i}$.

Заметим также, что по построению у нас справедливы равенства $\sigma_-(J_\varepsilon) = \sigma_-(J)$ и $\sigma_-(J_{-\varepsilon}) = \sigma_-(J) + \sigma_0(J)$.

Собираем полученные результаты вместе и получаем, что по теореме 47, которая может быть применена к возмущённым якобиевым матрицам в силу выбора ε, $\sigma_-(J_\varepsilon) = \sigma_-(J)$ равно числу перемен знака в последовательности Штурма

$$\{1, |J_{\varepsilon,1}|, \ldots, |J_{\varepsilon,n}| \equiv |J_\varepsilon|\}.$$

Аналогично получаем, что $\sigma_-(J_{-\varepsilon}) = \sigma_-(J) + \sigma_0(J)$ равно числу перемен знака в $\{1, |J_{-\varepsilon,1}|, \ldots, |J_{-\varepsilon,n}| \equiv |J_{-\varepsilon}|\}$.

Если $|J_i| \neq 0$ и $|J_{i+1}| \neq 0$, то перемена знака происходит или не происходит одновременно в обеих последовательностях Штурма для возмущённых матриц.

Если $|J_k| = 0,\ k \neq n$, то по теореме 49 $|J_{k-1}||J_{k+1}| < 0$, следовательно, в последовательностях

$$\{|J_{\pm\varepsilon,k-1}|, |J_{\pm\varepsilon,k}|, |J_{\pm\varepsilon,k+1}|\}$$

и

$$\{|J_{k-1}|, |J_k|, |J_{k+1}|\}$$

происходит ровно одна перемена знака.

Если $|J_n| \equiv |J| = 0$, то $\sigma_0(J) = 1$. Обратите внимание, что нулевое собственное число у якобиевой матрицы может быть только одно по теореме 50, значит, оно может появиться только при условии, что детерминант матрицы равен нулю. В результате мы получаем, что $|J_{\varepsilon,n-1}||J_{\varepsilon,n}| > 0,\ |J_{-\varepsilon,n-1}||J_{-\varepsilon,n}| < 0,\ |J_{n-1}||J_n| = 0$.

Таким образом, если $|J_k| = 0$ приписать знак $|J_{k-1}|$, то последовательности Штурма для матриц J_ε и J имеют одинаковые знаки. Теорема доказана. $\square$

О вычислении числа перемен знака на компьютере

Для того, чтобы посчитать число перемен знака в последовательности Штурма нам достаточно знать знак детерминанта главного минора. Поэтому можно использовать последовательность аналогичную рекурсивной формуле для вычисления детерминанта якобиевой матрицы, которую мы видели в теореме 49:

$$d_0 = 1,\ d_1 = |J_1| = a_1,$$

$$d_{i+1} = a_{i+1}d_i - b_i^2 d_{i-1},\ i = 1, \ldots, k-1,$$

где-то в середине подменяем значения, необходимые для вычисления следующей величины на нормированные на $t_k > 0$ величины

$$d_{k-1} := \frac{d_{k-1}}{t_k},\ d_k := \frac{d_k}{t_k}$$

и продолжаем считать дальше

$$d_{i+1} = a_{i+1}d_i - b_i^2 d_{i-1},\ i = k, \ldots, n-1.$$

Имеет смысл выбирать $t_k = \max\{|d_{k-1}|, |d_k|\}$. Непосредственная проверка показывает, что знаки в последовательности чисел $\{1, d_1, \ldots, d_n\}$ совпадают со знаками в последовательности Штурма $\{1, |J_1|, \ldots, |J_n| \equiv |J|\}$. Операцию нормировки можно применять неоднократно. С алгебраической точки зрения данный алгоритм увеличивает число операций. Однако на практике можно столкнуться со взрывным ростом абсолютной величины детерминантов главных миноров, что на компьютере приведёт к проблемам с вычислениями, например, переполнению. А потому для получения результата, а не ошибки программы, имеет смысл несколько увеличить количество операций и получить правильный ответ.

Вычисление собственного вектора якобиевой матрицы

Для того чтобы вычислить собственный вектор якобиевой матрицы, нам потребуется вспомогательный, но весьма любопытный результат.

Теорема 52. *Последняя компонента любого собственного вектора якобиевой матрицы не равна нулю.*

Доказательство. По определению собственного вектора $J\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$, $\vec{\mathbf{x}} \neq \vec{\mathbf{0}}$. Воспользуемся методом «от противного» и предположим, что $x_n = 0$. Тогда мы незамедлительно получаем, что

$$x_{n-1} = -\frac{(a_n - \lambda)x_n}{b_{n-1}} = 0,$$

$$x_{n-i} = -\frac{(a_{n-1} - \lambda)x_{n-i+1} + b_{n-i+1}x_{n-i+2}}{b_{n-i}} = 0,\ i = 2, \ldots, n-1.$$

т. е. $\vec{\mathbf{x}} = \vec{\mathbf{0}}$, что противоречит предположению, что вектор $\vec{\mathbf{x}}$ — собственный. Таким образом, теорема доказана. □

Используя полученный результат, мы легко можем вычислить собственный вектор якобиевой матрицы, зная собственное число λ и положив $x_n = 1$, а после, воспользовавшись формулами, которые мы видели в доказательстве теоремы 52. Непосредственные вычисления показывают, что решение задачи на поиск собственного вектора в этом случае сводится к решению системы

$$\begin{bmatrix} a_1 - \lambda & b_1 & & 0 \\ b_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-2} \\ 0 & & b_{n-2} & a_{n-1} - \lambda \end{bmatrix} \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \vdots \\ \mathbf{x_{n-1}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{-b_{n-1}} \end{bmatrix}.$$

Поскольку $|J-\lambda E|=0$, то по теореме 49 детерминант главного минора $n-1$-го порядка этой матрицы не равен нулю, а потому матрица системы невырождена, и у этой системы существует единственное решение, которое не равно нулю, как мы знаем из курса алгебры.

Лекция №13

Метод вращений (метод Якоби)

Обратите внимание, что название «метод Якоби» встречалось в итерационных методах, а название «метод вращений» — в разделе о прямых методах, а потому не следует путать разные методы! В данном случае мы будем изучать метод для поиска собственных чисел и собственных векторов матрицы. Как и прежде, мы предполагаем, что матрица у нас самосопряжена $A = A^*$.

Как известно из курса алгебры, для любой самосопряжённой матрицы существует унитарная матрица Q, столбцы которой состоят из собственных векторов матрицы A, — такая, что

$$Q^*AQ = \Lambda \equiv \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}.$$

При этом на главной диагонали матрицы Λ стоят собственные числа матрицы A.

Очевидно, что если мы определим функционал

$$\Phi(A) \equiv \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} |a_{ij}|^2,$$

то

$$\Phi(Q^*AQ) = \min_{T^*T=E} \Phi(T^*AT),$$

т. е. решение задачи о поиске минимума функционала даёт нам решение задачи о поиске собственных значений и собственных векторов. Нам нужно теперь указать итерационный алгоритм, который бы на каждом шаге уменьшал функционал $\Phi(\cdot)$. Для этого нам потребуется получить пару вспомогательных утверждений.

Определим другой функционал

$$S(A) \equiv \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2.$$

Обратите внимание, что его отличие от функционала $\Phi(A)$ состоит в том, что он суммирует квадраты модулей ВСЕХ элементов матрицы. Функционал $\Phi(A)$ суммирует квадраты модулей элементов матрицы, расположенных вне главной диагонали. Для проверки понимания материала докажите, что $\sqrt{S(A)}$ является нормой матрицы, но при этом эта норма НЕ будет мультипликативной.

Теорема 53. *Для любой квадратной матрицы A и унитарной матрицы T справедливы равенства*

$$S(TA) = S(AT) = S(A).$$

Доказательство. Представим нашу матрицу A в виде набора векторов-столбцов $A = [\vec{\mathbf{a}}_1 \ \ldots \ \vec{\mathbf{a}}_n]$. Непосредственные вычисления показывают, что

$$S(TA) = S([T\vec{\mathbf{a}}^1 \ \ldots \ T\vec{\mathbf{a}}^n]) =$$

по определению евклидова скалярного произведения и функционала $S(\cdot)$

$$= (T\vec{\mathbf{a}}^1, T\vec{\mathbf{a}}^1)+\ldots+(T\vec{\mathbf{a}}^n, T\vec{\mathbf{a}}^n) = (T^*T\vec{\mathbf{a}}^1, \vec{\mathbf{a}}^1)+\ldots+(T^*T\vec{\mathbf{a}}^n, \vec{\mathbf{a}}^n) =$$

по определению унитарной матрицы

$$= (\vec{\mathbf{a}}^1, \vec{\mathbf{a}}^1) + \ldots + (\vec{\mathbf{a}}^n, \vec{\mathbf{a}}^n) = S(A).$$

Аналогично доказываем $S(AT) = S(A)$, представив матрицу в виде набора векторов-строк. Для проверки понимания материала убедитесь, что вы можете выписать доказательство, а также доказать, что $S(T^*AT) = S(A)$. Теорема доказана. □

Пусть у нас, как и прежде в разделе о прямом методе вращений, Q_{ij} — элементарная матрица вращений.

Теорема 54. *Пусть у нас имеется самосопряжённая матрица* $A = A^* \equiv \{a_{ij}\}_{i,j=1}^{n,n}$ *и матрица* $\tilde{A} \equiv Q_{ij}^* A Q_{ij}^* \equiv \{\tilde{a}_{ij}\}_{i,j=1}^{n,n}$. *Тогда*

$$\Phi(\tilde{A}) = \Phi(A) + \left[|a_{ii}|^2 + |a_{jj}|^2 - |\tilde{a}_{ii}|^2 - |\tilde{a}_{jj}|^2\right].$$

Доказательство. Так как элементарная матрица вращений является унитарной, то по теореме 53 $S(A) = S(AQ) = S(Q^*AQ)$. По определению функционалов $S(\cdot)$ и $\Phi(\cdot)$ получаем

$$\Phi(A) + |a_{ii}|^2 + |a_{jj}|^2 + \sum_{k=1,k\neq i,j}^{n} |a_{kk}|^2 = S(A) = S(\tilde{A}) =$$

$$= \Phi(\tilde{A}) + |\tilde{a}_{ii}|^2 + |\tilde{a}_{jj}|^2 + \sum_{k=1,k\neq i,j}^{n} |\tilde{a}_{kk}|^2 =$$

$$= \Phi(\tilde{A}) + |\tilde{a}_{ii}|^2 + |\tilde{a}_{jj}|^2 + \sum_{k=1,k\neq i,j}^{n} |a_{kk}|^2.$$

Здесь мы также воспользовались тем, что матрица Q_{ij} — элементарная матрица вращений, а потому в матрице $\tilde{A}$ по сравнению с матрицей A изменились ТОЛЬКО строки и столбы с номерами i и j. Поэтому в $\sum_k$ мы можем убрать значок $\tilde{}$. Из приведённого равенства незамедлительно следует утверждение теоремы. $\square$

Выбор вращений

Теперь нам осталось определить каким способом выбирать матрицы вращений, чтобы сформулировать метод. Для упрощения выкладок будем предполагать, что работаем только с вещественными матрицами, хотя анлогичным способом возможно сформулировать метод и для комплексных матриц. Нам также потребуется несколько вспомогательных утверждений. И начнём мы с

утверждения, которое поможет вычислить разность $\Phi(A)-\Phi(\tilde{A})$. Именно эта разность определяет, насколько сильно нам удастся уменьшить функционал $\Phi(\cdot)$ за счёт одного элементарного вращения.

Теорема 55. *Пусть у нас имеется самосопряжённая матрица* $A = A^* \equiv \{a_{ij}\}_{i,j=1}^{n,n}$ *и матрица* $\tilde{A} \equiv Q_{ij}^* A Q_{ij}^* \equiv \{\tilde{a}_{ij}\}_{i,j=1}^{n,n}$ *с некоторой матрицей вращений, соответствующей некоторому углу поворота* α. *Тогда*

$$\Phi(A) - \Phi(\tilde{A}) = 2|a_{ij}|^2 - \frac{1}{2}\left[(a_{ii} - a_{jj})\sin 2\alpha + 2a_{ij}\cos 2\alpha\right]^2 = \\ = 2|a_{ij}|^2 - 2|\tilde{a}_{ij}|^2.$$

Доказательство. Воспользуемся теоремой 54 и тем, что нужные нам элементы матриц связаны соотношением

$$\begin{bmatrix} \tilde{a}_{ii} & \tilde{a}_{ij} \\ \tilde{a}_{ij} & \tilde{a}_{jj} \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{bmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}.$$

Доказательство завершает применение тригонометрических формул двойного угла. Убедитесь, что вы можете написать формулы перемножения матриц и тригонометрические формулы для проверки понимания материала. □

Теорема 55 даёт нам подсказку о том, каким образом выбирать очередную матрицу вращений, а именно, минимизируя разность $\Phi(A)-\Phi(\tilde{A})$. Однако в ней есть угол, величину которого нам потребуется выбрать для того, чтобы иметь возможность провести вычисления. Следующая теорема подскажет, что мы можем сделать

Теорема 56. *Пусть у нас имеется самосопряжённая матрица* $A = A^* \equiv \{a_{ij}\}_{i,j=1}^{n,n}$ *размера* n *и матрица* $\tilde{A} \equiv Q_{ij}^* A Q_{ij}^* \equiv \{\tilde{a}_{ij}\}_{i,j=1}^{n,n}$ *с матрицей вращений, выбираемой из условий* $|a_{ij}| = \max_{k\neq m}|a_{km}|$ *и* $(a_{ii} - a_{jj})\sin 2\alpha + 2a_{ij}\cos 2\alpha = 0$. *Тогда*

$$\Phi(\tilde{A}) \leq \left[1 - \frac{2}{n(n-1)}\right]\Phi(A).$$

Доказательство. По теореме 55 с учётом сделанного выбора угла α мы получаем $\Phi(\tilde{A}) = \Phi(A) - 2|a_{ij}|^2$. Далее, в силу того, что элемент a_{ij} выбран МАКСИМАЛЬНЫМ ПО МОДУЛЮ среди всех ВНЕДИАГОНАЛЬНЫХ элементов матрицы (их всего $n^2 - n = n(n-1)$ штук), то $\Phi(A) \leq n(n-1)|a_{ij}|^2$. Отсюда незамедлительно следует оценка

$$|a_{ij}|^2 \geq \frac{\Phi(A)}{n(n-1)}$$

из которой получается требуемая оценка, следовательно, теорема доказана. □

Теоремы 55 и 56 дают нам уравнение на угол α, с помощью которого мы максимально возможно уменьшаем функционал $\Phi(A)$ посредством одного элементарного вращения. Стоит заметить, однако, что сам угол нам не нужен, нам лишь нужно знать его косинус и синус, чтобы сформировать элементарную матрицу вращений. И здесь нам даёт ответ следующая теорема.

Теорема 57. *Решением уравнения* $(a_{ii} - a_{jj})\sin 2\alpha + 2a_{ij}\cos 2\alpha = 0$ *при условии* $a_{ij} \neq 0$ *является угол* α *такой, что*

$$\cos\alpha = \sqrt{\frac{1}{2}\left(1 - \frac{a_{ii} - a_{jj}}{r}\right)},\ \sin\alpha = \frac{2a_{ij}}{r - a_{ii} + a_{jj}}\cos\alpha,$$

$$r \equiv \sqrt{|a_{ii} - a_{jj}|^2 + 4|a_{ij}|^2}.$$

Доказательство. Для доказательства теоремы вспоминаем основное тригонометрическое тождество и с его помощью сводим уравнение к квадратному уравнению относительно $\cos 2\alpha$. Решаем это уравнение школьным методом и получаем два корня. Нам достаточно взять любой из них. □

Теперь мы можем сформулировать *метод вращений (метод Якоби)* для поиска собственных значений и собственных векторов самосопряжённой матрицы. $A_0 \equiv A$, $A_k = Q_k^* A_{k-1} Q_k$, где $Q_k \equiv Q_{i(k),j(k)}$ — элементарная матрица вращений, определяемая по формулам из теорем 56 и 57.

Теорема 58. *Последовательность матриц $\{A_k\}_{k=0}^{\infty}$ метода вращений (метода Якоби) для решения полной проблемы собственных значений самосопряжённой матрицы $A = A^*$ размера n сходится к диагональной матрице, причём*

$$\Phi(A_k) \le \left[1 - \frac{2}{n(n-1)}\right]^k \Phi(A).$$

Доказательство. По теореме 56 с применением метода математической индукции мы получаем оценку теоремы. Из этой оценки с помощью инструментов математического анализа следует, что $\Phi(A_k) \to 0$, $k \to \infty$. По определению функционала $\Phi(\cdot)$ это и означает, что последовательность матриц $\{A_k\}_{k=0}^{\infty}$ сходится к диагональной матрице, что завершает доказательство теоремы. □

Сходимость собственных значений

Итак, в теореме 58 мы выяснили, как быстро будет сходиться к нулю функционал $\Phi(\cdot)$, однако гораздо интереснее понять, насколько близко при этом мы будем находиться к собственным числам и собственным векторам матрицы. По определению предела $\forall \varepsilon > 0$ $\exists k$ — такое, что

$$\hat{A} \equiv A_k = (Q_1 \dots Q_k)^* A (Q_1 \dots Q_k) \equiv \hat{Q}^* A \hat{Q},$$

и при этом $\Phi(\hat{A}) \le \varepsilon^2$. Квадрат ε здесь взят для получения более простой формы записи некоторых формул в некоторых из последующих доказательств.

Обозначим $\hat{\Lambda} \equiv \operatorname{diag} \hat{A} = \operatorname{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_n\}$ и сравним полученные на диагонали матрицы $\hat{A}$ значения с точными собственными значениями матрицы A, выведенными в результате применения унитарного преобразования $Q^* A Q \equiv \Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$. Начнём с доказательства сходимости характеристических полиномов.

Теорема 59. *Характеристический полином* $\hat{P}(\lambda) \equiv |\hat{\Lambda} - \lambda E|$ *стремится к характеристическому полиному* $P(\lambda) \equiv |A - \lambda E|$ *при* $\varepsilon \to 0$.

Доказательство. Поскольку у нас есть только информация о том, что $\Phi(\hat{A}) \leq \varepsilon^2$, то этим и будем пользоваться при доказательстве теоремы. Заметим, что в силу унитарности матрицы $\hat{Q}$ и свойства определителя произведения матриц получаем

$$|\hat{\Lambda} - \lambda E| = |\hat{\Lambda} - \lambda E||E| = |\hat{\Lambda} - \lambda E||QQ^*| = |\hat{\Lambda} - \lambda E||Q||Q^*| =$$

$$= |Q||\hat{\Lambda} - \lambda E||Q^*| = |Q\hat{\Lambda}Q^* - \lambda QQ^*| = |Q\hat{\Lambda}Q^* - \lambda E|.$$

Далее обратим внимание, что

$$\hat{Q}\hat{\Lambda}\hat{Q}^* = \hat{Q}(\hat{A} - \hat{A} + \hat{\Lambda})\hat{Q}^* = A - \hat{Q}(\hat{A} - \hat{\Lambda})\hat{Q}^*.$$

Из сказанного выше, теоремы 53 и определения функционалов $S(\cdot)$ и $\Phi(\cdot)$ следует, что

$$S(\hat{Q}(\hat{A} - \hat{\Lambda})\hat{Q}^*) = S((\hat{A} - \hat{\Lambda})\hat{Q}^*) = S(\hat{A} - \hat{\Lambda}) = \Phi(\hat{A}).$$

Так как $\Phi(\hat{A}) \leq \varepsilon^2$ при $k \to \infty$, то коэффициенты матрицы $\hat{Q}\hat{\Lambda}\hat{Q}^*$ стремятся к коэффициентам матрицы A. Из курса алгебры известно, что это означает сходимость характеристических полиномов, или $\hat{P}(\lambda) \equiv |\hat{\Lambda} - \lambda E| = |\hat{Q}\hat{\Lambda}\hat{Q}^* - \lambda E| \to |A - \lambda E| \equiv P(\lambda)$ при $\varepsilon \to 0$. Теорема доказана. □

Теперь можно оценить, насколько хорошо мы приблизили искомые собственные значения.

Теорема 60 (оценка приближения собственных значений). *Справедливы следующие утверждения:*

- *для любого собственного числа λ_i матрицы A размера n существует диагональный элемент $\hat{\lambda}_{j(i)}$ матриц $\hat{A}$ и $\hat{\Lambda}$ — такой, что $|\lambda_i - \hat{\lambda}_{j(i)}| \le \sqrt{n}\varepsilon$;*
- *для любого диагонального элемента $\hat{\lambda}_j$ матриц $\hat{A}$ и $\hat{\Lambda}$ размера n существует собственное число $\lambda_{i(j)}$ матрицы A — такое, что $|\hat{\lambda}_j - \lambda_{i(j)}| \le \sqrt{n}\varepsilon$.*

Доказательство. Заметим, что

$$Q\Lambda Q^* = A = \hat{Q}\hat{A}\hat{Q}^* = \hat{Q}\hat{\Lambda}\hat{Q}^* - \hat{Q}\hat{\Lambda}\hat{Q}^* + \hat{Q}\hat{A}\hat{Q}^* = \hat{Q}\hat{\Lambda}\hat{Q}^* + \hat{Q}(\hat{A} - \hat{\Lambda})\hat{Q}^*.$$

Отсюда получаем, что матрица

$$C \equiv Q^*\hat{Q}(\hat{A} - \hat{\Lambda}) = \Lambda Q^*\hat{Q} - Q^*\hat{Q}\hat{\Lambda}$$

с элементами c_{ij}, удовлетворяющими неравенствам

$$|c_{ij}|^2 \le S(C) = S(\hat{Q}(\hat{A} - \hat{\Lambda})) = S(\hat{A} - \hat{\Lambda}) = \Phi(\hat{A} - \hat{\Lambda}) = \Phi(\hat{A}) \le \varepsilon^2$$

по теореме 53, определению функционалов $S(\cdot)$, $\Phi(\cdot)$ и предположению о малости внедиагональных элементов матрицы $\hat{A}$: $\Phi(\hat{A}) \le \varepsilon^2$.

Пусть $R \equiv Q^*\hat{Q}$. Непосредственная проверка показывает, что R — унитарная матрица. Обозначим элементы этой матрицы через r_{ij}. Тогда с помощью формулы перемножения матриц мы получаем, что

$$c_{ij} = \lambda_i r_{ij} - r_{ij}\hat{\lambda}_j.$$

Заметим далее, что для любого индекса i существует индекс $j(i)$ — такой, что $|r_{i,j(i)}|^2 = \max_{k=1,\dots,n} |r_{ik}|^2$. Так как матрица R

унитарная, то $RR^*= E$ или $|r_{i1}|^2 + \ldots + |r_{in}|^2 = 1$ для любого индекса $i = 1, \ldots, n$. Это означает, что $\max_{k=1,\ldots,n} |r_{ik}|^2 \geq \frac{1}{n}$. Отсюда следует первое утверждение теоремы

$$|\lambda_i - \hat{\lambda}_{j(i)}| = \left| \frac{c_{ij}}{r_{i,j(i)}} \right| \leq \sqrt{n}\varepsilon.$$

Аналогично доказываем второе утверждение теоремы. Следует убедиться, что вы можете проделать все выкладки с помощью равенства $R^*R = E$. Теорема доказана. □

Сходимость собственных векторов

Пусть у матрицы A размера n все собственные числа РАЗЛИЧНЫ. В противном случае возникнут сложности с получением скорости сходимости. Тогда с помощью перестановок столбцов матрицы Q мы можем получить упорядоченность этих собственных чисел по возрастанию $\lambda_1 < \ldots < \lambda_n$. Аналогичным образом мы можем упорядочить диагональные элементы матрицы $\hat{A}$ с помощью перестановок столбцов матрицы $\hat{Q}$: $\hat{\lambda}_1 \leq \ldots \hat{\lambda}_n$. Обратите внимание, что строгого неравенства для диагональных элементов матрицы $\hat{A}$ не требуется. Теперь получим вспомогательное утверждение.

Теорема 61. *Пусть* $a \equiv \min_{i \neq j,\ i,j=1,\ldots,n} |\lambda_i - \lambda_j|$, $\lambda_1 < \ldots < \lambda_n$, $\hat{\lambda}_1 \leq \ldots \hat{\lambda}_n$, *и* $\sqrt{n}\varepsilon < 0{,}5a$, *тогда*

$$|\lambda_i - \hat{\lambda}_i| \leq \sqrt{n}\varepsilon \ u \ |\lambda_i - \hat{\lambda}_j| > 0{,}5a \ \forall i \neq j, \ i, j = 1, \ldots, n.$$

Доказательство. Воспользуемся теоремой 60. Тогда в силу условия $\sqrt{n}\varepsilon < 0,5a$ и упорядоченности элементов λ_i и $\hat{\lambda}_i$ на расстоянии менее $0{,}5a$ от λ_i может находиться только одно число $\hat{\lambda}_i$ и его порядковый номер именно i. Для проверки понимания материала докажите, что в этом случае $\hat{\lambda}_i < \hat{\lambda}_{i+1}$, $i = 1, \ldots, n-1$. Теорема доказана. □

Собственные векторы матрицы A, составляющие столбцы ортогональной матрицы Q, определяются с точностью до направления, поэтому мы будем считать, что приближения к этим собственным векторам, посчитанные в матрице $\hat{Q}$, удовлетворяют неравенству $(\vec{\mathbf{q}}^{\,i}, \hat{\vec{\mathbf{q}}}^{\,\mathbf{i}}) \geq 0$, т. е. точный собственный вектор и его приближение направлены примерно в одном направлении (косинус угла положительный). Обратите внимание, что здесь предполагается, что матрицы Q и $\hat{Q}$ вещественны. Это означает, что диагональные элементы матрицы $R \equiv Q^*\hat{Q}$ из теоремы 60 неотрицательны. Для проверки понимания материала убедитесь, что вы можете это доказать по формуле перемножения матриц.

Теорема 62 (оценка приближения для собственных векторов). *В условиях теоремы 61*

$$S(Q - \hat{Q}) \leq \frac{8}{a^2}\varepsilon^2.$$

Доказательство. По теореме 53 $S(Q-\hat{Q}) = S(E-Q^*\hat{Q}) = S(E-R)$ в силу унитарности, точнее, ортогональности матрицы Q и по определению матрицы R. В доказательстве теоремы 60 мы получали $C \equiv R(\hat{A} - \hat{\Lambda}) = \Lambda R - R\hat{\Lambda}$. Отсюда с применением элементов доказательства теоремы 61 мы получаем

$$|r_{ij}| = \frac{|c_{ij}|}{|\lambda - \hat{\lambda}_i|} < \frac{|c_{ij}|}{0,5a} \ \forall i, j = 1, \ldots, n, \ i \neq j.$$

Используя определение функционалов $\Phi(\cdot)$ и $S(\cdot)$ получаем

$$\Phi(E - R) < \frac{4}{a^2} S(C) = \frac{4}{a^2}\Phi(\hat{A}) \leq \frac{4}{a^2}\varepsilon^2.$$

Нам осталось оценить только сумму квадратов диагональных элементов матрицы $E - R$, поскольку для суммы квадратов внедиагональных элементов оценка уже получена. Заметим, что

$$\sum_{i=1}^{n}(1 - r_{ii})^2 = \sum_{i=1}^{n}\left(1 - \sqrt{1 - \sum_{j=1,\ j\neq i}^{n}|r_{ij}|^2}\right)^2,$$

так как диагональные элементы матрицы R неотрицательны в силу выбора направления для приближений к собственным векторам.

Далее воспользуемся простым неравенством $(1-x)^2 \le 1-x,\ x \in [0,1]$ и получим

$$\sum_{i=1}^{n}\left(1-\sqrt{1-\sum_{j=1,\ j\neq i}^{n}|r_{ij}|^2}\right)^2 \le \sum_{i=1}^{n}\left(1-\sqrt{1-\sum_{j=1,\ j\neq i}^{n}|r_{ij}|^2}\right) =$$

$$= \sum_{i=1}^{n}\frac{\sum_{j=1,\ j\neq i}^{n}|r_{ij}|^2}{1+\sqrt{1-\sum_{j=1,\ j\neq i}^{n}|r_{ij}|^2}} \le \Phi(R) = \Phi(E-R) \le \frac{4}{a^2}\varepsilon^2.$$

Суммируя две оценки, получаем

$$S(Q-\hat{Q}) = S(E-R) = \Phi(E-R) + \sum_{i=1}^{n}(1-r_{ii})^2 \le \frac{8}{a^2}\varepsilon^2.$$

Теорема доказана. □

Заключение

На этом мы заканчиваем первую часть курса, посвящённую вычислительным методам линейной алгебры. Надеемся, что приведённые в данном учебно-методическом пособии материалы будут полезны как студентам при подготовке к экзамену, так и начинающим преподавателям для подготовки к лекциям.

Ваши замечания и пожелания просим отправлять по адресу *s.gololobov@g.nsu.ru*.

Список литературы

[1] Фаддеев Д. К. Фаддеева В. Н. Вычислительные методы линейной алгебры. М. ; Л. : Физматгиз, 1963.

[2] Коновалов А. Н. Введение в вычислительные методы линейной алгебры. Новосибирск : Наука, 1993.

[3] Воеводин В. В. Вычислительные основы линейной алгебры. М. : Наука, 1977.

[4] Годунов С. К. Решение систем линейных уравнений. Новосибирск : Наука, 1980.

[5] Бахвалов Н. С. Численные методы. М. : Наука, 1975.

[6] Самарский А. А. Гулин А. В. Численные методы. М. : Наука, 1989.

Учебное издание

Гололобов Сергей Владимирович,
Мацокин Александр Михайлович

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ АНАЛИЗА И ЛИНЕЙНОЙ АЛГЕБРЫ

Часть 1

Учебно-методическое пособие

Редактор *Я. О. Козлова*
Обложка *Е. В. Неклюдовой*

Подписано в печать 05.11.2019 г.
Формат 60 × 84 1/16. Уч.-изд. л. 10. Усл. печ. л. 9,3.
Тираж 150 экз. Заказ № 244
Издательско-полиграфический центр НГУ
630090, Новосибирск, ул. Пирогова, 2.