Лекции по вычислительным методам линейной алгебры 2013–2016 г.

Голубков А. Ю.

Лекция 1. Матричные нормы. Теория возмущений решений линейных систем.

Для начала следует определить предполагаемый объём предстоящей работы. Данный курс лекций и связанных с ним лабораторных работ будет посвящён двум основным прикладным аспектам вычислительной линейной алгебры: алгоритмам решения систем линейных алгебраических уравнений (СЛАУ) и алгоритмам алгебраической проблемы собственных значений. При этом основной аспект лекционного курса будет сделан на обосновании рассматриваемых алгоритмов, а точнее на демострации того, как теоретические положения базовых курсов линейной алгебры и функционального анализа трансформируются в конкретные алгоритмические идеи задач линейной алгебры. Естественно, будучи ограниченными рамками лекционных часов, мы не в состоянии ни охватить все имеющиеся алгоритмы, ни обсудить особенности их современных модификаций. Последнее невозможно и в силу того, что реализованные в существующих пакетах линейной алгебры вычислительные алгоритмы прошли весьма внушительный объём уточнений и оптимизаций, работавшими над ними авторскими коллективами, а это подрузамевает проведение значительного количества зачастую многолетних численных экспериментов, связанных с подбором различных параметров, не подпадающих под простое теоретическое описание (не указанные в базовых описаниях эвристически найденные параметры встречаются практически во всех коммерческих версиях алгоритмов). Поэтому мы вынуждены сконцентрироваться на несколько более скромной задаче обоснования природы построения интересующих нас алгоритмов линейной алгебры и обсуждения проблем, возникающих при их непосредственной реализации.

Сразу же оговоримся, что сам термин вычислительные методы подразумевает некоторую неоднозначность предлагаемых решений одной и той же задачи, поскольку фактически ни один из методов не является оптимальным во всех смыслах слова. Так, например, вычислительные методы в русле гауссова исключения (построения LU–разложения) не столь дорогостоящие (и в некоторых ситуациях наиболее эффективные) в плане численной трудоёмкости, в ряде случаев нуждаются в дополнительном уточнении получаемого результата, а их более точные модифицированные версии как, впрочем,

и алгоритмы триангуляции с использованием ортогональных или унитарных преобразований (алгоритмы, использующие QR–разложение) уже имеют значительно большую вычислительную сложность.

Отметим и то, что в современных вычислительных методах ситуация, когда предъявляемый алгоритм имеет теоретическое обоснование своей оптимальности на некотором классе задач является скорее приятным (и нетривиально доказываемым) исключением, нежели правилом. Более того, зачастую мы не располагаем и достаточным теоретическим обоснованием осуществимости алгоритма, хотя и имеем практические доказательства последнего (сходимость алгоритмов полной проблемы собственных значений установлена лишь для весьма идеалистической диагонализируемой ситуации, а их практическая осуществимость в общем случае по сути базируется на разрядных ограничениях). Отдельное место занимает и обоснование достоверности получаемого результата, что особенно важно в технических расчётах, где нужно иметь точную информацию о количестве правильных разрядов ответа и числе необходимых для этого разрядов в исходных данных.

Предваряя обсуждение конкретных алгоритмов, мы должны прежде всего сказать несколько слов о средствах измерения, используемых для оценки качества их работы. Поскольку мы имеем дело с конечномерными нормированными пространствами, то, естественно, речь пойдёт об их нормах, а точнее о наиболее употребимых среди них на практике.

Напомним, что под нормой на векторном пространстве $V_{\mathbb{k}}$ над полем $\mathbb{k}$, $\mathbb{k} = \mathbb{R}$, $\mathbb{C}$, понимается неотрицательный вещественный функционал $\|\,\|$, $\|\,\| : V \longrightarrow \mathbb{R}_+$, обладающий следующими свойствами:

1. $\|\,\|$ точен, $\|x\| = 0$, если и только если $x = 0$;
2. $\|\,\|$ однороден, $\|\lambda x\| = |\lambda|\|x\|$ для всех $x \in V$, $\lambda \in \mathbb{k}$, где в зависимости от $\mathbb{k} = \mathbb{R}$ или $\mathbb{k} = \mathbb{C}$ модуль $|\lambda|$ обозначает модуль вещественного или комплексного числа λ;
3. $\|\,\|$ полуаддитивен, $\|x + y\| \le \|x\| + \|y\|$ для любых $x, y \in V$.

При этом, полагая $\rho(x, y) = \|x - y\|$, $x, y \in V$, мы можем наделить множество V структурой метрического пространства с метрикой ρ, что в свою очередь позволяет применять к (V, ρ) весь известный понятийный аппарат теории метрических пространств.

В интересующей нас ситуации пространство $V_{\mathbb{k}}$ имеет конечную размерность $n \ge 1$ и может быть отождествлено с арифметическим пространством векторов-столбцов $\mathbb{k}^n_{\mathbb{k}}$. Наиболее важными для нас нормами на пространстве $\mathbb{k}^n_{\mathbb{k}}$ являются его p–нормы $\|\,\|_p$, $p \ge 1$, и предельная для них max–норма $\|\,\|_\infty$, которые определяются равенствами

$$\|x\|_p \;=\; \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}, \quad \|x\|_\infty \;=\; \max_{i=1,\ldots,n} |x_i| \quad (x = (x_1, \ldots, x_n)^t \in \mathbb{k}^n),$$

где символ t обозначает транспонирование и модули компонент вектора определяются вещественными или коплексными в соответствии с тем, каким является поле $\mathbb{k}$. Проверка свойств нормы для p–норм сводится к последовательному выведению неравенств Юнга, Гёльдера и Минковского (см. курс функционального анализа). Напомним также, что в случае если пространство $V_{\mathbb{k}} \cong \mathbb{k}^n_{\mathbb{k}}$ является евклидовым и обладает скалярным

произведением (для $\Bbbk = \mathbb{R}$) или эрмитовой формой (для $\Bbbk = \mathbb{C}$) $(\,,\,)$, оно естественным образом наделяется структурой нормированного пространства по отношению к евклидовой норме $\|x\| = \sqrt{(x,x)}$, $x \in V$ (следствие неравенства Коши). В частности, применительно к стандарным скалярному произведению

$$(x,y) \;=\; y^t x \;=\; \sum_{i=1} x_i y_i \quad (x = (x_1,\ldots,x_n)^t, y = (y_1,\ldots,y_n)^t \in \mathbb{R}^n)$$

и эрмитовой форме

$$(x,y) \;=\; y^* x \;=\; \sum_{i=1}^{n} x_i \overline{y_i} \quad (x = (x_1,\ldots,x_n)^t, y = (y_1,\ldots,y_n)^t \in \mathbb{C}^n)\,,$$

где $*$ — транспонирование с комплексным сопряжением, это позволяет получить обычную евклидову 2–норму $\|\;\|_2$. В общем случае произвольное скалярное проиведение (эрмитова форма) на пространстве $\Bbbk^n_{\Bbbk}$ определяется равенством $(x,y)_A = y^t A x$ ($(x,y)_A = y^* A x$), $x, y \in \Bbbk^n$, для соответствующей положительно определённой симметрической (самосопряжённой) матрицы $A \in M_n(\Bbbk)$. Из курса функционального анализа известно также, что любые две нормы $\|\;\|$ и $\|\;\|'$ на конечномерном векторном пространстве V являются эквивалентными в том смысле, что для некоторых положительных констант a и b справедливы неравенства

$$a\|x\| \le \|x\|' \le b\|x\| \quad (x \in V)\,.$$

Последнее означает, что метрическое пространство V полно относительно всех своих метрик, связанных с нормами, и сходимость последовательности элементов из V в одной из таких метрик влечёт за собой сходимость в любой другой из них. В частности, применительно к нормам $\|\;\|_1$, $\|\;\|_2$ и $\|\;\|_\infty$ пространства $\Bbbk^n$ коэффициенты a и b определяются следующим образом

$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2\,, \quad \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty\,, \quad \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty \quad (x \in \Bbbk^n)\,.$$

Определение 0.1. *Норма $\|\;\|$ на пространстве матриц $M_n(\Bbbk) \cong \Bbbk^{n^2}_{\Bbbk}$, $\Bbbk = \mathbb{R}$, $\mathbb{C}$, называется матричной, если $\|AB\| \le \|A\|\|B\|$ для любых $A, B \in M_n(\Bbbk)$.*

Заметим, что для всякой матричной нормы $\|\;\|$

1. $\|E\| = \|E^2\| \le \|E\|^2$, $\|E\| \ge 1$, где E — единичная матрица из $M_n(\Bbbk)$;
2. $1 \le \|E\| = \|AA^{-1}\| \le \|A\|\|A^{-1}\|$ для всех $A \in GL_n(\Bbbk)$.

Приведём несколько классических примеров матричных норм.

1. Норма Фробениуса $\|\;\|_F$, представляющая собой не что иное как обычную 2–норму пространства $M_n(\Bbbk) \cong \Bbbk^{n^2}_{\Bbbk}$, т.е.

$$\|A\|_F \;=\; \sqrt{\Big(\sum_{i,j=1}^{n} |a_{ij}|^2\Big)} \quad (A = (a_{ij}) \in M_n(\Bbbk))\,.$$

Норма $\|\;\|_F$ является матричной нормой, поскольку она является векторной нормой (см. выше) и в силу неравенства Коши для любых $A = (a_{ij}), B = (b_{ij}) \in M_n(\Bbbk)$

$$\|AB\|_F^2 \;=\; \sum_{i,j=1}^{n}\Big|\sum_{k=1}^{n}a_{ik}b_{kj}\Big|^2 \le \sum_{i,j=1}^{n}\Big(\sum_{k=1}^{n}|a_{ik}||b_{kj}|\Big)^2 \le$$
$$\sum_{i,j=1}^{n}\Big(\sum_{k=1}^{n}|a_{ik}|^2\Big)\Big(\sum_{m=1}^{n}|b_{mj}|^2\Big) \;=\; \Big(\sum_{i,k=1}^{n}|a_{ik}|^2\Big)\Big(\sum_{m,j=1}^{n}|b_{mj}|^2\Big) \;=\; \|A\|_F^2\|B\|_F^2\,.$$

2. В отличие от нормы Фробениуса ∞–норма матриц

$$\|A\|_{\max} \;=\; \max_{i,j=1,\dots,n}|a_{ij}| \quad (A=(a_{ij})\in M_n(\mathbb{k}))$$

не является матричной нормой при $n>1$, так как на матрице $B=(b_{ij}=1)$, заполненной единицами, $1=\|B\|_{\max}=\|B\|_{\max}^2<\|B^2\|_{\max}=n$ (в то время как для матричной нормы должно было бы выполняться обратное неравенство).

Вместе с тем данную векторную норму $\|\,\|_{\max}$ можно превратить в матричную, заменив её на норму $\|\,\|'$, $\|A\|'=n\|A\|_{\max}$, $A\in M_n(\mathbb{k})$ (очевидная проверка).

3. Важный класс матричных норм составляют операторные матричные нормы или матричные нормы, подчинённые векторным нормам. Всякая матрица $A\in M_n(\mathbb{k})$ определяет ограниченный линейный оператор $A: x\longmapsto Ax$, $x\in\mathbb{k}^n_{\mathbb{k}}$, пространства $\mathbb{k}^n_{\mathbb{k}}$, рассматриваемого с любой из имеющихся на нём норм $\|\,\|$ (ограниченные подмножества такого пространства суть то же самое, что подмножества векторов, координаты которых ограничены по модулю). Норма такого оператора

$$\|A\| \;=\; \sup_{0\ne x\in\mathbb{k}^n}\frac{\|Ax\|}{\|x\|} \;=\; \sup_{0\ne x\in\mathbb{k}^n,\ \|x\|\le 1}\frac{\|Ax\|}{\|x\|} \;=\; \sup_{x\in\mathbb{k}^n,\ \|x\|=1}\|Ax\| \;=\; \max_{x\in\mathbb{k}^n,\ \|x\|=1}\|Ax\|\,,$$

где переход от sup к max возможен ввиду компактности ограниченных замкнутых подмножеств конечномерного пространства, удовлетворяет неравенству

$$\|Ax\|\le\|A\|\|x\| \quad (x\in\mathbb{k}^n,\ A\in M_n(\mathbb{k}))$$

и, как следствие, неравенству $\|AB\|\le\|A\|\|B\|$, $A,B\in M_n(\mathbb{k})$, а потому является матричной нормой. Такая операторная матричная норма называется матричной нормой, подчинённой (индуцированной или согласованной с) векторной норме (нормой) $\|\,\|$.

Отметим, что по определению операторной нормы $\|\,\|$, $\|E\|=1$. Поэтому норма Фробениуса, для которой $\|E\|_F=\sqrt{n}$, не является операторной нормой при $n>1$.

Перечислим теперь стандартные факты относительно наиболее известных операторных матричных норм.

Замечание 0.2. *Максимальная столбцовая норма* $\|\,\|_1$ *(максимум* 1*–норм столбцов матрицы)*

$$\|A\|_1 \;=\; \max_{i=1,\dots,n}\sum_{k=1}^{n}|a_{ki}| \quad (A=(a_{ij})\in M_n(\mathbb{k}))$$

является операторной нормой, подчинённой векторной 1*–норме* $\|\,\|_1$.

Доказательство. Обозначим через $e_1,\dots,e_n$ элементы стандартного базиса пространства $\mathbb{k}^n_{\mathbb{k}}$, в котором вектор-столбец e_i имеет единичную i–ую координату и остальные координаты равные нулю, $i=1,\dots,n$. Тогда для всякой матрицы $A=(a_{ij})\in M_n(\mathbb{k})$

$$\|A\|_1 \;=\; \max_{i=1,\dots,n}\|Ae_i\|_1 \le \max_{x\in\mathbb{k}^n,\ \|x\|_1=1}\|Ax\|_1\,.$$

В то же время, обозначив i–ый столбец матрицы A через a_i, $i = 1, \ldots, n$, мы можем записать для любого $x = (x_1, \ldots, x_n)^t \in \mathbb{k}^n$

$$\|Ax\|_1 \;=\; \Big\|\sum_{i=1}^n a_i x_i\Big\|_1 \le \sum_{i=1}^n |x_i|\|a_i\|_1 \le \|A\|_1\|x\|_1\,.$$

Поэтому $\max_{x\in\mathbb{k}^n,\ \|x\|_1=1}\|Ax\|_1 \le \|A\|_1$ и, более того, в силу сказанного ранее данное неравенство является точным равенством, $\|A\|_1 = \max_{x\in\mathbb{k}^n,\ \|x\|_1=1}\|Ax\|_1$. □

Замечание 0.3. *Максимальная строчная норма* $\|\ \|_\infty$ *(максимум* 1*–норм строк матрицы)*

$$\|A\|_\infty \;=\; \|A^t\|_1 \;=\; \max_{i=1,\ldots,n}\sum_{k=1}^n |a_{ik}| \quad (A = (a_{ij}) \in M_n(\mathbb{k}))$$

является операторной нормой, подчинённой векторной ∞*–норме* $\|\ \|_\infty$.

Доказательство. В обозначениях предыдущего замечания

$$\|Ax\|_\infty \;=\; \max_{i=1,\ldots,n}\Big|\sum_{k=1}^n a_{ik}x_k\Big| \le \max_{i=1,\ldots,n}\sum_{k=1}^n |a_{ik}||x_k| \le \|A\|_\infty\|x\|_\infty$$

и, как следствие, $\max_{x\in\mathbb{k}^n,\ \|x\|_\infty=1}\|Ax\|_\infty \le \|A\|_\infty$. С другой стороны, пусть индекс $\hat{i}$ выбран таким образом, что

$$\|A\|_\infty \;=\; \sum_{k=1}^n |a_{\hat{i}k}|\,.$$

Тогда, полагая для всех $k = 1, \ldots, n$

$$y_k \;=\; \begin{cases} \frac{\overline{a_{\hat{i}k}}}{|a_{\hat{i}k}|} & \text{при } a_{\hat{i}k} \neq 0; \\ 0 & \text{иначе,} \end{cases}$$

где знак комплексного сопряжения в случае $\mathbb{k} = \mathbb{R}$ можно опустить, мы получим вектор $y = (y_1, \ldots, y_n)^t$, $\|y\|_\infty = 1$, для которого

$$\max_{x\in\mathbb{k}^n,\ \|x\|_\infty}\|Ax\|_\infty \ge \|Ay\|_\infty \;=\; \max_{i=1,\ldots,n}\Big|\sum_{k=1}^n a_{ik}y_k\Big| \ge \Big|\sum_{k=1}^n a_{\hat{i}k}y_k\Big| \;=\; \sum_{k=1}^n |a_{\hat{i}k}| \;=\; \|A\|_\infty\,.$$

Таким образом, $\|A\|_\infty = \max_{x\in\mathbb{k}^n,\ \|x\|_\infty=1}\|Ax\|_\infty$. □

Стоит сказать, что имея в своём распоряжении матричную норму $\|\ \|$ и автоморфизм (антиавтоморфизм) ϕ алгебры $M_n(\mathbb{k})$, мы можем определить новую матричную норму $\|\ \|_\phi$, положив $\|A\|_\phi = \|\phi(A)\|$, $A \in M_n(\mathbb{k})$.

Замечание 0.4. *Спектральная норма* $\|\ \|_2$

$$\|A\|_2 \;=\; \max_{\lambda\in\operatorname{Spec}A^*A}\sqrt{\lambda} \quad (A = (a_{ij}) \in M_n(\mathbb{k}))$$

является операторной нормой, подчинённой евклидовой векторной 2*–норме* $\|\ \|_2$.

Доказательство. Для краткости мы будем иметь дело с $\mathbb{k} = \mathbb{C}$ (в вещественном случае наши рассуждения останутся прежними с точностью до необходимых замен терминов).

Спектр $\operatorname{Spec} A^*A$ матрицы A^*A понимается в контексте функционального анализа, а точнее точечного спектра A^*A (в линейной алгебре в ряде случаев нуль не относится к собственным значениям, но мы будем рассматривать его как элемент спектра в случае вырожденной матрицы A^*A). Сразу же отметим, что $\operatorname{Spec} A^*A = \operatorname{Spec} AA^*$, поскольку в любом ассоциативном кольце с единицей (в нашем случае в $M_n(\mathbb{C})$) обратимость $1-ab$ равносильна обратимости $1-ba$, а потому A^*A и AA^* имеют равные резольвентные множества и, как следствие, идентичный спектр.

Из курса линейной алгебры известно, что матрица A^*A является самосопряжённой и неотрицательно определённой в том смысле, что $(xA^*A, x) = x^*A^*Ax = (Ax, Ax) \geq 0$ для всех $x \in \mathbb{C}^n$. Последнее также означает, что все собственные значения A^*A вещественны и неотрицательны. Кроме того, матрице A^*A отвечает ортонормированный базис $\{u_1, \ldots, u_n\}$ пространства $\mathbb{C}^n$, состоящий из её собственных векторов, $A^*Au_i = \lambda_i u_i$, $i = 1, \ldots, n$, где $\lambda_i \in \operatorname{Spec} A^*A \subset \mathbb{R}_+$. Обозначив через U матрицу со столбцами $u_1, \ldots, u_n$ и положив $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, мы можем записать $A^*AU = U\Lambda$ и $A^*A = U\Lambda U^*$ (матрица с ортонормированными столбцами U является унитарной, $U \in U_n(\mathbb{C})$). Тогда для любого $x = (x_1, \ldots, x_n)^t \in \mathbb{C}^n$, $\|x\|_2 = \sqrt{(x,x)} = 1$,

$$\|Ax\|_2 \;=\; \sqrt{(Ax, Ax)} \;=\; \sqrt{(A^*Ax, x)} \;=\; \sqrt{(U\Lambda U^*x, x)} \;=\; \sqrt{(\Lambda U^*x, U^*x)} \;=\; \sqrt{\sum_{i=1}^{n} \lambda_i |y_i|^2}\,,$$

где $y = U^*x = (y_1, \ldots, y_n)^t$, $\|y\|_2 = 1$. Следовательно,

$$\max_{x \in \mathbb{C}^n,\ \|x\|_2 = 1} \|Ax\|_2 \leq \max_{y \in \mathbb{C}^n,\ \|y\|_2 = 1} \sqrt{\sum_{i=1}^{n} \lambda_i |y_i|^2} \leq \max_{i=1,\ldots,n} \sqrt{\lambda_i} \;=\; \|A\|_2\,.$$

С другой стороны если номер $\hat{i}$ отвечает собственному вектору $u_{\hat{i}}$, для которого $\lambda_{\hat{i}} = \max_{i=1,\ldots,n} \lambda_i$, тогда

$$\max_{x \in \mathbb{C}^n,\ \|x\|_2 = 1} \|Ax\|_2 \geq \|Au_{\hat{i}}\|_2 \;=\; \|\lambda_{\hat{i}} u_{\hat{i}}\|_2 \;=\; \sqrt{\lambda_{\hat{i}}} \;=\; \|A\|_2\,.$$

Поэтому $\|A\|_2 = \max_{\lambda \in \operatorname{Spec} A^*A} \sqrt{\lambda}$. □

Напомним ещё одно связанное с матрицами важное понятие — понятие спектрального радиуса. Последний определяется равенством

$$\rho(A) \;=\; \max_{\lambda \in \operatorname{Spec} A} |\lambda|\,,$$

где A — произвольная матрица из $M_n(\mathbb{C})$ (мы всегда рассматриваем комплексный спектр). Из курса функционального анализа известно также, что для любой операторной нормы $\|\,\|$

$$\rho(A) \;=\; \lim_{n \to \infty} \|A^n\|^{\frac{1}{n}}\,.$$

При этом для всякой самосопряжённой матрицы A верно равенство $\rho(A) = \|A\|_2$.

Замечание 0.5. *Для любой матричной нормы* $\|\,\|$ *на* $M_n(\mathbb{C})$ *справедливо неравенство* $\rho(A) \leq \|A\|$ *для всех* $A \in M_n(\mathbb{C})$.

Доказательство. Достаточно заметить, что для всякого ненулевого собственного вектора x матрицы A, $Ax = \lambda x$, можно записать $AX = \lambda X$, обозначив через X матрицу со столбцами равными x. Поэтому $\|\lambda X\| = |\lambda|\|X\| = \|AX\| \leq \|A\|\|X\|$ и, как следствие, $|\lambda| \leq \|A\|$. □

Большинство известных матричных преобразований сводятся к умножениям исходной матрицы на матрицы какого-то определённого вида. Поэтому при выборе для оценки соответствующей матричной нормы было бы желательно получать в результате преобразования матрицу, норма которой не увеличивается относительно исходной. В частности, при использовании ортогональных и унитарных преобразований имеет смысл использовать для оценивания нормы, инвариантные относительно таких преобразований. Последние определяются следующим образом: матричная норма $\|\,\|$ на $M_n(\mathbb{R})$ ($M_n(\mathbb{C})$) называется ортогонально (соответственно унитарно) инвариантной, если $\|UAV\| = \|A\|$ для любых $A \in M_n(\mathbb{R})$, $U, V \in O_n(\mathbb{R})$ ($A \in M_n(\mathbb{C})$, $U, V \in U_n(\mathbb{C})$).

Замечание 0.6. *Спектральная и фробениусова нормы* $\|\,\|_2$ *и* $\|\,\|_F$ *унитарно инвариантны.*

Доказательство. В случае спектральной нормы достаточно заметить, что для всех $A \in M_n(\mathbb{C})$ и $U, V \in O_n(\mathbb{C})$

$$\operatorname{Spec}(UAV)^*(UAV) \;=\; \operatorname{Spec} V^*(A^*A)V \;=\; \operatorname{Spec} A^*A\,.$$

Если обозначить столбцы матрицы A через $a_1, \ldots, a_n$, а её строки — через $b_1^t, \ldots, b_n^t$, тогда

$$\|A\|_F \;=\; \sqrt{\sum_{i=1}^{n} \|a_i\|_2^2} \;=\; \sqrt{\sum_{i=1}^{n} \|b_i\|_2^2}$$

и потому

$$\|UA\|_F \;=\; \sqrt{\sum_{i=1}^{n} \|Ua_i\|_2^2} \;=\; \sqrt{\sum_{i=1}^{n} \|a_i\|_2^2} \;=\; \|A\|_F\,,$$

$$\|AV\|_F \;=\; \sqrt{\sum_{i=1}^{n} \|(b_i^tV)^t\|_2^2} \;=\; \sqrt{\sum_{i=1} \|V^tb_i\|_2^2} \;=\; \sqrt{\sum_{i=1}^{n} \|b_i\|_2^2} \;=\; \|A\|_F\,.$$

В последнем равенстве мы воспользовались также тем, что $V^t \in U_n(\mathbb{C})$ ($V^* = \overline{V}^t = \overline{V^t} = V^{-1}$, $V^{t*} = \overline{V} = V^{-1\,t} = V^{t\,-1}$). □

Несколько слов о сингулярном разложении.

Скажем теперь несколько слов о взаимосвязи между сингулярным разложением (или SVD–разложением) матрицы и её спектральной и фробениусовой нормами. Начнём с основной теоремы о SVD–разложении.

Теорема 0.7. *Для любой* $m \times n$ *матрицы* $A \in \mathbb{C}^{m\times n}$ *ранга* $\operatorname{rk} A = r$ *существуют положительные вещественные числа* $\sigma_1, \ldots, \sigma_r$, $\sigma_1 \geq \ldots \geq \sigma_r > 0$, *и унитарные матрицы* $U \in U_n(\mathbb{C})$, $V \in U_m(\mathbb{C})$ *такие, что* $A = V\Sigma U^*$, *где* Σ — *диагональная* $m \times n$*–матрица вида* $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_p)$, $\sigma_{r+1} = \ldots = \sigma_p = 0$, $p = \min\{m, n\}$.

Доказательство. Числа $\sigma_1, \ldots, \sigma_r$, участвующие в указанном разложении называются сингулярными значениями матрицы A, а столбцы матриц U и V, связанные соотношениями $AU = V\Sigma$ и $A^*V = U\Sigma^*$, принято называть правыми и соответственно левыми сингулярными векторами A. Приведённая нами теорема формулируется и в вещественном случае с заменой унитарных на ортогональные матрицы и эрмитова сопряжения на обычное транспонирование.

Заметим, что A^*A — самоспряжённая неотрицательно определённая матрица из $M_n(\mathbb{C})$, для которой имеется ортонормированный базис собственных векторов-столбцов составляющих унитарную матрицу $U \in U_n(\mathbb{C})$, $A^*AU = U\,\mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ для соответствующих вещественных $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$. Поскольку $\mathrm{rk}\, A^*A = \mathrm{rk}\, A = \mathrm{rk}\, A^* = r$ (вспоминая матрицу Грама системы векторов), мы имеем также $\sigma_{r+1} = \ldots = \sigma_n = 0$. Обозначим через Σ_r диагональную $r \times r$–матрицу $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ и через U_r — $n \times r$–матрицу, составленную из первых r столбцов $u_1, \ldots, u_r$ матрицы U. Тогда мы можем записать

$$U_r^* A^* A U_r \;=\; \Sigma_r^2\,, \quad (\Sigma_r^{-1} U_r^* A^*)(A U_r \Sigma_r^{-1}) \;=\; E_r\,,$$

где E_r — единичная матрица размера $r \times r$. Поэтому $m \times r$–матрица $V_r = AU_r\Sigma_r^{-1}$ имеет ортонормированные столбцы. Достроив матрицу V_r произвольным образом до унитарной матрицы $V \in U_m(\mathbb{C})$ и матрицу Σ_r до диагональной $m \times n$–матрицы Σ,

$$\Sigma \;=\; \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix},$$

где какие-то нулевые блоки могут и отсутствовать, мы получаем, что

$$\|Au_i\|_2^2 \;=\; u_i^* A^* A u_i \;=\; \sigma_i^2 \;=\; 0\,, \quad Au_i \;=\; 0 \quad (i = r+1, \ldots, n)\,,$$

а потому $V\Sigma = AU$ (см. равенство $V_r\Sigma_r = AU_r$). Таким образом, $A = V\Sigma U^*$. Заметим, что из последнего равенства следует, что $\mathrm{rk}\,\Sigma = \mathrm{rk}\, A = r$ и, следовательно, мы могли бы провести наше рассуждение, используя вместо r ранг A^*A, и прийти в итоге к тому же равенству.

Отметим также, что из определения SVD–разложения следует, что

$$A^*A \;=\; (V\Sigma U^*)^*(V\Sigma U^*) \;=\; U\Sigma^2 U^*\,.$$

Поэтому сингулярные значения являются не чем иным как положительными корнями из ненулевых собственных значений матрицы A^*A, что делает их выбор полностью однозначным. Кроме того, обозначив столбцы матриц U и V через $\{u_i\}$ и $\{v_j\}$, мы можем переписать разложение $A = V\Sigma U^*$ в виде

$$A \;=\; \sum_{i=1}^{r} \sigma_i v_i u_i^*\,,$$

поскольку в этой сумме произведений одностолбцовых матриц (v_i — матрица со столбцом v_i и нулевыми остальными столбцами) на однострочные (u_i^* — матрица со строкой u_i^* и нулевыми остальными строками) матрица с ненулевым i–ым столбцом нетривиально умножается только на матрицу с ненулевой i–ой строкой, причём каждое произведение $v_i u_i^*$ совпадает с обычным произведением столбца v_i на строку u_i^*.

Из равенства $AU = V\Sigma$ также следует, что образ $\operatorname{Ran} A$ оператора A, $A : \mathbb{C}^n \longrightarrow \mathbb{C}^m$ (т.е. линейная оболочка столбцов A) совпадает с подпространством, натянутым на $v_1, \ldots, v_r$, а ядро $\operatorname{Ker} A$ — с подпространством, натянутым на $u_{r+1}, \ldots, u_n$.

Кроме того, из определения сингулярных значений матрицы A следует, что наибольшее сингулярное значение σ_1 совпадает со спектральной нормой $\|A\|_2$ матрицы A, $\sigma_1 = \|A\|_2$, и, далее, каждое значение σ_i равно спектральной норме ограничения A на подпространство $\langle u_i, \ldots, u_n\rangle$, $\sigma_i = \|A|_{\langle u_i,\ldots,u_n\rangle}\|_2$. Вместе с тем учитывая унитарную инвариантность нормы Фробениуса (точнее используя доказательство этого факта) мы можем записать

$$\|A\|_F \;=\; \|\Sigma\|_F \;=\; \sqrt{\sigma_1^2 + \ldots + \sigma_r^2}\,.$$

□

Приведённые нами свойства сингулярных значений позволяют предложить другой путь отыскания сингулярного разложения матрицы.

Пусть мы располагаем способом нахождения вектора $u \in \mathbb{C}^n$, $\|u\|_2 = 1$, такого, что $Au = \sigma_1 v$, где $v \in \mathbb{C}^m$, $\sigma_1 = \|A\|_2$ (т.е. речь идёт о любом алгоритме отыскания собственного вектора, отвечающего максимальному по модулю собственному значению самосопряжённой матрицы). Достроим вектора u и v до унитарных матриц $U = (u\,U') \in U_n(\mathbb{C})$ и $V = (v\,V') \in U_m(\mathbb{C})$. Тогда

$$V^*AU \;=\; \begin{pmatrix} v^*Au & v^*AU' \\ V'^*Au & V'^*AU \end{pmatrix} \;=\; \begin{pmatrix} \sigma & w^* \\ 0 & V'^*AU' \end{pmatrix},$$

где $w = U'^*A^*v \in \mathbb{C}^{n-1}$. При этом,

$$\sigma \;=\; \|A\|_2 \;=\; \|V^*AU\|_2 \geq \frac{1}{\sqrt{\sigma^2 + \|w\|_2^2}}\|V^*AU(\sigma, w^*)^*\|_2 \geq$$
$$\frac{1}{\sqrt{\sigma^2 + \|w\|_2^2}}\|(\sigma^2 + \|w\|_2^2, (V'^*AU'w)^*)^*\|_2 \geq \sqrt{\sigma^2 + \|w\|_2^2}$$

и, следовательно, $\|w\|_2 \leq 0$, $\|w\|_2 = 0$, $w = 0$,

$$V^*AU \;=\; \begin{pmatrix} \sigma & 0 \\ 0 & V'^*AU' \end{pmatrix}.$$

Поэтому для продолжения построения мы можем перейти к меньшей подматрице V'^*AU' размера $(m-1)\times(n-1)$ (фактически, мы описали рекурсивную процедуру построения сингулярного разложения).

Ещё одно важное наблюдение, отосящееся к SVD–разложению, состоит в том, что имея в своём распоряжении такое разложение для невырожденной матрицы $A \in GL_n(\mathbb{C})$, $A = V\Sigma U^*$, $U, V \in U_n(\mathbb{C})$, мы без труда можем найти соответствующее разложение для обратной матрицы A^{-1}, $A^{-1} = U\Sigma^{-1}V^*$, и, следовательно, можем решить любую линейную систему $Ax = b$, полагая

$$x \;=\; \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i} u_i v_i^*\Big) b \;=\; \sum_{i=1}^{n} \frac{(b, v_i)}{\sigma_i} u_i\,,$$

где в правой части стоит линейная комбинация столбцов u_i.

Полезно будет отметить ещё одно весьма важное свойство сингулярных значений.

Теорема 0.8. *Пусть имеется SVD–разложение $m \times n$–матрицы $A \in \mathbb{C}^{m\times n}$ ранга* $\operatorname{rk} A = r$, $A = V\Sigma U^*$. *Тогда для всех* $k = 1, \ldots, r-1$

$$\min_{B\in\mathbb{C}^{m\times n},\ \operatorname{rk} B\le k} \|A - B\|_2 \;=\; \|A - A_k\|_2 \;=\; \sigma_{k+1}\,,$$

где

$$A_k \;=\; \sum_{i=1}^{k} \sigma_i v_i u_i^*\,.$$

Доказательство. Для начала заметим, что $A_k = V\Sigma(k)U^*$, $\operatorname{rk} A_k = k$, где матрица $\Sigma(k)$ получается из матрицы Σ в результате обнуления всех σ_i, $i \ge k+1$, а потому

$$\|A - A_k\|_2 \;=\; \|V(\Sigma - \Sigma(k))U^*\|_2 \;=\; \|\Sigma - \Sigma(k)\|_2 \;=\; \max_{i\ge k+1} \sigma_i \;=\; \sigma_{k+1}\,.$$

Пусть теперь $B \in \mathbb{C}^{m\times n}$ и $\operatorname{rk} B \le k$. Тогда ядро $\operatorname{Ker} B$ имеет размерность $\dim \operatorname{Ker} B \ge n-k$, подпространство $W_{k+1} = \langle u_1, \ldots, u_{k+1}\rangle$ имеет размерность $k+1$ и, следовательно, найдётся вектор $w = \alpha_1 u_1 + \ldots + \alpha_{k+1}u_{k+1} \in \operatorname{Ker} B \cap W_{k+1}$, $\|w\|_2 = 1$ (в противном случае $\dim \operatorname{Ker} B + W_{k+1} > n$?!). Тогда

$$\|A - B\|_2 \ge \|(A-B)w\|_2 \;=\; \|Aw\|_2 \;=\; \Big\|\sum_{i=1}^{r} \sigma_i (w, u_i) v_i\Big\|_2 \;=\; \Big\|\sum_{i=1}^{k+1} \sigma_i \alpha_i v_i\Big\|_2 \;=$$

$$\sqrt{\sum_{i=1}^{k+1} \sigma_i^2 |\alpha_i|^2} \ge \sigma_{k+1} \sqrt{\sum_{i=1}^{k+1} |\alpha_i|^2} \ge \sigma_{k+1}\|w\|_2 \;=\; \sigma_{k+1}\,.$$

Таким образом, $\sigma_{k+1} = \min_{B\in\mathbb{C}^{m\times n},\ \operatorname{rk} B\le k} \|A - B\|_2$. □

Следствие 0.9. *В условиях этой теоремы наименьшее сингулярное значение σ_r матрицы A совпадает с минимумом расстояний в спектральной норме от A до матриц меньшего ранга. В частности, для $r = \min\{m, n\}$ это соответствует расстоянию в спектральной норме от A до множества матриц неполного ранга.*

Следствие 0.10. *Множество матриц полного ранга открыто и всюду плотно в $\mathbb{C}^{m\times n}$.*

Доказательство. Достаточно заметить, что для всякой матрицы полного ранга $A \in \mathbb{C}^{m\times n}$, $\operatorname{rk} A = p = \min\{m, n\}$, с сингулярными значениями $\sigma_1 \ge \ldots \ge \sigma_p > 0$, любая матрица $B \in \mathbb{C}^{m\times n}$, $\|A - B\|_2 < \sigma_p$, имеет полный ранг. С другой стороны для всякой матрицы неполного ранга $C \in \mathbb{C}^{m\times n}$, $\operatorname{rk} C = r < p$, с сингулярным разложением $C = V'\Sigma' U'^*$ можно достроить Σ' до диагональной матрицы полного ранга $\Sigma'(\varepsilon)$, положив оставшиеся компоненты диагонали равными ε. Тогда

$$\|C - V'\Sigma'(\varepsilon)U'^*\|_2 \;=\; \|\Sigma' - \Sigma'(\varepsilon)\|_2 \;=\; |\varepsilon|\,.$$

Поскольку в роли ε может выступать любое число, отсюда следует всюду плотность множества матриц полного ранга. □

Число обусловленности матрицы и его свойства.

Нашей следующей целью станет получение классической, но далеко не самой точной оценки погрешности в решении линейных систем, использующей число обусловленности матрицы системы.

Для начала напомним, что пространство квадратных матриц $M_n(\Bbbk)$, $\Bbbk = \mathbb{R}$, $\mathbb{C}$, будучи конечномерным полно относительно любой из своих норм. Кроме того, оно может быть отождествлено с банаховой алгеброй ограниченных операторов при использовании в качестве основной любой операторной нормы. Для этого пространства классический результат функционального анализа об обратимости оператора близкого к обратимому можно переформулировать следующим образом.

Замечание 0.11. *Для любой матричной нормы $\|\,\|$ на пространстве $M_n(\Bbbk)$ и обратимой матрицы $A \in GL_n(\Bbbk)$ всякая матрица $A+B$, где $B \in M_n(\Bbbk)$, $\|B\| < \frac{1}{\|A^{-1}\|}$, обратима и обратная к ней матрица $(A+B)^{-1}$ представима в виде суммы сходящегося ряда*

$$(A+B)^{-1} \;=\; \sum_{k=0}^{\infty} (-1)^k (A^{-1}B)^k A^{-1}\,.$$

Доказательство. Исследуем сперва более частный вопрос об обратимости матрицы $E-C$, $\|C\| < 1$. Последовательность частичных сумм $\{D_n\}_{n=0}^{\infty}$, $D_n = C^0 + C + \ldots + C^n$, будучи фундаментальной,

$$\|D_n - D_m\| \;=\; \Big\| \sum_{i=\min\{n,m\}+1}^{\max\{n,m\}} C^i \Big\| \le \frac{\|C\|^{\min\{n,m\}+1}}{1-\|C\|} \quad (n,m \ge 0)\,,$$

она обязана сходиться к некоторой матрице D в банаховом пространстве $M_n(\Bbbk)$ с нормой $\|\,\|$. Вместе с тем определение матричной нормы гарантирует непрерывность всех имеющихся на алгебре $M_n(\Bbbk)$ операций и, как следствие,

$$D(E-C) \;=\; \lim_{n\to\infty} D_n(E-C) \;=\; \lim_{n\to\infty}(E-C^n) \;=\; E - \lim_{n\to\infty} C^n \;=\; E\,, \quad D \;=\; (E-C)^{-1}\,.$$

Возвращаясь к матрице $A+B = A(E-(-A^{-1}B))$, $\|-A^{-1}B\| \le \|A^{-1}\|\|B\| < 1$, нам остаётся лишь заметить, что по доказанному

$$(A+B)^{-1} \;=\; (E-(-A^{-1}B))^{-1}A^{-1} \;=\; \Big(\sum_{k=0}^{\infty}(-1)^k(A^{-1}B)^k\Big)A^{-1} \;=\; \sum_{k=0}^{\infty}(-1)^k(A^{-1}B)^kA^{-1}\,.$$

Возможность внесения множителя A^{-1} под знак суммы обусловлена непрерывностью операций алгебры $M_n(\Bbbk)$. □

Пусть теперь решается линейная система $Ax = b$ с невырожденной квадратной матрицей $A \in GL_n(\Bbbk)$, $\Bbbk = \mathbb{R}$, $\mathbb{C}$, и вектором правой части $0 \ne b \in \Bbbk^n$. Решение осуществляется при помощи некоторого алгоритма на некоторой вычислительной системе. При этом особенности представления данных и накапливающиеся ошибки округления говорят о том, что в действительности реализуемый алгоритм доставляет точное решение возмущённой системы $\hat{A}\hat{x} = \hat{b}$ с относительно малыми возмущениями в исходных данных $\hat{A} = A + \delta A$ и $\hat{b} = b + \delta b$. Естественно предположить, что характер возмущения не делает решаемую задачу некорректной (т.е. матрица $\hat{A}$ невырождена) и, более того, порядок возмущения в ней $\|\delta A\| < \frac{1}{\|A^{-1}\|}$, где $\|\,\|$ — выбранная операторная норма на $M_n(\Bbbk)$, согласованная с соответствующей векторной нормой $\|\,\|$ на $\Bbbk^n_{\Bbbk}$. Попробуем оценить относительную погрешность решения через относительные погрешности в исходных данных.

Теорема 0.12. *В указанных здесь ограничениях*

$$\frac{\|\hat{x}-x\|}{\|x\|} \le \frac{k(A)}{1-k(A)\frac{\|\delta A\|}{\|A\|}}\Big(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\Big),$$

где число $k(A)$, именуемое числом обусловленности матрицы A в операторной норме $\|\,\|$, определяется как $\|A\|\|A^{-1}\|$ для невырожденной A, и полагается формально равным $+\infty$ в противном случае.

Доказательство. Достаточно заметить, что в силу сказанного ранее

$$\begin{aligned}\hat{x}-x \;=\; (A+\delta A)^{-1}(b+\delta b)-A^{-1}b \;=\; ((A+\delta A)^{-1}-A^{-1})b+(A+\delta A)^{-1}\delta b \;=\\ \sum_{k=1}^{\infty}(-1)^k(A^{-1}\delta A)^kA^{-1}b+\sum_{k=0}^{\infty}(-1)^k(A^{-1}\delta A)^kA^{-1}\delta b \;=\\ \sum_{k=1}^{\infty}(-1)^k(A^{-1}\delta A)^kx+\sum_{k=0}^{\infty}(-1)^k(A^{-1}\delta A)^kA^{-1}\delta b\,,\end{aligned}$$

где помимо полноты пространства $M_n(\mathbb{k})$ используется ещё и полнота пространства $\mathbb{k}^n$ относительно рассматриваемой векторной нормы (сходимость ряда векторов вытекает из сходимости ряда операторов и известных свойств операторной нормы). Тогда, переходя к нормам левой и правой части и оценивая последнюю сверху суммой норм её составляющих, мы получаем

$$\|\hat{x}-x\| \le \frac{\|A^{-1}\|\|\delta A\|\|x\|}{1-\|A^{-1}\|\|\delta A\|}+\frac{\|A^{-1}\|\|\delta b\|}{1-\|A^{-1}\|\|\delta A\|}\,.$$

С учётом $\|Ax\|=\|b\|$, $\frac{1}{\|x\|}\le\frac{\|A\|}{\|b\|}$, мы можем записать

$$\frac{\|\hat{x}-x\|}{\|x\|} \le \frac{\|A^{-1}\|\|\delta A\|}{1-\|A^{-1}\|\|\delta A\|}+\Big(\frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\Big)\frac{\|\delta b\|}{\|b\|}\,.$$

Остаётся заменить $\|A^{-1}\|\|\delta A\|$ на $k(A)\frac{\|\delta A\|}{\|A\|}$. □

В действительности мы не располагаем вектором x и можем оценить относительную погрешность решения исходя лишь из оценки, использующей вектор невязки $r=b-A\hat{x}$,

$$\frac{\|\hat{x}-x\|}{\|x\|} \;=\; \frac{\|A^{-1}r\|}{\|x\|} \le \frac{\|A^{-1}\|\|A\|\|r\|}{\|b\|} \;=\; k(A)\frac{\|r\|}{\|b\|}\,.$$

Более того, как правило, у нас нет и самого числа обусловленности $k(A)$ и потому требуется ещё написать отдельную программу-оценщик величины $k(A)$.

Приведём весьма кратко основные свойства числа обусловленности матриц, а уже затем обсудим реальное качество, полученных нами оценок.

1. $k(A)=k(A^{-1})\ge 1$ для любых операторной нормы $\|\,\|$ и матрицы $A\in Gl_n(\mathbb{k})$; $k(AB)\le k(A)k(B)$, $A,B\in Gl_n(\mathbb{k})$.

2. В спектральной норме $\|\,\|_2$ число обусловленности $k_2(A)$ матрицы $A\in Gl_n(\mathbb{k})$ определяется отношением

$$k_2(A) \;=\; \frac{\sigma_1}{\sigma_n}\,,$$

где $\|A\|_2 = \sigma_1$ и $\|A^{-1}\|_2 = \frac{1}{\sigma_n}$, σ_1 и σ_n — наибольшее и наименьшее сингулярные значения A; в частности, если $A = A^*$ $(A = A^t)$, тогда

$$k_2(A) \;=\; \frac{|\lambda_{\max}|}{|\lambda_{\min}|}\,,$$

где $\lambda_{\max}$ и $\lambda_{\min}$ — наибольшее и наименьшее по модулю собственные значения A. При этом, для любой матричной нормы $\|\ \|$ имеет место неравенство

$$k(A) \geq \frac{|\lambda_{\max}|}{|\lambda_{\min}|}\,,$$

поскольку $\|A\| \geq \rho(A) = |\lambda_{\max}|$, $\|A^{-1}\| \geq \rho(A^{-1}) = \frac{1}{|\lambda_{\min}|}$. Отметим и то, что $k_2(A) = 1$ для любой $A \in U_n(\mathbb{C})$.

3. Число обусловленности матрицы в любой ортогонально (унитарно) инвариантной норме остаётся неизменным при умножении матрицы слева и справа на ортогональные (унитарные) матрицы. Поэтому при использовании таких норм выбор ортогональных (унитарных) преобразований является предпочтительным с точки зрения оценок погрешностей, в которых ключевую роль играет число обусловленности.

4. По аналогии с нормами числа обусловленности $k(A)$ и $k(A)'$ матрицы $A \in Gl_n(\mathbb{k})$ в различных нормах $\|\ \|$ и $\|\ \|'$ связаны между собой отношением эквивалентности, точнее, если для некоторых $a, b > 0$

$$a\|B\| \leq \|B\|' \leq b\|B\| \quad B \in M_n(\mathbb{k})\,,$$

тогда

$$a^2 k(A) \leq k(A)' \leq b^2 k(A)\,.$$

Для любых двух векторных норм $\|\ \|$ и $\|\ \|'$ на простанстве $\mathbb{k}^n$, связанных положительными константами $a', b' > 0$,

$$a'\|x\| \leq \|x\|' \leq b'\|x\| \quad (x \in \mathbb{k}^n)\,,$$

соответствующие операторные нормы связаны неравенствами

$$\frac{a'}{b'}\|B\| \leq \|B\|' \;=\; \max_{0 \neq x \in \mathbb{k}^n} \frac{\|Bx\|'}{\|x\|'} \leq \frac{b'}{a'}\|B\| \quad (B \in M_n(\mathbb{k})\,.$$

В частности, применительно к векторным нормам $\|\ \|_1$, $\|\ \|_2$ и $\|\ \|_\infty$ это даёт

$$\frac{1}{\sqrt{n}}\|B\|_2 \leq \|B\|_1 \leq \sqrt{n}\|B\|_2\,, \quad \frac{1}{\sqrt{n}}\|B\|_\infty \leq \|B\|_2 \leq \sqrt{n}\|B\|_\infty\,, \quad \frac{1}{n}\|B\|_\infty \leq \|B\|_1 \leq n\|B\|_\infty$$

для всех $B \in M_n(\mathbb{k})$ и, как следствие,

$$\frac{1}{n}k_2(A) \leq k_1(A) \leq nk_2(A)\,, \quad \frac{1}{n}k_\infty(A) \leq k_2(A) \leq nk_\infty(A)\,, \quad \frac{1}{n^2}k_\infty(A) \leq k_1(A) \leq n^2 k_\infty(A)$$

где $k_1(A)$, $k_2(A)$ и $k_\infty(A)$ — числа обусловленности матрицы A в столбцовой, спектральной и строчной нормах соответственно.

Отсюда следует, что плохо обусловленные матрицы (матрицы со значительным по величине числом обусловленности) являются таковыми во всех нормах.

5. Из сказанного ранее о максимальном и минимальном сингулярных значениях матрицы $A \in Gl_n(\mathbb{k})$ и выражении через их отношение её числа обусловленности $k_2(A) = \frac{\sigma_1}{\sigma_n}$ в спектральной норме $\| \|_2$ немедленно следует, что

$$\frac{1}{k_2(A)} \;=\; \min_{\delta A,\ \det(A+\delta A)=0} \frac{\|\delta A\|_2}{\|A\|_2}\,.$$

Другими словами, $\frac{1}{k_2(A)}$ — минимальное относительное расстояние в норме $\| \|_2$ от A до множества вырожденных матриц (матриц неполного ранга).

Можно предложить и непосредственное доказательство этого утверждения.

Поскольку $A + \delta A \in Gl_n(\mathbb{k})$ при всех $\delta A \in M_n(\mathbb{k})$, $\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2}$ (см. ранее), $\min_{\delta A,\ \det(A+\delta A)=0} \|\delta A\|_2 \geq \frac{1}{\|A^{-1}\|_2}$. Остаётся построить возмущение δA, $\|\delta A\|_2 = \frac{1}{\|A^{-1}\|_2}$, для которого $\det(A+\delta A) = 0$. Выберем вектор $x \in \mathbb{k}^n$, $\|x\|_2 = 1$, для которого $\|A^{-1}x\|_2 = \|A^{-1}\|_2$. Далее положим $y = \frac{1}{\|A^{-1}x\|_2} A^{-1}x = \frac{1}{\|A^{-1}\|_2} A^{-1}x$ и $\delta A = -\frac{xy^*}{\|A^{-1}\|_2}$. Тогда

$$\|\delta A\|_2 \;=\; \max_{0\neq z\in\mathbb{k}^n} \frac{\|xy^*z\|_2}{\|A^{-1}\|_2\|z\|_2} \;=\; \frac{1}{\|A^{-1}\|_2} \max_{0\neq z\in\mathbb{k}^n} \frac{|(z,y)|}{\|z\|_2} \;=\; \frac{1}{\|A^{-1}\|_2}\,.$$

Вместе с тем

$$(A+\delta A)y \;=\; \frac{x}{\|A^{-1}\|_2} - \frac{\|y\|_2^2 x}{\|A^{-1}\|_2} \;=\; 0\,.$$

6. Понятие числа обусловленности может быть введено также при помощи производной Фреше. Пусть имеется отображение $F : V \longrightarrow Y$, определённое на некотором открытом подмножестве V нормированного пространства X и принимающее значение в нормированном пространстве Y. Отображение F называется *дифференцируемым в точке* (*дифференцируемым по Фреше в точке*) $v \in V$, если имеется ограниченный линейный оператор $L_v \in \mathcal{L}(X,Y)$, для которого при всех $h \in X$, $\|h\|_X < \delta$ (при достаточно малом δ, $v + h \in V$),

$$F(v+h) - F(v) \;=\; L_v h + \alpha(v,h)\,,$$

где $\alpha(v,h) \in Y$, $\frac{\|\alpha(v,h)\|_Y}{\|h\|_X} \to 0$ при $\|h\|_X \to 0$ (т.е. $\|\alpha(v,h)\|_Y = o(\|h\|_X)$ при $\|h\|_X \to 0$). При этом выражение $L_v h$ называют *дифференциалом Фреше* отображения F в точке v, а сам оператор L_v — *производной Фреше* (*сильной производной*) отображения F в точке v. Последнюю принято также обозначать как $L_v = F'(v)$. Заметим, что производная Фреше определена однозначно, поскольку вместе с равенствами

$$L_v h + \alpha(v,h) \;=\; L'_v h + \alpha'(v,h)\,, \quad (L_v - L'_v)h \;=\; \alpha'(v,h) - \alpha(v,h)$$

мы должны были бы получить

$$\|L_v - L'_v\| \;=\; \sup_{0\neq h\in X} \frac{\|(L_v - L'_v)h\|_Y}{\|h\|_X} \;=\; \frac{\|\alpha'(v,h) - \alpha(v,h)\|_Y}{\|h\|_X}\,,$$

где в левой части равенства стоит константа, а в правой — функция вида $o(1)$ при $\|h\|_X \to 0$. Значит, $\|L_v - L'_v\| = 0$, $L_v = L'_v$.

Рассмотрим теперь отображение $F : A \longmapsto A^{-1}$, $A \in Gl_n(\mathbb{k})$, определённое на открытом подмножестве $Gl_n(\mathbb{k})$, $\mathbb{k} = \mathbb{R}, \mathbb{C}$, пространства матриц $M_n(\mathbb{k})$, на котором выбрана и зафиксирована некоторая матричная норма $\| \|$. Напомним, что для любых $A \in Gl_n(\mathbb{k})$ и $\delta A \in M_n(\mathbb{k})$, $\|\delta A\| < \frac{1}{\|A^{-1}\|}$, имеет место

$$(A+\delta A)^{-1} \;=\; \sum_{k=0}^{\infty}(-1)^k(A^{-1}\delta A)^k A^{-1}\,.$$

Поэтому, как и в приведённой нами выше оценке,

$$(A+\delta A)^{-1}-A^{-1} \;=\; \sum_{k=1}^{\infty}(-1)^k(A^{-1}\delta A)^k A^{-1} \;=$$

$$-\,A^{-1}\delta A A^{-1}+\sum_{k=2}^{\infty}(-1)^k(A^{-1}\delta A)^k A^{-1} \;=\; L_{A^{-1}}\delta A+\alpha(A^{-1},\delta A)\,,$$

где $L_{A^{-1}}: B\longmapsto -A^{-1}BA^{-1}$, $B\in M_n(\Bbbk)$,

$$\|\alpha(A^{-1},\delta A)\| \;=\; \Big\|\sum_{k=2}^{\infty}(-1)^k(A^{-1}\delta A)^k A^{-1}\Big\| \le \sum_{k=2}^{\infty}\|A^{-1}\|^{k+1}\|\delta A\|^k \;=\; \frac{\|A^{-1}\|^3\|\delta A\|^2}{1-\|A^{-1}\|\|\delta A\|}$$

и, как следствие, $\|\alpha(A^{-1},\delta A)\|=O(\|\delta A\|^2)$ при $\|\delta A\|\to 0$.

Отсюда следует, что число обусловленности $k(A)$ в рассматриваемой норме можно определить посредством равенства

$$k(A) \;=\; \lim_{\sigma\to 0}\sup_{\delta A,\ \|\delta A\|\le\sigma\|A\|}\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\sigma\|A^{-1}\|}\,.$$

Действительно, в силу предыдущего при достаточно малых σ

$$\sup_{\delta A,\ \|\delta A\|\le\sigma\|A\|}\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\sigma\|A^{-1}\|}\le\sup_{\delta A,\ \|\delta A\|\le\sigma\|A\|}\frac{\|L_{A^{-1}}\delta A\|}{\sigma\|A^{-1}\|}+o(1)\le k(A)+o(1)$$

($o(1)$ при $\sigma\to 0$), так как $\|L_{A^{-1}}\delta A\|\le\|\delta A\|\|A^{-1}\|^2\le\sigma\|A^{-1}\|k(A)$. С другой стороны при $\delta A=\sigma\|A\|E$ мы имеем при $\sigma\to 0$

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\sigma\|A^{-1}\|} \;=\; \frac{\|(E+\sigma\|A\|A^{-1})^{-1}-E\|}{\sigma} \;=$$

$$\|\|A\|A^{-1}+\|A\|^2A^{-2}\sigma(E+\sigma\|A\|A^{-1})^{-1}\|\ge k(A)-\frac{k(A)^2\sigma}{1-\sigma k(A)} \;=\; k(A)+o(1)\,.$$

Поэтому оценка снизу у нас также имеется и мы приходим к заявленному равенству.

Скажем теперь несколько слов о качестве полученной нами оценки относительной погрешности в решении линейной системы $Ax=b$,

$$\frac{\|\hat{x}-x\|}{\|x\|}\le\frac{k(A)}{1-k(A)\frac{\|\delta A\|}{\|A\|}}\Big(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\Big)\,.$$

В действительности данная оценка представляет собой наиболее пессимистический сценарий развития событий и в большинстве случаев является существенно завышенной. Последнее, к примеру, особенно заметно при решении линейных систем с диагональными матрицами. В этом случае решение сводится к выполнению операции деления,

что не может привести к существенному возрастанию погрешности вне зависимости от числа обусловленности матрица системы (последнее может быть сколь угодно большим). В то же время данная оценка может оказаться и вполне достижимой. В качестве иллюстрации можно рассмотреть следующий пример.

Пусть имеется матрица $A \in GL_n(\mathbb{C})$ с сингулярным разложением

$$A \;=\; \sum_{i=1}^{n} \sigma_i v_i u_i^* \,.$$

Тогда решение $Ax = b$ определяется равенством

$$x \;=\; \sum_{i=1}^{n} \frac{(b, v_i)}{\sigma_i} u_i \,.$$

Рассмотрим возмущение данных $A + \delta A = A - \varepsilon v_n u_n^*$ и $b + \delta b = b + \varepsilon (b, v_n)/|(b, v_n)| v_n$, $\varepsilon > 0$. Решение возмущённой системы $(A + \delta A)\hat{x} = b + \delta b$ определяется равенством

$$\hat{x} \;=\; \sum_{i=1}^{n-1} \frac{(b, v_i)}{\sigma_i} u_i + \frac{(b + \varepsilon (b, v_n)/|(b, v_n)| v_n, v_n)}{\sigma_n - \varepsilon} u_n \;=\; \sum_{i=1}^{n-1} \frac{(b, v_i)}{\sigma_i} u_i + \frac{(b, v_n)(1 + \varepsilon/|(b, v_n)|)}{\sigma_n - \varepsilon} u_n \,,$$

из которого следует, что

$$\|\hat{x} - x\|_2 \;=\; \Big| \frac{(b, v_n)(1 + \varepsilon/|(b, v_n)|)}{\sigma_n - \varepsilon} - \frac{(b, v_n)}{\sigma_n} \Big| \;=\; \frac{\varepsilon(|(b, v_n)| + \sigma_n)}{\sigma_n |\sigma_n - \varepsilon|} \,.$$

Поскольку

$$\frac{\|b\|_2}{\sigma_n} \geq \|x\|_2 = \sqrt{\sum_{i=1}^{n} \frac{|(b, v_i)|^2}{\sigma_i^2}} \geq \frac{\|b\|_2}{\sigma_1} \,,$$

мы получаем, что

$$\frac{\varepsilon(|(b, v_n)| + \sigma_n)}{\|b\|_2 |\sigma_n - \varepsilon|} \leq \frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \frac{\varepsilon \sigma_1 (|(b, v_n)| + \sigma_n)}{\sigma_n \|b\|_2 |\sigma_n - \varepsilon|} \,.$$

Более того, взяв $b = v_n$, мы получим $\|x\|_2 = \frac{1}{\sigma_n}$ и

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \;=\; \frac{\varepsilon(1 + \sigma_n)}{|\sigma_n - \varepsilon|} \,.$$

При этом $\|\delta A\|_2 = \|\delta b\|_2 = \varepsilon$ и

$$\frac{k_2(A)}{1 - k_2(A) \frac{\|\delta A\|_2}{\|A\|_2}} \;=\; \frac{\frac{\sigma_1}{\sigma_n}}{1 - \frac{\varepsilon}{\sigma_n}} \;=\; \frac{\sigma_1}{\sigma_n - \varepsilon} \,, \quad \frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta b\|_2}{\|b\|_2} \;=\; \varepsilon \Big(\frac{1}{\sigma_1} + \frac{1}{\|b\|_2} \Big)$$

и, следовательно,

$$\frac{k_2(A)}{1 - k_2(A) \frac{\|\delta A\|_2}{\|A\|_2}} \Big(\frac{\|\delta A\|_2}{\|A\|_2} + \frac{\|\delta b\|_2}{\|b\|_2} \Big) \;=\; \frac{\varepsilon}{\sigma_n - \varepsilon} \Big(1 + \frac{\sigma_1}{\|b\|_2} \Big) \,.$$

Заметим, что при $\sigma_1 = \sigma_n = 1$, $b = v_n$ и $\varepsilon < 1$ полученная здесь оценка равна в точности относительной погрешности решения возмущённой системы (этот случай отвечает $U = E$, $A + \delta A = V - \varepsilon v_n e_n^*$, $b + \delta b = (1 + \varepsilon)v_n$).

Следует отметить, что для реального использования оценки относительной погрешности решения линейной системы (естественно речь идёт об оценке через вектор невязки) необходимо располагать качественной программой-оценщиком выбранной матричной нормы.

Лекция 2. Гауссово исключение и его варианты.

В этой лекции мы напомним классическую схему метода Гаусса решения линейной системы $Ax = b$ с невырожденной матрицей $A = (a_{ij}) \in Gl_n(\mathbb{k})$ и вектором правой части $b = (b_1, \ldots, b_n)^t \in \mathbb{k}^n$, $\mathbb{k} = \mathbb{R}$, $\mathbb{C}$, которая не преполагает использование перестановок строк (столбцов) матрицы системы. Идея метода состоит в последовательном переходе от исходной системы $A^{(0)}x = b^{(0)}$ с $A^{(0)} = A$ и $b^{(0)} = b$ к равносильным ей системам $A^{(k)}x = b^{(k)}$, $k = 1, \ldots, n$, где

$$A^{(k)} \;=\; \begin{pmatrix} 1 & c_{12} & \ldots & c_{1k} & c_{1k+1} & \ldots & c_{1n} \\ 0 & 1 & \ldots & c_{2k} & c_{2k+1} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & c_{kk+1} & \ldots & c_{kn} \\ 0 & 0 & \ldots & 0 & a^{(k)}_{k+1k+1} & \ldots & a^{(k)}_{k+1n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a^{(k)}_{nk+1} & \ldots & a^{(k)}_{nn} \end{pmatrix}$$

и $b^{(k)} = (y_1, \ldots, y_k, b^{(k)}_{k+1}, \ldots, b^{(k)}_n)^t$, а последняя система имеет треугольный вид

$$A^{(n)} \;=\; \begin{pmatrix} 1 & c_{12} & \ldots & c_{1n} \\ 0 & 1 & \ldots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

и $b^{(n)} = (y_1, \ldots, y_n)^t$. При этом переход на каждом шаге осуществляется посредством обратимых элементарных преобразований строк (в рассматриваемом случае) типа умножение строки на ненулевое число ("масштабирование") и вычитание из i-ой строки j-ой, $i \neq j$, умноженной на некоторое число.

Точнее типичный k-ый шаг рассматриваемого нами процесса, реализующий переход $(A^{(k-1)}, b^{(k-1)}) \to (A^{(k)}, b^{(k)})$, $k = 1, \ldots, n-1$, в предположимости его осуществимости состоит в следующем:

1. k-ая строка матрицы $A^{(k-1)}$ и k-ая компонента вектора $b^{(k-1)}$ делятся на ведущий элемент $a^{(k-1)}_{kk}$, в результате чего вычисляются их новые компоненты

$$c_{ki} \;=\; a^{(k-1)}_{ki}/a^{(k-1)}_{kk} \quad (i = k+1, \ldots, n)\,, \quad y_k \;=\; b^{(k-1)}_k/a^{(k-1)}_{kk}\,.$$

2. из каждой i-ой строки матрицы $A^{(k-1)}$, $i = k+1, \ldots, n$, вычитается перевычисленная на предыдущем этапе k-ая строка, домноженная на $a^{(k-1)}_{ik}$, и из каждой i-ой компоненты вектора $b^{(k-1)}$, $i = k+1, \ldots, n$, вычитается новая k-ая компонента y_k, умноженная на $a^{(k-1)}_{ik}$, при этом появляются компоненты новых матрицы $A^{(k)}$ и вектора $b^{(k)}$,

$$a^{(k)}_{ij} \;=\; a^{(k-1)}_{ij} - a^{(k-1)}_{ik}c_{kj}\,, \quad b^{(k)}_i \;=\; b^{(k-1)}_i - a^{(k-1)}_{ik}y_k \quad (i, j = k+1, \ldots, n)\,.$$

Остальные компоненты матрицы $A^{(k-1)}$ и вектора $b^{(k-1)}$ переходят в матрицу $A^{(k)}$ и вектор $b^{(k)}$ без изменений. Нетрудно заметить, что осуществимость данного процесса

сводится к выполнению условия $a_{kk}^{(k-1)} \neq 0$ для всех $k = 1, \ldots, n-1$, что возможно далеко не всегда. К примеру, для весьма простой матрицы

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

наш процесс не может быть осуществим. Кроме того, понятно, что вся вычислительная работа k-го шага после масштабирования k-ой строки и k-ой компоненты вектора правой части сосредоточена в рамках нижней главной подматрицы размера $(n-k) \times (n-k)$ и последних $n-k$ компонент вектора правой части. На последнем n-ом шаге осуществляется лишь масштабирование на $a_{nn}^{(n-1)}$.

После перехода к треугольной системе $A^{(n)}x = b^{(n)}$ с верхней унитреугольной матрицей $A^{(n)}$ искомый вектор x находится посредством обратного хода метода Гаусса (обратной подстановкой): $x_n = y_n$ и далее для всех $j = n-1, \ldots, 1$

$$x_i \;=\; y_i - \sum_{j=i+1}^{n} c_{ij} x_j \,.$$

Понятно, что основная вычислительная процедура рассматриваемого процесса сводится к преобразованию ведущей квадратной подматрицы, которая осуществляется в рамках тройного цикла. Различные порядки выполнения последнего приводят к появлению различных версий метода Гаусса (в терминах скалярных произведений и т.п.). Несложно также оценить вычислительную сложность представленного здесь алгоритма: на шаге k, $k = 1, \ldots, n$, выполняется $n-k+1$ делений (на масштабирование) и после $2(n-k+1)(n-k)$ умножений и вычитаний (на перевычисление новой ведущей подматрицы и новых компонент вектора правой части с индексами $k+1, \ldots, n$). Поэтому суммарно мы получаем

$$n + \sum_{k=1}^{n-1} 3(n-k) + 2\sum_{k=1}^{n}(n-k)^2 \;=\; n + 3/2n(n-1) + 1/3(n-1)n(2n-1) \;=\; 2/3n^3 + O(n^2)$$

при $n \to \infty$. Обратый ход нам обойдётся суммарно в $2\sum_{i=1}^{n-1}(n-i) = n(n-1)$ умножений и вычитаний. Таким образом, наш алгоритм имеет сложность $2/3n^3 + O(n^2)$.

В случае если имеется необходимость решения нескольких линейных систем $Ax^i = b^i$, $i = 1, \ldots, s$, с общей матрицей системы A, тогда описанный нами процесс естественно может быть интерпретирован в терминах матричного уравнения $AX = B$, где $B = (b^1, \ldots, b^s)$ и $X = (x^1, \ldots, x^s)$ матрицы $n \times s$ векторов правой части векторов решений. Естественно, что наш процесс модифицируется внесением дополнительных перевычислений компонент матрицы $B^{(k)}$, $k = 0, \ldots, n$. Последнее может изметить суммарный вклад в сложность алгоритма на величину порядка $2sn^2 + O(n, s)$.

LU–разложение и его варианты

Описанный выше алгоритм метода Гаусса имеет следующую естественную интерпретацию в терминах матричных умножений. Переход от матрицы $A^{(k-1)}$ и вектора $b^{(k-1)}$ к матрице $A^{(k)}$ и вектору $b^{(k)}$, осуществляемый на k-ом шаге алгоритма, $k = 1, \ldots, n$, соответствует умножению $A^{(k-1)}$ и $b^{(k-1)}$ слева на нижнюю треугольную матрицу $L_k D_k$,

где $D_k = \mathrm{diag}(1, \ldots, 1, 1/a_{kk}^{(k-1)}, 1, \ldots, 1)$ — диагональная матрица с элементом $1/a_{kk}^{(k-1)}$ на позиции (k,k) и L_k — нижняя унитреугольная матрица вида

$$L_k = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & -a_{k+1k}^{(k-1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -a_{nk}^{(k-1)} & 0 & \ldots & 1 \end{pmatrix},$$

где все внедиагональные элементы кроме отмеченных компонент k-го столбца ниже главной (единичной) диагонали равны нулю. При этом умножение на D_k отвечает масштабированию, а умножение на L_k — перевычислению новой ведущей квадратной подматрицы и новых компонент вектора правой части, $A^{(k)} = L_k D_k A^{(k-1)}$ и $b^{(k)} = L_k D_k b^{(k-1)}$, причём $L_n = E$.

Поскольку финальная матрица $A^{(n)} = D_n L_{n-1} \cdots L_1 D_1 A$ является верхней унитреугольной матрицей, мы можем интерпретировать наш процесс метода Гаусса как процесс нахождения представления матрицы системы в виде произведения двух матриц $A = LU$, где $L = (D_n L_{n-1} \cdots L_1 D_1)^{-1} = D_1^{-1} L_1^{-1} \cdots L_{n-1}^{-1} D_n^{-1}$ — нижняя треугольная матрица, а $U = A^{(n)}$ — верхняя унитреугольная матрица. Представление невырожденной матрицы A в таком виде называется также LU–разложением матрицы A (разложение в произведение нижней унитреугольной и верхней треугольной матрицы именуют LR–разложением). Так как каждая матрица $D_k^{-1} L_k^{-1}$ имеет вид

$$D_k^{-1} L_k^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & a_{kk}^{(k-1)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & a_{k+1k}^{(k-1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a_{nk}^{(k-1)} & 0 & \ldots & 1 \end{pmatrix},$$

мы можем записать матрицу L в виде

$$L = \begin{pmatrix} a_{11}^{(0)} & 0 & 0 & \ldots & 0 \\ a_{21}^{(0)} & a_{22}^{(1)} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n-11}^{(0)} & a_{n-12}^{(1)} & a_{n-13}^{(2)} & \ldots & 0 \\ a_{n1}^{(0)} & a_{n2}^{(1)} & a_{n3}^{(2)} & \ldots & a_{nn}^{(n-1)} \end{pmatrix}.$$

Заметим, что при наличии LU–разложения $A = LU$ решение системы $Ax = b$ сводится к решению двух треугольных систем

$$\begin{cases} Ly = b \\ Ux = y. \end{cases}$$

Обсудим теперь вопросы однозначности и осуществимости LU–разложения, а также способы его построения.

Замечание 0.13. *Если невырожденная матрица A обладает LU–разложением, то это разложение единственно.*

Доказательство. Действительно, если имеется два подобных представления $A = LU = L'U'$, тогда матрица $L'^{-1}L = U'U^{-1}$ является одновременно как нижней треугольной, так и верхней унитреугольной, а потому может быть только единичной матрицей, т.е. $L = L'$ и $U = U'$. □

Относительно способов построения LU–разложения следует заметить, что в соответствии со сказанным ранее мы вполне можем повторить все шаги метода Гаусса с той лишь разницей, что на каждом шаге нам следует сохранить компоненты первого столбца ведущей квадратной подматрицы на их местах (поскольку матрица U унитреугольна, её единичные диагональные составляющие нас не интересуют). Иначе говоря, мы можем модифицировать метод Гаусса с целью нахождения LU–разложения следующим образом. На первом шаге мы переходим от матрицы $A = A^{(0)}$ к матрице $A'^{(1)}$ вида

$$A'^{(1)} = \begin{pmatrix} a_{11}^{(0)} & c_{12} & \dots & c_{1n} \\ a_{21}^{(0)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^{(0)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix},$$

а на шаге k должны получить матрицу $A'^{(k)}$,

$$A'^{(k)} = \begin{pmatrix} a_{11}^{(0)} & c_{12} & \dots & c_{1k} & c_{1k+1} & \dots & c_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{k-11}^{(0)} & a_{k-12}^{(1)} & \dots & c_{k-1k} & c_{k-1k+1} & \dots & c_{k-1n} \\ a_{k1}^{(0)} & a_{k2}^{(1)} & \dots & a_{kk}^{(k-1)} & c_{kk+1} & \dots & c_{kn} \\ a_{k+11}^{(0)} & a_{k+12}^{(1)} & \dots & a_{k+1k}^{(k-1)} & a_{k+1k+1}^{(k)} & \dots & a_{k+1n}^{(k)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1}^{(0)} & a_{n2}^{(1)} & \dots & a_{nk}^{(k-1)} & a_{nk+1}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix},$$

где все указанные здесь компоненты такие же, как и в методе Гаусса. В итоге после выполнения $n-2$ шагов мы получим матрицу $A'^{(n)}$

$$A'^{(n)} = \begin{pmatrix} a_{11}^{(0)} & c_{12} & \dots & c_{1n} \\ a_{21}^{(0)} & a_{22}^{(1)} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^{(0)} & a_{n2}^{(1)} & \dots & a_{nn}^{(n-1)} \end{pmatrix},$$

в которой верхний треугольник выше главной диагонали заполнен компонентами U–составляющей матрицы A, а нижний треугольник и диагональ — компонентами её L–составляющей. Понятно, что сложность такого алгоритма идентична сложности триангуляции в методе Гаусса и составляет $2/3n^3 + O(n^2)$ при $n \to \infty$.

Можно привести и иную вычислительную процедуру построения LU–разложения матрицы A. Итак, нам требуется построить матрицы $L = (l_{ij})$, $l_{ij} = 0$ при $1 \le i < j \le n$, и $U = (u_{ij})$, $u_{ij} = 0$ при $1 \le j < i \le n$, $u_{11} = \ldots = u_{nn} = 1$, связанные с исходной матрицей A равенством $A = LU$ или, что равносильно, системой из n^2 уравнений относительно n^2 неизвестных $\{l_{pq}, u_{ij}\}$,

$$a_{ij} \;=\; \sum_{k=1}^{n} l_{ik}u_{kj} \;=\; \sum_{k=1}^{\min\{i,j\}} l_{ik}u_{kj} \quad (i,j = 1,\ldots,n)\,.$$

Разобъём эту систему уравнений на две группы по $n(n+1)/2$ и $n(n-1)/2$ уравнений, отвечающим парам индексов в нижнем треугольнике и верхнем наддиагональном треугольнике, т.е.

$$a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}u_{kj} \;=\; l_{ij} + \sum_{k=1}^{j-1} l_{ik}u_{kj} \quad (1 \le j \le i \le n)\,, \tag{1}$$

и

$$a_{ij} \;=\; \sum_{k=1}^{i} l_{ik}u_{kj} \;=\; l_{ii}u_{ij} + \sum_{k=1}^{i-1} l_{ik}u_{kj} \quad (1 \le i < j \le n)\,, \tag{2}$$

где суммы при $i, j = 1$ опускаются. Первую группу соотношений можно переписать в виде

$$l_{ij} \;=\; a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \quad (1 \le j \le i \le n)\,,$$

а вторую — в виде

$$u_{ij} \;=\; \Big(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\Big)/l_{ii} \quad (1 \le i < j \le n)\,.$$

Последнее возможно, поскольку осуществимость нашего процесса равносильна $l_{ii} \ne 0$, $i = 1,\ldots,n$. Заметим, что в приведённых здесь равенствах компоненты j-го столбца L вычисляются через компоненты с теми же индексами матрицы A и компоненты первых $j-1$-го столбца матрицы L и первых $j-1$-ой строки матрицы U, компоненты i-ой строки матрицы U — через компоненты i-ой строки A и компоненты первых $i-1$-го столбца L и первых $i-1$-ой строки U. Поэтому мы можем организовать следующую вычислительную процедуру. На первом шаге

1. вычисляются компоненты первого столбца матрицы L, $l_{i1} = a_{i1}$, $i = 1,\ldots,n$;
2. вычисляются внедиагональные компоненты первой строки матрицы U, $u_{1j} = a_{1j}/l_{11}$, $j = 2,\ldots,n$.

На k-ом шаге после нахождения первых $k-1$-го столбца матрицы L и первых k-ой строки матрицы U

1. вычисляются компоненты k-го столбца L,

$$l_{ik} \;=\; a_{ik} - \sum_{s=1}^{k-1} l_{is}u_{sk} \quad (i = k,\ldots,n);$$

2. вычисляются внедиагональные компоненты k-ой строки U,

$$u_{kj} \;=\; \Big(a_{kj} - \sum_{t=1}^{k-1} l_{kt}u_{tj}\Big)/l_{kk} \quad (j = k+1, \ldots, n)\,.$$

Процесс завершается на n-ом шаге вычислением

$$l_{nn} \;=\; a_{nn} - \sum_{s=1}^{n-1} l_{ns}u_{sn}\,.$$

Поскольку на k-ом шаге для вычисления (i,k)-ой компоненты матрицы L или (k,j)-ой компоненты матрицы U мы однократно используем элемент a_{ik} или a_{kj} исходной матрицы, к которому больше не обращаемся, вычисляемые компоненты L и U на k-ом шаге имеет смысл хранить на месте соответствующих компонент k-го столбца и k-ой строки матрицы A. Несложно показать, что вычислительная сложность описанной процедуры по-прежнему составляет $2/3n^3 + O(n^2)$, $n \to \infty$.

Обсудим теперь условия осуществимости метода Гаусса для системы $Ax = b$ или, что то же самое, условия существования LU–разложения для матрицы A.

Теорема 0.14. *Невырожденная матрица $A = (a_{ij})$ обладает LU–разложением в том и только в том случае, если определители всех её главных квадратных подматриц $A_k = (a_{ij})_{i,j=1}^{k}$, $k = 1, \ldots, n$, отличны от нуля.*

Доказательство. Начнём с необходимости рассматриваемого нами условия. Если матрица A обладает LU–разложением $A = LU$, тогда, как нетрудно заметить, $A_k = L_kU_k$ и $\det A_k = \det L_kU_k = \det L_k = l_{11}\cdots l_{kk} \neq 0$, $k = 1, \ldots, n$.

Пусть теперь $\det A_k \neq 0$, $k = 1, \ldots, n$. Для доказательства существования LU–разложения матрицы A в этом случае можно воспользоваться индукцией по n с основанием $n = 1$ ($L = a_{11}$ и $U = 1$). Предположим, что при $1 \leq n < m$ наше утверждение уже доказано. Тогда для $n = m$ мы имеем

$$A \;=\; \begin{pmatrix} A_{m-1} & u \\ v^t & a_{mm} \end{pmatrix},$$

где u и v^t — блоки размера $(m-1)\times 1$ и $1\times(m-1)$, состоящие из всех компонент последнего столбца и последней строки матрицы A за исключением их общей компоненты a_{mm}. По предположению индукции матрица A_{m-1} обладает LU–разложением $A_{m-1} = L_{m-1}U_{m-1}$ и, как следствие,

$$\begin{pmatrix} L_{m-1}^{-1} & 0 \\ 0 & 1 \end{pmatrix} A \begin{pmatrix} U_{m-1}^{-1} & 0 \\ 0 & 1 \end{pmatrix} \;=\; \begin{pmatrix} E_{m-1} & u' \\ v'^t & a_{mm} \end{pmatrix} \;=\; \begin{pmatrix} E_{m-1} & 0 \\ v'^t & a'_{mm} \end{pmatrix} \begin{pmatrix} E_{m-1} & u' \\ 0 & 1 \end{pmatrix},$$

где $u' = L_{m-1}^{-1}u$, $v'^t = v^tU_{m-1}^{-1}$ и $a'_{mm} = a_{mm} - v'^tu' \neq 0$ в силу того, что $\det A \neq 0$ (E_{m-1} — единичная матрица размера $(m-1)\times(m-1)$). Остаётся положить

$$L \;=\; \begin{pmatrix} L_{m-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} E_{m-1} & 0 \\ v'^t & a'_{mm} \end{pmatrix} \;=\; \begin{pmatrix} L_{m-1} & 0 \\ v'^t & a'_{mm} \end{pmatrix},$$

$$U \;=\; \begin{pmatrix} E_{m-1} & u' \\ 0 & 1 \end{pmatrix} \begin{pmatrix} U_{m-1} & 0 \\ 0 & 1 \end{pmatrix} \;=\; \begin{pmatrix} U_{m-1} & u' \\ 0 & 1 \end{pmatrix}.$$

□

Можно также предложить иной путь обоснования приведённого здесь результата, который опирается на представление LU–разложения в терминах матричных миноров. Действительно, для выбранной нами матрицы $A = (a_{ij})$ обозначим через $a_j(i)$ и $a^j(i)$ вектора, составленные из первых i компонент её j-го столбца и j-ой строки,

$$a_j(i) \;=\; (a_{1j}, \ldots, a_{ij})^t\,, \quad a^j(i) \;=\; (a_{j1}, \ldots, a_{ji})\,.$$

Построим также матрицы

$$A_{ik} \;=\; (a_1(i)\ldots a_{i-1}(i)a_k(i)) \;=\; \begin{pmatrix} a_{11} & \ldots & a_{1i-1} & a_{1k} \\ a_{21} & \ldots & a_{2i-1} & a_{2k} \\ \ldots\ldots & \ldots & \ldots\ldots & \ldots\ldots \\ a_{i1} & \ldots & a_{ii-1} & a_{ik} \end{pmatrix}\,,$$

$$A^{ki} \;=\; \begin{pmatrix} a^1(i) \\ \vdots \\ a^{i-1}(i) \\ a^k(i) \end{pmatrix} \;=\; \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1i} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{i-11} & a_{i-12} & \ldots & a_{i-1i} \\ a_{k1} & a_{k2} & \ldots & a_{ki} \end{pmatrix}\,,$$

где $1 \le i \le k \le n$ и $A_{1k} = a_{1k}$, $A^{k1} = a_{k1}$ при $i = 1$. Предположим, что матрица A имеет LU–разложение $A = LU$. Тогда

$$\begin{cases} a_1(i) \;=\; l_1(i)\,, \\ a_2(i) \;=\; l_1(i)u_{12} + l_2(i)\,, \\ \qquad \ldots \\ a_{i-1}(i) \;=\; l_1(i)u_{1i-1} + \ldots + l_{i-2}(i)u_{i-2i-1} + l_{i-1}(i)\,, \end{cases}$$

а при $k \ge i$

$$a_k(i) \;=\; l_1(i)u_{1k}+\ldots+l_i(i)u_{ik}+l_{i+1}(i)u_{i+1k}+\ldots+l_{k-1}(i)u_{k-1k}+l_k(i) \;=\; l_1(i)u_{1k}+\ldots+l_i(i)u_{ik}\,.$$

Поэтому $A_{ik} = L_iU_{ik}$, $\det A_{ik} = l_{11}\cdots l_{ii}u_{ik}$, $1 < i \le k \le n$, и $\det A_{1k} = l_{11}u_{1k} = a_{1k}$. Следовательно,

$$u_{ik} \;=\; \frac{\det A_{ik}}{\det A_{ii}} \quad (1 \le i \le k \le n)\,.$$

Сходным образом,

$$\begin{cases} a^1(i) \;=\; l_{11}u^1(i)\,, \\ a^2(i) \;=\; l_{21}u^1(i) + l_{22}u^2(i)\,, \\ \qquad \ldots \\ a^{i-1}(i) \;=\; l_{i-11}u^1(i) + \ldots + l_{i-1i-1}u^{i-1}(i)\,, \end{cases}$$

а при $k \ge i$

$$a^k(i) \;=\; l_{k1}u^1(i) + \ldots + l_{ki}u^i(i) + l_{ki+1}u^{i+1}(i) + \ldots + l_{kk}u^k(i) \;=\; l_{k1}u^1(i) + \ldots + l_{ki}u^i(i)\,.$$

Значит, $A^{ki} = L^{ki}U_i$, $\det A^{ki} = l_{11}\cdots l_{i-1i-1}l_{ki}$, $1 < i \le k \le n$, и $\det A^{k1} = l_{k1} = a_{k1}$. Поэтому $l_{k1} = \det A^{k1}$, $k = 1, \ldots, n$, и

$$l_{ki} \;=\; \frac{\det A^{ki}}{\det A^{i-1i-1}} \quad (2 \le i \le k \le n)\,.$$

При этом $\det A^{ii} = \det A_{ii} = \det A_i$, $i = 1, \ldots, n$. Отсюда следует, что в случае $\det A_i \neq 0$, $i = 1, \ldots, n$, мы можем построить треугольные матрицы L и U по описанным здесь формулам и остаётся лишь проверить, что произведение таких матриц совпадает с исходной матрицей A. Последнее можно сделать непосредственно (прямой подстановкой), а можно свести к равенству многочленов на бесконечном множестве значений переменных (при малых изменениях коэффициентов LU–разложимой матрицы получаются LU–разложимые матрицы).

В общем случае произвольная невырожденная матрица обладает разложением, отличающимся от LU–разложения добавлением матриц-перестановок. Напомним, что под матрицей перестановкой понимается любая матрица P, полученная из единичной матрицы E в результате перестановки столбцов (строк). Матрица P такого вида может быть интерпретирована как матрица линейного преобразования, переставляющего элементы базиса. Она ортогональна, $P^t = P^{-1}$, и представима в виде произведения симметрических матриц перестановок P_{ij}, которые получаются из единичной матрицы E перестановкой i-го и j-го столбцов.

Теорема 0.15. *Для любой невырожденной матрицы A можно подобрать матрицы перестановки P, P' или P_1 и P_2, такие, что матрицы PA, AP' и P_1AP_2 обладают LU–разложением.*

Доказательство. Для определённости мы будем рассматривать случай двух матриц-перестановок. Случаи одной матрицы-перестановки получает простым упрощением приводимого ниже доказательства. Воспользуемся индукцией по размерности матрицы A с тривиальным основанием $n = 1$. Пусть для всех $1 \leq n < m$ наше утверждение доказано. Рассмотрим случай $n = m$. Домножая матрицу A при необходимости на некоторые матрицы-перестановки P' и P'' слева и справа (переставляя строки и столбцы A), мы можем перейти к матрице

$$P'AP'' = \begin{pmatrix} a_{11} & b^t \\ a & A_{m-1} \end{pmatrix},$$

где $a_{11} \neq 0$, b^t и a — строчный и столбцовый блоки размера $1 \times (m-1)$ и $(m-1) \times 1$. Заметим, что мы всегда можем добится этого, переставляя только строки или только столбцы матрицы A (т.е. одну из матриц P' или P'' можно считать единичной). Полученную матрицу можно записать в виде

$$\begin{pmatrix} a_{11} & b^t \\ a & A_{m-1} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ a & E_{m-1} \end{pmatrix} \begin{pmatrix} 1 & b'^t \\ 0 & A'_{m-1} \end{pmatrix},$$

где $b' = b/a_{11}$ и $A'_{m-1} = A_{m-1} - ab'^t$. Матрица A'_{m-1} невырождена и для неё по предположению индукции найдутся матрицы-перестановки H_1 и H_2, для которых $H_1A'_{m-1}H_2 = L_{m-1}U_{m-1}$. Тогда

$$\begin{pmatrix} 1 & 0 \\ 0 & H_1 \end{pmatrix} P'AP'' \begin{pmatrix} 1 & 0 \\ 0 & H_2 \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ H_1a & E_{m-1} \end{pmatrix} \begin{pmatrix} 1 & b'^tH_2 \\ 0 & L_{m-1}U_{m-1} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ H_1a & L_{m-1} \end{pmatrix} \begin{pmatrix} 1 & b'^tH_2 \\ 0 & U_{m-1} \end{pmatrix}.$$

Остаётся положить

$$P_1 \;=\; \begin{pmatrix} 1 & 0 \\ 0 & H_1 \end{pmatrix} P', \quad P_2 \;=\; P'' \begin{pmatrix} 1 & 0 \\ 0 & H_2 \end{pmatrix}.$$

□

Уместно будет привести пример класса матриц, обладающих LU–разложением.

Будем говорить, что матрица $A = (a_{ij})$ имеет *строгое диагональное преобладание по строкам* (или является *матрицей со строгим диагональным преобладанием по строкам*), если при каждом $i = 1, \ldots, n$ выполняется неравенство

$$|a_{ii}| > \sum_{1 \le j \le n,\ j \ne i} |a_{ij}|.$$

Аналогичным образом вводится понятие матрицы со строгим диагональным преобладанием по столбцам (как матрица, транспонированная к которой имеет строгое диагональное преобладание по строкам).

Замечание 0.16. *Любая матрица $A = (a_{ij})$ со строгим диагональным преобладанием по строкам (столбцам) обладает LU–разложением.*

Доказательство. Достаточно доказать это утверждение для произвольной матрицы со строгим диагональным преобладанием по строкам (поскольку в таком случае для всякой матрицы B со строгим диагональным преобладанием по столбцам матрица B^t является LU–разложимой или, что равносильно, имеет ненулевые определители $\det B_k^t = \det B_k$, $k \ge 1$, а это в свою очередь гарантирует LU–разложимость матрицы B).

Итак, пусть матрица $A = (a_{ij})$ имеет строгое диагональное преобладание по строкам. Покажем, что все ведущие квадратные подматрицы матрицы A, возникающие на этапах выполнения метода Гаусса, являются матрицами со строгим диагональным преобладанием по строкам. Достаточно проверить это на первом шаге, а затем воспользоваться индукцией по размерности матрицы. Коэффициенты ведущей подматрицы матрицы $A^{(1)}$ определяются соотношениями

$$a_{ij}^{(1)} \;=\; a_{ij} - a_{i1}a_{1j}/a_{11} \quad (i, j = 2, \ldots, n).$$

Поэтому при любом $i = 2, \ldots, n$ справедливы соотношения

$$\sum_{2 \le j \le n,\ j \ne i} |a_{ij}^{(1)}| \;=\; \sum_{2 \le j \le n,\ j \ne i} |a_{ij} - a_{i1}a_{1j}/a_{11}| \le$$

$$\sum_{2 \le j \le n,\ j \ne i} (|a_{ij}| + |a_{i1}||a_{1j}|/|a_{11}|) < \Big(\sum_{2 \le j \le n,\ j \ne i} |a_{ij}| \Big) + |a_{i1}|(|a_{11}| - |a_{1i}|)/|a_{11}| \;=$$

$$\Big(\sum_{1 \le j \le n, j \ne i} |a_{ij}| \Big) - |a_{i1}||a_{1i}|/|a_{11}| \le |a_{ii}| - |a_{i1}||a_{1i}|/|a_{11}| \le |a_{ii} - a_{i1}a_{1i}/a_{11}| \;=\; |a_{ii}^{(1)}|,$$

при выводе которых мы воспользовались неравенствами

$$\sum_{2 \le j \le n,\ j \ne i} |a_{1j}| < |a_{11}| - |a_{1i}|, \qquad \sum_{1 \le j \le n,\ j \ne i} |a_{ij}| < |a_{ii}|.$$

□

Завершая разговор о вариантах представления невырожденных матриц в виде произ-

ведения нижних и верхних треугольных матриц и матриц перестановок, стоит сказать несколько слов о разложении Брюа. Не вдаваясь в подробности точных определении из теории линейных алгебраических групп (без обращения к понятиям тора, борелевской подгруппы и группы Вейля), мы будем понимать по разложением Брюа элемента A группы $Gl_n(\Bbbk)$, $\Bbbk = \mathbb{R}$, $\mathbb{C}$, его представление в виде $A = LPL'$, где L и L' — невырожденные нижние треугольные матрицы, P — матрица-перестановка. Впрочем для нас в дальнейшем при обсуждении алгоритмов поиска собственных значений более существенным будет нахождение не самого разложения Брюа матрицы A, а его вычисление его модифицированного варианта: $A = LPU$, в котором L и U — невырожденные нижняя и верхняя треугольные матрицы и P — матрица перестановка. Между раздложением Брюа и модифицированным разложением Брюа имеется тесная взаимосвязь. Точнее если известно модифицированное разложение Брюа матрицы $AQ = LPU$, где Q — матрица-перестановка вида

$$Q = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix},$$

тогда разложение Брюа матрицы A можно определить как $A = L(PQ)L'$, полагая $L' = QUQ$. Сходным образом на основе разложения Брюа матрицы AQ можно построить модифицированное разложение Брюа матрицы A.

Теорема 0.17. *Любая невырожденная матрица A обладает модифицированным разложением Брюа $A = LPU$, в котором матрица-перестановка P определена однозначно.*

Доказательство. Начнём с осуществимости данного разложения. Для его нахождения можно предложить следующий алгоритм, позволяющий построить в результате выполнения n шагов, где n — размерность исходной матрицы $A^{(0)} = A = (a_{ij}) \in Gl_n(\Bbbk)$, последовательность матриц $A^{(i)} = L_i A^{(i-1)} U_i$, $i = 1, \dots, n$, в которой матрица $A^{(i)}$ получается из своей предшественницы $A^{(i-1)}$ домножением слева и справа на подходящие невырожденные нижнюю и верхнюю треугольные матрицы L_i и U_i, а последняя матрица $A^{(n)}$ является матрицей-перестановкой. При этом требумое разложение определяется равенствами $P = A^{(n)}$, $L = (L_n \cdots L_1)^{-1} = L_1^{-1} \cdots L_n^{-1}$, $U = (U_1 \cdots U_n)^{-1} = U_n^{-1} \cdots U_1^{-1}$. Опишем типичный k-ый шаг этого процесса, соответствующий переходу от матрицы $A^{(k-1)}$ к матрице $A^{(k)} = L_k A^{(k-1)} U_k$ и построению матриц L_k и U_k. Он состоит в следующем: в k-ой строке матрицы $A^{(k-1)}$ находится первый ненулевой элемент $a_{ki_k}^{(k-1)}$ (т.е. $a_{ki}^{(k-1)} = 0$ при всех $1 \leq i < i_k$), затем осуществляется масштабирование k-ой строки матрицы $A^{(k-1)}$ делением её элементов на a_{ki_k} и вычитание из каждой строки с номером $s = k+1, \dots, n$ и каждого столбца с номером $t = i_k + 1, \dots n$, k-ой строки и i_k-го столбца масштабированной матрицы $A^{(k-1)'}$ домноженных на элементы $a_{si_k}^{(k-1)'} = a_{si_k}^{(k-1)}$ и $a_{kt}(k-1)' = a_{kt}^{(k-1)} / a_{ki_k}^{(k-1)}$ с целью обнуления всех компонент i_k-го столбца, начиная с $k+1$-ой, и всех компонент k-строки, начиная с $i_k + 1$-ой. В терминах матричных умножений эти действия реализуются в виде перехода от $A^{(k-1)}$ к $A^{(k)} = L_k A^{(k-1)} U_k$, где матрицы L_k и U_k отличаются от единичной только своими k-ым столбцом и i_k-ой строкой, соответственно,

$$L_k \;=\; \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/a_{ki_k}^{(k-1)} & 0 & \dots & 0 \\ 0 & 0 & \dots & -a_{k+1i_k}^{(k-1)}/a_{ki_k}^{(k-1)} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -a_{ni_k}^{(k-1)}/a_{ki_k}^{(k-1)} & 0 & \dots & 1 \end{pmatrix},$$

$$U_k \;=\; \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -a_{ki_k+1}^{(k-1)}/a_{ki_k}^{(k-1)} & \dots & -a_{kn}^{(k-1)}/a_{ki_k}^{(k-1)} \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Заметим, что по построению в матрице $A^{(k-1)}$ во всех строках с номерами $1,\dots,k-1$ и столбцах с номерами $i_1,\dots,i_{k-1}$ стоит один единственный ненулевой элемент $a_{si_s}^{(k-1)} = a_{si_s}^{(s)} = 1$, $s = 1,\dots,k-1$. Поэтому номер i_k отличен от всех номеров $i_1,\dots,i_k$ и потому $a_{si_k}^{(k-1)} = 0$, $s = 1,\dots,k-1$. Это также означает, что преобразования k-го шага не затрагивают первые $k-1$ строку и столбцы с номерами $i_1,\dots,i_{k-1}$ матрицы $A^{(k-1)}$. Последний шаг нашего построения сводится к масштабированию последней строки матрицы $A^{(n-1)}$, содержащей единственный ненулевой элемент на позиции (n,i_n), т.е. в данном случае $L_n = \mathrm{diag}(1,\dots,1,1/a_{ni_n}^{(n-1)})$, $U_n = E$.

Отметим, что выбор матриц L и U вообще говоря не является однозначным. Например, матрица U из рассматримаего нами алгоритма является унитреугольной, а матрица L имеет вообще говоря неединичные диагональные элементы. Если бы мы использовали вместо масштабирования строк масштабирование столбцов, то получили бы наоборот унитреугольную матрицу L и матрицу U с диагональными элементами, которые могут быть неединичными.

Обсудим теперь единственность матрицы-перестановки P. Для каждой невырожденной матрицы $A \in Gl_n(\mathbb{k})$ и любого $i = 1,\dots,n$ положим

$$\sigma_i(A) \;=\; \min\{j \mid a^i(j) \notin \langle a^1(j),\dots,a^{i-1}(j)\rangle_{\mathbb{k}}\},$$

т.е. минимальный номер j, для которого строка $a^i(j) = (a_{i1},\dots,a_{ij})$, составленная из первых j компонент i-ой строки A линейно независима со строками $a^l(j) = (a_{l1},\dots,a_{lj})$, $l = 1,\dots,i-1$, составленными из первых j компонент её строк с меньшими номерами. В частности, для матрицы перестановки $P = (p_{ij})$, $p_{si_s} = 1$, $p_{sj} = 0$ для $1 \le j \ne i_s \le n$, $1 \le i_1 \ne \dots \ne i_n \le n$, мы сразу получаем, что $\sigma_s(P) = i_s$, $s = 1,\dots,n$.

Покажем теперь, что для любых невырожденных нижней и верхней треугольных матриц L и U верны равенства $\sigma_i(A) = \sigma_i(LA) = \sigma_i(AU) = \sigma_i(LAU)$, $i = 1,\dots,n$. Действительно, в матрице $B = LA$

$$b^s(j) \;=\; \sum_{t=1}^{s} l_{st}a^t(j) \quad (s,j = 1,\dots,n),$$

где $l_{ss} \neq 0$. Поэтому $\langle b^1(j), \ldots, b^{i-1}(j)\rangle = \langle a^1(j), \ldots, a^{i-1}(j)\rangle$ и $b^i(j)$ входит в это подпространство тогда и только и только тогда, когда в него входит $a^i(j)$, т.е. $\sigma_i(B) = \sigma_i(A)$, $i = 1, \ldots, n$. Соответственно в матрице $C = AU$ имеет место равенство $c^i(j) = a^i(j)U_j$, $U_j = (u_{pq})_{p,q=1}^j$, $i, j = 1, \ldots, n$. Следовательно, $\langle c^1(j), \ldots, c^{i-1}(j)\rangle = \langle a^1(j), \ldots, a^{i-1}(j)\rangle U_j$ включает в себя $c^i(j) = a^i(j)U_j$, если и только если $\langle a^1(j), \ldots, a^{i-1}(j)\rangle$ содержит $a^i(j)$, т.е. $\sigma_i(C) = \sigma_i(A)$, $i = 1, \ldots, n$.

Применяя данное наблюдение к матрице A с модифицированным разложением Брюа $A = LPU$, мы немедленно получаем, что $\sigma_i(A) = \sigma_i(P)$, $i = 1, \ldots, n$, и $P = (p_{ij})$, где $p_{i\sigma_i(A)} = 1$, $p_{ij} = 0$, $1 \leq j \neq \sigma_i(A) \leq n$. □

Особенности реализации метода Гаусса и пути улучшения его численной устойчивости

Обсудим теперь вопросы численной устойчивости метода Гаусса и связаного с ним алгоритма построения LU–разложения. Для начала следует отметить следующий результат, который доказывается несложным анализом погрешности шага алгоритма построения LU–разложения в модели арифметики с плавающей запятой с использованием индукции по размерности.

Предложение 0.18. *Пусть имеется невырожденная матрица $A \in Gl_n(\mathbb{R})$, заданная в арифметике с плавающей запятой, для которой осуществим алгоритм построения LU–разложения. Тогда вычисленные в результате его выполнения компоненты разложения $\hat{L}$ и $\hat{U}$ удовлетворяют соотношениям*

$$\hat{L}\hat{U} \;=\; A + \delta A\,, \quad |\delta A| \leq 3(n-1)\varepsilon(|A| + |\hat{L}||\hat{U}|) + O(\varepsilon^2)\,,$$

где ε — используемая машинная точность, $|B| = (|b_{ij}|)$ для любой матрицы $B = (b_{ij})$ и матричное неравенство $C \leq D$ подразумевает выполнение неравенств $c_{ij} \leq d_{ij}$ для всех компонент матриц $C = (c_{ij})$ и $D = (d_{ij})$.

Приведённая здесь оценка погрешности является в достаточной степени точной. Она показывает, что реализация метода Гаусса или алгоритма построения LU–разложения без использования дополнительных преобразований матрицы системы может привести к значительному росту абсолютных значений коэффициентов матриц треугольных факторов, что в свою очередь может привести к значительным потерям в точности получаемого результата. При этом основной причиной подобного возрастания модулей коэффициентов является малость ведущего элемента, используемого в процессе масштабирования, осуществляемого на каждом шаге метода. Последнее имеет отношение не только к плохо обусловленным матрицам, но и к матрицам с малым числом обусловленности. Например, матрица

$$A \;=\; \begin{pmatrix} \epsilon & 1 \\ 1 & 0 \end{pmatrix},$$

где $\epsilon > 0$, имеет собственные значения $\alpha_{1,2} = 1/2(\epsilon \pm \sqrt{\epsilon^2 + 4})$ и число обусловленности

$$k_2(A) \;=\; \frac{\epsilon + \sqrt{\epsilon^2+4}}{\sqrt{\epsilon^2+4} - \epsilon} \;=\; \frac{(\epsilon + \sqrt{\epsilon^2+4})^2}{4}\,,$$

которое стремится к единице при $\epsilon \to 0$. Вместе с тем LU–разложение такой матрицы имеет вид

$$A = \begin{pmatrix} \epsilon & 0 \\ 1 & -1/\epsilon \end{pmatrix} \begin{pmatrix} 1 & 1/\epsilon \\ 0 & 1 \end{pmatrix}.$$

Поэтому в данном случае при малых ϵ компоненты факторов LU–разложения могут быть весьма значительными.

В качестве выхода из создавшегося положения используются модернизации метода Гаусса, использующие три основных стратегии выбора ведущего элемента на k-ом шаге алгоритма — это стратегии частичного выбора по первым строке или столбцу ведущей подматрицы и стратегия полного выбора по всей ведущей подматрице. Идея этих модернизаций состоит в следующем изменении процедуры вычисления матрицы $A^{(k)}$ на k-ом, $1 \le k \le n-1$, шаге:

1. среди элементов первого столбца $a_{ik}^{(k-1)}$, $i = k, \ldots, n$, ведущей подматрицы находится элемент $a_{i_k k}^{(k-1)}$ с наибольшим модулем, после чего осуществляется переименование: k-ая строка матрицы $A^{(k-1)}$ становится i_k-ой, а i_k-ая строка — k-ой (т.е. осуществляется перестановка местами k-ой и i_k-ой строк матрицы $A^{(k-1)}$), а затем выполняется обычный (k-ый) шаг метод Гаусса (стратегия частичного выбора по столбцу);

2. среди элементов первой строки $a_{kj}^{(k-1)}$, $j = k, \ldots, n$, ведущей подматрицы находится элемент $a_{kj_k}^{(k-1)}$ с наибольшем модулем, после чего осуществляется перенумерация переменных (столбцов): x_k и x_{j_k} меняются местами (т.е. осуществляется перестановка местами k-го и j_k-го столбцов матрицы $A^{(k-1)}$), вслед за этим выполняется k-ый шаг метода Гаусса в привычной форме (стратегия частичного выбора по строке);

3. среди всех элементов $a_{ij}^{(k-1)}$, $i, j = k, \ldots, n$, ведущей подматрицы находится элемент $a_{i_k j_k}^{(k-1)}$ с наибольшим модулем и осуществляется перестановка местами строк с номерами k и i_k и столбцов с номерами k и j_k матрицы $A^{(k-1)}$, после чего выполняется обычный k-ый шаг метода Гаусса (стратегия полного выбора по всей матрице).

Шаг с номером n (масштабирование последней строки матрицы $A^{(n-1)}$) остаётся без изменений. Заметим, что метод Гаусса с выбором ведущего элемента по строке, столбцу или всей матрице применим для решения линейных систем с любой невырожденной матрицей, а не только с матрицей, обладающей LU–разложением. При решении линейной системы перестановка (перенумерование) строк матрицы, используемая в рамках выбора по столбцу или всей матрице, фактически не осуществляется (она просто становится ведущей строкой). Вместе с тем перенумерация переменных в алгоритме с выбором по строке и всей матрице является нефиктивным действием с точки зрения решения линейной системы. Поэтому возникает необходимость хранить информацию о перестановках такого рода в виде отдельного вектора, по которому можно восстановить выполненную нами перестановку переменных. Отметим и то, что стратегии частичного выбора вносят несущественный вклад в усложнение метода Гаусса, так как $n-k$ сравнений на k-ом шаге добавляют в совокупности $O(n^2)$ операций. В отличие от них стратегия полного выбора предполагает выполнение $(n-k)^2$ сравнений на k-ом шаге, что добавляет $n^3/3 + O(n^2)$ операций к уже имеющимся $2/3n^3 + O(n^2)$.

Скажем несколько слов об интерпретации описанных алгоритмов в терминах матричных умножений. Начнём с метода Гаусса с выбором по столбцу. Перенумерация строк, осуществляемая на k-ом шаге, соответствует умножению слева матрицы $A^{(k-1)}$ на матрицу перестановку P_{ki_k}, которая получается перестановкой столбцов с номерами k и i_k, $k \le i_k \le n$, единичной матрицы E (мы предполагаем, что $P_{kk} = E$). Затем выполняется стандартный шаг метода Гаусса, т.е. переход от $P_{ki_k}A^{(k-1)} = A'^{(k-1)}$ к матрице $L'_k D'_k A'^{(k-1)} = A^{(k)}$, где

$$L'_k D'_k \;=\; \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/a'^{k-1}_{kk} & 0 & \dots & 0 \\ 0 & 0 & \dots & -a'^{(k-1)}_{k+1k}/a'^{(k-1)}_{kk} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -a'^{(k-1)}_{nk}/a'^{(k-1)}_{kk} & 0 & \dots & 1 \end{pmatrix},$$

$a'^{(k-1)}_{kk} = a^{(k-1)}_{i_k k}$, $a'^{(k-1)}_{i_k k} = a^{(k-1)}_{kk}$ и $a'^{(k-1)}_{sk} = a^{(k-1)}_{sk}$, $k \le s \ne k, i_k \le n$ (мы указываем здесь компонентны k-го столбца, указанные в записи $L'_k D'_k$). Таким образом, мы приходим к следущему результату

$$U \;=\; A^{(n)} \;=\; D'_n L'_{n-1} D'_{n-1} P_{n-1i_{n-1}} \cdots L'_k D'_k P_{ki_k} \cdots L'_1 D'_1 P_{1i_1} A\,.$$

Остаётся заметить, что

$$P_{n-1i_{n-1}} \cdots P_{k+1i_{k+1}} L'_k D'_k \;= \\ (P_{n-1i_{n-1}} \cdots P_{k+1i_{k+1}})(L'_k D'_k)(P_{k+1i_{k+1}} \cdots P_{n-1i_{n-1}})(P_{n-1i_{n-1}} \cdots P_{k+1i_{k+1}}) \;= \\ L''_k (P_{n-1i_{n-1}} \cdots P_{k+1i_{k+1}})\,,$$

где матрица L''_k представляет собой результат перестановки компонент k-го столбца (на и ниже главной диагонали) матрицы $L'_k D'_k$, отвечающей последовательному применению транспозиций $(k+1\, i_{k+1}), \dots, (n-1\, i_{n-1})$, при всех $k = 1, \dots, n-2$. Следовательно,

$$U \;=\; (D'_n L'_{n-1} D'_{n-1} L''_{n-2} \cdots L''_1)(P_{n-1i_{n-1}} \cdots P_{1i_1})A\,, \quad PA \;=\; LU\,,$$

где $L = (D'_n L'_{n-1} D'_{n-1} L''_{n-2} \cdots L''_1)^{-1}$ и $P = P_{n-1i_{n-1}} \cdots P_{1i_1}$.

Метод Гаусса с выбором по строке выполняет на k-ом шаге перестановку столбцов с номерами k и j_k, $k \le j_k \le n$, что соответствует умножению матрицы $A^{(k-1)}$ слева на матрицу P_{kj_k} и переходу к матрице $A'^{(k-1)} = A^{(k-1)} P_{kj_k}$. После этого выполняется привычный шаг метода Гаусса, т.е. переход к матрице $A^{(k)} = L'_k D'_k A'^{(k-1)}$, где в отличие от предыдущего алгоритма следует положить $a'^{(k-1)}_{sk} = a^{(k-1)}_{sj_k}$, $s = k, \dots, n$. Поэтому в данном случае

$$U \;=\; A^{(n)} \;=\; D'_n L'_{n-1} D'_{n-1} \cdots L'_1 D'_1 A P_{1j_1} \cdots P_{n-1j_{n-1}}\,, \quad AP \;=\; LU\,,$$

где $L = (D'_n L'_{n-1} D'_{n-1} \cdots L'_1 D'_1)^{-1}$, $P = P_{1j_1} \cdots P_{n-1j_{n-1}}$.

Метод Гаусса с полным выбором (выбором по всей матрице) представляет собой по сути комбинацию описанных двух методов. Его k-ый шаг представляет собой переход от $A^{(k-1)}$ к матрице $A'^{(k-1)} = P_{ki_k} A^{(k-1)} P_{kj_k}$ с последущим выполнением шага

метода Гаусса, состоящим в переходе к матрице $A^{(k)} = L'_k D'_k A'^{(k-1)}$, где матрица $L'_k D'_k$ строится как и ранее с той лишь разницей, что в данном случае $a'^{(k-1)}_{sk} = a^{(k-1)}_{sj_k}$, $k \le s \ne k, i_k \le n$, $a'^{(k-1)}_{kk} = a^{(k-1)}_{i_k j_k}$, $a'^{(k-1)}_{i_k k} = a^{(k-1)}_{kj_k}$. Поэтому

$$U = A^{(n)} = D'_n L'_{n-1} D'_{n-1} P_{n-1 i_{n-1}} \cdots L'_k D'_k P_{k i_k} \cdots L'_1 D'_1 P_{1 i_1} A P_{1 j_1} \cdots P_{n-1 j_{n-1}} = (D'_n L'_{n-1} D'_{n-1} L''_{n-2} \cdots L''_1)(P_{n-1 i_{n-1}} \cdots P_{1 i_1}) A (P_{1 j_1} \cdots P_{n-1 j_{n-1}}), \quad LU = PAP',$$

где $L = (D'_n L'_{n-1} D'_{n-1} L''_{n-2} \cdots L''_1)^{-1}$, $P = P_{n-1 i_{n-1}} \cdots P_{1 i_1}$ и $P' = P_{1 j_1} \cdots P_{n-1 j_{n-1}}$.

Сказанное означает, что рассматриваемые здесь модификации метода Гаусса представляют собой алгоритмы построения описанных ранее разложений невырожденных матриц в виде произведений двух треугольных факторов и матриц-перестановок.

Отметим также, что решение линейной системы $Ax = b$ с использованием метода Гаусса с выбором по столбцу сводится к решению двух треугольных систем

$$\begin{cases} Ly = Pb \\ Ux = y, \end{cases}$$

где $PA = LU$ (в стандартной реализации (без поиска самого разложения) мы решаем на заключительном этапе систему $Ux = L^{-1}Pb$), её решение с использованием метода Гаусса с выбором по строке соответствует решению системы

$$\begin{cases} Ly = b \\ Uz = y \\ x = Pz, \end{cases}$$

где $AP = LU$ (без нахождения разложения мы решаем систему $Uz = L^{-1}b$ и находим $x = Pz$), и, наконец, её решению с использованием метода Гаусса с полным выбором отвечает решение системы

$$\begin{cases} Ly = Pb \\ Uz = y \\ x = P'z, \end{cases}$$

где $PAP' = LU$ (здесь на заключительном этапе мы решаем систему $Uz = L^{-1}Pb$, $x = P'z$). Заметим, что информацию о матрицах-перестановках P и P' можно хранить в виде вектора значений соответствующих им перестановок из группы $\mathfrak{S}_n$. К примеру, для матрицы P' (P для метода с выбором по строке) такой перестановкой является $(1 j_1) \cdots (n-1 j_{n-1})$, а для матрицы P для метода с выбором по столбцу — это перестановка $(n-1 i_{n-1}) \cdots (1 i_1)$ (в нашем случае $(\sigma\sigma')(i) = \sigma(\sigma'(i))$, $\sigma, \sigma' \in \mathfrak{S}_n$, $i = 1, \ldots, n$).

По построению метод Гаусса с выбором элемента по столбцу гарантирует нам, что в матрице L, участвующей в связанной с реализацией данного метода факторизации $PA = LU$, диагональные компонентны имеют наибольшие модули среди всех элементов своих столбцов, метод Гаусса с выбором главного элемента по строке обеспечивает ограниченность единицей модулей всех компонент матрицы U из факторизации $AP = LU$ (в каждой её строке модули компонент не превышают модуль диагонального элемента, который равен единице), а метод Гаусса с выбором по всей матрице доставляет матрицы L и U в факторизации $PAP' = LU$, удовлетворяющие указанным выше

условиям первых двух методов с частичным выбором (модули компонент столбцов L не превосходят модули соответствующих диагональных элементов, модули элементов U не превышают единицу). Многочисленные численные эксперименты показали высокую численную устойчивость данным алгоритмов. Более того, в большинстве случаев найденное в результате реализации данных алгоритмов решение $\hat{x}$ удовлетворяет соотношению $(A + \delta A)\hat{x} = b$ с возмущением δA

$$\|\delta A\|_\infty \approx \varepsilon \|A\|_\infty \,,$$

где $\varepsilon = \beta^{-t}$ при использовании t-разрядной арифметики с плавающей запятой и основанием β. Данная оценка верна во всяком случае для методов с выбором по столбцу и всей матрице. Вместе с тем данная оценка обеспечивает лишь малость относительной невязки, поскольку

$$\|r\|_\infty \;=\; \|b - (A + \delta A)\hat{x} + \delta A\hat{x}\|_\infty \le \|\delta A\|_\infty \|\hat{x}\|_\infty \approx \varepsilon \|A\|_\infty \|\hat{x}\|_\infty \,.$$

При этом относительная погрешность решения оценивается как

$$\frac{\|\hat{x} - x\|_\infty}{\|x\|_\infty} \;=\; \frac{\|A^{-1}(A\hat{x} - b)\|_\infty}{\|x\|_\infty} \;=\; \frac{\|A^{-1}\delta A\hat{x}\|_\infty}{\|x\|_\infty} \le$$
$$\frac{\|A^{-1}\|_\infty \|\delta A\|_\infty \|\hat{x}\|_\infty}{\|x\|_\infty} \approx \varepsilon k_\infty(A) \frac{\|\hat{x}\|_\infty}{\|x\|_\infty} \approx \varepsilon k_\infty(A) \,,$$

где мы воспользовались также тем, что

$$\|\hat{x}\|_\infty \;=\; \|(A + \delta A)^{-1} b\|_\infty \;=\; \Big\| \Big(\sum_{k=0}^{\infty} (-1)^k (A^{-1}\delta A)^k \Big) (A^{-1} b) \Big\|_\infty \le \frac{\|x\|_\infty}{1 - \varepsilon k_\infty(A)} \,.$$

Существует несколько стандартных приёмов улучшения качества полученного решения системы $Ax = b$. Одним из них является уравновешивание рассматриваемой системы, идея которого состоит в том, чтобы перейти от системы $Ax = b$ к системе $D_1 A D_2 y = D_1 b$, связанной с исходной посредством диагональных матриц с ненулевыми диагональными элементами D_1 и D_2, выбор которых должен быть осуществлён из соображений уменьшения числа обусловленности $k_\infty(D_1 A D_2)$. В таком случае мы можем расчитывать на уменьшение относительной погрешности

$$\frac{\|\hat{y} - D_2^{-1} x\|_\infty}{\|D_2^{-1} x\|_\infty} \approx \varepsilon k_\infty(D_1 A D_2) \,,$$

а вместе с тем и на качественное приближение $D_2\hat{y}$ к x в D_2^{-1}-норме $\| \;\|_{D_2^{-1}}$, где $\|z\|_{D_2^{-1}} = \|D_2^{-1} z\|_\infty$ ($\hat{y}$ — решение, найденное в результате реализации метода Гаусса с выбором для масштабированной (уравновешенной) системы). При этом матрицы D_1 и D_2 не известны заранее, т.е. на каждом шаге алгоритма осуществляется необходимое домножение на диагональные матрицы, а указанные здесь D_1 и D_2 возникают в результате накопления множителей (с учётом их сопряжений с перестановками).

Предлагались различные варианты использования такого подхода, за каждым из которых стоят исключительно эвристические соображения, а не какая-либо серьёзная

аналитика. Одной из таких стратегий является строчное масштабирование, предполагающее выбор только одной матрицы D_1, призванный обеспечить близкие $\|\ \|_\infty$-нормы строк матрицы для которой выполняется гауссово исключение (матрица D_2 полагается равной единичной матрице E) с целью снизить вероятность сложения малого числа с большим в процессе его реализации. Естественно речь идёт о масштабировании всей ведущей подматрицы на каждом шаге алгоритма (т.е. замене D_k на диагональную матрицу с единицами в качестве первых $k-1$ элементов диагонали). Впрочем, следует сразу сказать, что и упомянутая здесь стратегия и любые другие известные на сегоднящней день стратегии уравновешивания могут привести не к улучшению, а к ухудшению качества решения.

Другой известный приём улучшения качества решения состоит в использовании итерационного уточнения, которое выполняется следующим образом:

1. исходная система $Ax = b$ решается с помощью какого-либо из описанных ранее алгоритмов и в результате вычисляется решение $\hat{x}$;

2. применяя тот же алгоритм к системе $Az = r$ с вектором невязки $r = b - A\hat{x}$, найдём вектор $\hat{z}$ и положим в качестве уточнённого решения $\hat{x}_{new} = \hat{x} + \hat{z}$.

Естественно, что в точной арифметике мы не получили бы в результате ничего нового, поскольку в этом случае $r = z = 0$. В арифметике с плавающей запятой такого не происходит, но, тем не менее, найденный вектор $\hat{x}_{new}$ может оказаться не сильно лучше вектора $\hat{x}$, если тот уже был близок к решению, как это имеет место для методов Гаусса с выбором. Вместе с тем применительно к другим стратегиям выбора несколько шагов такого уточнения могут значительно улучшить качество решения (т.е. количество верных разрядов) $\hat{x}$.

Значительно более эффективной версией процесса итерационного уточнения является версия этого алгоритма, в которой вектор невязки вычисляется с удвоенной точностью. Известны качественные оценки работы такого алгоритма итерационного уточнения с переменной точностью, оправдывающего его использование. Стоит также отметить, что при его реализации мы фактически решаем две треугольные системы с факторами, найденными в процессе реализации основного алгоритма, и выполняем необходимые перестановки, что добавляет лишь порядка $O(n^2)$ операций.

Помимо обсуждавшихся нами проблем вычислительной устойчивости метода Гаусса существует весьма важная проблема его реализации, связанная с заполнением ведущей подматрицы на шаге исключения, что особенно актуально при решении систем с разреженными матрицами, вся значимая информация о которых хранится в относительно небольшом векторе, содержащем значения ненулевых компонент и необходимую информацию об их расположении. В частности, простым примером актуальности такой проблематики может служить ситуация со стреловидной матрицей

$$\begin{pmatrix} * & * & * & \dots & * \\ * & * & 0 & \dots & 0 \\ * & 0 & * & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ * & 0 & 0 & \dots & * \end{pmatrix}$$

для которой уже первый шаг метода Гаусса может привести к полному заполнению ведущей подматрицы.

Нашей следующей целью станет получение исчерпывающей информации о структуре заполнения ведущих подматриц на этапах выполнения метода Гаусса с выбором ведущего элемента. Под *локальным заполнением* на k-ом шаге метода Гаусса мы будем понимать количество нулевых элементов матрицы $A^{(k-1)}$, которые стали ненулевыми после выполнения данного шага. Естественно интерес представляет не сама матрица $A^{(k-1)}$, а её нижний блок $(n-k)\times(n-k)$, преобразуемый в новую ведущую подматрицу. Обозначим через B_k матрицу, полученную из матрицы $(a_{ij}^{(k-1)})_{i,j=k}^{n}$ заменой ненулевых элементов единицами и через $\overline{B_k}$ — матрицу, полученную заменой единичных элементов матрицы B_k нулями, а нулей — единицами, $\overline{B_k} = M - B_k$, где M — матрица размера $(n-k+1)\times(n-k+1)$, состоящая из одних единиц. Следующий результат носит название теоремы Тьюарсона.

Теорема 0.19. *Если элемент $a_{i+k-1j+k-1}^{(k-1)} \neq 0$, $i,j = 1, n-k+1$, выбирается в качестве ведущего на k-ом шаге метода Гаусса, тогда локальное заполнение на этом шаге совпадает с $(i+k-1, j+k-1)$-ым элементом матрицы $G_k = B_k\overline{B_k}^t B_k = (g_{st}^{(k)})_{s,t=k}^{n}$.*

Доказательство. В соответствии с имеющимся описанием алгоритма нам следует понять сколько элементов $a_{p+k-1q+k-1}^{(k)}$, $p,q = 1,\ldots,n-k+1$, $p \neq i$, $q \neq j$, стали ненулевыми из числа тех, для которых элемент $a_{p+k-1q+k-1}^{(k-1)}$ был равен нулю. Поскольку

$$a_{p+k-1q+k-1}^{(k)} = a_{p+k-1q+k-1}^{(k-1)} - a_{i+k-1q+k-1}^{(k-1)} a_{p+k-1j+k-1}^{(k-1)} / a_{i+k-1j+k-1}^{(k-1)} = \\ - a_{i+k-1q+k-1}^{(k-1)} a_{p+k-1j+k-1}^{(k-1)} / a_{i+k-1j+k-1}^{(k-1)},$$

подобный переход возможен лишь в случае, когда $a_{i+k-1q+k-1}^{(k-1)} \neq 0$ и $a_{p+k-1j+k-1}^{(k-1)} \neq 0$, т.е. если $b_{i+k-1q+k-1}^{(k)} \neq 0$, $b_{p+k-1j+k-1}^{(k)} \neq 0$ при $b_{p+k-1q+k-1}^{(k)} = 0$, где $B_k = (b_{st}^{(k)})$. Таким образом, искомое заполнение

$$g_{i+k-1j+k-1}^{(k)} = \sum_{1\leq p\leq n-k+1,\ p\neq i} \sum_{1\leq q\leq n-k+1,\ q\neq j} b_{i+k-1q+k-1}^{(k)}(1-b_{p+k-1q+k-1}^{(k)}) b_{p+k-1j+k-1}^{(k)} = \\ \sum_{1\leq p,q\leq n-k+1} b_{i+k-1q+k-1}^{(k)}(1-b_{p+k-1q+k-1}^{(k)}) b_{p+k-1j+k-1}^{(k)} = (B_k\overline{B_k}^t B_k)_{i+k-1j+k-1},$$

поскольку $(1-b_{p+k-1j+k-1}^{(k)})b_{p+k-1j+k-1}^{(k)} = (1-b_{i+k-1q+k-1}^{(k)})b_{i+k-1q+k-1}^{(k)} = 0$. □

Это позволяет модифицировать метод Гаусса следующим образом. Зафиксируем пороговое $\epsilon > 0$. Выберем в качестве ведущего элемента на k-ом шаге элемент $a_{i+k-1j+k-1}^{(k-1)}$, $i,j = 1,\ldots,n-k+1$, для которого элемент $g_{ij}^{(k)}$ является минимальным среди всех элементов $|a_{i+k-1j+k-1}^{(k-1)}| > \epsilon$ (если таких элементов несколько, то мы выберем среди них элемент с наибольшим модулем). Затем выполним шаг метода Гаусса. Полученный вариант метода Гаусса с выбором ведущего элемента обеспечивает оптимальное заполнение среди всех возможных вариантов выбора ведущего элемента по модулю большего ϵ, хотя и имеет весьма значительную вычислительную сложность. Он также соответствует факторизации $PAP' = LU$, как и описанный выше метод Гаусса с полным выбором.

Говоря о заполнении имеет смысл упомянуть ситуации, когда оно имеет предсказуемую природу.

1. В случае с невырожденной матрицей

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n-p-1} & a_{1n-p} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n-p-1} & a_{2n-p} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{p+11} & a_{p+12} & \dots & a_{p+1n-p-1} & a_{p+1n-p} & \dots & a_{p+1n} \\ 0 & a_{p+22} & \dots & a_{p+2n-p-1} & a_{p+2n-p} & \dots & a_{p+2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & a_{nn-p} & \dots & a_{nn} \end{pmatrix},$$

имеющей $1 \leq p < n$ диагоналей ниже главной (при $p = 1$ такая матрица называется *верхней хессенберговой* или *матрицей в форме Хессенберга*), на каждом шаге метода Гаусса обрабатывается верхний прямоугольный блок ведущей подматрицы размера $(p+1) \times n$ и потому LU–разложение такой матрицы A (при наличии последнего) имеет в качестве L ленточную нижнюю треугольную матрицу с p диагоналями ниже главной. При этом в случае $p << n$ наш алгоритм потребует $O(n^2)$ операций при $n \to \infty$. Изменение методов Гаусса с выбором по строке или всей матрице приведёт к изменению структуры заполнения матрицы (произойдёт заполнение нижнего треугольника) и потому в данном случае целесообразно использовать стратегию выбора по столбцу, а точнее по первому столбцу верхнего блока размера $(p+1) \times n$ ведущей подматрицы, что обеспечит сохранение структуры заполнения матрицы.

Естественно, что для структуры с p диагоналями выше главной уместно будет использовать выбор по строке (первой строке вертикального блока $n \times (p+1)$ ведущей подматрицы).

2. Для невырожденной ленточной матрицы A с p и q диагоналями ниже и выше главной, $1 \leq p, q < n$ (с шириной ленты $p+q+1$), на каждом шаге Гауссова исключения обрабатывается верхний угловой блок размера $(p+1) \times (q+1)$ ведущей подматрицы. Поэтому LU–разложение такой матрицы имеет в качестве L и U ленточные нижнюю и верхнюю треугольные матрицы с p и q диагоналями ниже и выше главной соответственно. При $p, q << n$ реализация метода Гаусса (построения LU–разложенния) такой матрицы обойдётся в $O(n)$ при $n \to \infty$ операций (сходная оценка будет и для решения ленточных треугольных систем). Хранение в данном случае естественно организовать по диагоналям. Применение в данном случае метода Гаусса с выбором по первому столбцу (первой строке) указанного блока ведущей подматрицы приводит к заполнению верхнего треугольника дополнительными p диагоналями (нижнего треугольника дополнительными q диагоналями). Выбор по всей матрице нецелесообразен, поскольку может превратить нашу матрицу в заполненную.

Естественно указанные здесь изменения структур заполнения равносильны аналогичным изменениям структуры разложений $PA = LU$ и $AP = LU$ (добавлениям в U или L дополнительных ненулевых диагоналей).

Оценщик строчной матричной нормы, построенный на основе LU–разложения

Напомним, что качественная оценка числа обусловленности является необходимым условием получения оценок погрешности решения линейных систем (см. ранее). Применительно к строчной матричной норме $\| \|_\infty$ наиболее сложной частью оценки величины

$k_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$ является нахождение оценки для $\|A^{-1}\|_\infty$ без явного вычисления обратной матрицы A^{-1}. Задача формулируется так: при известном разложении $PA = LU$ (естественно полученным в результате примения описанных выше ранее алгоритмов) построить алгоритм нахождения оценки $\|A^{-1}\|_\infty$, требующий $O(n^2)$ операций. Идея подобной оценки состоит в интерпретации $\|A\|$ для любой операторной матричной нормы $\|\ \|$, согласованной с векторной нормой $\|\ \|$, как решения задачи $\frac{\|Ax\|}{\|x\|} \to \max$, что применительно к $\|A^{-1}\|$ может быть сведено к задаче $\frac{\|A^{-1}y\|}{\|y\|} = \frac{\|x\|}{\|y\|} \to \max$, где $Ax = y$ (при этом отношение в левой части всегда не превосходит $\|A^{-1}\|$). Наша цель подобрать y таким образом, чтобы x был наибольшим по выбранной норме $\|\ \|$ относительно y. В нашем случае речь пойдёт о норме $\|\ \|_\infty$, а вектор y можно выбирать с единичной нормой. Мы будем иметь дело с вещественными системами.

Для начала заметим, что алгоритм обратной подстановки для верхней треугольной системы $Tx = y$ может быть переписан в виде: на исходном этапе вспомогательный вектор столбец $p(n) = (p_1, \ldots, p_n)^t$ полагается равным нулю, затем для всех $k = n, \ldots, 1$ вычисляется $x_k = (y_k - p_k)/t_{kk}$ и вектор $p(k-1) = p(k-1) + x_k t_k(k-1)$, $p(k-1) = (p_1, \ldots, p_{k-1})^t$, $t_k(k-1) = (t_{1k}, \ldots, t_{k-1k})^t$. Алгоритм оценки строится следующим образом. На шаге $k = n, \ldots, 1$ выбирается $y_k \in \{\pm 1\}$ из следующих соображений: вычисляются

$$\begin{aligned} x_k^+ &= (1 - p_k)/t_{kk}\,, \quad s_k^+ = |x_k^+| + \|p(k-1) + x_k^+ t_k(k-1)\|_1\,, \\ x_k^- &= (-1 - p_k)/t_{kk}\,, \quad s_k^- = |x_k^-| + \|p(k-1) + x_k^- t_k(k-1)\|_1\,, \end{aligned}$$

затем при $s_k^+ \geq s_k^-$ мы полагаем $x_k = x_k^+$ и в противном случае — $x_k = x_k^-$. Норма $\|x\|_\infty$ найденного в итоге вектора x объявляется приближённым значением $\|T^{-1}\|_\infty$.

В соответствии с этим алгоритмом предлагается использовать следующий алгоритм оценки $k_\infty(A)$ при известном разложении $PA = LR$, найденном в методе Гаусса с частичным выбором по столбцу (это обеспечивает ограниченность модулей коэффициентов нижней унитреугольной матрицы L единицей, что в свою очередь может служить основанием для предположения о хорошей обусловленности L). Итак,

1. применим описанный выше алгоритм к системе $R^t x = y$ и найдём соответствующие векторора $\hat{x}$ и $\hat{y}$;
2. решим системы $L^t u = \hat{x}$, $Lw = Pu$ и $Rz = w$ ($A^t P^t u = R^t L^t u = \hat{y}$, $LRz = Pu$, $Az = u$);
3. положим $\hat{k_\infty}(A) = \|A\|_\infty \|z\|_\infty / \|u\|_\infty$.

Идея этого алгоритма состоит в том, что для матрицы A с сингулярным разложением $A = V\Sigma U^t$ решение $u = P(A^t)^{-1}\hat{y}$ сильно смещено в сторону правых сингулярных (векторов-столбцов U), отвечающих σ_n (минимальное сингулярное значение A), поскольку характер нашего приближения позволяет считать, что основной вклад в рост числа обусловленности вносит множитель R. Поэтому мы вправе расчитывать на то, что решение $z = A^{-1}u$ будет иметь евклидову норму порядка $1/\sigma_n$.

Лекция 3. Метод Гаусса — Жордана и другие модификации метода Гаусса для систем специального вида

Мы начнём со следующей модификации метода Гаусса, принадлежащей Жордану. Как и прежде, речь идёт о вещественной или комплексной линейной системе $Ax = b$ с невырожденной матрицей A. Предлагаемый процесс состоит в последовательном переходе от исходной системы $A^{(0)}x = b^{(0)}$, $A^{(0)} = A$, $b^{(0)} = b$, к эквивалентным ей системам $A^{(k)}x = b^{(k)}$, $k = 1, \ldots, n$, где

$$A^{(k)} = \begin{pmatrix} 1 & 0 & \ldots & 0 & a^{(k)}_{1k+1} & \ldots & a^{(k)}_{1n} \\ 0 & 1 & \ldots & 0 & a^{(k)}_{2k+1} & \ldots & a^{(k)}_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & a^{(k)}_{kk+1} & \ldots & a^{(k)}_{kn} \\ 0 & 0 & \ldots & 0 & a^{(k)}_{k+1k+1} & \ldots & a^{(k)}_{k+1n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a^{(k)}_{nk+1} & \ldots & a^{(k)}_{nn} \end{pmatrix}$$

и $b^{(k)} = (b^{(k)}_1, \ldots, b^{(k)}_n)^t$, причём последняя система имеет единичную матрицу $A^{(n)} = E$ и в качестве правой части вектор-решение $b^{(n)} = x$. Как и в методе Гаусса, переход $A^{(k-1)} \to A^{(k)}$, $b^{(k-1)} \to b^{(k)}$ на каждом шаге k осуществляется посредством обратимых элементарных преобразований строк типа умножение строки на ненулевое число ("масштабирование") и вычитание из i-ой строки j-ой, $i \neq j$, умноженной на некоторое число, но в отличие от метода Гаусса масшабированная k-ая ведущая строка вычитается с домножением на k-ый коэффициент i-ой строки для всех $i \neq k$ и аналогичное преобразование осуществляется с вектором правой части (напомним, что в методе Гаусса подобным образом обрабатываются строки и компоненты вектора правой части с номерами $i > k$). Поэтому в данном случае основной объём вычислительной работы сосредоточен в блоке $n \times (n-k)$ ($n \times (n-k+1)$ в терминах расширенной матрицы системы, если быть более точным), а не только в рамках ведущей подматрицы и масштабируемой строки.

Шаг нашего процесса с номером k, реализующий переход $(A^{(k-1)}, b^{(k-1)}) \to (A^{(k)}, b^{(k)})$, $k = 1, \ldots, n$, при условии его осуществимости состоит в следующем:

1. k-ая строка матрицы $A^{(k-1)}$ и k–ая компонента вектора $b^{(k-1)}$ делятся на ведущий элемент $a^{(k-1)}_{kk}$, в результате чего вычисляются их новые компоненты

$$a^{(k)}_{ki} = a^{(k-1)}_{ki} / a^{(k-1)}_{kk} \quad (i = k+1, \ldots, n), \quad b^{(k)}_k = b^{(k-1)}_k / a^{(k-1)}_{kk}.$$

2. из каждой i-ой строки матрицы $A^{(k-1)}$, $i \neq k$, вычитается перевычисленная на предыдущем этапе k-ая строка, домноженная на $a^{(k-1)}_{ik}$, и соответственно из каждой i-ой компоненты вектора $b^{(k-1)}$, $i \neq k$, вычитается новая k-ая компонента $b^{(k)}_k$, умноженная на $a^{(k-1)}_{ik}$, при этом появляются компоненты новых матрицы $A^{(k)}$ и вектора $b^{(k)}$,

$$a^{(k)}_{ij} = a^{(k-1)}_{ij} - a^{(k-1)}_{ik} a^{(k)}_{kj}, \quad b^{(k)}_i = b^{(k-1)}_i - a^{(k-1)}_{ik} b^{(k)}_k,$$

где $i = 1, \ldots, n$, $i \neq k$, $j = k+1, \ldots, n$.

При этом первый шаг данного алгоритма совпадает в точности с первым шагом методом Гаусса, а его осуществимость сводится к выполнению условия $a_{kk}^{(k-1)} \neq 0$ для всех $k = 1, \ldots, n$. Поскольку нижняя квадратная подматрица размера $(n-k) \times (n-k)$ матрицы $A^{(k)}$, $k = 0, \ldots, n-1$, совпадает с ведущей подматрицей метода Гаусса, а элемент $a_{kk}^{(k-1)}$, участвующий в масштабировании на k-ом шаге есть тот же ведущий элемент, что и в методе Гаусса, методы Гаусса и Гаусса — Жордана осуществимы при одном и том же условии наличия LU–разложения у матрицы системы A или что равносильно при условии невырожденности всех её главных квадратных подматриц.

Метод Гаусса — Жордана несколько более сложен в вычислительном отношении, чем метод Гаусса. Точнее выполнение k-го шага этого метода требует $n-k+1$ делений (на масштабирование) и после $2(n-k+1)(n-1)$ умножений и вычитаний (на перевычисление новых компонент матрицы и новых компонент вектора правой части с индексами $i \neq k$). Таким образом, общее число операций $n^3 + O(n^2)$ при $n \to \infty$.

Как и метод Гаусса, метод Гаусса — Жордана может быть применён для решения нескольких линейных систем $Ax^i = b^i$, $i = 1, \ldots, s$, с общей матрицей системы A. В этом случае речь идёт о решении матричного уравнения $AX = B$, где $B = (b^1, \ldots, b^s)$ и $X = (x^1, \ldots, x^s)$ матрицы $n \times s$ векторов правой части векторов решений, и весь процесс осуществляется последовательными перевычислениями расширенной матрицы системы, полученной в результате добавления s столбцов векторов правых частей ($(A^{(0)}, B^{(0)}) \to (A^{(1)}, B^{(1)}) \to \ldots \to (A^{(n)}, B^{(n)}) = (E, X)$). Естественно, что наш процесс усложняется перевычислениями компонент матрицы $B^{(k)}$, $k = 0, \ldots, n$, вносящим вклад порядка $2sn^2 + O(n, s)$. В частности, при решении подобной системы с матрицей $B = E$ мы получим в ответе $B^{(n)} = A^{-1}$.

Привлекательной стороной метода Гаусса — Жордана является возможность осуществления его последовательно-параллельных реализации со стандартным разделением матрицы системы на строчные блоки, не требующей дополнительных балансировок работы на отдельных узлах (в отличие от метода Гаусса в рассматриваемом алгоритме объём вычислительной работы на узлах распределён относительно равномерно: после перевычисления на одном из них ведущей строки и пересылки её компонент остальным они загружены практически одинаковым объёмом работы по перевычислению новых компонент).

В терминах матричных умножении метод Гаусса — Жордана может быть интерпретирован как последовательность переходов от матрицы $A^{(k-1)}$ и вектора $b^{(k-1)}$ к матрице $A^{(k)}$ и вектору $b^{(k)}$, $k = 1, \ldots, n$, полученная в результате умножения $A^{(k-1)}$ и $b^{(k-1)}$ слева на матрицу $T_k D_k$, где $D_k = \mathrm{diag}(1, \ldots, 1, 1/a_{kk}^{(k-1)}, 1, \ldots, 1)$ — диагональная матрица с элементом $1/a_{kk}^{(k-1)}$ на позиции (k, k) и T_k — матрица вида

$$T_k = \begin{pmatrix} 1 & 0 & \ldots & 0 & -a_{1k}^{(k-1)} & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & -a_{2k}^{(k-1)} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & -a_{k+1k}^{(k-1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & -a_{nk}^{(k-1)} & 0 & \ldots & 1 \end{pmatrix},$$

где все внедиагональные элементы кроме отмеченных компонент k-го столбца равны

нулю. Таким образом, $A^{(k)} = T_k D_k A^{(k-1)}$ и $b^{(k)} = T_k D_k b^{(k-1)}$, $k = 1, \ldots, n$, и

$$A^{(n)} \;=\; T_n D_n \cdots T_1 D_1 A \;=\; E\,, \quad (T_n D_n \cdots T_1 D_1)^{-1} \;=\; D_1^{-1} T_1^{-1} \cdots D_n^{-1} T_n^{-1} \;=\; A^{-1}\,.$$

Другими словами, наше построение соответствует представлению обратной матрицы A^{-1} к матрице A в виде произведения матриц $D_k^{-1} T_k^{-1}$,

$$D_k^{-1} T_k^{-1} \;=\; \begin{pmatrix} 1 & 0 & \ldots & 0 & a_{1k}^{(k-1)} & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & a_{2k}^{(k-1)} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a_{kk}^{(k-1)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & a_{k+1k}^{(k-1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a_{nk}^{(k-1)} & 0 & \ldots & 1 \end{pmatrix} \qquad (k = 1, \ldots, n)\,.$$

Очевидное сходство методов Гаусса и Гаусса — Жордана имеет своим следствием и общие проблемы их реализации связанные с численной устойчивостью, а точнее с её отсутствием без использования дополнительных преобразований. В частности, для преодоления роста модулей элементов матриц $A^{(k)}$ в процессе выполнения метода Гаусса — Жордана, связанные с делением на малый по модулю ведущий элемент на этапе масштабирования, используют более устойчивые в вычислительном отношении модификации этого метода, базирующиеся на частичном или полном выборе ведущего элемента в нижней квадратной подматрице аналогично методам Гаусса с выбором по столбцу, строке или всей матрице. Реализуются данные модернизации по следующей схеме: на k-ом шаге осуществляется выбор ненулевого ведущего элемента $a_{i_k j_k}^{(k-1)}$ среди всех элементов подматрицы $(a_{ij}^{(k-1)})_{i,j=k}^{n}$ согласно выбранной стротегии, затем i_k-ая строка меняется местами с k-ой и k-ый столбец меняется местами с j_k-ым (осуществляется перенумерация строк и переменных) и осуществляется шаг метода Гаусса — Жордана.

Нашей следующей целью станет обсуждение вариантов гауссова исключения и LU–разложения для симметрических систем. Наше обсуждение мы начнём с ряда предварительных сведений о положительно определённых матрицах.

Для начала напомним, что матрица $A \in M_n(\Bbbk)$, $\Bbbk = \mathbb{R}$, $\mathbb{C}$ называется *положительно (неотрицательно) определённой* (обозначение $A > 0$ ($A \geq 0$)), если (Ax, x) — положительное (неотрицательное) вещественное число для любого $0 \neq x \in \Bbbk^n$, где $(\ ,\ )$ — стандартное скалярное произведение для $\Bbbk = \mathbb{R}$ или стандартная эрмитова форма для $\Bbbk = \mathbb{C}$, $(x, y) = y^* x = x_1 \overline{y_1} + \ldots + x_n \overline{y_n}$, $x = (x_1, \ldots, x_n)^t, y = (y_1, \ldots, y_n)^t \in \Bbbk^n$, где в вещественном случае знак комплексного сопряжения можно опустить.

Матрица $A \in M_n(\mathbb{C})$ называется *самосопряжённой*, если $A^* = A$. В вещественном случае это условие соответствует понятию *симметрической* матрицы $A = A^t$. Для самосопряжённой матрицы A очевидным образом выполняются равенства $(Ax, y) = y^* Ax = (Ay)^* x = (x, Ay)$ и, как следствие, $(Ax, x) = (x, Ax) = \overline{(Ax, x)} \in \mathbb{R}$, $x, y \in \mathbb{C}^n$. Напомним, что на основе этих соотношений в курсе линейной алгебре устанавливалась вещественность спектра $\operatorname{Spec} A$ самосопряжённой матрицы $A = A^*$ (аналогичный вывод был получен и в курсе функционального анализа для ограниченных самосопряжённых операторов на гильбертовом пространстве).

Заметим, что для симметрической вещественной матрицы $A = A^t$ вещественное и комплексное условия положительной (неотрицательной) определённости равносильны, поскольку $(Ax, x) = Re(Ax, x) + iIm(Ax, x)$, где

$$Re(Ax, x) \;=\; (Ax_{re}, x_{re}) + (Ax_{im}, x_{im})\,, \quad Im(Ax, x) \;=\; (Ax_{im}, x_{re}) - (Ax_{re}, x_{im}) \;=\; 0$$

для любого $x = x_{re} + ix_{im}$, $x_{re}, x_{im} \in \mathbb{R}^n$, $i^2 = -1$.

Замечание 0.20. *Пусть матрица $A = (a_{ij}) \in M_n(\Bbbk)$, $\Bbbk = \mathbb{R}, \mathbb{C}$, положительно определена. Тогда матрица A невырождена, все её главные подматрицы $A_k = (a_{ij})_{i,j=1}^k$, $k = 1, \ldots, n$, положительно определены и невырождены и, как следствие, A обладает LU–разложением.*

Доказательство. Для начала заметим, что вместе с равенством $Ax = 0$ будет выполняться и равенство $(Ax, x) = 0$, которое возможно лишь для $x = 0$ ($A > 0$). Возьмём теперь произвольный вектор $x = (x_1, \ldots, x_k)^t \in \Bbbk^k$, $k = 1, \ldots, n$, и определим на его основе вектор $\hat{x} = (x_1, \ldots, x_k, 0, \ldots, 0) \in \Bbbk^n$. Тогда $(A\hat{x}, \hat{x}) = (A_k x, x) > 0$ при $x \neq 0$ и значит, $A_k > 0$, $\det A_k \neq 0$. Следовательно, для матрицы A осуществимо LU–разложение. □

Теорема 0.21. *Для любой самомосопряжённой матрицы $A = A^* \in M_n(\mathbb{C})$ равносильны следующие условия:*

1. *A — положительно определена;*
2. *все собственные значения A положительны,* $\operatorname{Spec} A \subset \mathbb{R}_+$;
3. $\det A_k \in \mathbb{R}_+$ *при всех* $k = 1, \ldots, n$.

Доказательство. Как уже отмечалось, $\operatorname{Spec} A \subset \mathbb{R}$, поскольку $A = A^*$. Из курса линейной алгебры известно существование для матрицы A ортонормированного базиса собственных векторов $\{v_1, \ldots, v_n\}$, $Av_i = \lambda_i v_i$, $i = 1, \ldots, n$, где $\operatorname{Spec} A = \{\lambda_1, \ldots, \lambda_n\}$. Поэтому, обозначив через V унитарную матрицу со столбцами $v_1, \ldots, v_n$ и через Λ — диагональную матрицу $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, мы можем записать $A = V\Lambda V^*$. Если $A > 0$, то $(Av_i, v_i) = \lambda_i > 0$, $i = 1, \ldots, n$. С другой стороны если $\lambda_i > 0$, $i = 1, \ldots, n$, тогда любой ненулевой вектор $x \in \mathbb{C}^n$ может быть записан как $x = y_1 v_1 + \ldots + y_n v_n = Vy$ для подходящего $0 \neq y = (y_1, \ldots, y_n)^t \in \mathbb{C}^n$ и, следовательно,

$$(Ax, x) \;=\; (V\Lambda V^* x, x) \;=\; (\Lambda V^* x, V^* x) \;=\; (\Lambda y, y) \;=\; \sum_{i=1}^n \lambda_i |y_i|^2 > 0\,.$$

Поэтому в этом случае $A > 0$. Таким образом, первые два условия равносильны.

Отметим, что $\det A = \det \Lambda = \lambda_1 \cdots \lambda_n$ и потому при $A > 0$ должно выполняться $\det A > 0$. Более того, в случае $A > 0$ мы получаем также, что $A_k = A_k^* > 0$ (см. предыдущее замечание) и $\det A_k > 0$, $k = 1, \ldots, n$.

Остаётся показать, что условие $\det A_k > 0$, $k = 1, \ldots, n$, обеспечивает $A > 0$. Действительно, в таком случае мы располагаем LU–разложением $A = LU = L'DU$, где $D = \operatorname{diag}(l_{11}, \ldots, l_{nn})$ — диагональ матрицы L, причём $\det A_k = \det L_k U_k = \det D_k = l_{11} \cdots l_{kk} > 0$, $k = 1, \ldots, n$, и, следовательно, $l_{11}, \cdots, l_{nn} > 0$. Более того, мы имеем $A = A^* = U^* DL'^*$, что в силу однозначности определения LU–разложения гарантирует нам $L' = U^*$, $A = U^* DU$. Поэтому для всякого $0 \neq x \in \mathbb{C}^n$, $z = Ux = (z_1, \ldots, z_n)^t \neq 0$ и

$$(Ax,x) \;=\; (U^*DUx,x) \;=\; (DUx,Ux) \;=\; (Dz,z) \;=\; \sum_{i=1}^{n} l_{ii}|z_i|^2 > 0\,,$$

т.е. $A > 0$. □

Развивая идею заключительной части этого доказательства, можно прийти к следующему выводу.

Замечание 0.22. *Любая самосопряжённая матрица $A = A^*$, обладающая LU–разложением, представима в виде $A = R^*DR$, где R — верхняя треугольная матрица, на диагонали которой стоят положительные вещественные числа, и D — диагональная матрица с диагональными элементами ± 1. При этом в случае $A > 0$ можно считать, что $D = E$.*

Доказательство. По аналогии с предыдущим мы можем выделить диагональную составляющую LU–разложения матрицы A, $A = LU = L'DU$, $D = \operatorname{diag}(l_{11},\dots,l_{nn}) = \operatorname{diag}(d_1,\dots,d_n)$, а затем, используя единственность LU–разложения и равенство $A = A^* = U^*D^*L'^*$, прийти к равенствам $D = D^*$, $L' = U^*$. Согласно первого из них $0 \neq d_i = \overline{d_i} \in \mathbb{R}$, $i = 1,\dots,n$. Следовательно, полагая $D' = \operatorname{diag}(\operatorname{sign} d_1,\dots,\operatorname{sign} d_n)$ и $D'' = \operatorname{diag}(\sqrt{|d_1|},\dots,\sqrt{|d_n|})$, мы приходим к искомому равенству $A = R^*D'R$ с $R = D''U$. Напомним, что $A > 0$, если и только если $d_1,\dots,d_n > 0$ (см. ранее). Поэтому в данном случае $D' = E$. □

Следствие 0.23. *Самосопряжённая матрица A тогда и только тогда положительно определена, когда она обладает представлением вида $A = B^*B$ для некоторой невырожденной матрицы B. Более того, можно считать, что матрица B является верхней треугольной и имеет положительную вещественную диагональ.*

Представление LU–разложимой самосопряжённой матрицы A в виде $A = R^*DR$ с диагональной матрицей $D = \operatorname{diag}(d_1,\dots,d_n)$, $d_i \in \{\pm 1\}$, и верхней треугольной матрицей $R = (r_{ij})$, $r_{ii} \in \mathbb{R}_+$, называют *разложением Холесского* матрицы A (в форме, использующей корни). Представление $A = R^*DR$ с вещественной диагональной матрицей D и верхней унитреугольной матрицей R также называют *разложением Холесского* матрицы A в форме свободной от корней. Заметим, что вне зависимости от формы разложение Холесского определено однозначно в силу единственности LU–разложения.

Приведём несколько стандартных способов построения разложения Холесского. Для начала остановимся на схеме построения разложения Холесского в форме свободной от корней в терминах матричных умножений. Пусть имеется матрица $A = A^*$, для которой осуществимо LU–разложение. Мы построим на её основе последовательность матриц $A^{(0)} = A, A^{(1)},\dots,A^{(n-1)}$, в которой каждая матрица $A^{(k)} = (A^{(k)})^*$ имеет вид

$$A^{(k)} \;=\; \begin{pmatrix} d_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_k & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & a^{(k)}_{k+1k+1} & \dots & a^{(k)}_{k+1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & a^{(k)}_{nk+1} & \dots & a^{(k)}_{nn} \end{pmatrix},$$

а последняя матрица $A^{(n-1)}$ — диагональная, $A^{(n-1)} = \mathrm{diag}(d_1, \ldots, d_n)$. При этом переход $A^{(k-1)} \to A^{(k)}$, реализуемый на k-ом шаге, представляет собой подобие $A^{(k)} = L_k A^{(k-1)} L_k^*$ с нижней унитреугольной матрицей L_k,

$$L_k \;=\; \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & -a_{k+1\,k}^{(k-1)}/a_{kk}^{(k-1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -a_{nk}^{(k-1)}/a_{kk}^{(k-1)} & 0 & \ldots & 1 \end{pmatrix},$$

которая отличается от единичной матрицы лишь компонентами k-го столбца, стоящими ниже главной диагонали. Осуществимость подобного шага равносильна выполнению неравенства $a_{kk}^{(k-1)} \neq 0$. Здесь следует отметить, что умножение $L_k A^{(k-1)}$ имеет своим результатом аннулирование поддиагональных элементов k-го столбца $A^{(k-1)}$ посредством вычитания из каждой строки с номером $i = k+1, \ldots, n$, строки с номером k с подходящим множителем. Самосопряжённая матрица $A^{(k)} = L_k A^{(k-1)} L_k^*$ имеет те же первые k столбцов, что и матрица $L_k A^{(k-1)}$, а потому имеет тот вид, который был указан выше. Описанное здесь преобразование ведущей нижней подматрицы

$$(a_{ij}^{(k-1)})_{i,j=k}^{n} \;=\; \begin{pmatrix} a_{kk}^{(k-1)} & u_{k-1}^* \\ u_{k-1} & A' \end{pmatrix}, \quad A' \;=\; (a_{ij}^{(k-1)})_{i,j=k+1}^{n}, \quad u_k \;=\; (a_{2k}^{(k-1)}, \ldots, a_{nk}^{(k-1)})^t,$$

представляет собой сопряжение

$$\begin{pmatrix} 1 & 0 \\ -u_{k-1}/a_{kk}^{(k-1)} & E_{n-k} \end{pmatrix} \begin{pmatrix} a_{kk}^{(k-1)} & u_{k-1}^* \\ u_{k-1} & A' \end{pmatrix} \begin{pmatrix} 1 & -u_{k-1}^*/a_{kk}^{(k-1)} \\ 0 & E_{n-k} \end{pmatrix} = \\ \begin{pmatrix} a_{kk}^{(k-1)} & 0 \\ 0 & A' - 1/a_{kk}^{(k-1)} u_{k-1} u_{k-1}^* \end{pmatrix}$$

где $a_{kk}^{(k-1)} = d_k$ и $A' - 1/a_{kk}^{(k-1)} u_{k-1} u_{k-1}^* = (a_{ij}^{(k)})_{i,j=k+1}^{n}$,

$$a_{ij}^{(k)} \;=\; a_{ij}^{(k-1)} - a_{ik}^{(k-1)} a_{kj}^{(k-1)}/a_{kk}^{(k-1)} \;=\; a_{ij}^{(k-1)} - a_{ik}^{(k-1)} \overline{a_{jk}^{(k-1)}}/a_{kk}^{(k-1)} \quad (i,j = k+1, \ldots, n).$$

Итоговым результатом нашего построения является матрица

$$D \;=\; A^{(n-1)} \;=\; L_{n-1} \cdots L_1 A L_1^* \cdots L_{n-1}^*,$$

что позволяет записать $A = LDL^*$ для

$$D \;=\; \mathrm{diag}(a_{11}^{(0)}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}), \quad L \;=\; R^* \;=\; (L_{n-1} \cdots L_1)^{-1} \;=\; L_1^{-1} \cdots L_{n-1}^{-1}.$$

Поскольку матрица L_k^{-1} отличается от матрицы L_k лишь сменой знаков поддиагональных элементов k-го столбца, мы получаем также, что

$$L \;=\; \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ a_{21}^{(0)}/a_{11}^{(0)} & 1 & 0 & \ldots & 0 \\ a_{31}^{(0)}/a_{11}^{(0)} & a_{32}^{(1)}/a_{22}^{(1)} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1}^{(0)}/a_{11}^{(0)} & a_{n2}^{(1)}/a_{22}^{(1)} & a_{n3}^{(2)}/a_{33}^{(2)} & \ldots & 1 \end{pmatrix}.$$

Заметим, что в подобной версии разложения Холесского нам достаточно хранить только нижний треугольник матрицы. При этом на первом шаге мы оставляет без изменения первый столбец и пересчитываем компонентны всех столбцов, начиная со второго, на втором шаге мы оставляем без изменения первые два столбца и пересичитываем компоненты столбцов, начиная с третьего, и т.д. Итоговая нижняя треугольная матрица, по которой можно восстановить разложение Холесского матрицы A, будет иметь вид

$$\begin{pmatrix} a_{11}^{(0)} & 0 & \dots & 0 \\ a_{21}^{(0)} & a_{22}^{(1)} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1}^{(0)} & a_{n2}^{(1)} & \dots & a_{nn}^{(n-1)} \end{pmatrix}.$$

Точно также можно было бы хранить верхний треугольник матрицы A и организовать процесс пересчёта компонент строк.

Процесс поиска разложения Холесского может быть реализован и несколько иным способом в духе приведённого ранее алгоритма построения LU–разложения. Итак, наша цель состоит в нахождении матриц $D = \mathrm{diag}(d_1, \dots, d_n)$, $d_i \in \{\pm 1\}$, и $R = (r_{ij})$, $r_{ij} = 0$ при $i > j$, $r_{ii} \in \mathbb{R}_+$, связанных с матрицей $A = A^*$ соотношением $A = R^* D R$ из разложения Холесского в форме с корнями. Данное соотношение может быть переписано в виде системы равенств

$$\sum_{k=1}^{n} \overline{r_{ki}} d_k r_{kj} \;=\; a_{ij} \quad (i, j = 1, \dots, n)\,,$$

из которых ввиду $a_{ij} = \overline{a_{ji}}$ достаточно оставить лишь $n(n+1)/2$ соотношений на верхний треугольник и диагональ

$$\sum_{k=1}^{n} \overline{r_{ki}} d_k r_{kj} \;=\; a_{ij} \quad (1 \le i \le j \le n)\,.$$

В свою очередь эти соотношения можно переписать в виде

$$\sum_{k=1}^{i} \overline{r_{ki}} d_k r_{kj} \;=\; \sum_{k=1}^{i-1} \overline{r_{ki}} d_k r_{kj} + r_{ii} d_i r_{ij} \;=\; a_{ij} \quad (1 \le i \le j \le n)$$

и, как следствие,

$$d_i \;=\; \mathrm{sign}\Big(a_{ii} - \sum_{k=1}^{i-1} d_k |r_{ki}|^2\Big)\,, \quad r_{ii} \;=\; \sqrt{\Big|a_{ii} - \sum_{k=1}^{i-1} d_k |r_{ki}|^2\Big|} \quad (i = 1, \dots, n)\,,$$

$$r_{ij} \;=\; \Big(a_{ij} - \sum_{k=1}^{i-1} \overline{r_{ki}} d_k r_{kj}\Big) / (r_{ii} d_i) \quad (1 \le i < j \le n)\,,$$

где сумма по k, участвующая во всех выражениях, опускается при $i = 1$. В случае $A > 0$ компоненты d_k не вычисляются (они все равны 1) и знак модуля в подкоренном выражении может быть опущен. Данные формулы наводят мысль на следующей организации вычислений. Матрица A хранится в виде своего верхнего треугольника. На

первом шаге вычисляются элементы $d_1 = \operatorname{sign} a_{11}$, $r_{11} = \sqrt{|a_{11}|}$ и все $r_{1j} = a_{1j}/(d_1 r_{11})$, $j = 2, \ldots, n$, элементы $\{r_{1i}\}$ замещают элементы $\{a_{1i}\}$. На втором шаге мы сперва находим d_2, r_{22} и затем r_{2j}, $j = 2, \ldots, n$, $\{r_{2i}\}$ храним на месте $\{a_{2i}\}$, и т.д. На k-ом шаге последовательно вычисляем d_k, r_{kk} и r_{kj}, $j = k+1, \ldots, n$, найденные элементы $\{r_{ki}\}$ замещают элементы $\{a_{ki}\}$. Процесс завершается вычислением d_n и r_{nn} на шаге с номером n, при этом элемент r_{nn} помещается на место элемента a_{nn}. Элементы $d_1, \ldots, d_n$ хранятся в виде отдельного вектора. Несложно проверить, что вычислительная сложность описанного алгоритма оценивается как $n^3/3 + O(n^2)$ при $n \to \infty$, т.е. практически в половину меньше вычислительной сложности метода Гаусса.

Как и в случае LU–разложения, любой из найденных вариантов разложения Холесского сводит задачу решения системы $Ax = b$ к решению двух треугольных систем.

Стоит также заметить, что описанные нами два способа поиска разложения Холесского являются по сути лишь двумя способами записи одной вычислительной процедуры (в форме внешних произведений в первом случае и в форме скалярных — во втором).

Разложение Холесского и базирующиеся на нём алгоритмы решения линейных систем с симметрическими матрицами обладает весьма интересными особенностями своей численной реализации. Основная проблема возникает с делением на элемент r_{kk} на k-ом шаге, малость которого может стать причиной возрастания погрешности. При этом в отличие от LU–разложения выполнение такого деления для случая $A = A^* > 0$ не вызывает больших проблем в силу неравенства $|r_{ki}|^2 \leq a_{ii}$ (в случае неопределённой матрицы системы A мы, естественно, не располагаем таким неравенством). Вместе с тем, накопление ошибок округления может привести к ситуации $r_{kk} = 0$, что делает невозможным дальнейшее продолжение вычислений. Для преодоления этих трудностей различными авторами были предложены разнообразные варианты усложнения разложения Холесского, имеющие своей целью получить близкое к нему по виду разложение, которое строится алгоритмом сходной сложности $n^3/3 + O(n^2)$, но обладающим более высокой численной устойчивостью. Остановимся кратко на основных идеях подобных модернизаций.

Начнём с разложения Холесского с диагональным выбором. Для начала следует отметить, что сопряжение матрицы A с матрицей-перестановкой P приводит к матрице PAP^t, диагональ которой представляет собой результат перестановки диагональных компонент A, точнее если $P = (p_{ij})$, $p_{i\sigma(i)} = 1$, $\sigma \in \mathfrak{S}_n$, то $(PAP^t)_{ii} = a_{\sigma(i)\sigma(i)}$, $i = 1, \ldots, n$. Это соображение позволяет изменить описанную нами первую схему построения разложения Холесского следующим образом: на k-ом шаге при переходе от $A^{(k-1)}$ к $A^{(k)}$, $k = 1, \ldots, n-1$,

1. находим наибольший по модулю элемент $a^{(k-1)}_{i_k i_k}$ среди всех элементов $a^{(k-1)}_{ii}$, $i = k, \ldots, n$, сопрягаем матрицу $A^{(k-1)}$ с матрицей перестановкой P_{ki_k}, в которой переставлены k-ая и i_k-ая строки, и получаем матрицу $A'^{(k-1)} = P_{ki_k} A^{(k-1)} P_{ki_k}$;

2. выполняем переход $A'^{(k-1)} \to A^{(k)} = L'_k A'^{(k-1)} L'^*_k$ по описанной выше схеме.

В итоге мы получаем матрицу $D = L'_{n-1} P_{n-1 i_{n-1}} \cdots L'_1 P_{1 i_1} A P'_{1 i_1} L'^*_1 \cdots P_{n-1 i_{n-1}} L'^*_{n-1}$, которая, как и в случае метода Гаусса с выбором по столбцу, может быть записана в виде $D = L^{-1}(P_{n-1 i_{n-1}} \cdots P_{1 i_1}) A (P_{1 i_1} \cdots P_{n-1 i_{n-1}}) L^{*-1}$, что приводит нас к равенству $A = PLDL^*P^t$, где $P = P_{1 i_1} \cdots P_{n-1 i_{n-1}}$.

Естественно, что такой подход не может быть гарантированно успешным во всех

случаях, поскольку малыми могут оказаться все диагональные элементы. Поэтому более естественно будет перейти от разложения с диагональной матрицей D к разложению, в котором в роли D выступает трёхдиагональная матрица T. В терминах матриц это соответствует представлению $A = PLTL^*P^t$ для некоторых нижней унитреугольной матрицы L и матрицы-перестановки P. При этом процесс строится таким образом, что модули коэфициентов матрицы L ограничены сверху 1 или некоторой константой. Известны несколько способов организации процессов подобного рода с трёхдиагональной и трёхдиагональной блочно-диагональной с блоками 1×1 и 2×2 матрицей (алгоритмы Аазена, Парлетта — Рейда, Банча — Кауфмана и Банча — Парлетта). В качестве иллюстрации мы приведём алгоритмы Парлетта — Рейда и Банча — Парлетта.

Алгоритм Парлетта — Рейда представляет собой алгоритм симметрической трёхдиагонализации матрицы с выбором по столбцу ниже побочной диагонали. Другими словами, на основе матрицы $A = A^* = A^{(0)}$ последовательно строятся подобные ей матрицы $A^{(1)}, \ldots, A^{(n-2)}$, последняя из которых трёхдиагональна, $A^{(n-2)} = T$. Матрица $A^{(k)}$ имеет вид

$$A^{(k)} = \begin{pmatrix} a_1 & b_1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ b_1 & a_2 & b_2 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & b_2 & a_3 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_k & b_k & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & b_k & a^{(k)}_{k+1k+1} & a^{(k)}_{k+1k+2} & \ldots & a^{(k)}_{k+1n} \\ 0 & 0 & 0 & \ldots & 0 & a^{(k)}_{k+2k+1} & a^{(k)}_{k+2k+2} & \ldots & a^{(n)}_{k+2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & & 0 & a^{(k)}_{nk+1} & a^{(k)}_{nk+2} & \ldots & a^{(k)}_{nn} \end{pmatrix}.$$

Переход $A^{(k-1)} \to A^{(k)}$, составляющий k-ый шаг алгоритма выполняется следующим образом: среди компонент $a^{(k-1)}_{k+1k}, \ldots, a^{(k-1)}_{nk}$ (первого столбца ведущей подматрицы ниже главной диагонали) выбирается компонента $a^{(k-1)}_{i_k k}$ с наибольшим модулем и если тот отличен от нуля, то выполняется переход к матрице $A'^{(k-1)} = P_{k+1i_k} A^{(k-1)} P_{k+1i_k}$, в результате которого в матрице $A^{(k-1)}$ переставляются строки и столбцы с номерами $k+1$ и i_k, а затем вычисляется матрица $A^{(k)} = L_k A'^{(k-1)} L_k^*$ для нижней унитреугольной матрицы L_k отличной от единичной матрицы лишь компонентами своего $k+1$-го столбца

$$L_k = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & -a'^{(k-1)}_{k+2k} / a'^{(k-1)}_{k+1k} & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -a'^{(k-1)}_{nk} / a'^{(k-1)}_{k+1k} & 0 & 0 & \ldots & 1 \end{pmatrix},$$

что обеспечивает аннулирование компонент k-ых столбца и строки матрицы $A'^{(k-1)}$ с номерами $k+2, \ldots, n$. В результате этих преобразований после выполнения $n-2$ шагов

будет получена матрица

$$T \;=\; A^{(n-2)} \;=\; L_{n-2}P_{n-1i_{n-2}}\cdots L_1P_{2i_1}AP_{2i_1}L_1^*\cdots P_{n-1i_{n-2}}L_{n-2}^* \;=\; L^{-1}P^tAPL^{*-1},$$

где нижняя унитреугольная матрица L собирается по схеме аналогичной методу Гаусса с выбором по столбцу, $P = P_{2i_1}\cdots P_{n-1i_{n-2}}$. Таким образом, данный алгоритм позволяет получить представление $A = PLTL^*P^t$, в котором

$$T \;=\; \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(1)} & 0 & \dots & 0 & 0 & 0 \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(2)} & \dots & 0 & 0 & 0 \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{n-2n-2}^{(n-3)} & a_{n-2n-1}^{(n-2)} & 0 \\ 0 & 0 & 0 & \dots & a_{n-1n-2}^{(n-2)} & a_{n-1n-1}^{(n-2)} & a_{n-1n}^{(n-2)} \\ 0 & 0 & 0 & \dots & 0 & a_{nn-1}^{(n-2)} & a_{nn}^{(n-2)} \end{pmatrix}.$$

К сожалению алгоритм Парлетта — Рейда не вполне отвечает нашим требованиям по трудоёмкости, поскольку его сложность оценивается как $2/3n^3 + O(n^2)$ при $n \to \infty$ (улучшенная версия этого алгоритма известная как алгоритм Аазена требует всего лишь $n^3/3 + O(n^2)$ операций). Вместе с тем он заведомо обеспечивает требуемые границы модулей матрицы L.

Алгоритм Банча — Парлетта и родственный ему алгоритм Банча — Кауфмана построены по несколько иной схеме. Она подразумевает построение на базе матрицы $A = A^* = A^{(0)}$ матриц $A^{(1)}, \dots, A^{(m)}$, где последняя матрица $T = A^{(m)}$ является трёхдиагональной матрицей блочно-диагональной структуры $T = \mathrm{diag}(T_1, \dots, T_m)$, где каждый блок T_i имеет размер $n_i \times n_i$, $n_i \in \{1, 2\}$. Матрица $A^{(k)}$ может быть представлена как блочно-диагональная матрица вида $A^{(k)} = \mathrm{diag}(T_1, \dots, T_k, \tilde{A}^{(k)})$, где $\tilde{A}^{(k)}$ — ведущая подматрица размера $(n - m_k) \times (n - m_k)$, $m_k = n_1 + \dots + n_k$, $\tilde{A}^{(k)} = (a_{ij}^{(k)})_{i,j=m_k+1}^{n}$. На k-ом шаге переход $A^{(k-1)} \to A^{(k)}$ выполняется по следующему правилу: в соответствии с некоторой стратегией выбирается матрица перестановка $\tilde{P}_k$ и размер n_k нового невырожденного диагонального блока T_k, где

$$\tilde{P}_k\tilde{A}^{(k-1)}\tilde{P}_k^t \;=\; \begin{pmatrix} T_k & B_k^* \\ B_k & C_k \end{pmatrix}$$

и блок $C_k = C_k^*$ имеет размер $(n - m_k) \times (n - m_k)$, $m_k = m_{k-1} + n_k$. Затем, полагая

$$\tilde{L}_k \;=\; \begin{pmatrix} E_{n_k} & 0 \\ -B_kT_k^{-1} & E_{n-m_k} \end{pmatrix},$$

мы выполняем блочное преобразование

$$\tilde{L}_k\tilde{P}_k\tilde{A}^{(k-1)}\tilde{P}_k^t\tilde{L}_k^* \;=\; \begin{pmatrix} T_k & 0 \\ 0 & C_k - B_kT_k^{-1}B_k^* \end{pmatrix} \;=\; \begin{pmatrix} T_k & 0 \\ 0 & \tilde{A}^{(k)} \end{pmatrix}.$$

Иначе говоря, осуществляемая нами процедура может быть записана как

$$A^{(k)} \;=\; \mathrm{diag}(T_1, \dots, T_k, \tilde{A}^{(k)}) \;=\; L_kP_kA^{(k-1)}P_k^tL_k^*,$$

где $P_k = \mathrm{diag}(E_{m_{k-1}}, \tilde{P}_k)$ и $L_k = \mathrm{diag}(E_{m_{k-1}}, \tilde{L}_k)$. Как и в предыдущем алгоритме, результатом её выполнения является представление $A = PLTL^*P^t$. Стратегия выбора $\tilde{P}_k$ и n_k, используемая в алгоритме Банча — Парлетта на k-ом шаге алгоритма, состоит в следующем:

1. действуя в рамках ведущей подматрицы $\tilde{A}^{(k-1)}$, находим

$$\mu_0(k) \;=\; \max_{i,j=m_{k-1}+1,\ldots,n} |a_{ij}^{(k-1)}|\,, \quad \mu_1(k) \;=\; \max_{i=m_{k-1}+1,\ldots,n} |a_{ii}^{(k-1)}|\,;$$

2. если $\mu_1(k) \geq \alpha\mu_0(k)$, где выбор параметра $0 < \alpha < 1$ оговаривается дополнительно, тогда мы полагаем $n_k = 1$ и подбираем $\tilde{P}_k$ таким образом, что

$$|(\tilde{P}_k \tilde{A}^{(k-1)} \tilde{P}_k^t)_{m_{k-1}+1 m_{k-1}+1}| \;=\; \mu_1(k)\,,$$

   т.е. если $a_{i_k i_k}^{(k-1)}$ — элемент диагонали матрицы $\tilde{A}^{(k-1)}$ с наибольшим модулем, то $P_k = P_{m_{k-1}+1 i_k}$; в противном случае $n_k = 2$ и матрица $\tilde{P}_k$ подбирается с тем, чтобы

$$|(\tilde{P}_k \tilde{A}^{(k-1)} \tilde{P}_k^t)_{m_{k-1}+2 m_{k-1}+1}| \;=\; \mu_0(k)\,,$$

   точнее в данном случае наибольший по модулю элемент $a_{i_k j_k}^{(k-1)}$ матрицы $\tilde{A}^{(k-1)}$ лежит вне диагонали (можно считать, что $j_k < i_k$) и для того, чтобы его перевести на позицию $(m_{k-1}+2, m_{k-1}+1)$ нам достаточно взять $P_k = P_{m_{k-1}+2 i_k} P_{m_{k-1}+1 j_k}$.

Заметим, что невырожденность блока T_k во втором случае обеспечивается тем, что его диагональные элементы (i_k-ый и j_k-ый диагональные элементы $\tilde{A}^{(k-1)}$) строго меньше по модулю оставшихся двух равных между собой элементов (модуль определителя этой матрицы больше или равен $\mu_0(k)^2 - \mu_1(k)^2 \geq (1-\alpha^2)\mu_0(k)^2$). Исходя из соображений уменьшения оценки роста модулей в L–составляющей получаемого разложения, в этом алгоритме предлагается использовать $\alpha = (1+\sqrt{17})/8$. Оценка числа операций в алгоритме Банча — Парлетта не вполне соответствует нашим требованиям: $n^3/3 + O(n^2)$ (на вычисления), к которым добавляется от $n^3/6$ до $n^3/12$ сравнений, необходимых для выбора диагональных блоков. Известно также, что данный алгоритм довольно устойчив (точнее по устойчивости он сопоставим с методом Гаусса с полным выбором).

В алгоритме Банча — Кауфмана используется иная стратегия выбора n_k и $\tilde{P}_k$:

1. будем считать, что элемент $a_{m_{k-1}+1 m_{k-1}+1}^{(k-1)}$ (первый первых строки и столбца) матрицы $\tilde{A}^{(k-1)}$ имеет наибольший модуль среди всех её диагональных элементов (при необходимости этого можно добиться сопряжением (подобием) с подходящей матрицей перестановкой);

2. найдём элемент с наибольшим модулем в первом столбце подматрицы $\tilde{A}^{(k-1)}$, т.е. определим i_k, $m_{k-1}+1 \leq i_k \leq n$,

$$\lambda \;=\; |a_{i_k m_{k-1}+1}^{(k-1)}| \;=\; \max_{i=m_{k-1}+1,\ldots,n} |a_{i m_{k-1}+1}^{(k-1)}|$$

   (для простоты мы предполагаем, что матрица невырождена и $\lambda > 0$);

3. если $|a^{(k-1)}_{m_{k-1}+1 m_{k-1}+1}| \geq \alpha\lambda$ (α определяется, как и в предыдущем алгоритме, то мы полагаем $n_k = 1$ и $P_k = E$; в противном случае мы находим j_k, $m_{k-1}+1 \leq j_k \leq n$,

$$\sigma \;=\; |a^{(k-1)}_{j_k i_k}| \;=\; \max_{j=m_{k-1}+1,\ldots,n,\ j\neq i_k} |a^{(k-1)}_{j i_k}|$$

(наибольший по модулю внедиагональный элемент i_k–столбца матрицы $\tilde{A}^{(k-1)}$), и в ситуации $\sigma|a^{(k-1)}_{m_{k-1}+1 m_{k-1}+1}| \geq \alpha\lambda^2$ вновь полагаем $n_k = 1$ и $P_k = E$, а иначе

- при $|a^{(k-1)}_{i_k i_k}| \geq \alpha\sigma$ положим $n_k = 1$ и возьмём $P_k = P_{m_{k-1}+1 i_k}$, обеспечив тем самым перестановку первого и i_k-го столбцов матрицы $\tilde{A}^{(k-1)}$ и вместе с тем перемешение её i_k-го диагонального элемента на первую позицию:

$$(\tilde{P}_k \tilde{A}^{(k-1)} \tilde{P}_k^t)_{m_{k-1}+1 m_{k-1}+1} \;=\; a^{(k-1)}_{i_k i_k}\,;$$

- при $|a^{(k-1)}_{i_k i_k}| < \alpha\sigma$ положим $n_k = 2$ и возьмём $P_k = P_{m_{k-1}+2 i_k} P_{m_{k-1}+1 j_k}$, что даёт

$$(\tilde{P}_k \tilde{A}^{(k-1)} \tilde{P}_k^t)_{m_{k-1}+2 m_{k-1}+1} \;=\; a^{(k-1)}_{i_k j_k} \;=\; a^{(k-1)}_{j_k i_k}\,.$$

Заметим, что невырожденность блока T_k в последнем случае следует из неравенства $|\det T_k| = |a^{(k-1)}_{i_k i_k} a^{(k-1)}_{j_k j_k} - (a^{(k-1)}_{i_k j_k})^2| \geq \sigma^2 - |a^{(k-1)}_{i_k i_k}|^2 \geq \sigma^2(1-\alpha^2)$, которое устанавливается следующим образом: в рассматриваемой ситуации: $|a^{(k-1)}_{m_{k-1}+1 m_{k-1}+1}| < \alpha|a^{(k-1)}_{i_k m_{k-1}+1}| \leq \alpha\sigma = \alpha|a^{(k-1)}_{i_k j_k}|$, что в силу п. 1 даёт нам $|a^{(k-1)}_{i_k i_k}|, |a^{(k-1)}_{j_k j_k}| \leq |a^{(k-1)}_{m_{k-1}+1 m_{k-1}+1}| \leq \alpha\sigma$ (этот пункт в книге Дж. Голуб, Ч. Ван Лоун "Матричные вычисления" отсутствует). Алгоритм Банча — Кауфмана несколько более экономен, чем алгоритм Банча — Парлетта ($n^3/3 + O(n^2)$ при $n \to \infty$), поскольку при выборе нового диагонального блока он оперирует не со всей ведущей подматрицей, а лишь с её диагональю и двумя столбцами.

Заметим, что в отличие от разложения Холесского алгоритмы Парлетта — Рейда и Банча — Кауфмана применимы для всех симметрических систем (из организации их вычислений следует, что соответствующее разложение может быть построено даже в вырожденном случае).

Построив разложение $A = PLTL^*P^t$ или $P^tAP = LTL^*$ по любому из представленных нами алгоритмов, мы можем свести линейную систему $Ax = b$ к системам

$$\begin{cases} Lz \;=\; P^t b\,, \\ Tw \;=\; z\,, \\ L^* y \;=\; w\,, \\ x \;=\; Py\,, \end{cases}$$

где решение треугольных систем осуществляется стандартным обратным ходом метода Гаусса, а решение трёхдиагональной симметрической системы может быть осуществлено по любому из известных алгоритмов для ленточных систем, причём в случае с блочно-диагональной системой с блоками 1×1 и 2×2 решение находятся прямым обращением блоков.

Естественным развитием предложенных здесь методов являются их разнообразные блочные аналоги.

Вариант домашнего задания. Первая часть.

Всюду ниже ставится задача программной реализации соответствующего алгоритма и предполагается выполнение проверки качества реализации на плохо обусловленных матрицах (таких, например, как матрица Гильберта).

1. Реализовать метод Гаусса с частичным выбором по столбцу, построить с его помощью разложение $PA = LU$, написать связанный с этим разложением оценщик числа обусловленности матрицы в строчной норме и оценить с его помощью относительную погрешность решения системы с использованием вектора невязки.

2. Построить разложение $PAP' = LU$ для разреженной матрицы на основе метода Гаусса с оптимальным заполнением и реализовать метод Гаусса с частичным строчным выбором для ленточных систем, построив для них разложение $PA = LU$. Реализовать для обоих методов итерационные уточнения простой и переменной точности и сравнить качество получаемых результатов.

3. Реализовать методы Холесского и Холесского с диагональным выбором для симметрических систем, обладающих LU–разложением. Построить с их помощью разложения $A = LDL^*$ и $PAP^t = LDL^*$. Сравнить полученные результаты и реализовать для обоих алгоритмов итерационное уточнение переменной точности.

4. Реализовать метод Парлетта — Рейда для получения разложения $PAP^t = LTL^*$ и решения на его основе симметрических линейных систем. Для решения трёхдиагональной системы использовать метод Гаусса со строчным выбором и диагональным хранением данных

5. Решить аналогичную задачу для метода Аазена.

6. Реализовать метод Банча — Парлетта для получения разложения $PAP^t = LTL^*$ и решения с его помощью симметрических линейных систем. Проанализировать роль параметра α для качества получаемых решений.

7. Решить аналогичную задачу для алгоритма Банча — Кауфмана.

8. Построить QR–разложение методом вращений и написать с его помощью оценщик строчной матричной нормы. Оценить с помощью последнего относительную погрешность решения линейной системы методом вращений. Протестировать качество полученных результатов (сравнить с оценкой для треугольной системы, полученной в ходе построения разложения).

Лекция 4. QR–разложение и алгоритмы его построения.

Нашей следующей целью станет исследование представления произвольной матрицы A размера $n \times m$ в виде $A = QR$, где Q — ортогональная или унитарная в комплексном случае матрица размера $n \times n$ и R — верхняя треугольная матрица размера $n \times m$. Мы также рассмотрим классические способы построения такого QR–разложения, построенные по схеме близкой гауссову исключению с той лишь разницей, что роль элементов системы порождающих полной линейной группы, умножения на которые соответствуют элементарным преобразованиям, в данном случае будут выполнять элементы систем порождающих ортогональной (специальной ортогональной) и унитарной групп, отвечающие элементарным движениям конечномерного евклидова пространства.

Естественно, что выбор подобных преобразований мотивируется равенством спектральных норм (или норм Фробениуса) исходной матрицы и преобразованной с их помощью (умноженной на них) матрицы, что, к примеру, позволяет осуществлять переход от системы $Ax = b$ к системе $Rx = Q^*b$, $A = QR$, без потери качества в решении (без изменения числа обусловленности в спектральной норме для матрицы системы). При этом решение треугольной системы $Rx = Q^*b$ с учётом умножения матрицы Q^* на вектор b будет стоить порядка $O(n^2)$ операций. Впрочем, с точки зрения сложности построение QR–разложения намного более затратно чем алгоритмы круга гауссова исключения. Имеются и другие важные приложения QR–разложения, о которых мы будем говорить позднее.

Замечание 0.24. *Для любой матрицы A размера $n \times m$ существует представление $A = QR$, где Q — ортогональная или унитарная матрица размера $n \times n$, R — верхняя треугольная матрица размера $n \times m$. Более того, можно считать, что на диагонали R стоят неотрицательные вещественные числа.*

Доказательство. Достаточно выбрать в матрице A столбцы с номерами $i_1, \ldots, i_r$, $r = \operatorname{rk} A$, где i_1 — номер первого ненулевого столбца, i_2 — номер первого столбца линейно независимого с i_1-ым столбцом, $i_2 > i_1$, i_3 — номер первого столбца линейно независимого с i_1-ым и i_2-ым столбцами, $i_3 > i_2 > i_1$, и т.д. Затем применить к выбранным столбцам стандартный процесс последовательной ортогонализации Грама — Шмидта и получить на их основе ортонормированные столбцы $u_{i_1}, \ldots, u_{i_r}$. Дополним эти столбцы до ортогональной или унитарной матрицы U так, чтобы u_{i_j} был столбцом последней с номером i_j для всех j. Остаётся построить R размера $n \times m$, составленную из столбцов-координат разложения столбцов A по базису $\{u_{i_1}, \ldots, u_{i_r}\}$. □

Замечание 0.25. *Для любой невырожденной матрицы A её QR–разложение $A = QR$ с ортогональной или унитарной матрицей Q и верхней треугольной матрицей R, которая имеет положительную вещественную диагональ, определено однозначно. При этом матрица R совпадает с верхним треугольным фактором разложения Холесского матрицы A^*A.*

Доказательство. Наличие двух подобных разложений $A = QR = Q'R'$ означает, что $Q'^*Q = R'R^{-1} = R''$ — верхняя треугольная матрица с положительной вещественной диагональю, обратная к которой совпадает с нижней треугольной матрицей R''^*. Из единственности LU–разложения $E = R''^*R''$ следует, что R'' — диагональная матрица с квадратами диагональных элементов равными единице. Поэтому $R'' = E$.

Остаётся напомнить однозначность разложения Холесского, которая следует из однозначности LU–разложения. $\square$

Сказанное наводит на мысль о естественном способе построения QR–разложения на базе процесса ортогонализации столбцов. Применительно к невырожденной матрице $A \in Gl_n(\Bbbk)$, $\Bbbk = \mathbb{R}, \mathbb{C}$, столбцы которой мы обозначим через $a_1, \ldots, a_n$ такой процесс строится следующим образом: найдём $u_1 = \frac{a_1}{\|a_1\|_2}$ и далее

$$u_k \;=\; \frac{a_k - \sum\limits_{i=1}^{k-1}(a_k, u_i)u_i}{\left\|a_k - \sum\limits_{i=1}^{k-1}(a_k, u_i)u_i\right\|_2} \;=\; \frac{a_k - \sum\limits_{i=1}^{k-1} r_{ik}u_i}{r_{kk}} \quad (k = 2, \ldots, n)\,,$$

где $r_{ik} = (a_k, u_i)$, $r_{kk} = \left\|a_k - \sum\limits_{i=1}^{k-1} r_{ik}u_i\right\|_2$, затем положим $U = (u_1 \ldots u_n)$ и $R = (r_{ij})$, полагая $r_{ij} = 0$, $1 \leq j < i \leq n$. Сооветствующая вычислительная процедура CGS выглядит следующим образом:

1. $U := (0 \ldots 0)$ (нулевая матрица $m \times n$);
2. для всех $j = 1, \ldots, n$
   - $r_j := U^* a_j$;
   - $t := a_j - U r_j$ (т.е. $(E - U^*U)a_j$);
   - $r_{jj} := \|t\|_2$;
   - $u_j := t/r_{jj}$.

Немного более качественная версия этого алгоритма (модифицированная процедура Грама — Шмидта (MGS)), эквивалентная ему в точной арифметике, выглядит так: для всех $j = 1, \ldots, n$

1. $t := a_j$;
2. для всех $i = 1, \ldots, j-1$
   - $r_{ij} := u_i^* t$;
   - $t := t - u_i r_{ij}$;
3. $r_{jj} := \|t\|_2$;
4. $u_j := t/r_{jj}$,

где единственное принципиальное отличие по сравнению с классическим алгоритмом состоит в п. 2 и процессом вычисления вектора t, в котором при каждом i выполняется фактически следующее: вместо привычного $r_{ij} = u_i^* a_j$

$$r_{ij} \;=\; u_i^*\big(a_j - \sum_{k=1}^{i-1} u_k r_{kj}\big)\,.$$

К сожалению ни классический, ни модифицированный процесс не являются численно устойчивыми. Точнее нормирование, выполняемое на каждом шаге, может привести к нарушению ортогональности столбцов $u_1, \ldots, u_n$ (деление на малый элемент приводит к значительному росту погрешности). Поэтому данные процедуры применяются с дополнительными модификациями наподобие встроенной переортогонализации столбцов.

Мы рассмотрим два других способа построения QR–разложения (или ортогональной (унитарной) триангуляции) посредством элементарных ортогональных и унитарных преобразований, в роли которых выступают вращения двумерных плоскостей и отражения относительно гиперплоскостей.

Метод Гивенса (метод вращений)

Пусть вектора $e_1, \ldots, e_n$ формируют ортонормированный базис n-мерного вещественного пространства, которое мы будем отождествлять с пространством столбцов $\mathbb{R}^n$. Под элементарным вращением понимается поворот двумерного подпространства $\langle e_i, e_j \rangle$, $1 \leq i < j \leq n$, на угол ϕ, оставляющий неподвижными все вектора подпространства $\langle e_k \mid k \neq i, j \rangle$, т.е. преобразование действующее на элементах выбранного нами базиса по правилу: $e_i \longmapsto e_i \cos\phi + e_j \sin\phi$, $e_j \longmapsto e_i(-\sin\phi) + e_j \cos\phi$, $e_k \longmapsto e_k$, $k \neq i, j$. Матрица этого преобразования $T_{ij}(\phi)$ называется *матрицей элементарного вращения*. В базисе $\{e_1, \ldots, e_n\}$ матрица $T_{ij}(\phi)$ отличается от единичной лишь элементами $\cos\phi$, $-\sin\phi$, $\sin\phi$ и $\cos\phi$, стоящими на позициях (i,i), (i,j), (j,i) и (j,j) соответственно. В рамках плоскости $\langle e_i, e_j \rangle$ матрица данного преобразования представляет собой обычную матрицу Якоби (из его работы 1842 г.)

$$T(\phi) \;=\; \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}.$$

Матрица $T_{ij}(\phi)$ входит в специальную ортогональную группу $SO_n(\mathbb{R})$, причём $T_{ij}(\phi)^{-1} = T_{ij}(\phi)^t = T_{ij}(-\phi)$.

Идею метода вращений или метода Гивенса содержит в себе следующее наблюдение.

Замечание 0.26. *Для всякого вектора $x = (x_1, \ldots, x_n)^t \in \mathbb{R}^n$, $n \geq 2$, можно подобрать $n-1$ матрицу элементарного вращения $T_{1i+1}(\phi_i)$, $i = 1, \ldots, n-1$, для которых*

$$T_{1n}(\phi_{n-1}) \cdots T_{12}(\phi_1) x \;=\; (\|x\|_2, 0, \ldots, 0)^t \;=\; \|x\|_2 (1, 0, \ldots, 0)^t.$$

Доказательство. Поскольку нашей целью является построение алгоритма, мы предложим конструктивное доказательство этого факта без использования индуктивных соображений. В случае $x = 0$ мы можем считать, что $\phi_i = 0$, $T_{1i+1}(\phi_i) = E$, $i = 1, \ldots, n-1$. Поэтому можно считать, что $x \neq 0$. Положим $x(k) = (x_1, \ldots, x_k)^t$, $k = 1, \ldots, n$, $x^{(0)} = x$ и далее

$$x^{(i)} \;=\; (\|x(i+1)\|_2, 0, \ldots, 0, x_{i+2}, \ldots, x_n)^t \quad (i = 1, \ldots, n-2),$$

$x^{(n-1)} = (\|x\|_2, 0, \ldots, 0)^t$. Пусть s — индекс первой ненулевой координаты вектора x, т.е. $x(s) \neq 0$ и $x(i) = 0$ при всех $i < s$. В случае $s > 2$ положим $\phi_1 = \ldots = \phi_{s-2} = 0$. При $s > 1$ мы возьмём $\phi_{s-1} = -\operatorname{sign} x_s \pi/2$. Это гарантирует нам, что $\cos\phi_{s-1} = 0$, $\sin\phi_{s-1} = -\operatorname{sign} x_s$ и

$$T_{1s}(\phi_{s-1}) x^{(s-2)} \;=\; x^{(s-1)} \;=\; (|x_s|, 0, \ldots, 0, x_{s+1}, \ldots, x_n)^t,$$

где в данной ситуации $x^{(s-2)} = x^{(0)} = x$. При $s = 1$ определим ϕ_1 из соотношений

$$\cos\phi_1 \;=\; \frac{x_1}{\|x(2)\|_2}, \quad \sin\phi_1 \;=\; -\frac{x_2}{\|x(2)\|_2},$$

в соответствии с которыми $T_{12}(\phi_1)x^{(0)} = x^{(1)}$, так как

$$T(\phi_1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} \frac{x_1}{\|x(2)\|_2} & \frac{x_2}{\|x(2)\|_2} \\ -\frac{x_2}{\|x(2)\|_2} & \frac{x_1}{\|x(2)\|_2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;=\; \begin{pmatrix} \|x(2)\|_2 \\ 0 \end{pmatrix}.$$

Далее определим $\phi_{s+1}, \ldots, \phi_{n-1}$ следующим образом: если $x_{s+j+1} = 0$, то $\phi_{s+j} = 0$; в противном случае ϕ_{s+j} находится из соотношений

$$\cos\phi_{s+j} \;=\; \frac{\|x(s+j)\|_2}{\|x(s+j+1)\|_2}, \quad \sin\phi_{s+j} \;=\; -\frac{x_{s+j+1}}{\|x(s+j+1)\|_2}.$$

Такой выбор обеспечивает выполнение равенства $T_{1s+j+1}(\phi_{s+j})x^{(s+j-1)} = x^{(s+j)}$, поскольку

$$T_{1s+j+1}(\phi_{s+j})\begin{pmatrix} \|x(s+j)\|_2 \\ x_{s+j+1} \end{pmatrix} \;=$$

$$\begin{pmatrix} \frac{\|x(s+j)\|_2}{\|x(s+j+1)\|_2} & \frac{x_{s+j+1}}{\|x(s+j+1)\|_2} \\ -\frac{x_{s+j+1}}{\|x(s+j+1)\|_2} & \frac{\|x(s+j)\|_2}{\|x(s+j+1)\|_2} \end{pmatrix}\begin{pmatrix} \|x(s+j)\|_2 \\ x_{s+j+1} \end{pmatrix} \;=\; \begin{pmatrix} \|x(s+j+1)\|_2 \\ 0 \end{pmatrix}.$$

Остаётся заметить, что такой выбор $\phi_1, \ldots, \phi_{n-1}$ обеспечивает нам выполнение равенств

$$T_{1i+1}(\phi_i)x^{(i-1)} \;=\; x^{(i)} \quad (i = 1, \ldots, n-1).$$

□

Опишем теперь алгоритм метода отражений решения линейной системы $Ax = b$, $A \in Gl_n(\mathbb{R})$, и связанный с ним алгоритм построения QR–разложения. Как и в методах гауссова исключения, мы построим последовательсть эквивалентных систем $A^{(k)}x = b^{(k)}$, $k = 0, 1, \ldots, n-1$, которая начинается исходной системой ($A^{(0)} = A$, $b^{(0)} = b$) и завершается верхней треугольной системой $A^{(n-1)}x = b^{(n-1)}$.

При каждом k матрица системы $A^{(k)}x = b^{(k)}$ имеет вид

$$A^{(k)} \;=\; \begin{pmatrix} \|a_1^{(0)}\|_2 & c_{12} & \ldots & c_{1k} & c_{1k+1} & \ldots & c_{1n} \\ 0 & \|a_1^{(1)}\|_2 & \ldots & a_{2k} & c_{2k+1} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \|a_1^{(k-1)}\|_2 & c_{kk+1} & \ldots & c_{kn} \\ 0 & 0 & \ldots & 0 & a_{k+1k+1}^{(k)} & \ldots & a_{kn}^{(k)} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a_{nk+1}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix},$$

а $b^{(k)} = (y_1, \ldots, y_k, b_{k+1}^{(k)}, \ldots, b_n^{(k)})^t$, где $a_1^{(i)} = (a_{i+1i+1}^{(i)}, \ldots, a_{ni+1}^{(i)})^t$ — первый столбец ведущей подматрицы, полученной после выполнения i-ого шага. Вычислительная работа на k-ом шаге, осуществляющим переход $A^{(k-1)} \to A^{(k)}$, $b^{(k-1)} \to b^{(k)}$, состоит в перевычислении компонент ведущей подматрицы $A'^{(k-1)} = (a_{ij}^{(k-1)})_{i,j=k}^n$ и компонент вектора правой части с индексами $k, k+1, \ldots, n$, выполняемой следующим образом:

1. в соответствии с доказанным выше замечанием выберем $n-k$ углов $\phi_{kk+1},\dots,\phi_{kn}$, для которых

$$T'_{1n-k+1}(\phi_{kn})\cdots T'_{12}(\phi_{kk+1})a_1^{(k-1)} \;=\; (\|a_1^{(k-1)}\|_2,0,\dots,0)^t\,,$$

где $T'_{1s+1}(\phi_{kk+s})$, $s=1,\dots,n-k$, — матрицы вращения $n-k+1$–мерного пространства (размера $(n-k+1)\times(n-k+1)$);

2. для каждого $t=2,\dots,n-k$ столбец $a_t^{(k-1)}$ ведущей подматрицы $A'^{(k-1)}$ заменяется столбцом

$$T'_{1n-k+1}(\phi_{kn})\cdots T'_{12}(\phi_{kk+1})a_t^{(k-1)} \;=\; (c_{kt},a_{k+1t}^{(k)},\cdots,a_{nt}^{(k)})^t\,,$$

её первый столбец $a_1^{(k-1)}$ заменяется столбцом $(\|a_1^{(k-1)}\|_2,0,\dots,0)^t$, что соответствует переходу

$$T'_{1n-k+1}(\phi_{kn})\cdots T'_{12}(\phi_{kk+1})A'^{(k-1)} \;=\; \begin{pmatrix} \|a_1^{(k-1)}\|_2 & c_{kk+1} & \dots & c_{kn} \\ 0 & a_{k+1k+1}^{(k)} & \dots & a_{k+1n}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{nk+1}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix},$$

компонентны правой части перевычисляются по схеме

$$T'_{1n-k+1}(\phi_{kn})\cdots T'_{12}(\phi_{kk+1})(b_k^{(k-1)},\cdots,b_n^{(k-1)})^t \;=\; (y_k,b_{k+1}^{(k)},\dots,b_n^{(k)})^t\,.$$

Естественно, что данный процесс перевычисления столбцов ведущей подматрицы и компонент вектора правой части может осуществляется согласовано с нахождением очередного угла ϕ_{kk+s} (после его вычисления или вычисления матрицы $T'_{1s+1}(\phi_{kk+s})$ (речь идёт, естественно о нахождении только $\cos\phi_{kk+s}$ и $\sin\phi_{kk+s}$) эта матрица умножается на столбцы ведущей подматрицы, начиная со второго, и на соответствующую часть вектора правой части). Другими словами, суть k-го шага состоит в вычислении

$$A^{(k)} \;=\; T_{kn}(\phi_{kn})\cdots T_{kk+1}(\phi_{kk+1})A^{(k-1)}\,,\quad b^{(k)} \;=\; T_{kn}(\phi_{kn})\cdots T_{kk+1}(\phi_{kk+1})b^{(k-1)}\,.$$

В результате выполнения $n-1$ шага мы приходим к треугольной системе $A^{(n-1)}x=b^{(n-1)}$, где

$$A^{(n-1)} \;=\; \begin{pmatrix} \|a_1^{(0)}\|_2 & c_{12} & \dots & c_{1n-1} & c_{1n} \\ 0 & \|a_1^{(1)}\|_2 & \dots & a_{2n-1} & c_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \|a_1^{(n-2)}\|_2 & c_{n-1n} \\ 0 & 0 & \dots & 0 & a_{nn}^{(n-1)} \end{pmatrix}$$

и $b^{(n-1)}=(y_1,\dots,y_{n-1},b_n^{(n-1)})^t$. Решение данной системы осуществляется уже обычным обратным ходом метода Гаусса. Заметим, что по построению $A^{(n-1)}=TA$, $b^{(n-1)}=Tb$ для

$$T \;=\; T_{n-1n}(\phi_{n-1n})(T_{n-2n}(\phi_{n-2n})T_{n-2n-1}(\phi_{n-2n-1}))\cdots \\ (T_{kn}(\phi_{kn})\cdots T_{kk+1}(\phi_{kk+1}))\cdots(T_{1n}(\phi_{1n})\cdots T_{12}(\phi_{12}))\,.$$

Заметим, что фактически описанный процесс триангуляции применим к любой квадратной системе (вопрос разрешимости треугольной системы (проверка отсутствия нулевых элементов на диагонали)) может проводиться отдельно перед началом выполнения обратного хода.

Следует подчеркнуть, что рассматриваемый нами процесс триангуляции посредством элементарных вращений содержит в себе достаточный запас ресурсов для распараллеливания (превращения в последовательно-параллельный процесс). Это обусловлено не столько с возможностью эффективной параллельной реализации матричного умножения (в нашем случае умножение матрицы вращения на вектор сводится к перевычислению лишь двух его компонент), а с возможностью одновременных умножений на матрицы вращения с целью одновременного аннулирования компонент столбца ведущей подматрицы.

Оценим объём вычислительной работы на k-ом шаге нашего алгоритма (триангуляции вращениями). Нам потребуется:

1. построить $n-k$ матриц $T'_{1s+1}(\phi_{kk+s})$, $s=1,\ldots,n-k$; поскольку нахождение $\cos\phi = \frac{x}{\sqrt{x^2+y^2}}$ и $\sin\phi = \frac{-y}{\sqrt{x^2+y^2}}$ требует 6 операции, это требует $6(n-k)$ операций;

2. вычислить $\|a_1^{(k-1)}\|_2$ требует $n-k+1+n-k+1 = 2(n-k+1)$ операций;

3. перевычисление $n-k$ столбцов ведущей подматрицы требует выполнить умножение каждого из них на матрицы $T'_{1s+1}(\phi_{kk+s})$, $s=1,\ldots,n-k$ (умножение каждой матрицы $T_{ij}(\phi)$ на вектор z, т.е. перевычисление его i-ой и j-ой компонент по формулам $z_i\cos\phi - z_j\sin\phi$ и $z_i\sin\phi + z_j\cos\phi$, обходится в 6 операций), потребует в общей сложности $6(n-k)^2$ операций, к которым добавляется ещё $6(n-k)$ операций на аналогичные преобразования вектора правой части.

Следовательно, описанный алгоритм обходится в $2n^3+O(n^2)$ операций, к которым добавляется ещё $O(n^2)$ операций на решение треугольной системы, $n\to\infty$.

Рассмотренный нами процесс триангуляции может быть применим к нахожению QR–разложения любой прямоугольной вещественной матрицы A размера $n\times m$. Для этого достаточно повторить $\min\{n,m\}-1$ шаг описанного алгоритма и получить верхнюю треугольную матрицу R' размера $n\times m$, у которой все компоненты диагонали неотрицательны, за исключением быть может последнего $r'_{\min\{n,m\},\min\{n,m\}}$. Другими словами, в результате мы построим матрицы элементарного вращения $\{T_{kk+s}(\phi_{kk+s}) \mid k=1,\ldots,\min\{n,m\}-1, s=1,\ldots,n-k\}$ и матрицу

$$T^{-1} = (T_{12}(-\phi_{12})\cdots T_{1n}(-\phi_{1n}))\cdots(T_{kk+1}(-\phi_{kk+1})\cdots T_{kn}(-\phi_{kn}))\cdots \\ T_{\min\{n,m\}-1\,\min\{n,m\}}(-\phi_{\min\{n,m\}-1\,\min\{n,m\}}),$$

хранимую отдельно, такие, что $T^tA = R'$. Остаётся взять

$$D = \mathrm{diag}(\underbrace{1,\ldots,1}_{\min\{n,m\}-1}, \operatorname{sign} r'_{\min\{n,m\}\,\min\{n,m\}}, \underbrace{1,\ldots,1}_{n-\min\{n,m\}})$$

и положить $DT^tA = R$, $Q = TD$, обеспечив выполнение равенства $A = QR$ для $Q\in O_n(\mathbb{R})$ и верхней треугольной матрицы R с неотрицательной диагональю. Количество

операций, необходимых для построения матриц Q и R в случае $n = m$ оценивается как $3n^3 + O(n^2) + 2n^3 + O(n^2) = 5n^3 + O(n^2)$, $n \to \infty$. При этом матрица R хранится на месте верхнего треугольника матрицы A. Матрица Q хранится отдельно (или мы строим матрицы T и R', сохраняя иформацию о T на месте нижнего треугольника A: ϕ_{ij}, $i < j$ ($\cos\phi_{ij}$ или $\sin\phi_{ij}$), на позиции (j, i)). Построение матрицы T занимает $n - 1$ шаг: на начальном этапе считаем $T = E$, k-ый шаг построения сводится к переходу $T = TT_{kk+1}(\phi_{kk+1}) \cdots T_{kn}(\phi_{kn})$, требущему выполнения $6n(n-k)$ операций (отсюда собственно и выводится оценка $3n^3 + O(n^2)$). Затем вычисляется матрица $Q = TD$ (при необходимости знаки элементов последнего столбца T меняются на противоположные).

Следует оговориться, что если не требовать однозначности QR–разложения (ограничиться однозначностью с точностью до умножения Q–составляющей на диагональную матрицу с коэффициентами ± 1 на диагонали), то можно обойтись построением матриц T и R'.

Естественным следствием описанного нами процесса является следующий вывод.

Замечание 0.27. *Специальная ортогональная группа $SO_n(\mathbb{R})$ порождается матрицами элементарных вращений $T_{ij}(\phi)$, $1 \le i < j \le n$, $\phi \in [0, 2\pi)$.*

Доказательство. В терминах наших предыдущих рассуждений для всякой матрицы $A \in SO_n(\mathbb{R})$ можно подобрать такое произведение матриц элементарных вращений T, что $TA = R'$ — верхняя треугольная матрица, у которой все компоненты диагонали кроме быть может последней положительны. Поскольку $R' \in SO_n(\mathbb{R})$, $\det R' = r'_{11} \cdots r'_{nn} = 1$ и, как следствие, $r'_{ii} > 0$ при всех i. Более того, ${R'}^t = {R'}^{-1}$ и потому в силу единственности LU–разложения ($E = {R'}^t R'$) $R' = \operatorname{diag}(r'_{11}, \ldots, r'_{nn})$, ${R'}^2 = {R'}^t R' = E$, ${r'_{ii}}^2 = 1$, $r'_{ii} = 1$, $i = 1, \ldots, n$, т.е. $R' = E$, $A = T^t = T^{-1}$. □

Метод Хаусхолдера (метод отражений)

Для начала стоит сказать несколько слов об операторах отражения относительно гиперплоскости ортогональной некоторому ненулевому вектору.

Пусть имеется евклидово пространство E (для определённости над полем действительных или комплексных чисел) со скалярным произведением или эрмитовой формой $(\,,\,)$ и ненулевой вектор $x \in E$. Тогда оператор отражения U_x относительно гиперплоскости $\langle x\rangle^\perp = \{y \in E \mid (x, y) = 0\}$ ортогональной вектору x может быть записан в виде

$$U_x \;=\; \mathrm{Id}_E - 2\frac{(\,,x)}{\|x\|_2}x \;:\; z \longmapsto U_x(z) \;=\; z - 2\frac{(z,x)}{\|x\|_2}x \quad (z \in E)\,,$$

где $(\,,x)$ — отвечающий x ограниченный линейный функционал на пространстве E, $\|(\,,x)\|_{E^*} = \|x\|_2 = \sqrt{(x,x)}$. Действительно, поскольку мы располагаем ортогональным разложением $E = \langle x\rangle \oplus \langle x\rangle^\perp$, в соответствии с которым любой вектор $z \in E$ может быть единственным образом записан как $z = \frac{(z,x)}{\|x\|_2}x + y$, где y — проекция z на гиперплоскость $\langle x\rangle^\perp$, мы сразу получаем, что

$$U_x(z) \;=\; z - 2\frac{(z,x)}{\|x\|_2}x \;=\; -\frac{(z,x)}{\|x\|_2}x + y\,.$$

В частности, если вектор x имеет единичную норму $\|x\|_2 = 1$, то оператор U_x имеет вид $U_x = \mathrm{Id}_x - 2(\,,x)x$.

Понятно, что разложение $E = \langle x\rangle \oplus \langle x\rangle^\perp$ соответствует разложению пространства E в прямую сумму инвариантных собственных подпространств оператора U_x, отвечающих собственным значениям -1 и 1.

Замечание 0.28. *Оператор U_x унитарен, самосопряжён и, как следствие, $U_x^2 = \mathrm{Id}_E$.*

Доказательство. Действительно, из описания оператора U_x немедленно следует, что $(U_x(z), U_x(z')) = (z, z')$ для всех $z, z' \in E$. Поэтому он является унитарным. Кроме того,

$$(U_x(z), z') \;=\; (z, z') - 2\frac{(z,x)(x,z')}{\|x\|_2} \;=\; (z, z') - 2\frac{(z,x)\overline{(z',x)}}{\|x\|_2} \;=\; (z, U_x(z')) \quad (z, z' \in E)$$

(в вещественном случае комплексное сопряжение можно опустить) и, следовательно, оператор U_x самосопряжён. □

Оператор U_x является фредгольмовым (тождественный плюс оператор единичного ранга). Поэтому в случае гильбертова пространства E отсюда сразу следует, что $\mathrm{Spec}\, U_x = \{\pm 1\}$.

В интересующей нас конечномерной ситуации $E = \mathbb{C}^n$, $(x, y) = y^*x$, $x, y \in \mathbb{C}^n$ (для большей общности мы будем рассматривать комплексную ситуацию) оператор U_x представляет собой матрицу $U_x = E - \frac{2}{\|x\|_2}xx^*$, которая называется *матрицей отражения* относительно гиперплоскости $\langle x\rangle^\perp$.

Идею метода отражений (метода Хаусхолдера) несложно вывести из следующего наблюдения.

Замечание 0.29. *Для любых векторов $e, y \in \mathbb{C}^n$, $\|e\|_2 = 1$, можно подобрать вектор $x \in \mathbb{C}^n$, $\|x\|_2 = 1$, и число $\alpha \in \mathbb{C}$, $|\alpha| = 1$, такие, что $U_x(y) = \alpha\|y\|_2 e$.*

Доказательство. Случай $y = 0$ не представляет интерес (можно взять любой x, $\|x\|_2 = 1$). Пусть $y \neq 0$ и $(y, e) \in \mathbb{R}$. Если $y = \pm\|y\|_2 e$, тогда при $y = \|y\|_2 e$ можно взять любой $x \in \langle e\rangle^\perp$, $\|x\|_2 = 1$ ($U_x(y) = y = \|y\|_2 e$), а при $y = -\|y\|_2 e$ достаточно взять $x = e$ ($U_e(y) = -y = \|y\|_2 e$). В остальных случаях условие $(y, e) \in \mathbb{R}$ гарантирует линейную независимость векторов y и e. Понятно также, что в данном случае

$$\langle y, e\rangle \;=\; \langle y - \|y\|_2 e\rangle \oplus \langle y + \|y\|_2 e\rangle\,,$$

причём $\langle y + |y\|_2 e\rangle = \langle y, e\rangle \cap \langle y - \|y\|_2 e\rangle^\perp$, поскольку $(y - \|y\|_2 e, y + \|y\|_2 e) = 0$ (последнее равенство имеет место тогда и только тогда, когда $(y, e) \in \mathbb{R}$). Так как

$$y \;=\; 1/2(y - \|y\|_2 e) + 1/2(y + \|y\|_2 e)\,, \quad \|y\|_2 e \;=\; -1/2(y - \|y\|_2 e) + 1/2(y + \|y\|_2 e)\,,$$

Нам остаётся лишь положить $x = \frac{y - \|y\|_2 e}{\|y - \|y\|_2 e\|_2}$, $U_x = U_{y - \|y\|_2 e}$. Найденный таким образом вектор x определён однозначно с точностью до умножения на $\beta \in \mathbb{C}$, $|\beta| = 1$.

В случае $(y, e) \in \mathbb{C} \setminus \mathbb{R}$ достаточно заменить e на вектор αe, полагая $\alpha = \frac{(y,e)}{|(y,e)|}$, и положить $x = \frac{y - \|y\|_2 \alpha e}{\|y - \|y\|_2 \alpha e\|_2}$. Число α определено однозначно с точностью до умножения на ± 1. □

Метод отражений решения линейной системы $Ax = b$, $A \in Gl_n(\mathbb{C})$, и связанный с ним алгоритм построения QR–разложения строятся по схеме аналогичной методу вращений, описанному нами ранее. Как и прежде, строится последовательсть эквивалентных систем $A^{(k)}x = b^{(k)}$, $k = 0, 1, \ldots, n-1$, в которой $A^{(0)} = A$, $b^{(0)} = b$ и при каждом k

$$A^{(k)} = \begin{pmatrix} \|a_1^{(0)}\|_2\alpha_1 & c_{12} & \ldots & c_{1k} & c_{1k+1} & \ldots & c_{1n} \\ 0 & \|a_1^{(1)}\|_2\alpha_2 & \ldots & a_{2k} & c_{2k+1} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \|a_1^{(k-1)}\|_2\alpha_k & c_{kk+1} & \ldots & c_{kn} \\ 0 & 0 & \ldots & 0 & a_{k+1k+1}^{(k)} & \ldots & a_{kn}^{(k)} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & a_{nk+1}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix},$$

$b^{(k)} = (y_1, \ldots, y_k, b_{k+1}^{(k)}, \ldots, b_n^{(k)})^t$, где, как и ранее, $a_1^{(i)} = (a_{i+1i+1}^{(i)}, \ldots, a_{ni+1}^{(i)})^t$ — первый столбец ведущей подматрицы, полученной после выполнения i-ого шага, $\alpha_i \in \mathbb{C}$, $|\alpha_i| = 1$, и при $k = n-1$

$$A^{(n-1)} = \begin{pmatrix} \|a_1^{(0)}\|_2\alpha_1 & c_{12} & \ldots & c_{1n-1} & c_{1n} \\ 0 & \|a_1^{(1)}\|_2\alpha_2 & \ldots & a_{2n-1} & c_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \|a_1^{(n-2)}\|_2\alpha_{n-1} & c_{n-1n} \\ 0 & 0 & \ldots & 0 & a_{nn}^{(n-1)} \end{pmatrix}$$

и $b^{(n-1)} = (y_1, \ldots, y_{n-1}, b_n^{(n-1)})^t$. На k-ом шаге при переходе $A^{(k-1)} \to A^{(k)}$, $b^{(k-1)} \to b^{(k)}$, выполняются перевычисления компонент ведущей подматрицы $A'^{(k-1)} = (a_{ij}^{(k-1)})_{i,j=k}^n$ и компонент вектора правой части с индексами $k, k+1, \ldots, n$, в следующей последовательности:

1. в соответствии с замечанием выберем вектор $z'^{(k)} = (z_k^{(k)}, \ldots, z_n^{(k)})^t \in \mathbb{C}^{n-k+1}$, $\|z'^{(k)}\|_2 = 1$, и число $\alpha_k \in \mathbb{C}$, $|\alpha_k| = 1$, для которых

$$U'_{z'^{(k)}}(a_1^{(k-1)}) = a_1^{(k-1)} - 2(a_1^{(k-1)}, z^{(k)})z^{(k)} = (\|a_1^{(k-1)}\|_2\alpha_k, 0, \ldots, 0)^t,$$

где $U'_{z'^{(k)}}$ — матрица отражения размера $(n-k+1) \times (n-k+1)$, т.е. $\alpha_k = \frac{a_{kk}^{(k-1)}}{|a_{kk}^{(k-1)}|}$ и в случае $a_1^{(k-1)} = (a_{kk}^{(k-1)}, 0, \ldots, 0)^t$, $\|a_1^{(k-1)}\|_2 = |a_{kk}^{(k-1)}|$ мы полагаем $U'_{z^{(k)}} = E' = E_{n-k+1}$ (если условие невырожденности A не учитывается, то вполне может оказаться, что $a_1^{(k-1)} = 0$, но и в этой ситуации следует положить $U'_{z^{(k)}} = E'$), в противном случае следует взять $z'^{(k)} = \frac{a_1^{(k-1)} - \alpha_k\|a_1^{(k-1)}\|_2 e_1^{(k)}}{\|a_1^{(k-1)} - \alpha_k\|a_1^{(k-1)}\|_2 e_1^{(k)}}$, где $e_1^{(k)} = (1, 0, \ldots, 0)^t \in \mathbb{C}^{n-k+1}$.

2. для каждого $t = 2, \ldots, n-k$ столбец $a_t^{(k-1)}$ ведущей подматрицы $A'^{(k-1)}$ заменяется столбцом

$$U'_{z'^{(k)}}a_t^{(k-1)} = (c_{kt}, a_{k+1t}^{(k)}, \cdots, a_{nt}^{(k)})^t,$$

её первый столбец $a_1^{(k-1)}$ заменяется столбцом $(\|a_1^{(k-1)}\|_2\alpha_k, 0, \ldots, 0)^t$, что соответствует переходу

$$U'_{z'^{(k)}}A'^{(k-1)} \;=\; \begin{pmatrix} \|a_1^{(k-1)}\|_2\alpha_k & c_{kk+1} & \dots & c_{kn} \\ 0 & a^{(k)}_{k+1k+1} & \dots & a^{(k)}_{k+1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a^{(k)}_{nk+1} & \dots & a^{(k)}_{nn} \end{pmatrix},$$

компонентны правой части перевычисляются по схеме

$$U'_{z'^{(k)}}(b_k^{(k-1)},\cdots,b_n^{(k-1)})^t \;=\; (y_k, b_{k+1}^{(k)},\dots,b_n^{(k)})^t\,.$$

Другими словами, суть k-го шага состоит в вычислении

$$A^{(k)} \;=\; U_{z(k)}A^{(k-1)}\,,\quad b^{(k)} \;=\; U_{z(k)}b^{(k-1)}\,,$$

где $z(k) = (\underbrace{0,\dots,0}_{k-1}, z'^{(k)})^t$. Решение треугольной системы $A^{(n-1)}x = b^{(n-1)}$ выполняется обратным ходом метода Гаусса. При этом, как нетрудно заметить, $A^{(n-1)} = UA$ и $b^{(n-1)} = Ub$ для $U = U_{z^{(n-1)}}\cdots U_{z^{(1)}}$.

Аналогично методу отражений описанный здесь процесс триангуляции применим ко всякой квадратной системе (вопрос разрешимости треугольной системы (проверка отсутствия нулевых элементов на диагонали) может быть проведён перед началом выполнения обратного хода.

Превращение в последовательно-параллельный процесс этого процесс базируется в большей степени не на параллельных реализациях матричных умножений, а на специфике умножения матрицы отражения на вектор, распараллеливание которого сводится по сути к параллельному вычислению скалярного произведения векторов.

Заметим, что на вычисление $U_x y = y - 2(y,x)x$, $\|x\|_2 = 1$, $x, y \in \mathbb{R}^n$, требуется $4n$ операций ($1 + n + n$ умножений и $n - 1 + n$ сложений и вычитаний).

Вычислительная работа на k-ом шаге алгоритма (триангуляции отражениями) предполагает:

1. вычислить $\alpha_k = \frac{a_{kk}^{(k-1)}}{|a_{kk}^{(k-1)}|}$, $s_k = |a_{k+1k}^{(k-1)}|^2 + \ldots + |a_{nk}^{(k-1)}|^2$, $\|a_1^{(k-1)}\|_2 = \sqrt{|a_{kk}^{(k-1)}|^2 + s_k}$, $t_k = \alpha_k\|a_1^{(k-1)}\|_2$, $\tilde{z}^{(k)} = (a_{kk}^{(k-1)} - t_k, a_{k+1k}^{(k-1)},\dots,a_{nk}^{(k-1)})^t$ и $g_k = \frac{2}{\sqrt{|a_{kk}^{(k-1)}-t_k|^2+s_k}} = \frac{2}{\|\tilde{z}^{(k)}\|_2}$ (в совокупности это составит $O(n^2)$ операций);

2. вычислить вектора $U'_{z'^{(k)}}a_i^{(k-1)} = a_i^{(k-1)} - g_k(a_i^{(k-1)},\tilde{z}^{(k)})\tilde{z}^{(k)}$, $i = 2,\dots,n-k+1$, и вектор $U'_{z'^{(k)}}(b_k^{(k-1)},\dots,b_n^{(k-1)})^t$ (это требует $4(n-k+1)^2$ операций на шаг и, следовательно, всего порядка $4/3n^3 + O(n^2)$ операций).

Таким образом, данный алгоритм обходится в $4/3n^3 + O(n^2)$ операций, к которым добавляется ещё $O(n^2)$ операций на решение треугольной системы, $n \to \infty$.

Изложенный нами процесс триангуляции применим к нахожению QR–разложения любой прямоугольной комплексной матрицы A размера $n \times m$. Для этого достаточно повторить $\min\{n,m\} - 1$ шаг описанного алгоритма и получить верхнюю треугольную матрицу R' размера $n \times m$, $R' = UA$, хранимую на месте верхнего треугольника матрицы A, и хранимую отдельно матрицу $U^* = U_{z^{(1)}}\cdots U_{z^{(\min\{n,m\}-1)}}$. Остаётся взять

$$D \;=\; \mathrm{diag}(d_1,\ldots,d_{\min\{n,m\}},\;\underbrace{1,\ldots,1}_{n-\min\{n,m\}}\;),$$

где $d_i = 1$ при $r'_{ii} = 0$ и $d_i = \frac{\overline{r'_{ii}}}{|r'_{ii}|}$ иначе, и положить $R = DR'$, $Q = U^*\overline{D}$, обеспечив тем самым выполнение равенства $A = QR$ для $Q \in U_n(\mathbb{C})$ и верхней треугольной матрицы R с неотрицательной действительной диагональю. Количество операций, необходимых для построения матриц Q и R в случае $n = m$ оценивается как $2n^3 + O(n^2) + 4/3n^3 + O(n^2) = 10/3n^3 + O(n^2)$, $n \to \infty$ (на построение $U^* = U_{z^{(n-1)}} \cdots U_{z^{(1)}}$ требуется $\sum\limits_{k=2}^{n-2} 4n(n-k+1) = 2n^3 + O(n^2)$ операций, затем вычисляется $U^*\overline{D}$, что добавляет порядка $O(n^2)$ операций).

Отказавшись от требования однозначности QR–разложения (ограничиться однозначностью с точностью до умножения Q–составляющей на диагональную матрицу с единичными по модулю диагональными коэффициентами), можно обойтись построением матриц U и R'.

Отметим также, что информация о матрице U может быть полностью восстановлена по внедиагональным компонентам первых столбцов ведущих подматриц на этапах выполнения метода отражений. Поэтому в принципе мы можем хранить только эти компоненты на соответствующих позициях нижнего треугольника матрицы A.

В качестве следствия из сказанного можно получить следующий вывод.

Замечание 0.30. *Ортогональная группа $O_n(\mathbb{R})$ порождается матрицами отражений U_x, $x \in \mathbb{R}^n$, $\|x\|_2 = 1$.*

Доказательство. По предыдущему для любой матрицы $A \in O_n(\mathbb{R})$ можно подобрать произведение матриц отражений U, что $UA = R'$ — верхняя треугольная матрица. Так как $R' \in O_n(\mathbb{R})$, $R'^t = R'^{-1}$, из единственности LU–разложения следует, что $R' = \mathrm{diag}(r'_{11},\ldots,r'_{nn})$, $R'^2 = R'^t R' = E$, ${r'_{ii}}^2 = 1$, $r'_{ii} \in \{\pm 1\}$, $i = 1,\ldots,n$. Остаётся заметить, что любая диагональная матрица с коэффициентами ± 1 на диагонали является произведением диагональных матриц отражения $D_i = U_{e_i} = \mathrm{diag}(\underbrace{1,\ldots,1}_{i-1},-1,\underbrace{1,\ldots,1}_{n-i})$, где $e_i = (\underbrace{0,\ldots,0}_{i-1},1,\underbrace{0,\ldots,0}_{n-i})^t$, $i = 1,\ldots,n$. □

Заметим, что методы вращений и отражений требуют почти в два раза больше операций, чем модифицированный процесс Грамма — Шмидта (MGS). Тем не менее, полученные в результате их реализации матрицы Q значительно ближе к ортогональным нежели в MGS (в последнем величина $\|Q^*Q - E\|_2$ в $k_2(A)$ хуже, чем в первых двух алгоритмах).

QR–разложение с выбором и задача наименьших квадратов

Построение QR–разложения для матрицы A размера $m \times n$ неполного столбцового ранга $\mathrm{rk}\,A = r < n$ не обязательно приводит к построению ортогональной или унитарной матрицы Q, линейная оболочка первых r столбцов которой совпадает с оболочкой столбцов матрицы A (т.е. образом A как линейного преобразования). Тем не менее, имеется естественный выход из создавшегося положения, позволяющий получить разложение

$AP = QR$, где P — матрица перестановка и R — верхняя треугольная матрица размера $m \times n$

$$R \;=\; \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$

с блоками R_{11} и R_{12} размера $r \times r$ и $r \times (n-r)$, первый из которых имеет ранг $\operatorname{rk} R_{11} = r$. Нетрудно заметить, что в данном случае линейная оболочка первых r столбцов матрицы Q совпадает с линейной оболочкой первых r столбцов матрицы QR и, более того, с линейной оболочкой всех столбцов этой матрицы, т.е. равна образу преобразования A. Процесс строится следующим образом: пусть нами уже построены ортогональные (унитарные) матрицы $Q_1, \ldots, Q_{k-1}$ и матрицы перестановки $P_1, \ldots, P_{k-1}$, а вместе с ними и матрица

$$Q_{k-1} \cdots Q_1 A P_1 \cdots P_{k-1} \;=\; \begin{pmatrix} R_{11}^{(k-1)} & R_{12}^{(k-1)} \\ 0 & R_{22}^{(k-1)} \end{pmatrix}$$

с выделенными блоками $R_{11}^{(k-1)}$, $R_{12}^{(k-1)}$ и $R_{22}^{(k-1)}$ размеров $(k-1) \times (k-1)$, $(k-1) \times (n-k+1)$ и $(m-k+1) \times (m-k+1)$ соответственно, причём блок $R_{11}^{(k-1)}$ представляет собой невырожденную верхнюю треугольную матрицу. Среди столбцов матрицы $R_{22}^{(k-1)}$ выделим столбец с максимальной нормой имеющий в основной матрице номер i_k, $k \le i_k \le n$. Если норма указанного столбца равна нулю, то процесс завершён и $\operatorname{rk} A = k-1$. В противном случае возьмём $P_k = P_{ki_k}$ и определим ортогональную (унитарную) матрицу Q'_k размера $(m-k+1) \times (m-k+1)$, для которой

$$Q'_k (r_{ki_k}^{(k-1)}, \ldots, r_{mi_k}^{(k-1)})^t \;=\; \alpha_k \sqrt{|r_{ki_k}^{(k-1)}|^2 + \ldots + |r_{mi_k}^{(k-1)}|^2} (1, 0, \ldots, 0)^t \,,$$

и положим $Q_k = \operatorname{diag}(E_{k-1}, Q'_k)$. Построение этой матрицы осуществляется по схеме метода вращений (или метода отражений). Затем перейдём к матрице

$$Q_k \begin{pmatrix} R_{11}^{(k-1)} & R_{12}^{(k-1)} \\ 0 & R_{22}^{(k-1)} \end{pmatrix} P_k \;=\; \begin{pmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{pmatrix}$$

в которой невырожденный верхний треугольный блок $R_{11}^{(k)}$ представляет собой блок $R_{11}^{(k-1)}$, дополненный нулевыми компонентами последней строки, с присоединённым к нему k-ым столбцом, первые $k-1$ компонент которого суть компоненты $i_k - k + 1$-го столбца матрицы $R_{12}^{(k-1)}$, а последняя k-ая компонента совпадает с евклидовой нормой $i_k - k + 1$-го столбца матрицы $R_{22}^{(k-1)}$, умноженной на число α_k (последнее относится к методу отражений). В итоге после выполнения r-го шага мы получим R–фактор требуемого вида

$$\begin{pmatrix} R_{11}^{(r)} & R_{12}^{(r)} \\ 0 & R_{22}^{(r)} \end{pmatrix} \;=\; \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} .$$

Далее, используя метод вращений или отражений, мы можем подобрать r ортогональных унитарных матриц $U_1, \ldots, U_r$ размера $n \times n$, для которых

$$U_r \cdots U_1 \begin{pmatrix} R_{11}^t \\ R_{12}^t \end{pmatrix} \;=\; \begin{pmatrix} T_{11}^t \\ 0 \end{pmatrix} ,$$

где T_{11}^t — невырожденная верхняя треугольная матрица размера $r \times r$. Таким образом, для подходящих ортогональных (унитарных) матриц Q и $U = PU_1^t \cdots U_r^t$

$$Q^* A U \;=\; \begin{pmatrix} T_{11} & 0 \\ 0 & 0 \end{pmatrix}.$$

Разложения такого вида называют *полными ортогональными разложениями.*

Наличие полного ортогонального разложения позволяет находить решение задачи наименьших квадратов $\|Ax - b\|_2 \to \min$ с наименьшей евклидовой нормой (среди всех её решений x) следующим образом. Для начала заметим, что

$$\|Ax - b\|_2 \;=\; \|Q^* A x - Q^* b\|_2 \;=\; \|Q^* A U (U^* x) - Q^* b\|_2 .$$

Поэтому, полагая $U^* x = y = (y_1, y_2)^t$, $Q^* b = (a_1, a_2)^t$, где y_1 и y_2 — вектора из первых r и оставшихся $n - r$ компонент y, a_1 и a_2 — вектора из первых r и оставшихся $m - r$ компонент $Q^* b$, мы получаем, что

$$\|Ax - b\|_2 \;=\; \|T_{11} y_1 - a_1\|_2 + \|a_2\|_2$$

и, следовательно, решением этой задачи с наименьшей евклидовой нормой служит вектор $x_{LS} = U\hat{y}$, $\hat{y} = (\hat{y}_1, \hat{y}_2)^t$ с $\hat{y_1}^t = T_{11}^{-1} a_1$, $\hat{y}_2 = 0$.

Естественно приведённый нами алгоритм решения задачи наименьших квадратов на основе QR–разложения с выбором столбца и связаного с ним построения полного ортогонального разложения является лишь одним из возможных подходов решения данной задачи, на детальном рассмотрении которой мы не будем останавливаться.

Приведение к верхней форме Хессенберга ортогональным (унитарным) подобием и двухдиагонализация вращениями и отражениями

К числу весьма важных приложений методов вращений и отражений относятся алгоритмы приведения к форме Хессенберга (трёхдиагонализации в самосопряжённом случае) методами ортогонального (унитарного) подобия, а также родственные им алгоритмы двухдиагонализации.

Идея использования ортогонального (унитарного) подобия для приведения квадратной $n \times n$ матрицы A к верхней хессенберговой форме (т.е. к форме матрицы, отличной от верхней треугольной лишь наличием одной побочной диагонали ниже главной) состоит в следующем. Строится последовательность матриц $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-3)}$, в которой каждая матрица $A^{(k)}$ имеет вид

$$A^{(k)} \;=\; \begin{pmatrix}
c_{11} & c_{12} & c_{13} & \ldots & c_{1k} & c_{1k+1} & a^{(k)}_{1k+2} & \ldots & a^{(k)}_{1n} \\
c_{21} & c_{22} & c_{23} & \ldots & c_{2k} & c_{2k+1} & a^{(k)}_{2k+2} & \ldots & a^{(k)}_{2n} \\
0 & c_{23} & c_{33} & \ldots & c_{3k} & c_{3k+1} & a^{(k)}_{3k+2} & \ldots & a^{(k)}_{3n} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & c_{kk} & c_{kk+1} & a^{(k)}_{kk+2} & \ldots & a^{(k)}_{kn} \\
0 & 0 & 0 & \ldots & c_{k+1k} & c_{k+1k+1} & a^{(k)}_{k+1k+2} & \ldots & a^{(k)}_{k+1n} \\
0 & 0 & 0 & \ldots & 0 & a^{(k)}_{k+2k+1} & a^{(k)}_{k+2k+2} & \ldots & a^{(k)}_{k+2n} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & 0 & a^{(k)}_{nk+1} & a^{(k)}_{nk+2} & \ldots & a^{(k)}_{nn}
\end{pmatrix},$$

а переход $A^{(k-1)} \to A^{(k)}$ осуществляется подбором ортогональной (унитарной) матрицы U'_k размера $(n-k)\times(n-k)$, для которой

$$U'_k(a^{(k-1)}_{k+1k},\ldots,a^{(k-1)}_{nk})^t \;=\; \alpha_k\sqrt{|a^{(k-1)}_{k+1k}|^2+\ldots+|a^{(k-1)}_{nk}|^2}(1,0,\ldots,0)^t\,,$$

и выполнением подобия $U_kA^{(k-1)}U_k^* = A^{(k)}$ для $U_k = \mathrm{diag}(E_k, U'_k)$. При этом U'_k подбирается по схеме, используемой в методе вращений (или отражений). Финальная матрица $A^{(n-3)}$ будет искомой верхней хессенберговой матрицей. В случае самосопряжённой (симметрической) матрицы A подобные ей матрицы $A^{(k)}$ будут также самосопряжёнными (симметрическими) и потому результирующая матрица $A^{(n-3)}$ будет трёхдиагональной.

Стоит отметить, что QR–разложение невырожденной хессенберговой матрицы имеет в качестве Q–составляющей хессенбергову матрицу (единственность QR–разложения и процесс ортогонализации Грама — Шмидта).

Родственный описанному процессу процесс двухдиагонализации сводится к построению цепочки матриц $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-1)}$, где матрица $A^{(k)} = Q_kA^{(k-1)}U_k$ имеет вид

$$A^{(k)} \;=\; \begin{pmatrix} c_{11} & c_{12} & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & c_{22} & c_{23} & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & c_{33} & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & c_{k-1k} & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & c_{kk} & c_{kk+1} & a^{(k)}_{kk+2} & \ldots & a^{(k)}_{kn} \\ 0 & 0 & 0 & \ldots & 0 & a^{(k)}_{k+1k+1} & a^{(k)}_{k+1k+2} & \ldots & a^{(k)}_{k+1n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & a^{(k)}_{nk+1} & a^{(k)}_{nk+2} & \ldots & a^{(k)}_{nn} \end{pmatrix},$$

где ортогональные (унитарные) матрицы $Q_k = \mathrm{diag}(E_{k-1}, U'_k)$ и $U_k = \mathrm{diag}(E_k, U'_k)$ подбираются по схеме метода вращений (отражений) с целью аннулирования всех поддиагональных компонент k-го и компонент с номерами вне двух диагоналей k-ой строки матрицы $A^{(k-1)}$.

Лекция 5. Итерационные методы решения линейных систем. Основные стационарные итерационные процессы.

В отличие от точных методов решения линейной системы $Ax = b$, гарантирующей получение точного решения (в отсутствие ошибок округления) за конечное число шагов, итерационные методы решения линейных систем предполагают построение приближения к точному решению с заданной точностью за конечное, предсказуемое число шагов. Основная идея рассматриваемых методов состоит в построении последовательности векторов $\{x^{(k)}\}_{k\geq 0}$, которые сходятся к точному решению системы $Ax = b$ и строятся на основе начального приближения $x^{(0)}$ по правилу: $x^{(k)} = C^{(k)}x^{(k-1)} + z^{(k)}$ для некоторых матрицы $C^{(k)}$ и вектора $z^{(k)}$, $k \geq 1$. Реализация подобных итерационных процессов, область их применимости и скорость сходимости к точному решению $x = A^{-1}b$, зависящая в том числе и от выбранной для её оценки матричной нормы, может быть различной. Наше обсуждение мы начнём с обсуждения стационарных процессов с постоянными матрицами $R = C^{(k)}$ и векторами сдвига $c = b^{(k)}$, $k \geq 1$.

Расщеплением невырожденной матрицы A называется любое её представление в виде $A = M - K$ с невырожденной матрицей M. Расщепление $A = M - K$ называется M-*регулярным*, если матрица $M + K$ положительно определена, $M + K > 0$. Для всякого расщепления $A = M - K$ мы можем переписать систему $Ax = Mx - Kx = b$ как $Mx = Kx + b$ и $x = Rx + c$, где $R = M^{-1}K$ и $c = M^{-1}b$. Поэтому, выбрав начальное приближение $x^{(0)}$, мы можем построить итерационный процесс $x^{(k)} = Rx^{(k-1)} + c$, $k \geq 1$, предполагая сходимость последнего к неподвижной точке отображения $T : z \longmapsto Rz + c$.

Замечание 0.31. *Пусть* $\|\ \|$ *— векторная норма, для которой значение подчинённой ей матричной нормы* $\|R\| = \max\limits_{x,\ \|x\|=1} \|Rx\|$ *матрицы* R *меньше единицы,* $\|R\| < 1$. *Тогда для любого начального приближения* $x^{(0)}$ *последовательсть* $\{x^{(k)}\}_{k=0}^{\infty}$, *где* $x^{(k)} = Rx^{(k-1)} + c$, $k \geq 1$, *сходится к* $x = Rx + c$, $\lim\limits_{k\to\infty} x^{(k)} = x$.

Доказательство. Достаточно заметить, что в данном случае отображение $T : z \longmapsto Rz + c$ является сжимающим с коэффициентом сжатия $\|R\| < 1$. При этом

$$\|x^{(k)} - x\| \ = \ \|(Rx^{(k-1)} + c) - (Rx + c)\| \leq \|R\|\|x^{(k-1)} - x\| \leq \|R\|^k\|x^{(0)} - x\| \quad (k \geq 0)\,.$$

□

Напомним, что в курсе функционального анализа на основе формулы спектрального радиуса (формулы Коши — Адамара) ограниченного оператора A на банаховом пространстве нами было доказано существование для любого $\varepsilon > 0$ нормы пространства $\|\ \|_\varepsilon$, которая эквивалентна его исходной норме $\|\ \|$ и удовлетворяет условию $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$, где $\|A\|_\varepsilon$ — норма оператора A на рассматриваемом пространстве с нормой $\|\ \|_\varepsilon$ и $\rho(A)$ — его спектральный радиус. Применительно к интересующей нас конечномерной ситуации можно доказать следующий вариант этого утверждения.

Замечание 0.32. *Для любых* $A \in M_n(\mathbb{C})$ *и* $\varepsilon > 0$ *существует векторная норма* $\|\ \|_\varepsilon$, *для которой* $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$.

Доказательство. Найдётся $C \in Gl_n(\mathbb{C})$, для которой

$$C^{-1}AC \ = \ J \ = \ \mathrm{diag}(J_{\lambda_1,n_1}, \ldots, J_{\lambda_s,n_s})$$

— жорданова нормальная форма матрицы A. Положим $D = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$ и

$(CD)^{-1}ACD = D^{-1}JD = \mathrm{diag}(J_{\lambda_1,n_1}(\varepsilon),\ldots,J_{\lambda_s,n_s}(\varepsilon))$, где

$$J_{\lambda,m}(\varepsilon) \;=\; \begin{pmatrix} \lambda & \varepsilon & 0 & \ldots & 0 & 0 \\ 0 & \lambda & \varepsilon & \ldots & 0 & 0 \\ 0 & 0 & \lambda & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda & \varepsilon \\ 0 & 0 & 0 & \ldots & 0 & \lambda \end{pmatrix} \;=\; \lambda E_m + \varepsilon E'_m$$

(E'_m — матрица с единичной побочной диагональю выше главной). Введём векторную норму $\|x\|_\varepsilon = \|(CD)^{-1}x\|_\infty$. Тогда

$$\|A\|_\varepsilon \;=\; \max_{x,\ \|x\|_\varepsilon=1} \|Ax\|_\varepsilon \;=\; \max_{y,\ \|y\|_\infty=1} \|(CD)^{-1}ACDy\|_\infty \;=$$

$$\|(CD)^{-1}ACD\|_\infty \le \varepsilon + \max_{i=1,\ldots,s} |\lambda_i| \le \varepsilon + \rho(A)\,.$$

□

Следствие 0.33. *Итерационный процесс $x^{(k)} = Rx^{(k-1)} + c$, $k \ge 1$, сходится к точному решению $x = Rx + c$ при любом начальном приближении $x^{(0)}$ в том и только в том случае, если $\rho(R) < 1$.*

Доказательство. Действительно, если $\rho(R) < 1$, то найдётся $\varepsilon > 0$ и связанная с ним векторная норма $\|\ \|_\varepsilon$, для которых $\|R\|_\varepsilon < 1$ и, следовательно, по первому из доказанных нами замечаний отображение $T : z \longmapsto Rz + c$ пространства $\mathbb{C}^n$ с нормой $\|\ \|_\varepsilon$ является сжимающим и любой из рассматриваемых процессов сходится к неподвижной точке этого отображения.

С другой стороны если $\rho(R) \ge 1$ (в предположении обратимости $E - R$ и существования $x = Rx + c$), тогда найдётся ненулевой собственный вектор x_λ, $Rx_\lambda = \lambda x_\lambda$, для собственного значения λ, $|\lambda| = \rho(R) \ge 1$, и значит, для начального приближения $x^{(0)} = x + x_\lambda$ процесс $x^{(k)} = x + \lambda^k x_\lambda$, $k \ge 0$, не сходится к $x = Rx + c$, $\|x^{(k)} - x\| = |\lambda|^k \|x_\lambda\| \not\to 0$.

Вместе с тем это не означает, что такая сходимость не имеет место для какого-то другого начального приближения. □

Отметим, что на случай банаховых пространств данный результат переносится с наложением требования компактности оператора R для обратной импликации.

Стоит также сказать, что данный вывод далеко не всегда обеспечивает нас хорошей скоростью сходимости процесса $x^{(k)} = Rx^{(k-1)} + c$, $k \ge 1$, с $\rho(R) < 1$ в доступных нам для оценивания этой скорости матричных нормах (это необходимо для оценки числа необходимых итераций для получения приближения заданной точности). Действительно, пусть речь идёт об оценке сходимости в норме $\|\ \|_\infty$. По предыдущему при некотором $\varepsilon > 0$ мы располагаем нормой $\|\ \|_\varepsilon$, $\|x\|_\varepsilon = \|Bx\|_\infty$ для подходящей невырожденной матрицы B (см. выше), которая обеспечивает нас неравенством $\|R\|_\varepsilon < 1$. При этом $\|B^{-1}x\|_\varepsilon = \|x\|_\infty$, $\|x\|_\varepsilon \le \|B\|_\infty \|x\|_\infty$ и, следовательно,

$$(1/\|B\|_\infty)\|x\|_\varepsilon \le \|x\|_\infty \le \|B^{-1}\|_\varepsilon \|x\|_\varepsilon \;=\; \|B^{-1}\|_\infty \|x\|_\varepsilon\,.$$

Поэтому $\|A\|_\infty \le k(B)_\infty \|A\|_\varepsilon$ и доступная нам оценка скорости сходимости $x^{(k)}$ к $x =$

$Rx + c$ может быть записана как

$$\|x^{(k)} - x\|_\infty \le \|R^k\|_\infty \|x^{(0)} - x\|_\infty \le k_\infty(B)\|R\|_\varepsilon^k \|x^{(0)} - x\|_\infty \le k_\infty(B)(\rho(R)+\varepsilon)^k \|x^{(0)} - x\|_\infty ,$$

причём если следовать описанной выше схеме $B = (CD)^{-1}$, то $k_\infty(B) \le k_\infty(C)k_\infty(D) = k_\infty(C)\varepsilon^{1-n}$. Таким образом, порядок убывания величины $k_\infty(B)(\rho(R) + \varepsilon)^k$ с ростом k может быть медленным на протяжении весьма значительного интервала изменения k (заметим, что рассматриваемые оценки фактически не могут быть улучшены), а это не позволит получить качественное приближение за разумное время реализации данного процесса.

Уместно будет привести также следующее замечание.

Замечание 0.34. *Пусть положительно определённая матрица $A \in M_n(\mathbb{C})$ имеет M-регулярное расщепление $A = M - K$. Тогда матрица $M^{-1}K$ имеет вещественный спектр и $\rho(M^{-1}K) < 1$.*

Доказательство. Поскольку в данном случае $M \pm K > 0$, мы сразу получаем, что $2M > 0$. Для всякого ненулевого собственного вектора x_λ, $M^{-1}Kx_\lambda = \lambda x_\lambda$, мы имеем $((M \pm K)x_\lambda, x_\lambda) = ((1 \pm \lambda)Mx_\lambda, x_\lambda) = (1 \pm \lambda)(Mx_\lambda, x_\lambda) > 0$, $(Mx_\lambda, x_\lambda) > 0$ и значит, $1 \pm \lambda > 0$, $\lambda \in \mathbb{R}$, $|\lambda| < 1$. □

Для вещественных матриц сходный вывод потребует наложения дополнительных условий на симметричность A и M (положительная определённость вещественных симметрических матриц равносильна их положительной определённости как комплексных матриц).

Основной целью при выборе расщепления $A = M - K$ невырожденной матрицы A для организации стационарного процесса $x^{(k)} = Rx^{(k-1)} + c$, $k \ge 1$, где $R = M^{-1}K$ и $c = M^{-1}b$, является согласование двух почти противоположных друг другу требований: простота вычисления произведения R на вектора и малость величины $\rho(R)$.

Во всех приводимых ниже методах организации подобных процессов используются следующие обозначения: $A = D - \tilde{L} - \tilde{U}$, $L = D^{-1}\tilde{L}$ и $U = D^{-1}\tilde{U}$, где D — диагональ матрицы A, $\tilde{L}$ и $\tilde{U}$ — её нижний и верхний треугольники со знаком минус.

Метод простой итерации (или метод Якоби) состоит в использовании диагонального расщепления $A = D - D(L+U)$ в предположении обратимости диагонали D матрицы A. Данный итерационный процесс строится по схеме $x^{(k)} = R_J x^{(k-1)} + c_J$, где $R_J = L + U$ и $c_J = D^{-1}b$. В координатах векторов $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^t$ это записывается как

$$x_i^{(k)} \;=\; 1/a_{ii}\Big(-\sum_{j\ne i} a_{ij}x_j^{(k-1)} + b_i\Big) \quad (i = 1, \ldots, n)\,.$$

Метод Гаусса — Зейделя представляет собой модификацию метода Якоби, которая состоит в выборе некоторого порядка вычисления координат вектора $x^{(k)}$ (порядок выбирается и фиксируется изначально) с заменой в формуле для вычисления $x_i^{(k)}$ компонент $x_j^{(k)}$ на компоненты вычисленные на предыдущих этапах реализации k-го шага. Например, если компоненты вычисляются в естественном порядке от 1 до n, тогда этот процесс записывается как

$$x_i^{(k)} \;=\; 1/a_{ii}\Big(-\sum_{j<i} a_{ij}x_j^{(k)} - \sum_{j>i} a_{ij}x_j^{(k-1)} + b_i\Big) \quad (i = 1, \ldots, n)$$

или в терминах треугольного расщепления $(D-\tilde{L})x^{(k)} = \tilde{U}x^{(k-1)}+b$, $x^{(k)} = R_{GS}x^{(k-1)} + c_{GS}$ для $R_{GS} = (D-\tilde{L})\tilde{U} = (E-L)^{-1}U$ и $c_{GS} = (D-\tilde{L})^{-1}b$. Естественно, что данный процесс может быть реализован $n!$ способами в соответствии с выбранным порядком перевычисления координат векторов-приближений.

Метод последовательной верхней релаксации с параметром ω ($SOR(\omega)$, где, как выяснится позднее, $\omega \in (0,2)$) представляет собой уравновешенную посредством параметра ω версию метода Гаусса — Зейделя. Точнее выбрав и зафиксировав один из возможных $n!$ порядков перевычисления компонент вектора $x^{(k)}$, мы будем вычислять очередную его компоненту $x_i^{(k)}$ как $(1-\omega)x_i^{(k-1)} + \omega x_i^{(k)}$, где в последнем выражении $x_i^{(k)}$ вычислена в соответствии с процедурой метода Гаусса — Зейделя с выбранным порядком его реализации, т.е. применительно к естественному порядку перевычисления от 1 до n это выглядит следующим образом

$$x_i^{(k)} \;=\; (1-\omega)x_i^{(k-1)} + (\omega/a_{ii})\Big(-\sum_{j<i} a_{ij}x_j^{(k)} - \sum_{j>i} a_{ij}x_j^{(k-1)} + b_i\Big) \quad (i=1,\ldots,n)\,.$$

Последнюю формулу можно переписать следующим образом

$$a_{ii}x_i^{(k)} + \omega\sum_{j<i} a_{ij}x_j^{(k)} \;=\; (1-\omega)a_{ii}x_i^{(k-1)} - \omega\sum_{j>i} a_{ij}x_j^{(k-1)} + \omega b_i \quad (i=1,\ldots,n)$$

или, что равносильно как,

$$(D-\omega\tilde{L})x^{(k)} \;=\; ((1-\omega)D + \omega\tilde{U})x^{(k-1)} + \omega b\,.$$

Другими словами, описанный процесс соответствует стационарному процессу $x^{(k)} = R_{SOR(\omega)}x^{(k-1)} + c_{SOR(\omega)}$, $k \geq 1$, где

$$R_{SOR(\omega)} \;=\; (D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{U}) \;=\; (E-\omega L)^{-1}((1-\omega)E+\omega U)\,,$$
$$c_{SOR(\omega)} \;=\; \omega(D-\omega\tilde{L})^{-1}b\,,$$

что соответствует расщеплению

$$A \;=\; 1/\omega(D-\omega\tilde{L}) - 1/\omega((1-\omega)D+\omega\tilde{U})\,.$$

Согласно своего описания данный процесс имеет $n!$ реализаций в зависимости от выбранного порядка перевычисления компонент вектора-приближения, вычисляемого на очередном шаге. При $\omega = 1$ метод $SOR(\omega) = SOR(1)$ превращается в описанный ранее метод Гаусса — Зейделя. В случаях $\omega < 1$ и $\omega > 1$ метод $SOR(\omega)$ принято называть нижней и верхней последовательной релаксацией соответственно.

Метод симметрической последовательной верхней релаксации с параметром ω (или $SSOR(\omega)$) представляет собой усложнённую версию метода $SOR(\omega)$, которая строится по следующей схеме: фиксируем порядок перевычисления (один из возможных) перед началом выполнения и разбиваем k-ый шаг на два этапа:

1. вычисляем вектор $x^{(k/2)}$ на основе вектора $x^{(k-1)}$ по методу $SOR(\omega)$ для выбранного порядка вычисления компонент;

2. вычисляем вектор $x^{(k)}$ на основе вектора $x^{(k/2)}$ по методу $SOR(\omega)$ с обратным выбранному порядком перевычисления компонент.

В частности, при выборе естественного порядка перевычисления от 1 до n указанные этапы сводятся к двум действиям:

$$x^{(k/2)} = (D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{U})x^{(k-1)}+\omega(D-\omega\tilde{L})^{-1}b\,,$$
$$x^{(k)} = (D-\omega\tilde{U})^{-1}((1-\omega)D+\omega\tilde{L})x^{(k/2)}+\omega(D-\omega\tilde{U})^{-1}b\,,$$

т.е. $x^{(k)} = R_{SSOR(\omega)}x^{(k-1)} + c_{SSOR(\omega)}$, $k \geq 1$, где

$$R_{SSOR(\omega)} = (D-\omega\tilde{U})^{-1}((1-\omega)D+\omega\tilde{L})(D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{U}) =$$
$$(E-\omega U)^{-1}((1-\omega)E+\omega L)(E-\omega L)^{-1}((1-\omega)E+\omega U)\,,$$
$$c_{SSOR(\omega)} = \omega((D-\omega\tilde{U})^{-1}((1-\omega)D+\omega\tilde{L})(D-\omega\tilde{L})^{-1}+E)b =$$
$$\omega(2-\omega)(D-\omega\tilde{U})^{-1}D(D-\omega\tilde{L})^{-1}b\,.$$

Это соответствует расщеплению $A = M - K$, в котором

$$M = \frac{1}{\omega(2-\omega)}(D-\omega\tilde{L})D^{-1}(D-\omega\tilde{U}) = \frac{1}{\omega(2-\omega)}D(E-\omega L)(E-\omega U)\,,$$

и

$$K = \frac{1}{\omega(2-\omega)}((1-\omega)D+\omega\tilde{L})D^{-1}((1-\omega)D+\omega\tilde{U}) =$$
$$\frac{1}{\omega(2-\omega)}D((1-\omega)E+\omega L)((1-\omega)E+\omega U)\,.$$

Замечание 0.35. *Если A — самосопряжённая положительно определённая матрица, $A = A^* > 0$, тогда матрица $R_{SSOR(\omega)}$ имеет неотрицательные вещественные собственные значения, т.е.* $\operatorname{Spec} R_{SSOR(\omega)} \subset \mathbb{R}_+$. *Более того, при всех ω, $0 < \omega < 2$, $\rho(R_{SSOR(\omega)}) < 1$ и, как следствие,* $\operatorname{Spec} R_{SSOR(\omega)} \subset (0,1)$. *Последнее также означает сходимость метода $SSOR(\omega)$, $0 < \omega < 2$, при любом начальном приближении.*

Доказательство. В данном случае $D = D^* > 0$, $\tilde{L}^* = \tilde{U}$. Матрица

$$R_{SSOR(\omega)} = (E-\omega U)^{-1}((1-\omega)E+\omega L)(E-\omega L)^{-1}((1-\omega)E+\omega U)$$

сопряжена с матрицей

$$(E-\omega U)R_{SSOR(\omega)}(E-\omega U)^{-1} =$$
$$((1-\omega)E+\omega L)(E-\omega L)^{-1}((1-\omega)E+\omega U)(E-\omega U)^{-1} =$$
$$(E-\omega L)^{-1}((1-\omega)E+\omega L)((1-\omega)E+\omega U)(E-\omega U)^{-1} =$$
$$(D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{L})D^{-1}((1-\omega)D+\omega\tilde{U})(D-\omega\tilde{U})^{-1}D =$$
$$(D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{L})D^{-1}((D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{L}))^*D = BD^{-1}B^*D\,.$$

где $B = (D-\omega\tilde{L})^{-1}((1-\omega)D+\omega\tilde{L})$. Поскольку $D = D^* > 0$, мы можем записать $D = T^2$ для $T^{-1} = (T^*)^{-1} = \sqrt{D^{-1}} > 0$ и значит, $T(BT^{-2}B^*T^2)T^{-1} = (TBT^{-1})(TBT^{-1})^*$. Таким образом, $\operatorname{Spec} R_{SSOR(\omega)} = \operatorname{Spec}(TBT^{-1})(TBT^{-1})^* \subset \mathbb{R}_+$.

Из сказанного ранее следует, что в данном случае расщепление $A = M - K$ для метода $SSOR(\omega)$, $0 < \omega < 2$, отвечает матрицам $M = M^* > 0$ и $K = K^* \geq 0$ ($K > 0$ при $\omega \neq 1$). Поэтому данное расщепление M–регулярно и в силу доказанного ранее замечания $\rho(R_{SSOR(\omega)}) < 1$. □

Условия сходимости методов Якоби, Гаусса — Зейделя, $SOR(\omega)$ и $SSOR(\omega)$

Напомним, что матрица $A = (a_{ij})$ имеет строгое диагональное преобладание по строкам (или является матрицей с таким преобладанием), если

$$|a_{ii}| > \sum_{1 \le j \ne i \le n} |a_{ij}| \quad (i = 1, \ldots, n)\,.$$

Сходным образом определяется условие строго диагонального преобладания по столбцам (как строчное для транспонированной матрицы).

Теорема 0.36. *Для всякой матрицы A со строгим диагональным преобладанием по строкам (столбцам) методы Якоби и Гаусса — Зейделя сходятся при любом начальном приближении, причём в случае матрицы со строгим диагональным преобладанием по строкам $\|R_{GS}\|_\infty \le \|R_J\|_\infty < 1$ (для столбцового диагонального преобладания верно аналогичное неравенство с точностью до замены строчной нормы $\|\ \|_\infty$ на столбцовую норму $\|\ \|_1$).*

Доказательство. Рассмотрим случай матрицы A со строгим диагональным преобладанием по строкам, обеспечивающим, в частности, обратимость её диагонали. По предыдущему $R_J = L + U$ и при реализации в естественном порядке $R_{GS} = (E - L)^{-1}U$. Строчная норма любой матрицы $B = (b_{ij})$ может быть записана как $\|B\|_\infty = \||B|e\|_\infty$, где в правой части стоит векторая max–норма вектора $|B|e$, $|B| = (|b_{ij}|)$, $e = (1, \ldots, 1)^t$. Для краткости далее неравенство $a \ge b$ применительно к вещественным векторам одной размерности a и b подразумевает выполнение неравенств $a_i \ge b_i$ для всех i. Условие строчного диагонального преобладания можно переписать как

$$(E - |L| - |U|)e \;=\; (E - |R_J|)e > 0\,, \quad e > |R_J|e\,,$$

откуда следует, что $\|R_J\|_\infty < 1 = \|e\|_\infty$. Вместе с тем мы получаем

$$0 \le |L|(E - |L| - |U|)e\,, \quad |U|e \le |U|e + |L|(E - |L| - |U|)e \;=\; (E - |L|)(|L| + |U|)e$$

и, как следствие, положительности коэффициентов матрицы $(E - |L|)^{-1} = \sum\limits_{i=0}^{n-1} |L|^i$

$$(E - |L|)^{-1}|U|e \le (|L| + |U|)e\,.$$

Остаётся заметить, что

$$|(E - L)^{-1}U|e \le |(E - L)^{-1}||U|e \;=\; \Big|\sum_{i=0}^{n-1} L^i\Big||U|e \le$$
$$\sum_{i=0}^{n-1} |L|^i|U|e \;=\; (E - |L|)^{-1}|U|e \le (|L| + |U|)e\,,$$

а потому $\|R_{GS}\|_\infty = \||(E - L)^{-1}U|e\|_\infty \le \|(|L| + |U|)e\|_\infty = \|R_J\|_\infty < 1$.

Для столбцового диагонального преобладания доказательство проводится практически аналогично с использованием выражения $\|B\|_1 = \|e^t|B|\|_\infty$.

При реализации метода Гаусса — Зейделя с использованием порядка $\sigma(1),\dots,\sigma(n)$, $\sigma\in\mathfrak{S}_n$, вычисления на k-ом шаге записываются как

$$x^{(k)}_{\sigma(i)} \;=\; 1/a_{\sigma(i)\sigma(i)}\Big(-\sum_{j<i}a_{\sigma(i)\sigma(j)}x^{(k)}_{\sigma(j)}-\sum_{j>i}a_{\sigma(i)\sigma(j)}x^{(k-1)}_{\sigma(j)}+b_{\sigma(i)}\Big)\quad(i=1,\dots,n)\,,$$

а потому в данном случае, полагая $P=P_\sigma=(e_{\sigma(1)}\dots e_{\sigma(n)})$ (e_i — i-ый столбец единичной матрицы), мы можем записать $(P^tAP)_{ij}=a_{\sigma(i)\sigma(j)}$, $(P^ty)_i=y_{\sigma(i)}$, $1\le i,j\le n$, и, как следствие,

$$(D_1-\tilde L_1)P^tx^{(k)} \;=\; \tilde U_1P^tx^{(k-1)}+P^tb\,,\quad x^{(k)} \;=\; P(D_1-\tilde L_1)^{-1}\tilde U_1P^tx^{(k-1)}+P(D_1-\tilde L_1)^{-1}P^tb\,,$$

где $P^tAP=D_1-\tilde L_1-\tilde U_1$, $D_1=P^tDP$. Условие строчного диагонального преобладания для матрицы P^tAP также имеет место, поскольку

$$\begin{aligned}(|D_1|-|\tilde L_1|-|\tilde U_1|)e \;&=\; (2|D_1|-|P^tAP|)e \;=\; (2P^t|D|P-P^t|A|P)e \;=\\ &P^t(2|D|-|A|)Pe \;=\; P^t(|D|-|\tilde L|-|\tilde U|)Pe \;=\; P^t(|D|-|\tilde L|-|\tilde U|)e>0\,.\end{aligned}$$

Поэтому в силу сказанного ранее

$$\begin{aligned}\|P(D_1-\tilde L_1)^{-1}\tilde U_1P^t\|_\infty\le k_\infty(P)\|(D_1-\tilde L_1)^{-1}\tilde U_1\|_\infty \;&=\; \|(D_1-\tilde L_1)^{-1}\tilde U_1\|_\infty\le\\ \|D_1^{-1}(\tilde L_1+\tilde U_1)\|_\infty \;=\; \|P^tD^{-1}P(P^tAP-P^tDP)\|_\infty \;&=\; \|P^t(D^{-1}A-E)P\|_\infty \;=\\ \|P^t(L+U)P\|_\infty\le k_\infty(P)\|L+U\|_\infty \;&=\; \|L+U\|_\infty<1\,.\end{aligned}$$

□

Известно также следующее усиление данного результата, которое мы приведём без доказательства.

Предложение 0.37. *Пусть неразложимая матрица A имеет слабое диагональное преобладание по строкам в том смысле, что*

$$|a_{ii}|\ge\sum_{1\le j\ne i\le n}|a_{ij}|\quad(i=1,\dots,n)\,,$$

причём для по меньшей мере одного i данное неравенство является строгим, тогда $\rho(R_{GS})\le\rho(R_J)<1$.

Неразложимость матрицы A подразумевает следующее: не существует ни одной матрицы-перестановки P, сопряжение с которой матрицы A даёт блочную верхнюю треугольную матрицу с квадратными диагональными блоками

$$PAP^t \;=\; \begin{pmatrix}A_{11}&A_{12}\\0&A_{22}\end{pmatrix}.$$

К сказанному можно добавить также то, что в случае самосопряжённой положительно определённой матрицы $A=A^*>0$, удовлетворяющей условию $D+\tilde L+\tilde U=D+\tilde L+\tilde L^*>0$, сходимость метода Якоби при любом начальном приближении обеспечивается неравенством $\rho(R_J)<1$, причём в данном случае $\operatorname{Spec}R_J\subset(-1,1)$ (см. ранее).

Лемма 0.38. *Условие $0<\omega<2$ необходимо для сходимости метода $SOR(\omega)$ при любом начальном приближении, поскольку $\rho(R_{SOR(\omega)})\ge|\omega-1|$.*

Доказательство. В силу сказанного ранее можно считать, что речь идёт о реализации $SOR(\omega)$ в естественном порядке и значит, $R_{SOR(\omega)} = (E - \omega L)^{-1}((1-\omega)E + \omega U)$. В таком случае мы можем записать характеристический многочлен матрицы $R_{SOR(\omega)}$ как

$$\chi_{R_{SOR(\omega)}}(\lambda) \;=\; (-1)^n \det(\lambda E - R_{SOR(\omega)}) \;=$$
$$(-1)^n \det((E - \omega L)(\lambda E - R_{SOR(\omega)})) \;=\; (-1)^n \det((\lambda + \omega - 1)E - \omega\lambda L - \omega U)\,.$$

Поэтому

$$|\chi_{R_{SOR(\omega)}}(0)| \;=\; \prod_{i=1}^{n} |\lambda_i| \;=\; |\det((\omega-1)E - \omega U)| \;=\; |\omega - 1|^n\,,$$

$\rho(R_{SOR(\omega)}) = \max\limits_{i=1,\ldots,n} |\lambda_i| \geq |\omega - 1|$, $\{\lambda_1, \ldots, \lambda_n\}$ — полный список собственных значений матрицы $R_{SOR(\omega)}$, записанных с учётом кратности. Отсюда следует, что выполнение необходимого и достаточного условия сходимости рассматриваемого метода при любом начальном приближении $\rho(R_{SOR(\omega)}) < 1$ влечёт за собой неравенство $0 < \omega < 2$. □

Теорема 0.39. *Для любых самосопряжённой положительно определённой матрицы $A = A^* > 0$ и параметра ω, $0 < \omega < 2$, выполняется неравенство $\rho(R_{SOR(\omega)}) < 1$ и, как следствие, метод $SOR(\omega)$ сходится при любом начальном приближении. В частности, это справедливо для метода Гаусса — Зейделя ($SOR(1)$).*

Доказательство. По аналогии с предыдущим нам достаточно рассматривать реализацию метода $SOR(\omega)$ в естественном порядке с $R_{SOR(\omega)} = M^{-1}K$, где $M = (D - \omega\tilde{L})/\omega$ и $K = ((1-\omega)D + \omega\tilde{L}^*)/\omega$. Положим $Q = A^{-1}(2M - A)$. Тогда для всякого $\lambda \in \operatorname{Spec} Q \subset \mathbb{C}$ и $0 \neq x_\lambda$, $Qx_\lambda = \lambda x_\lambda$, мы имеем $(2M - A)x_\lambda = \lambda A x_\lambda$ и $x_\lambda^*(2M^* - A) = \overline{\lambda} x_\lambda^* A$,

$$Re\lambda\; x_\lambda^* A x_\lambda \;=\; 1/2(\lambda + \overline{\lambda}) x_\lambda^* A x_\lambda \;=\; x_\lambda^*(M + M^* - A)x_\lambda\,.$$

При этом $x_\lambda^* A x_\lambda > 0$ и $x_\lambda^*(M + M^* - A)x_\lambda > 0$, поскольку $A = A^* > 0$, $D = D^* > 0$ и

$$M + M^* - A \;=\; (D - \omega\tilde{L} + D - \omega\tilde{L}^*)/\omega - A \;=\; (2/\omega - 1)D > 0\,.$$

Таким образом, $Re\lambda > 0$ для любого $\lambda \in \operatorname{Spec} Q$. Заметим, что

$$(Q - E)(Q + E)^{-1} \;=\; 2(A^{-1}M - E)(2A^{-1}M)^{-1} \;=\; (A^{-1}M - E)M^{-1}A \;=$$
$$E - M^{-1}A \;=\; E - M^{-1}(M - K) \;=\; M^{-1}K \;=\; R_{SOR(\omega)}\,.$$

Напомним, что для любой рациональной функции f, не имеющей полюсов на спектре $\operatorname{Spec} B$ матрицы, справедливо равенство $\operatorname{Spec} f(B) = f(\operatorname{Spec} B)$. Поэтому применительно к $f(x) = (x-1)/(x+1)$ и матрице Q

$$\operatorname{Spec} R_{SOR(\omega)} \;=\; \operatorname{Spec} f(Q) \;=\; f(\operatorname{Spec} Q) \;=\; \{(\lambda - 1)/(\lambda + 1) \mid \lambda \in \operatorname{Spec} Q\}\,,$$

где в силу неравенства $Re\lambda > 0$

$$\left|\frac{\lambda - 1}{\lambda + 1}\right| \;=\; \sqrt{\frac{(Re\lambda - 1)^2 + (Im\lambda)^2}{(Re\lambda + 1)^2 + (Im\lambda)^2}} < 1\,.$$

Следовательно, $\rho(R_{SOR(\omega)}) < 1$. □

Напомним также, что в условиях данной теоремы сходящимся при любом начальном приближении будет и метод $SSOR(\omega)$.

Одним из следствий данной теоремы является тот факт, что области применимости методов Якоби и Гаусса — Зейделя различны. Для этого достаточно заметить, что применительно к самосопряжённым положительно определённым матрицам условие сходимости метода Якоби равносильно регулярности их диагонального расщепления (см. далее). Последнее всегда имеет место для матриц 2×2, но в больших размерностях может нарушаться. Например, симметрические матрицы

$$A \;=\; \begin{pmatrix} \alpha & 1 & 1 \\ 1 & \alpha & 1 \\ 1 & 1 & \alpha \end{pmatrix}, \quad A' \;=\; 2D - A \;=\; \begin{pmatrix} \alpha & -1 & -1 \\ -1 & \alpha & -1 \\ -1 & -1 & \alpha \end{pmatrix}$$

имеют характеристические многочлены

$$\begin{aligned} \chi_A(\lambda) \;&=\; \det(A - \lambda E) \;=\; (\alpha - \lambda)^3 - 3(\alpha - \lambda) + 2 \;=\; (\alpha - 1 - \lambda)^2(\alpha + 2 - \lambda)\,, \\ \chi_{A'}(\lambda) \;&=\; \det(A' - \lambda E) \;=\; (\alpha - \lambda)^3 - 3(\alpha - \lambda) - 2 \;=\; (\alpha + 1 - \lambda)^2(\alpha - 2 - \lambda)\,. \end{aligned}$$

Поэтому при любом α, $1 < \alpha < 2$, матрица A положительно определена, а матрица A' таковой не является.

Не столь простое в общем случае определение согласовано упорядоченной матрицы мы приведём в следующем довольно упрощённом виде. Будем говорить, что матрица $A = D - \tilde{L} - \tilde{U}$ с обратимой диагональю D является *согласовано упорядоченной*, если спектр $\operatorname{Spec} R_J(\alpha)$ матрицы $R_J(\alpha) = \alpha D^{-1}\tilde{L} + \alpha^{-1}D^{-1}\tilde{U} = \alpha L + \alpha^{-1}U$ не зависит от ненулевого комплексного параметра α.

Естественным примером матриц такого рода может служить матрицы вида

$$A \;=\; \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

с диагональными обратимыми матрицами A_{11} и A_{22}. Для этого достаточно заметить, что

$$R_J(\alpha) \;=\; -\begin{pmatrix} 0 & \alpha^{-1}A_{11}^{-1}A_{12} \\ \alpha A_{22}^{-1}A_{21} & 0 \end{pmatrix} \;=\; \begin{pmatrix} E_1 & 0 \\ 0 & \alpha E_2 \end{pmatrix} R_J(1) \begin{pmatrix} E_1 & 0 \\ 0 & \alpha^{-1}E_2 \end{pmatrix}$$

для соответствующих единичных матриц E_1 и E_2.

Теорема 0.40. *Пусть A — согласовано упорядоченная матрица и $\omega \neq 0$. Тогда*

1. *$-\lambda \in \operatorname{Spec} R_J$ для любого $\lambda \in \operatorname{Spec} R_J$, $R_J = R_J(1)$;*
2. *если $\mu \in \operatorname{Spec} R_J$, то любое решение τ квадратного уравнения $(\tau + \omega - 1)^2 = \tau\omega^2\mu^2$ входит в $\operatorname{Spec} R_{SOR(\omega)}$;*
3. *если $0 \neq \tau \in \operatorname{Spec} R_{SOR(\omega)}$, то всякое решение μ квадратного уравнения из предыдущего пункта входит в $\operatorname{Spec} R_J$.*

Доказательство. Первое утверждение следует из равенств $R_J(-1) = -R_J = -R_J(1)$ и $\operatorname{Spec} R_J(-1) = \operatorname{Spec} -R_J = -\operatorname{Spec} R_J = \{-\lambda \mid \lambda \in \operatorname{Spec} R_J\} = \operatorname{Spec} R_J$.

Рассмотрим два оставшихся пункта. Для начала заметим, что нулевое решение τ

уравнения $(\tau+\omega-1)^2=\tau\omega^2\mu^2$ соответствует случаю $\omega=1$ и методу Гаусса — Зейделя $SOR(1)=GS$ с вырожденной матрицей R_{GS}. Остаётся заметить, что ненулевые собственные значения τ матрицы $R_{SOR(\omega)}$ (без ограничения общности рассуждений мы будем предполагать, что речь идёт о реализации метода $SOR(\omega)$ в естественном порядке для $R_{SOR(\omega)}=(E-\omega L)^{-1}((1-\omega)E+\omega U)$) удовлетворяют уравнению

$$\begin{aligned}0 \;=\; \chi_{R_{SOR(\omega)}}(\tau) \;=\; (-1)^n\det(\tau E-R_{SOR(\omega)}) \;=\; \det((E-\omega L)(\tau E-R_{SOR(\omega)}) \;=\\ \det((\tau+\omega-1)E-\omega\tau L-\omega U) \;=\; \det\Big(\sqrt{\tau}\omega\Big(\frac{\tau+\omega-1}{\sqrt{\tau}\omega}E-R_J(\sqrt{\tau})\Big)\Big) \;=\\ \det\Big(\frac{\tau+\omega-1}{\sqrt{\tau}\omega}E-R_J(\sqrt{\tau})\Big),\end{aligned}$$

которое в силу равенства $\operatorname{Spec}R_J=\operatorname{Spec}R_J(\sqrt{\tau})$ (речь идёт о любом из возможных значений $\sqrt{\tau}$) равносильно включению

$$\frac{\tau+\omega-1}{\sqrt{\tau}\omega} \;=\; \mu\in\operatorname{Spec}R_J\,.$$

□

Следствие 0.41. *Для любой согласовано упорядоченной матрицы A имеет место равенство $\rho(R_{GS})=\rho(R_{SOR(1)})=(\rho(R_J))^2$.*

В заключение мы приведём без доказательства следующий результат об одном из способов выбора параметра релаксации близкого к оптимальному, т.е. обеспечивающему минимизацию величины $\rho(R_{SOR(\omega)})$ на интервале $0<\omega<2$.

Предложение 0.42. *Пусть A — согласовано упорядоченная матрица, для которой матрица R_J имеет вещественный спектр и спектральный радиус $\rho(R_J)=\mu<1$. Тогда*

$$\rho(R_{SOR(\omega)}) \;=\; \begin{cases}1-\omega+\frac{\omega^2\mu^2}{2}+\omega\mu\sqrt{1-\omega+\frac{\omega^2\mu^2}{2}} & \text{при } 0<\omega\le\hat{\omega},\\ \\ \sqrt{(1-\omega)^2+2(1-\omega)\omega^2\mu^2+\frac{\omega^4\mu^4}{2}} & \text{при } \hat{\omega}<\omega<2,\end{cases}$$

где $\hat{\omega}=\frac{2}{1+\sqrt{1-\mu^2}}$. Поэтому в качестве приближения к $\omega_{opt}=\operatorname{arginf}_{0<\omega<2}\rho(R_{SOR(\omega)})$ можно использовать величину $\hat{\omega}$, для которой $\rho(R_{SOR(\hat{\omega})})=\hat{\omega}-1=\frac{\mu^2}{(1+\sqrt{1-\mu^2})^2}$.

Последнее, в частности, применимо к случаю матрицы $A=A^*>0$, имеющей D–регулярное расщепление.

К сказанному следует добавить, что перечисленные нами методы имееют естественные блочные аналоги, на которые также переносятся доказанные нами результаты о сходимости.

Лекция 6. Ускорение сходимости стационарных итерационных процессов.

Начнём со следующей весьма частной схемы ускорения сходимости стационарного процесса $x^{(k)} = Rx^{(k-1)} + c$, $k \geq 1$, в предположении $\operatorname{Spec} R \subset [m, M] \subset \mathbb{R}$, где $m = \lambda_{\min}$ и $M = \lambda_{\max}$ — наименьшее и наибольшее собственные значения матрицы R. Рассмотрим экстраполированный процесс $y^{(k)} = \gamma(Ry^{(k-1)} + c) + (1-\gamma)y^{(k-1)}$, $k \geq 1$, с ненулевым вещественным параметром γ (фактически это релаксация с параметром γ исходного процесса или параметризованный метод градиентного спуска (Ричардсона) для системы с матрицей $E - R$). Оба процесса (исходный и экстраполированный) в случае своей сходимости сходятся, очевидно, к одному вектору $x = Rx + c$. Нашей целью является выбор γ, обеспечивающий наилучшую сходимость процесса $y^{(k)} = (\gamma R + (1-\gamma)E)y^{(k-1)} + \gamma c = R_\gamma y^{(k-1)} + \gamma c$ в некоторой норме, а точнее минимизирующий спектральный радиус $\rho(R_\gamma)$ матрицы R_γ. Поскольку $\operatorname{Spec} R_\gamma = \{\gamma\lambda + 1 - \gamma \mid \lambda \in \operatorname{Spec} R\}$ и линейная функция принимает на отрезке наибольшее и наименьшее значение в его концах, мы имеем

$$\rho(R_\gamma) \;=\; \max\{|\gamma m + 1 - \gamma|, |\gamma M + 1 - \gamma|\} \;=\; \max\{|\gamma(m-1) + 1|, |\gamma(M-1) + 1|\}\,.$$

Отрезок $[m-1, M-1]$ параметризуется элементами отрезка $[-1, 1]$ посредством отображения $x \longmapsto \frac{M+m-2}{2} + x\frac{M-m}{2}$, $x \in [-1, 1]$. Поэтому исследование семейства функций $\{\gamma t + 1\}$ на отрезке $[m-1, M-1]$ сводится к изучению функций $\{f_\gamma\}$ на отрезке $[-1, 1]$, где

$$f_\gamma(x) \;=\; x\gamma\frac{(M-m)}{2} + \Big(\gamma\frac{M+m-2}{2} + 1\Big)\,.$$

Наша задача состоит в нахождении среди функций $\{f_\gamma\}$ функции с наименьшей чебышевской нормой $\|\ \|_{C[-1,1]}$ на отрезке $[-1, 1]$, $\|f\|_{C[-1,1]} = \max_{x\in[-1,1]} |f(x)|$, $f \in C[-1, 1]$.

Пусть $M + m \neq 2$ и при этом имеется $0 \neq \gamma$, для которого

$$\|f_\gamma\|_\infty < \Big|\frac{M-m}{2-M-m}\Big| \;=\; \|f_{\gamma_0}\|_\infty\,,$$

где $\gamma_0 = \frac{2}{2-M-m}$ и $f_{\gamma_0}(x) = x\frac{M-m}{2-M-m}$. Тогда функция $f_\gamma - f_{\gamma_0}$ принимает в точках ± 1 ненулевые значения разных знаков (знак разности определяется знаком $f_{\gamma_0}(\pm 1)$). Вместе с тем

$$f_\gamma(x) - f_{\gamma_0}(x) \;=\; x\frac{M-m}{M+m-2}\Big(\gamma\frac{M+m-2}{2} + 1\Big) + \Big(\gamma\frac{M+m-2}{2} + 1\Big) \;=\; \Big(\gamma\frac{M+m-2}{2} + 1\Big)\Big(x\frac{M-m}{M+m-2} + 1\Big)\,,$$

причём $\delta = \gamma\frac{M+m-2}{2} + 1 \neq 0$ и

$$f_\gamma(-1) - f_{\gamma_0}(-1) \;=\; 2\delta\frac{m-1}{M+m-2}\,, \quad f_\gamma(1) - f_{\gamma_0}(1) \;=\; 2\delta\frac{M-1}{M+m-2}\,.$$

Таким образом, в случае $(m-1)(M-1) \geq 0$ существование такого γ исключается, т.е. в данной ситуации ($m \geq 1$ или $M \leq 1$)

$$\rho(R_{\gamma_0}) \;=\; \Big|\frac{M-m}{2-m-M}\Big| \;=\; \min_{0\neq\gamma\in\mathbb{R}} \rho(R_\gamma)\,.$$

Заметим, что условие $(m-1)(M-1) > 0$ обеспечивает выполнение обоих неравенств $M+m-2 \neq 0$ и $\rho(R_{\gamma_0}) < 1$. Последнее обстоятельство гарантирует также сходимость экстраполированного процесса с параметром γ_0 при любом начальном приближении (даже в случае если исходный процесс не обладал этим свойством (при $m > 1$ или $m \leq -1$ $(M < 1)$).

В случае $(m-1)(M-1) < 0$ с учётом $m-1 < M-1$ мы имеем $m < 1 < M$ и, как следствие матрица, $E - R$ может быть вырождена. Более того,

$$\rho(R_\gamma) \;=\; \begin{cases} \max\{|\gamma(m-1)+1|, \gamma(M-1)+1\} \geq 1 & \text{при } \gamma > 0, \\ \max\{\gamma(m-1)+1, |\gamma(M-1)+1|\} \geq 1 & \text{при } \gamma < 0. \end{cases}$$

Поэтому здесь ни один из возможных экстраполированных процессов не является сходящимся при любом начальном приближении.

Итак, имеется одна единственная возможность реализовать схему нашего ускорения: $(m-1)(M-1) > 0$ для $\gamma_0 = \frac{M-m}{2-M-m}$, $\rho(R_{\gamma_0}) = \left|\frac{M-m}{2-M-m}\right|$. При этом в подходящей норме скорость сходимости соответствующего оптимального экстраполированного процесса будет определяться множителем $\rho(R_{\gamma_0})^k = \left|\frac{M-m}{2-M-m}\right|^k$.

Последнее имеет непосредственное отношение к случаю симметризуемой матрицы R, $R = VBV^{-1}$ для некоторых самосопряжённой и обратимой матриц $B = B^*$ и V. Тогда для векторной нормы $\|\ \|_{V^{-1}}$, $\|x\|_{V^{-1}} = \|V^{-1}x\|_2$,

$$\|y^{(k)} - x\|_{V^{-1}} \;=\; \|R_{\gamma_0}^k (y^{(0)} - x)\|_{V^{-1}} \leq \|R_{\gamma_0}\|^k \|y^{(0)} - x\|_{V^{-1}} \;=\; \left|\frac{M-m}{2-M-m}\right|^k \|y^{(0)} - x\|_{V^{-1}},$$

где $\|R_{\gamma_0}\|_{V^{-1}} = \|V^{-1}R_{\gamma_0}V\|_2 = \|\gamma_0 B + (1-\gamma_0)E\|_2 = \rho(R_{\gamma_0}) = \left|\frac{M-m}{2-M-m}\right|$. К сказанному можно лишь добавить, что в исходной евклидовой норме соответствующая оценка записывается в виде

$$\|y^{(k)} - x\|_2 \leq k_2(V) \left|\frac{M-m}{2-M-m}\right|^k \|y^{(0)} - x\|_2 .$$

Заметим, что в случае $m = M \neq 1$ наше рассуждение приводит к получению нильпотентной матрицы R_{γ_0}, $\rho(R_{\gamma_0}) = 0$, $\gamma_0 = \frac{1}{1-m}$, что обеспечивает гарантированную сходимость экстраполированного процесса к точному решению за n шагов (случай $m = M = 1$ соответствует вырожденной матрице $E - R$).

Следует отметить и то, что в действительности нам доступны лишь приближения m и M к границам спектра $\operatorname{Spec} R$ матрицы R, но, тем не менее, это не меняет существенно приведённые выше рассуждения: при $(m-1)(M-1) > 0$

$$\rho(R_\gamma) \leq \|f_\gamma\|_{C[-1,1]}, \quad \min_{0 \neq \gamma \in \mathbb{R}} \|f_\gamma\|_{C[-1,1]} \;=\; \|f_{\gamma_0}\|_{C[-1,1]} \;=\; \left|\frac{M-m}{2-m-M}\right|,$$

а потому выбор γ_0 по-прежнему вполне оправдан.

Экстраполированный процесс $y^{(k)} = R_\gamma y^{(k-1)} + \gamma c$, $k \geq 1$, может быть переписан для $y^{(0)} = x^{(0)}$ как

$$y^{(k)} \;=\; R_\gamma^k (x^{(0)} - x) + x \;=\; (\gamma R + (1-\gamma)E)^k (x^{(0)} - x) + x \;=$$

$$\sum_{i=0}^{k} \binom{k}{i} \gamma^i (1-\gamma)^{k-i} R^i (x^{(0)} - x) + x \;=$$

$$\sum_{i=0}^{k}\binom{k}{i}\gamma^i(1-\gamma)^{k-i}(x^{(i)}-x)+x \;=\; \sum_{i=0}^{k}\binom{k}{i}\gamma^i(1-\gamma)^{k-i}x^{(i)}\,.$$

Поэтому естественным обобщением описанной здесь схемы улучшения сходимости исходного процесса $x^{(k)} = Rx^{(k-1)} + c$, $k \geq 1$, может служить построение на его основе процесса

$$y^{(k)} \;=\; \sum_{i=0}^{k}\gamma_{ki}x^{(i)} \quad (k \geq 0)\,,$$

в котором вещественные параметры $\{\gamma_{ki}\}$ связаны ограничением $\sum\limits_{i=0}^{k}\gamma_{ki} = 1$. Необходимость выполнения последнего обусловлена тем, что последовательности $x^{(i)} = x = Rx + c$, $i \geq 0$, должна соответствовать последовательность $y^{(i)} = x$, $i \geq 0$. В таком случае, полагая $p_k(t) = \sum\limits_{i=0}^{k}\gamma_{ki}t^i$, $k \geq 0$, мы можем записать

$$y^{(k)} - x \;=\; \sum_{i=0}^{k}\gamma_{ki}(x^{(i)}-x) \;=\; p_k(R)(x^{(0)}-x) \;=\; p_k(R)(y^{(0)}-x)\,.$$

Основной объём известных стратегий организации подобных процессов полиномиального ускорения (стратегий выбора систем полиномов $\{p_k\}$) сосредоточен на решении одной из следующих задач

- минимизация спектрального радиуса $\rho(p_k(R))$, т.е. выбор p_k должен доставлять решение задачи $\rho(p(R)) = \max\{|p(\lambda)| \mid \lambda \in \mathrm{Spec}(R)\} \to \inf$ для всех возможных действительных полиномов p степени k, $p(1) = 1$;
- минимизация величины $\|p_k(R)(x^{(0)} - x)\|$ в конкретной норме $\|\ \|$.

В действительности решение первой задачи заменяется её более грубой версией поиска $\mathrm{arginf}_{\deg p=k,\ p(1)=1}\|p\|_{C[m,M]}$, где $\|f\|_{C([m,M])} = \|f\|_\infty$, $f \in C[m,M]$. Точное решение такой задачи находится в рамках соответствующих полиномов Чебышева (см. далее).

Заметим, что в симметризуемой ситуации $R = VBV^{-1}$, $B = B^*$, на которую в основном и нацелено такое решение, справедливы оценки

$$\|y^{(k)} - x\|_{V^{-1}} \leq \|p_k(R)\|_{V^{-1}}\|y^{(0)} - x\|_{V^{-1}} \;= \\ \|p_k(B)\|_2\|y^{(0)} - x\|_{V^{-1}} \;= \rho(p_k(B))\|y^{(0)} - x\|_{V^{-1}} \;=\; \rho(p_k(R))\|y^{(0)} - x\|_{V^{-1}}\,,$$

и $\|y^{(k)} - x\|_2 \leq \rho(p_k(R))k_2(V)\|y^{(0)} - x\|_2$ (см. выше).

В круг второй задачи входят многие известные методы, среди которых в первую очередь стоит отметить метод сопряжённых градиентов или точнее ускорение по этому методу, если речь идёт о рассматриваемой здесь проблематике.

Уместно будет привести также следующую трёхчленную схему организации полиномиального ускорения, используемую в большинстве известных случаев.

Предложение 0.43. *Пусть последовательность полиномов $\{p_k\}$ строится по правилу*

$$p_0(t) \;=\; 1\,, \quad p_1(t) \;=\; \gamma_1 t + 1 - \gamma_1\,,$$
$$p_{i+1}(t) \;=\; \rho_{i+1}(\gamma_{i+1}t + 1 - \gamma_{i+1})p_i(t) + (1-\rho_{i+1})p_{i-1}(t) \quad (i \geq 1)$$

для некоторых $\{\gamma_i\}$ и $\{\rho_j\}$. Тогда соответствующий итерационный процесс может быть описан как: $y^{(1)} = \gamma_1(Ry^{(0)} + c) + (1-\gamma_1)y^{(0)}$ и далее

$$y^{(i+1)} \;=\; \rho_{i+1}(\gamma_{i+1}(Ry^{(i)} + c) + (1-\gamma_{i+1})y^{(i)}) + (1-\rho_{i+1})y^{(i-1)} \quad (i \geq 1)\,.$$

Доказательство. Действительно, $y^{(i+1)} = p_{i+1}(R)(y^{(0)} - x) + x$, где $x = Rx + c$ и

$$p_{i+1}(R)(y^{(0)}-x) \;=\; \rho_{i+1}(\gamma_{i+1}R+(1-\gamma_{i+1})E)p_i(R)(y^{(0)}-x)+(1-\rho_{i+1})p_{i-1}(R)(y^{(0)}-x) \;=$$
$$\rho_{i+1}(\gamma_{i+1}R + (1-\gamma_{i+1})E)(y^{(i)} - x) + (1-\rho_{i+1})(y^{(i-1)} - x) \;=$$
$$\rho_{i+1}(\gamma_{i+1}(Ry^{(i)} + c) + (1-\gamma_{i+1})y^{(i)}) - \rho_{i+1}x + (1-\rho_{i+1})y^{(i-1)} - (1-\rho_{i+1})x \;=$$
$$\rho_{i+1}(\gamma_{i+1}(Ry^{(i)} + c) + (1-\gamma_{i+1})y^{(i)}) + (1-\rho_{i+1})y^{(i-1)} - x\,.$$

Более того, не составляют труда показать, что итерационный процесс указанного вида соответствует в точности заданной системе полиномов. □

Заметим, что описанный ранее экстраполированный процесс вписывается в данную схему с постоянными параметрами $\gamma_i = \gamma$, $\rho_{i+1} = 1$, $i \geq 1$.

Многочлены Чебышева и чебышевское ускорение.

Рассмотрим последовательность полиномов $\{t_k\}_{k\geq 0}$, определённых согласно следующего рекуррентного соотношения: $t_0 = 1$, $t_1 = x$ и далее

$$t_{i+1} \;=\; t_{i+1}(x) \;=\; 2xt_i(x) - t_{i-1}(x) \quad (i \geq 1)\,,$$

представляющей собой частный случай приведённой ранее схемы для $\gamma_i = 1$ и $\rho_{i+1} = 2$, $i \geq 1$. Несложно показать по индукции, что t_i — целочисленный многочлен степени i со старшим коэффициентом 2^{i-1}, $i \geq 1$. На отрезке $[-1, 1]$ данному рекуррентному соотношению удовлетворяет система функций $t_i(x) = \cos(i \arccos x)$, $i \geq 0$. В общем случаем её решением (в $\mathbb{C}$) служит система функций

$$t_i(z) \;=\; 1/2\big((z + \sqrt{z^2-1})^i + (z + \sqrt{z^2-1})^{-i}\big) \;=\; 1/2\big((z + \sqrt{z^2-1})^i + (z - \sqrt{z^2-1})^i\big)\,,$$

где используется одна из возможных ветвей квадратного корня. Многочлены $\{t_i\}$ составляют систему *многочленов Чебышева* (на отрезке $[-1, 1]$).

В действительсти наличие подобного трёхчленного соотношения между многочленами Чебышева является следствием того, что они получаются в результате процесса переортогонализации степенных функций $\{x^i\}$ в соответствующем интегральном гильбертовом пространстве.

Заметим, что каждый многочлен t_n, $n \geq 1$, имеет n нулей в интервале $(-1, 1)$ в точках $x_m = \cos\frac{\pi(2m-1)}{2n}$, $m = 1, \ldots, n$, и потому может быть записан в виде

$$t_n(x) \;=\; 2^{n-1}\prod_{m=1}^{n}(x - x_m)\,.$$

Кроме того, в точках $y_k = \cos\frac{\pi k}{n}$, $k = 0, 1, \ldots, n$, полином t_n принимает равные по модулю значения переменных знаков: $t_n(y_k) = (-1)^k$, причём

$$1 \;=\; |t_n(y_k)| \;=\; \max_{x\in[-1,1]} |t_n(x)| \;=\; \|t_n\|_{C[-1,1]} \;=\; \|\cos(n\arccos x)\|_{C[-1,1]}\,,$$

Система точек $\{y_k\}$ называется *чебышевским альтернансом* многочлена t_n. В дальнейшем символом $\overline{t_n}$ мы будем обозначать нормализованный многочлен Чебышева $2^{1-n}t_n$, $n \geq 1$.

Интерес к многочленам Чебышева связан прежде всего с тем, что они позволяют решить задачу о наилучшем равномерном приближении нуля нормализованными многочленами заданной степени (стоит сказать, что последняя в неевклидовом пространстве непрерывных функций на отрезке не может быть решена в духе задачи наименьших квадратов). Точнее имеет место

Теорема 0.44. *Нормализованный многочлен $\overline{t_n}$ является наименее отклоняющимся от нуля в равномерной (чебышевской) метрике $\|\ \|_{C[-1,1]}$ многочленом среди всех нормализованных действительных многочленов степени $n \geq 1$, т.е. для любого $p_n(x) = x^n + p_{n-1}x^{n-1} + \ldots + p_0 \in \mathbb{R}[x]$,*

$$\|p_n\|_{C[-1,1]} \geq \|\overline{t_n}\|_{C[-1,1]} \;=\; 2^{1-n}\,.$$

Доказательство. Предположим противное, т.е. имеется нормализованный многочлен $p \in \mathbb{R}[x]$, $\deg p = n$, такой, что $\|p\|_{C[-1,1]} < 2^{1-n} = \|\overline{t_n}\|_{C[-1,1]}$. В таком случае в точках чебышевского альтернанса $\{y_k\}_{k=0}^n$ многочлена t_n мы имеем

$$\operatorname{sign}(\overline{t_n}-p)(y_k) \;=\; \operatorname{sign}(\overline{t_n}(y_k)-p(y_k)) \;=\; \operatorname{sign}((-1)^k 2^{1-n}-p(y_k)) \;=\; (-1)^k \quad (k=0,\ldots,n)\,.$$

Последнее означает, что, как и у многочлена t_n, у многочлена $t_n - p$ между каждой соседней парой точек y_k и y_{k+1} (в интервале (y_{k+1}, y_k)) имеется корень, т.е. многочлен $t_n - p$ степени не выше $n-1$ имеет в интервале $(-1, 1)$ не менее n корней?! □

Теорема 0.45. *Для любого действительного числа d, $|d| \geq 1$, и любого многочлена действительного многночлена $f \in \mathbb{R}[x]$, $\deg f = n \geq 1$, $f(d) = 1$,*

$$\|f\|_{C[-1,1]} \geq \frac{1}{|t_n(d)|} \;=\; \|\tilde{t_n}\|_{C[-1,1]}\,,$$

где $\tilde{t_n} = t_n/t_n(d)$. Другими словами, многочлен $\tilde{t_n}$ является многочленом, наименее отклоняющимся от нуля на отрезке $[-1, 1]$ в равномерной метрике среди всех действительных многочленов степени n, принимающих в точке d единичное значение.

Доказательство. Как и в предыдущем случае, доказательство проводится от противного. Пусть имеется многочлен $g \in \mathbb{R}[x]$, $\deg g = n$, $g(d) = 1$, для которого $\|g\|_{C[-1,1]} < \frac{1}{|t_n(d)|}$. Тогда в точках чебышевского альтернанса $\{y_k\}_{k=0}^n$ многочлена t_n

$$\operatorname{sign}(\tilde{t_n}-g)(y_k) \;=\; \operatorname{sign}(\tilde{t_n}(y_k)-g(y_k)) \;=\; \operatorname{sign}\Big(\frac{(-1)^k}{t_n(d)}-g(y_k)\Big) \;=\; (-1)^k \quad (k=0,\ldots,n)\,,$$

что гарантирует наличие у многочлена $\tilde{t_n} - g$ по меньшей мере n нулей в интервале

$(-1,1)$. Вместе с тем он имеет ещё один нуль в точке d, лежащей вне этого интервала. Поэтому многочлен $\tilde{t}_n - g$ степени не выше n имеет по меньшей мере $n+1$ корень?!

Заметим, что мы не стали здесь выделять в отдельное рассмотрение случай $|d| = 1$, поскольку в данной ситуации точка d входит в альтернанс (d совпадает с y_0 или y_n), $\frac{1}{|t_n(d)|} = |t_n(d)| = 1 = |g(d)| \le \|g\|_{C[-1,1]}$. Поэтому в предыдущем рассуждении случай $|d| = 1$ автоматически отбрасывается. □

Если возникает вопрос о построении нормализованного вещественного полинома степени n, наименее отклоняющегося от нуля на заданном отрезке $[a,b]$, тогда решение этой задачи состоит использовании взаимосвязи между пространствами $C[-1,1]$ и $C[a,b]$ (посредством замен $\beta : [a,b] \longrightarrow [-1,1]$, $\beta(x) = \frac{2x-a-b}{b-a}$, и $\beta^{-1}(y) = \frac{a+b}{2} + y\frac{b-a}{2}$), что позволяет взять в качестве такого многочлена

$$(b-a)^n 2^{1-2n} t_n\Big(\frac{2x-a-b}{b-a}\Big) \;=\; (b-a)^n 2^{1-2n} t_n(\beta(x))\,.$$

Действительно, всякий нормализованный многочлен $p = p(x) \in \mathbb{R}[x]$, $\deg p = n$, на отрезке $[a,b]$ может быть представлен как многочлен $g(y) = p(\beta^{-1}(y))$ от $y \in [-1,1]$ со старшим коэффициентом $(b-a)^n 2^{-n}$, причём по доказанному

$$\|p\|_{C[a,b]} \;=\; \|g\|_{C[-1,1]} \ge (b-a)^n 2^{-n} \|\overline{t_n}\|_{C[-1,1]} \;=$$
$$(b-a)^n 2^{-n} \|\overline{t_n}(\beta(x))\|_{C[a,b]} \;=\; (b-a)^n 2^{1-2n}\,,$$

где $(b-a)^n 2^{-n}\overline{t_n}(\beta(x)) = (b-a)^n 2^{1-2n} t_n(\beta(x))$ — нормализованный многочлен степени n.

Следствие 0.46. *Пусть $m < M$ и $(m-1)(M-1) > 0$ ($1 \notin [m,M]$). Тогда для любого действительного многочлена $f \in \mathbb{R}[x]$, $\deg f = n$, $f(1) = 1$,*

$$\|f\|_{C[m,M]} \ge \frac{1}{\Big|t_n\Big(\frac{2-m-M}{M-m}\Big)\Big|} \;=\; \Big\|t_n\Big(\frac{2x-m-M}{M-m}\Big)\Big/t_n\Big(\frac{2-m-M}{M-m}\Big)\Big\|_{C[m,M]}\,.$$

Доказательство. Сперва заметим, что условие $1 \notin [m,M]$ равносильно неравенству

$$\Big|\frac{2-m-M}{M-m}\Big| \;=\; \frac{|2-m-M|}{M-m} > 1\,,$$

поскольку последнее эквивалентно неравенствам $2-m-M > M-m$, $1 > M$, и $m+M-2 > M-m$, $m > 1$.

При замене $\beta^{-1} : y \longmapsto \frac{m+M}{2} + y\frac{M-m}{2}$, $y \in [-1,1]$ (параметризации отрезка $[m,M]$ элементами отрезка $[-1,1]$), решением $\beta^{-1}(y) = 1$ является $\hat{y} = \beta(1) = \frac{2-m-M}{M-m}$, причём $|\hat{y}| > 1$ (см. выше).

Каждому многочлену $f = f(x) \in \mathbb{R}[x]$, $\deg f = n$, $f(1) = 1$, соответствует многочлен $g = g(y) = f(\beta^{-1}(y)) \in \mathbb{R}[y]$, $\deg g = n$, $g(\hat{y}) = f(1) = 1$. Поэтому согласно доказанной ранее теореме

$$\|f\|_{C[m,M]} \;=\; \|g\|_{C[-1,1]} \ge \frac{1}{|t_n(\hat{y})|} \;=\; \|t_n/t_n(\hat{y})\|_{C[-1,1]} \;=$$
$$\Big\|t_n\Big(\frac{2x-m-M}{M-m}\Big)\Big/t_n\Big(\frac{2-m-M}{M-m}\Big)\Big\|_{C[m,M]}\,.$$

□

Поэтому применительно к нашей исходной задаче с заданными границами (оценками границ) спектра m и M матрицы R, $m < M$, и естественным условием $1 \notin [m, M]$, обеспечивающим, в частности, обратимость матрицы $E - R$ (впрочем включение $1 \in [m, M]$ вместе с условием $p(1) = 1$ означало бы, что $\|p\|_{C[m,M]} \geq 1$, а это в свою очередь делает бессмысленным представленную схему улучшения сходимости), в качестве системы полиномов $\{p_i\}$, $\deg p_i = i$, $p_i(1) = 1$,

$$\|p_i\|_{C[m,M]} \;=\; \min_{f \in \mathbb{R}[x],\ \deg f = i,\ f(1)=1} \|f\|_{C[m,M]}\,,$$

можно использовать полиномы $p_i(x) = t_i\Big(\frac{2x-m-M}{M-m}\Big)/d_i$, где $d_i = t_i\Big(\frac{2-m-M}{M-m}\Big)$, $i \geq 1$. При этом

$$p_1(x) \;=\; \frac{2x-m-M}{2-m-M} \;=\; x\frac{2}{2-m-M} + \Big(1 - \frac{2}{2-m-M}\Big)$$

и далее в соответствии с рекуррентным соотношением для системы $\{t_i\}$

$$p_{i+1}(x) \;=\; \frac{1}{d_{i+1}}\Big(2\Big(\frac{2x-m-M}{M-m}\Big)t_i\Big(\frac{2x-m-M}{M-m}\Big) - t_{i-1}\Big(\frac{2x-m-M}{M-m}\Big)\Big) \;=$$
$$\frac{1}{d_{i+1}}\Big(2d_i\Big(\frac{2x-m-M}{M-m}\Big)p_i(x) - d_{i-1}p_{i-1}(x)\Big)\,,$$

где

$$d_{i+1} \;=\; 2\Big(\frac{2-m-M}{M-m}\Big)d_i - d_{i-1}\,.$$

Следовательно, полагая $\gamma_i = \frac{2}{2-m-M}$, $i \geq 1$, мы можем записать

$$p_{i+1}(x) \;=\; 2\Big(\frac{2-m-M}{M-m}\Big)\frac{d_i}{d_{i+1}}(\gamma_{i+1}x + 1 - \gamma_i)p_i(x) - \frac{d_{i-1}}{d_{i+1}}p_{i-1}(x)$$

и нам остаётся лишь взять $\rho_{i+1} = 2\Big(\frac{2-m-M}{M-m}\Big)\frac{d_i}{d_{i+1}} = 1 + \frac{d_{i-1}}{d_{i+1}}$, $i \geq 1$, чтобы рассматривать систему $\{p_i\}$ в рамках описанной выше трёхчленной схемы.

Построенный по данной системе процесс $\{y^{(i)}\}$, называемый *чебышевским ускорением процесса* $x^{(k)} = Rx^{(k-1)} + c$, $k \geq 1$, имеет следуюший вид:

$$y^{(1)} \;=\; \frac{2}{2-m-M}(Ry^{(0)} + c) - \frac{m+M}{2-m-M}y^{(0)}$$

и далее

$$y^{(k+1)} \;=\; \rho_{k+1}(\gamma_{k+1}(Ry^{(k)} + c) + (1-\gamma_{k+1})y^{(k)}) + (1-\rho_{k+1})y^{(k-1)} \;=$$
$$\frac{4}{M-m}\frac{d_k}{d_{k+1}}(Ry^{(k)} + c) - \frac{2(M+m)}{M-m}\frac{d_k}{d_{k+1}}y^{(k)} - \frac{d_{k-1}}{d_{k+1}}y^{(k-1)} \;=$$
$$\frac{2}{M-m}\frac{d_k}{d_{k+1}}((2R - (m+M)E)y^{(k)} + 2c) - \frac{d_{k-1}}{d_{k+1}}y^{(k-1)}\,.$$

Напомним, что в данном случае $y^{(k)} - x = p_k(R)(y^{(0)} - x)$, где

$$\rho(p_k(R)) \leq \|p_k\|_{C[m,M]} \;=\; \frac{1}{|d_k|} \;=\; \frac{1}{\Big|t_k\Big(\frac{2-m-M}{M-m}\Big)\Big|}\,.$$

Выясним теперь насколько более качественным будет улучшение сходимости в рамках чебышевского ускорения по сравнению с описанным ранее оптимальным экстраполиро-

ванным процессом. Для этого следует оценить порядок роста величины d_k с ростом k. Мы имеем

$$d_k \;=\; t_k\Big(\frac{2-m-M}{M-m}\Big) \;=$$

$$\frac{1}{2}\left(\left(\frac{2-m-M}{M-m}+\sqrt{\Big(\frac{2-m-M}{M-m}\Big)^2-1}\right)^k+\left(\frac{2-m-M}{M-m}-\sqrt{\Big(\frac{2-m-M}{M-m}\Big)^2-1}\right)^k\right),$$

где

$$\frac{2-m-M}{M-m}\pm\sqrt{\Big(\frac{2-m-M}{M-m}\Big)^2-1} \;=\; \frac{2-m-M\pm 2\sqrt{(1-m)(1-M)}}{M-m} \;=$$

$$\frac{(1-m)+(1-M)\pm 2\sqrt{(1-m)(1-M)}}{(1-m)-(1-M)} \;=$$

$$\begin{cases} \frac{\sqrt{1-m}\pm\sqrt{1-M}}{\sqrt{1-m}\mp\sqrt{1-M}} & \text{при } M<1,\\[2ex] \frac{\sqrt{m-1}\mp\sqrt{M-1}}{\sqrt{m-1}\pm\sqrt{M-1}} & \text{при } m>1. \end{cases}$$

Положим

$$\mu \;=\; \begin{cases} \frac{\sqrt{1-m}+\sqrt{1-M}}{\sqrt{1-m}-\sqrt{1-M}} & \text{при } M<1,\\[2ex] \frac{\sqrt{m-1}+\sqrt{M-1}}{\sqrt{M-1}-\sqrt{m-1}} & \text{при } m>1. \end{cases}$$

Тогда, суммируя сказанное, мы получаем

$$d_k \;=\; \begin{cases} \frac{\mu^k+\mu^{-k}}{2} & \text{при } M<1,\\[2ex] (-1)^k\,\frac{\mu^k+\mu^{-k}}{2} & \text{при } m>1 \end{cases}$$

и, как следствие, $|d_k|\geq\frac{\mu^k}{2}$ (по определению $\mu>1$). Таким образом,

$$\rho(p_k(R))\leq\frac{1}{|d_k|}\leq\frac{2}{\mu^k}\,.$$

Остаётся заметить, что

$$\Big|\frac{M-m}{2-m-M}\Big| \qquad = \qquad \begin{cases} \frac{M-m}{(\sqrt{1-m})^2+(\sqrt{1-M})^2} > \frac{M-m}{(\sqrt{1-m}+\sqrt{1-M})^2} \;=\; \frac{1}{\mu} & \text{при } M<1,\\[2ex] \frac{M-m}{(\sqrt{m-1})^2+(\sqrt{M-1})^2} > \frac{M-m}{(\sqrt{m-1}+\sqrt{M-1})^2} \;=\; \frac{1}{\mu} & \text{при } m>1. \end{cases}$$

Поэтому с ростом k чебышевское ускорение имеет более качественную сходимость чем оптимальный экстраполированный процесс. Впрочем, применительно к ситуации, когда $\mu\approx 1$, оба этих способа улучшения сходимости не дадут особых результатов. Вместе с тем в остальных случаях их использование является весьма эффективным и позволяет заранее планировать количество итераций необходимых для получения приближения заданной точности к решению $x=Rx+c$. Естественно, что речь идёт прежде

всего о симметризуемых процессах, к которым относятся методы Якоби и $SSOR(\omega)$ для $A = A^* > 0$.

Известен также следующий вариант использования процедуры чебышеского ускорения: фиксируется $k \geq 1$ и затем строятся приближения $\{z^{(i)}\}_{i\geq 0}$, где $z^{(i+1)}$ получается на основе $z^{(i)}$ в результате выполнения k шагов чебышевского ускорения. Такой процесс позволяет получить оценку

$$\|z^{(i)} - x\| \leq \|p_k(R)\|\|z^{(i-1)} - x\| \leq \|p_k(R)\|^i \|z^{(0)} - x\|\,,$$

уточнение которой может быть выполнено согласно приведённым выше результатам.

Лекция 7. Методы минимизации.

Алгоритмы решения систем линейных алгебраических уравнений, именуемые методами минимизации, сводят решение этой задачи (задачи нахождения $x = A^{-1}b$) к задачи поиска минимума некоторого квадратичного функционала, связанного с матрицей системы A и вектором правой части b, достигаемого в точке $x = A^{-1}b$. Точнее речь идёт о методах, строящих последовательные приближения к x в соответствии со стратегией минимизации выбранного функционала. Основная масса рассматриваемых нами здесь алгоритмов относится к случаю самосопряжённой положительно определённой матрицы $0 < A = A^* \in GL_n(\mathbb{C})$ и связанному с ней и вектором правой части $b \in \mathbb{C}^n$ функционалу

$$Q(y) \;=\; 1/2(Ay, y) - Re(b, y) \;=\; 1/2y^*Ay - Re(y^*b) \;=\; 1/2(y^*Ay - y^*b - b^*y) \quad (y \in \mathbb{C}^n)\,.$$

Замечание 0.47. *Глобальный минимум функционала Q достигается в одной единственной точке $x = A^{-1}b$.*

Доказательство. Действительно, для любого $y \in \mathbb{C}^n$

$$Q(y + A^{-1}b) \;=\; Q(y) + Q(A^{-1}b) + 1/2(y^*A(A^{-1}b) + (A^{-1}b)^*Ay) \;=$$
$$Q(y) + Q(A^{-1}b) + Re(y^*b) \;=\; 1/2y^*Ay + Q(A^{-1}b) \geq Q(A^{-1}b) \;=\; -1/2b^*A^{-1}b\,.$$

Поскольку $y^*Ay > 0$ при всех $0 \neq y \in \mathbb{C}^n$, отсюда следует, что минимум функционала Q достигается в одной точке $x = A^{-1}b$. □

Естественно ничто не мешает использовать и другие квадратичные функционалы с единственным глобальным минимумом в точке $x = A^{-1}b$. В частности, можно рассматривать функционал невязки в евклидовой норме

$$F_2(y) \;=\; \|b - Ay\|_2^2 \;=\; (b - Ay, b - Ay) \;=\; \|b\|_2 + \|Ay\|_2^2 - 2Re(b, Ay)$$

или в какой-либо другой норме, причём в данном случае матрица не обязана удовлетворять условию $A = A^* > 0$. Заметим, что в случае $A = A^* > 0$ для евклидовой нормы $\|\ \|_{A^{-1}}$, отвечающей скалярному произведению с матрицей A^{-1}, $(z, y)_{A^{-1}} = y^*A^{-1}z$, $z, y \in \mathbb{C}^n$, функционал невязки $F_{A^{-1}}$ совпадает с точностью до сдвига и умножения на константу с функционалом Q, точнее

$$F_{A^{-1}}(y) \;=\; \|b\|_{A^{-1}}^2 + \|Ay\|_{A^{-1}}^2 - 2Re(b, Ay)_{A^{-1}} \;=$$
$$\|b\|_{A^{-1}}^2 + \|y\|_A^2 - 2Re(b, y) \;=\; \|b\|_{A^{-1}}^2 + 2Q(y) \quad (y \in \mathbb{C}^n)\,.$$

Большинство рассматриваемых здесь методов построения последовательных приближений к $\operatorname{arginf} Q = A^{-1}b = x$ состоят в следующем: выбирается начальное приближение $x^{(0)}$ и строятся вектора $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$, $k \geq 1$, каждый из которых получается из своего предшественника в результате сдвига с коэффициентом α_k вдоль некоторого направления $p^{(k)}$. При этом стратегии выбора системы направлений $\{p^{(k)}\}$ могут быть различными, что и определяет во многом многообразие методов такого рода. Заметим, что приближение $x^{(k)}$, $k \geq 1$, является элементом алгебраического многообразия $x^{(0)} + \langle p^{(1)}, \ldots, p^{(k)} \rangle$. Относительно выбора коэффицента сдвига α_k следует сказать, что в большинстве случаев он выбирается как решение одномерной задачи минимизации функционала Q вдоль выбранного направления $p^{(k)}$, $\alpha_k = \operatorname{arginf}_{\alpha \in \mathbb{C}} Q(x^{(k-1)} + \alpha p^{(k)})$.

Замечание 0.48. *В описанном процессе выбора коэффициента сдвига:*

$$\alpha_k \;=\; \frac{(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2} \;=\; \frac{(p^{(k)})^* r^{(k-1)}}{(p^{(k)})^* A p^{(k)}} \quad (k\geq 1)\,,$$

где $r^{(k-1)} = b - Ax^{(k-1)}$ *— вектор невязки на* $k-1$*-ом шаге. При этом*

$$Q(x^{(k)}) \;=\; Q(x^{(k-1)}) - \frac{|(r^{(k-1)},p^{(k)})|^2}{2\|p^{(k)}\|_A^2}\,,$$

а потому уменьшение функционала Q *на* k*-ом шаге построения возможно только при условии* $r^{(k-1)} \not\perp p^{(k)}$*.*

Доказательство. Достаточно заметить, что $Re(x,y) = Re(y,x)$ и

$$Re(ix,y) \;=\; 1/2(iy^*x - ix^*y) \;=\; i/2((x,y)-(x,y)^*) \;=\; -Im(x,y) \;=\; Im(y,x)$$

при всех $x, y \in \mathbb{C}^n$, а потому для всякого $\alpha \in \mathbb{C}$

$$\begin{gathered}
Q(x^{(k-1)}+\alpha p^{(k)}) \;=\; Q(x^{(k-1)}) + Q(\alpha p^{(k)}) + Re(\alpha p^{(k)}, x^{(k-1)})_A \;= \\
Q(x^{(k-1)}) + 1/2|\alpha|^2 (p^{(k)},p^{(k)})_A - Re(\alpha p^{(k)}, b) + Re(\alpha p^{(k)}, Ax^{(k-1)}) \;= \\
Q(x^{(k-1)}) + 1/2((Re\alpha)^2 + (Im\alpha)^2)\|p^{(k)}\|_A^2 - Re((Re\alpha + iIm\alpha)p^{(k)}, r^{(k-1)}) \;= \\
Q(x^{(k-1)}) + 1/2((Re\alpha)^2\|p^{(k)}\|_A^2 - (Re\alpha)(Re(p^{(k)}, r^{(k-1)})) + \\
(1/2(Im\alpha)^2\|p^{(k)}\|_A^2 - (Im\alpha)(Re(ip^{(k)}, r^{(k-1)}))) \;= \\
Q(x^{(k-1)}) - 1/2\frac{(Re(r^{(k-1)},p^{(k)}))^2 + (Im(r^{(k-1)},p^{(k)}))^2}{\|p^{(k)}\|_A^2} + \\
1/2\|p^{(k)}\|_A^2\Big(\Big(Re\alpha - \frac{Re(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2}\Big)^2 + \Big(Im\alpha - \frac{Im(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2}\Big)^2\Big) \geq \\
Q(x^{(k-1)}) - \frac{|(r^{(k-1)},p^{(k)})|^2}{2\|p^{(k)}\|_A^2}\,,
\end{gathered}$$

причём равенство достигается при $\alpha = \alpha_k = \frac{(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2}$. □

Поэтому одной из наиболее общих схем организации подобных процессом является так называемый метод *произвольных направлений*, который состоит в следующем: выберем начальное приближение $x^{(0)}$; после выполнения $k-1$-го шага вычисляем невязку $r^{(k-1)} = b - Ax^{(k-1)}$ и в случае если последняя отлична от нуля или больше по выбранной норме некоторого порогового значения (если не используется какое-либо иное условие остановки процесса), то выбираем очередное направление $p^{(k)}$ из условия $p^{(k)} \not\perp r^{(k-1)}$ и вычисляем $\alpha_k = \frac{(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2}$ и $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$, $k \geq 1$.

Остановимся на некоторых классических примерах методов близких к методу произвольных направлений с той лишь разницей, что используемые в них стратегии выбора очередного направления не подразумевают выполнение обязательной проверки условия $p^{(k)} \not\perp r^{(k-1)}$.

Покоординатная релаксация

Одним из самых простых методов такого рода является *покоординатная релаксация*, суть которой состоит в использовании в качестве направлений $\{p^{(k)}\}$ циклически перебираемой системы базисных векторов $\{e_1, \ldots, e_n\}$, где $e_i = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{n-i})^t$, $i = 1, \ldots, n$, т.е. $p^{(k)} = e_i$, если $k \equiv i \ (n)$. Вычислительная процедура метода покоординатной релаксации выглядит следующим образом: выбираем начальное приближение $x^{(0)}$; на шаге с номером k вычисляем остаток r от деления k на n и полагаем $i = r$ при $1 \le i \le n-1$ и $i = n$ при $r = 0$, находим коэффициент сдвига α_k по общему правилу

$$\alpha_k \ = \ \frac{(r^{(k-1)}, p^{(k)})}{\|p^{(k)}\|_A^2} \ = \ \frac{(b - Ax^{(k-1)}, e_i)}{\|e_i\|_A^2} \ = \ \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k-1)}\Big)$$

и вычисляем новое приближение $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)} = x^{(k-1)} + \alpha_k e_i$, которое отличается от $x^{(k-1)}$ лишь i-ой компонентой

$$x_i^{(k)} \ = \ \frac{1}{a_{ii}}\Big(b_i - \sum_{j \ne i} a_{ij} x_j^{(k-1)}\Big) .$$

В качестве условия выхода можно использовать условие $\|r^{(k)}\|_2 < \varepsilon$.

Заметим, что полное перевычисление компонент вектора $x^{(0)}$ будет осуществлено после выполнения первых n шагов описанного построения. При этом,

$$x_i^{(n)} \ = \ x_i^{(i)} \ = \ \frac{1}{a_{ii}}\Big(b_i - \sum_{j \ne i} a_{ij} x_j^{(i-1)}\Big) \ =$$

$$\frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij} x_j^{(i-1)} - \sum_{j>i} a_{ij} x_j^{(0)}\Big) \ = \ \frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij} x_j^{(j)} - \sum_{j>i} a_{ij} x_j^{(0)}\Big) \ =$$

$$\frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij} x_j^{(n)} - \sum_{j>i} a_{ij} x_j^{(0)}\Big) \quad (i = 1, \ldots, n) .$$

Следовательно, выполнение n шагов метода покоординатной релаксации равносильно выполнению одного шага метода Гаусса — Зейделя, реализованного в естественном порядке от 1 до n (перенумеровав элементы базиса, мы получим аналогичный результат для соответствующей реализации метода Гаусса — Зейделя). Последнее обстоятельство вместе с условием $A = A^* > 0$ обеспечивает также сходимость метода покоординатной релаксации при любом начальном приближении.

Метод Ричардсона или наискорейшего градиентного спуска

Идея этого метода состоит в использовании классической стратегии градиентного спуска примительно к функционалу Q в случае вещественной матрицы $A = A^t > 0$ и вещественного вектора правой части b. В данном случае функционал Q имеет вид

$$Q(y) \ = \ 1/2 \sum_{i,j=1}^{n} a_{ij} y_i y_j - \sum_{i=1}^{n} y_i b_i \quad (y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n) ,$$

При этом для всякого $i = 1, \ldots, n$ в силу симметричности матрицы A

$$\frac{\partial Q}{\partial y_i} \;=\; a_{ii}y_i + 1/2\sum_{j\neq i}(a_{ij}+a_{ji})y_j - b_i \;=\; (Ay)_i - b_i$$

и, как следствие, градиент функционала Q в точке y имеет вид: $\nabla Q(y) = Ay - b$ (подчеркнём ещё раз, что сказанное относится только к вещественной ситуации и не применимо к комплексной записи Q).

Градиентный спуск базируется на следующем соображении: для любого $y \in \mathbb{R}^n$ и произвольного направления p (для определённости $\|p\|_2 = 1$) при достаточно малом положительном δ можно записать

$$Q(y + p\delta) \;=\; Q(y) + (\nabla Q(y), p)\delta + O(\delta^2)$$

для некоторой ограниченной функции O при $\delta \to 0$, где $(\nabla Q(y), p)$ — производная функционала Q в точке y вдоль направления p. Поэтому функционал Q убывает в точке y наилучшим образом в направлении $p = -\nabla Q(y)/\|\nabla Q(y)\|_2$ (при этом случай $\nabla Q(y) = 0$ отвечает точке глобального минимума $y = A^{-1}b$ функционала Q).

Поэтому, исходя из стратегии выбора наибольшего убывания функционала или, что равносильно, в направлении невязки, мы получаем следующий алгоритм, именуемый *методом Ричардсона*: выбираем начальное приближение $x^{(0)}$; если уже найдено $k-1$-ое приближение $x^{(k-1)}$, тогда либо $r^{(k-1)} = 0$ или меньше в евклидовой норме некоторого заданного ε и процесс останавливается, либо вычисляется новое k-ое приближение по правилу $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$, где $p^{(k)} = r^{(k-1)} = b - Ax^{(k-1)} = -\nabla Q(x^{(k-1)})$ и

$$\alpha_k \;=\; \frac{(r^{(k-1)}, p^{(k)})}{\|p^{(k)}\|_A^2} \;=\; \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_A^2} \;=\; \frac{(r^{(k-1)})^t r^{(k-1)}}{(r^{(k-1)})^t A r^{(k-1)}}\,.$$

Заметим, что условие $r^{(k-1)} \not\perp p^{(k)} = r^{(k-1)}$ в данном случае обеспечивается автоматически. Стоит отметить и то, что в описанном процессе

$$x^{(k)} \;=\; x^{(k-1)} + \alpha_k p^{(k)} \;=\; x^{(k-1)} + \alpha_k(b - Ax^{(k-1)}) \;=$$
$$(E - \alpha_k A)x^{(k-1)} + \alpha_k b \;=\; ((1-\alpha_k)E + \alpha_k(E-A))x^{(k-1)} + \alpha_k b\,,$$

т.е. по сути речь идёт об экстраполированном процессе с переменным параметром для стационарного итерационного процесса $x^{(k)} = (E - A)x^{(k-1)} + b$, $k \geq 1$.

Сходимость метода Ричардсона обеспечивает следующее наблюдение.

Замечание 0.49. *В методе Ричардсона после выполнения k шагов*

$$2Q(x^{(k)}) + b^t A^{-1} b \;=\; \|r^{(k)}\|_{A^{-1}}^2 \leq (1 - 1/k_2(A))\|r^{(k-1)}\|_{A^{-1}}^2 \leq (1 - 1/k_2(A))^k \|r^{(0)}\|_{A^{-1}}^2\,,$$

где $r^{(i)} = b - Ax^{(i)}$ — невязка на i-ом шаге и $k_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$ — число обусловленности матрицы $A = A^t > 0$ в спектральной норме.

Доказательство. Сперва заметим, что при всех $y \in \mathbb{R}^n$

$$2Q(y) + b^t A^{-1} b \;=\; y^t A y - 2b^t y + b^t A^{-1} b \;=\; (b - Ay)^t A^{-1}(b - Ay) \;=\; \|b - Ay\|_{A^{-1}}^2\,.$$

Согласно описания нашего алгоритма

$$r^{(k)} \;=\; b - Ax^{(k)} \;=\; b - Ax^{(k-1)} - \alpha_k Ap^{(k)} \;=\; r^{(k-1)} - \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_A^2} Ar^{(k-1)}$$

и потому

$$\|r^{(k)}\|_{A^{-1}}^2 \;=\; \Bigl(r^{(k-1)} - \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_A^2} Ar^{(k-1)}\Bigr)^t A^{-1} \Bigl(r^{(k-1)} - \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_A^2} Ar^{(k-1)}\Bigr) \;=$$

$$\|r^{(k-1)}\|_{A^{-1}}^2 - 2\frac{\|r^{(k-1)}\|_2^4}{\|r^{(k-1)}\|_A^2} + \frac{\|r^{(k-1)}\|_2^4}{\|r^{(k-1)}\|_A^2} \;=\; \|r^{(k-1)}\|_{A^{-1}}^2 \Bigl(1 - \frac{\|r^{(k-1)}\|_2^4}{\|r^{(k-1)}\|_A^2 \|r^{(k-1)}\|_{A^{-1}}^2}\Bigr)\,.$$

Остаётся заметить, что $\|y\|_B^2 = y^tBy \le \|B\|_2\|y\|_2^2$ для любой матрицы $B = B^t > 0$ ($B = B^* > 0$) и, следовательно, $1/\|B\|_2 \le \|y\|_2^2/\|y\|_B^2$ при всех $y \ne 0$. Поэтому

$$\frac{\|r^{(k-1)}\|_2^4}{\|r^{(k-1)}\|_A^2 \|r^{(k-1)}\|_{A^{-1}}^2} \ge \frac{1}{k_2(A)}\,, \quad \|r^{(k)}\|_{A^{-1}}^2 \le \Bigl(1 - \frac{1}{k_2(A)}\Bigr)\|r^{(k-1)}\|_{A^{-1}}^2\,.$$

Последнюю оценку можно переписать для евклидовой нормы в виде

$$\|r^{(k)}\|_2 \le \sqrt{k_2(A)}\sqrt{1 - \frac{1}{k_2(A)}}\|r^{(k-1)}\|_2 \le \sqrt{k_2(A)}\Bigl(1 - \frac{1}{k_2(A)}\Bigr)^{k/2}\|r^{(0)}\|_2\,,$$

поскольку $\|y\|_B \le \sqrt{\|B\|_2}\|y\|_2$ и $\|B^{1/2}y\|_{B^{-1}} = \|y\|_2 \le \sqrt{\|B^{-1}\|_2}\|B^{1/2}y\|_2 = \sqrt{\|B^{-1}\|_2}\|y\|_B$, $(1/\sqrt{\|B^{-1}\|_2})\|y\|_2 \le \|y\|_B$, $B = B^t > 0$ ($B = B^* > 0$). Отсюда сразу следует, что для плохо обусловленной матрицы A сходимость метода Ричардсона является довольно медленной. □

В силу сказанного ранее метод Ричардсона представляет собой вариант процесса близкий к полиномиальному ускорению стационарного процесса $x^{(k)} = (E-A)x^{(k-1)} + b$, $k \ge 0$, а потому подпадает под общую для всех процессов такого рода схему получения оценки $\|x^{(k)} - x\|$, точнее с учётом очевидной симметричности матриц $E - \alpha_k A$, $k \ge 1$, мы имеем

$$\|x^{(k)} - x\|_2 \;=\; \|(E - \alpha_k A)(x^{(k-1)} - x)\|_2 \;=\; \|(E - \alpha_k A)\cdots(E - \alpha_1 A)(x^{(0)} - x)\|_2 \le$$

$$\|p_k(A)\|_2\|x^{(0)} - x\|_2 \;=\; \rho(p_k(A))\|x^{(0)} - x\|_2\,,$$

где $p_k(t) = \prod\limits_{i=1}^{k}(1 - \alpha_i t)$, $\rho(p_k(A)) = \max\limits_{\lambda \in \mathrm{Spec}\, A} |p_k(\lambda)|$. Заметим, что при всех $i \ge 1$

$$\frac{1}{\lambda_{\min}} \ge \alpha_i \;=\; \frac{\|r^{(i-1)}\|_2^2}{\|r^{(i-1)}\|_A^2} \ge \frac{1}{\|A\|_2} \;=\; \frac{1}{\rho(A)} \;=\; \frac{1}{\lambda_{\max}}\,,$$

поскольку мы можем записать $A = B^2$ для подходящей $B = B^t > 0$ и получить для любого $0 \ne x$ неравества

$$\frac{1}{\|B\|_2} \;=\; \frac{1}{\sqrt{\lambda_{\max}}} \le \frac{\|x\|_2}{\|x\|_A} \;=\; \frac{\|x\|_2}{\|Bx\|_2} \le \|B^{-1}\|_2 \;=\; \frac{1}{\sqrt{\lambda_{\min}}}\,,$$

и, следовательно, $\lambda_{\min}/\lambda_{\max} = 1/k_2(A) \le \alpha_i t \le 1$ при $\lambda_{\min} \le t \le \lambda_{\max}$. Поэтому

$$\rho(p_k(A)) \le \|p_k(t)\|_{C[\lambda_{\min},\lambda_{\max}]} \le (1 - 1/k_2(A))^k\,,$$

$$\|x^{(k)} - x\|_2 \le (1 - 1/k_2(A))^k\|x^{(0)} - x\|_2\,.$$

Метод сопряжённых направлений

Допустим, что мы располагаем n линейно независимыми направлениями $\{p^{(1)}, \ldots, p^{(n)}\}$ и хотим построить последовательность приближений $x^{(k)}$, $k = 0, \ldots, n$, к решению нашей минимизационной задачи $x = A^{-1}b = \operatorname{arginf}_{y \in \mathbb{C}^n} Q(y)$, в которой каждый $x^{(k)}$, $k = 1, \ldots, n$, является решением k-мерной минимизационной задачи

$$x^{(k)} \;=\; \operatorname{arginf}_{y \in x^{(0)} + \langle p^{(1)}, \ldots, p^{(k)} \rangle} Q(y)$$

и, в частности, $x^{(n)} = x$. При этом нам бы хотелось сохранить схему метода произвольных направлений, т.е. осуществлять построение $x^{(k)}$ по привычной схеме $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$ с $\alpha_k = \operatorname{arginf}_{\alpha \in \mathbb{C}} Q(x^{(k-1)} + \alpha p^{(k)}) = \frac{(r^{(k-1)}, p^{(k)})}{\|p^{(k)}\|_A^2}$.

Одним из способов реализовать эту идею является использование в качестве таких направлений полной системы A–сопряжённых или A–ортогональных направлений $\{p^{(1)}, \ldots, p^{(n)}\}$, $p^{(i)} \perp_A p^{(j)}$, $1 \le i \ne j \le n$, в том смысле, что $(p^{(i)}, p^{(j)})_A = (p^{(j)})^t A p^{(i)} = 0$. Подобная система может быть получена из любой системы линейно независимых векторов посредством процесса переортогонализации с формой $(\ ,\ )_A$. Понятно также, что A–сопряжённые направления являются линейно независимыми ($\|p^{(i)}\|_A^2 > 0$ и потому любая линейная комбинация $\sum\limits_i \beta_i p^{(i)} = 0$ имеет нулевые коэффициенты $\beta_i = 0$, $(\sum\limits_i \beta_i p^{(i)}, p^{(i)})_A = \beta_i \|p^{(i)}\|_A^2$).

Итак, пусть мы имеем A–сопряжённые направления $\{p^{(1)}, \ldots, p^{(n)}\}$. Тогда для любых $k \ge 1$ и

$$y \;=\; x^{(0)} + \sum_{i=1}^{k-1} \beta_i p^{(i)} + \alpha p^{(k)} \;=\; y' + \alpha p^{(k)}$$

мы можем записать

$$\begin{aligned} Q(y) \;&=\; 1/2\|y\|_A^2 - Re(b, y) \;= \\ &Q(y') + Q(\alpha p^{(k)}) + Re(\alpha p^{(k)}, y')_A \;=\; Q(y') + Q(\alpha p^{(k)}) + Re(\alpha p^{(k)}, x^{(0)})_A \,. \end{aligned}$$

Поэтому с учётом доказанного нами ранее замечания (см. вывод α_k с точностью до замены $r^{(k-1)}$ на $r^{(0)}$)

$$\begin{aligned} x^{(k)} \;&=\; \operatorname{arginf}_{y \in x^{(0)} + \langle p^{(1)}, \ldots, p^{(k)} \rangle} Q(y) \;= \\ &\operatorname{arginf}_{y' \in x^{(0)} + \langle p^{(1)}, \ldots, p^{(k-1)} \rangle} Q(y') + p^{(k)} \operatorname{arginf}_\alpha (Q(\alpha p^{(k)}) + Re(\alpha p^{(k)}, x^{(0)})_A) \;= \\ &x^{(k-1)} + \frac{(r^{(0)}, p^{(k)})}{\|p^{(k)}\|_A^2} p^{(k)} \,, \end{aligned}$$

причём $(r^{(k-1)}, p^{(k)}) = (b - Ax^{(k-1)}, p^{(k)}) = (r^{(0)}, p^{(k)})$ и, как следствие, $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$, где $\alpha^{(k)}$ вычисляется по прежней формуле

$$\alpha_k \;=\; \frac{(r^{(0)}, p^{(k)})}{\|p^{(k)}\|_A^2} \;=\; \frac{(r^{(k-1)}, p^{(k)})}{\|p^{(k)}\|_A^2} \,.$$

Кроме того, мы сразу получаем, что

$$Q(x^{(k+1)}) \;=\; Q(x^{(k)}) - \frac{|(r^{(0)}, p^{(k)})|^2}{2\|p^{(k)}\|_A^2} \,.$$

Поэтому для системы A–сопряжённых направлений $\{p^{(1)},\dots,p^{(n)}\}$ каждое последовательное приближение $x^{(k)}$, $k\geq 1$, построенное по стандартной схеме с выбором в качестве параметра сдвига решения одномерной минимизационной задачи, является решением k–мерной минимизационной задачи.

Заметим также, что включение $A^{-1}\in\langle E,A,\dots,A^{n-1}\rangle$ гарантирует принадлежность $x=A^{-1}b$ подпространству Крылова $K_n(A,b)=\langle b,Ab,\dots,A^{n-1}b\rangle$, $x\in K_n(A,b)$. Поэтому $A^{-1}b-x^{(0)}=A^{-1}(b-Ax^{(0)})=A^{-1}r^{(0)}\in K_n(A,r^{(0)})$ и $x=A^{-1}b\in x^{(0)}+K_n(A,r^{(0)})$. Следовательно, выбор A–сопряжённых направлений имеет смысл осуществлять в рамках крыловского подпространства $K_n(A,r^{(0)})$, что может существенно ограничить количество шагов для получения точного решения размерностью данного подпространства. Конечно в действительсти выход из процесса осуществляется из условия малости в евклидовой норме (или иной норме) вектора невязки, но, поставив своей целью построение вектора $x^{(n)}$, мы превратим такой процесс в один из точных методов решения системы $Ax=b$.

Стоит сказать и то, что одной из систем A–сопряжённых направлений является система попарно ортогональных собственных векторов матрицы A. Впрочем использование последней довольно затруднительно ввиду понятных трудностей её построения.

В соответствии со сказанным можно предложить следующий алгоритм построения приближений к решению рассматриваемой минимизационной задачи с последовательным выбором A–сопряжённых направлений сдвига, называемый *методом сопряжённых направлений*. Выбираем начальное приближение $x^{(0)}$ и полагаем $p^{(1)}=r^{(0)}=b-Ax^{(0)}$ в случае если $\|r^{(0)}\|_2>\varepsilon$ (иначе выходим из процесса, взяв в качестве искомого приближения $x^{(0)}$). После выполнения $k-1$-го шага и нахождения $x^{(i)}$, $p^{(i)}$ и $r^{(i)}=b-Ax^{(i)}$, $i=1,\dots,k-1$, в случае $\|r^{(k-1)}\|_2<\varepsilon$ осуществляем выход из процесса, взяв в качестве искомого приближения $x^{(k-1)}$, а в случае $\|r^{(k-1)}\|_2>\varepsilon$ находим новое направление $p^{(k)}\perp_A\langle p^{(1)},\dots,p^{(k-1)}\rangle$, для которого $p^{(k)}\not\perp r^{(k-1)}$, а затем вычисляем $\alpha_k=\frac{(r^{(k-1)},p^{(k)})}{\|p^{(k)}\|_A^2}$, $x^{(k)}=x^{(k-1)}+\alpha_k p^{(k)}$ и $r^{(k)}=r^{(k-1)}-\alpha_k Ap^{(k)}$.

Следует обосновать возможность выбора $p^{(k)}\perp_A p^{(i)}$, $i=1,\dots,k-1$, $p^{(k)}\not\perp r^{(k-1)}$. Действительно, если такой выбор невозможен, то для всякого $p\perp_A\langle p^{(1)},\dots,p^{(k-1)}\rangle$ или, что то же самое, $p\perp A\langle p^{(1)},\dots,p^{(k-1)}\rangle=\langle Ap^{(1)},\dots,Ap^{(k-1)}\rangle$, должно быть $p\perp r^{(k-1)}$, т.е. $r^{(k-1)}\in\langle Ap^{(1)},\dots,Ap^{(k-1)}\rangle$. Последнее означает, что

$$A^{-1}r^{(k-1)}\;=\;A^{-1}\Bigl(r^{(0)}-\sum_{i=1}^{k-1}\alpha_i Ap^{(i)}\Bigr)\;=\;A^{-1}r^{(0)}-\sum_{i=1}^{k-1}\alpha_i p^{(i)}\in\langle p^{(1)},\dots,p^{(k-1)}\rangle\,,$$

а потому $A^{-1}r^{(0)}\in\langle p^{(1)},\dots,p^{(k-1)}\rangle$, $x=A^{-1}b=x^{(0)}+A^{-1}r^{(0)}\in x^{(0)}+\langle p^{(1)},\dots,p^{(k-1)}\rangle$ и

$$x^{(k-1)}\;=\;\mathrm{arginf}_{y\in x^{(0)}+\langle p^{(1)},\dots,p^{(k-1)}\rangle}Q(y)\;=\;x\,,\quad r^{(k-1)}\;=\;0\,.$$

Другими словами, при этом предположении на шаге $k-1$ было бы найдено точное решение $x^{(k-1)}=x$ и осуществлён выход из процесса.

Метод сопряжённых градиентов

В описанном нами методе сопряжённых направлений мы практически не использовали возможность их поиска в рамках крыловского подпростанства $K_n(A,r^{(0)})$, отвечающего

вектору нулевой невязки $r^{(0)}$ (кроме того, что мы взяли данный вектор в качестве первого направления $p^{(1)}$). Исправить этот недостаток призван следующий вариант метода сопряжённых направлений — *метод сопряжённых градиентов.*

В отличие от метода сопряжённых направлений в методе сопряжённых градиентов методе k-ое направление $p^{(k)}$ выбирается A–сопряжённым к векторам $p^{(1)}, \ldots, p^{(k-1)}$, $k > 1$, т.е. в подпространстве $\langle Ap^{(1)}, \ldots, Ap^{(k-1)}\rangle^{\perp}$, с тем, чтобы оно было наиболее близким в евклидовой норме к вектору невязки $r^{(k-1)}$ (в предположении $\|r^{(k-1)}\|_2 > \varepsilon$),

$$p^{(k)} \;=\; \operatorname{arginf}_{p\in\langle Ap^{(1)},\ldots,Ap^{(k-1)}\rangle^{\perp}} \|p - r^{(k-1)}\|_2 \,.$$

Это оправдывается тем, что в действительном случае направление $r^{(k-1)} = -\nabla Q(x^{(k-1)})$ является направлением наболее быстрого убывания функционала Q в точке $x^{(k-1)}$ (к слову последнее обстоятельство объясняет и само название метода). Другим словами, направление $p^{(k)}$ есть не что иное как проекция вектора $r^{(k-1)}$ на подпространство $\langle Ap^{(1)}, \ldots, Ap^{(k-1)}\rangle^{\perp}$. При этом равенство нулю такой проекции равносильно включению $r^{(k-1)} \in \langle Ap^{(1)}, \ldots, Ap^{(k-1)}\rangle$, которое в свою очередь гарантирует равенство $r^{(k-1)} = 0$ (см. выше) и подразумевает выход из процесса на шаге $k-1$. Более того, по этой причине взаимная ортогональность векторов $r^{(k-1)}$ и $p^{(k)}$ равносильна равенству нулю вектора $p^{(k)}$, что в свою очередь влечёт за собой равенство нулю и вектора $r^{(k-1)}$. Поэтому данная стратегия выбора направления $p^{(k)}$ обеспечивает ещё и выполнение условия $p^{(k)} \not\perp r^{(k-1)}$.

Основные свойства метода сопряжённых градиентов являются следствиями теоремы Пифагора и могут быть записаны в форме серии наблюдений. Начнём с наиболее элементарного.

Замечание 0.50. *В методе сопряжённых градиентов k-ое направление $p^{(k)}$, $k > 1$, выбираемое в случае $r^{(k-1)} \neq 0$ из условия*

$$p^{(k)} \;=\; \operatorname{arginf}_{p\in\langle Ap^{(1)},\ldots,Ap^{(k-1)}\rangle^{\perp}} \|p - r^{(k-1)}\|_2$$

можно записать в виде

1. $p^{(k)} = r^{(k-1)} - AP_{k-1}z^{(k-1)}$, *где* $P_{k-1} = (p^{(1)} \ldots p^{(k-1)})$ — *матрица размера* $n\times(k-1)$ *со столбцами* $p^{(1)}, \ldots, p^{(k-1)}$ *и* $z^{(k-1)}$ — *вектор из* $\mathbb{k}^{k-1}$, $\mathbb{k} = \mathbb{R}, \mathbb{C}$ *(в зависимости от условий исходной задачи), определяемый из условия*

$$z^{(k-1)} \;=\; \operatorname{arginf}_{z\in\mathbb{k}^{k-1}} \|r^{(k-1)} - AP_{k-1}z\|_2 \,;$$

2. $p^{(k)} = \operatorname{arginf}_{p=r^{(k-1)}-AP_{k-1}z,\; z\in\mathbb{k}^{k-1}} \|p\|_2$;

3. $p^{(k)} = r^{(k-1)} - AP_{k-1}z^{(k-1)}$, *где* $r^{(k-1)} - AP_{k-1}z^{(k-1)} \in \langle Ap^{(1)}, \ldots, Ap^{(k-1)}\rangle^{\perp}$.

Доказательство. Как уже отмечалось, полагая $V = \langle Ap^{(1)}, \ldots, Ap^{(k-1)}\rangle = AP_{k-1}\mathbb{k}^{k-1}$, мы можем записать $r^{(k-1)} = r_V^{(k-1)} + r_{V^{\perp}}^{(k-1)}$, где

$$p^{(k)} \;=\; r_{V^{\perp}}^{(k-1)} \;=\; \operatorname{arginf}_{p\in r^{(k-1)}+V} \|p\|_2 \,, \quad r_V^{(k-1)} \;=\; AP_{k-1}z^{(k-1)} \;=\; \operatorname{arginf}_{q\in r^{(k-1)}+V^{\perp}} \|q\|_2$$

и

$$z^{(k-1)} \;=\; \operatorname{arginf}_{z\in\mathbb{k}^{k-1}} \|r^{(k-1)} - AP_{k-1}z\|_2 \,.$$

□

Теорема 0.51. *После выполнения $k \geq 1$ шагов метода сопряжённых градиентов при условии их осуществимости (т.е. при условии $r^{(i)} \neq 0$, $i = 0, \ldots, k-1$)*

1. $P_j^* r^{(j)} = 0$, $j = 1, \ldots, k$;
2. *при всех* $j = 1, \ldots, k$

$$\langle p^{(1)}, \ldots, p^{(j)} \rangle \;=\; \langle r^{(0)}, \ldots, r^{(j-1)} \rangle \;=\; K_j(A, r^{(0)}) \;=\; \langle r^{(0)}, Ar^{(0)}, \ldots, A^{j-1} r^{(0)} \rangle ;$$

3. $r^{(i)} \perp r^{(j)}$, $0 \leq i \neq j \leq k-1$.

Доказательство. Для большей общности мы будем рассматривать комплексную ситуацию. Как и в любом методе сопряжённых направлений, при любом $m = 1, \ldots, k$

$$Q(x^{(m)}) \;=\; \min_{x \in x^{(0)} + \langle p^{(1)}, \ldots, p^{(m)} \rangle} Q(x) \;=\; \min_{y \in \mathbb{C}^m} Q(x^{(0)} + P_m y) ,$$

где

$$Q(x^{(0)} + P_m y) \;=\; Q(x^{(0)}) + 1/2(AP_m y, P_m y) + Re(P_m y, Ax^{(0)} - b) \;=$$
$$Q(x^{(0)}) + 1/2(P_m^* A P_m y, y) - Re(P_m y, r^{(0)}) .$$

Поскольку направления $\{p^{(i)}\}$ являются A–ортогональными,

$$P_m^* A P_m \;=\; \mathrm{diag}(\|p^{(1)}\|_A^2, \ldots, \|p^{(m)}\|_A^2) .$$

Поэтому по аналогии с замечанием о выборе α_k

$$1/2(P_m^* A P_m y, y) - Re(P_m y, r^{(0)}) \;=$$
$$1/2 \sum_{i=1}^{m} (|y_i|^2 \|p^{(i)}\|_A^2 - y_i (p^{(i)}, r^{(0)}) - \overline{y_i} (r^{(0)}, p^{(i)})) \;=$$

$$\sum_{i=1}^{m} ((Re y_i)^2 (1/2 \|p^{(i)}\|_A^2) - (Re y_i) Re(p^{(i)}, r^{(0)}) + (Im y_i)^2 (1/2 \|p^{(i)}\|_A^2) - (Im y_i) Re(ip^{(i)}, r^{(0)})) \;=$$

$$\sum_{i=1}^{m} ((Re y_i)^2 (1/2 \|p^{(i)}\|_A^2) - (Re y_i) Re(r^{(0)}, p^{(i)}) + (Im y_i)^2 (1/2 \|p^{(i)}\|_A^2) - (Im y_i) Im(r^{(0)}, p^{(i)})) \;=$$

$$\sum_{i=1}^{m} \frac{\|p^{(i)}\|_A^2}{2} \Big(\Big(Re y_i - \frac{Re(r^{(0)}, p^{(i)})}{\|p^{(i)}\|_A^2} \Big)^2 + \Big(Im y_i - \frac{Im(r^{(0)}, p^{(i)})}{\|p^{(i)}\|_A^2} \Big)^2 - \frac{|(r^{(0)}, p^{(i)})|^2}{\|p^{(i)}\|_A^4} \Big) .$$

Следовательно, наша m-мерная минимизационная задача $Q(x^{(0)} + P_m y) \to \inf$ сводится к набору из m одномерных задач

$$\Big(Re y_i - \frac{Re(r^{(0)}, p^{(i)})}{\|p^{(i)}\|_A^2} \Big)^2 + \Big(Im y_i - \frac{Im(r^{(0)}, p^{(i)})}{\|p^{(i)}\|_A^2} \Big)^2 \to \inf \quad (i = 1, \ldots m)$$

и мы получаем

$$\mathrm{arginf}_{y \in \mathbb{C}^m} Q(x^{(0)} + P_m y) \;=\; \mathrm{arginf}_{y \in \mathbb{C}^m} (1/2(P_m^* A P_m y, y) - Re(P_m y, r^{(0)})) \;=$$
$$y^{(m)} \;=\; \Big(\frac{(r^{(0)}, p^{(1)})}{\|p^{(1)}\|_A^2}, \ldots, \frac{(r^{(0)}, p^{(m)})}{\|p^{(m)}\|_A^2} \Big)^t \;=\; (P_m^* A P_m)^{-1} P_m^* r^{(0)} .$$

Кроме того, мы немедленно получаем, что $x^{(m)} = x^{(0)} + P_m y^{(m)}$ и

$$P_m^* r^{(m)} \;=\; P_m^*(b - Ax^{(m)}) \;=\; P_m^*(r^{(0)} - AP_m y^{(m)}) \;=\; P_m^* r^{(0)} - P_m^* r^{(0)} \;=\; 0\,.$$

Доказательство второго пункта мы проведём при помощи индукции по j с очевидным основанием $\langle r^{(0)}\rangle = \langle p^{(1)}\rangle$. Пусть для некоторого $1 \leq j < k$ уже доказано, что

$$\langle p^{(1)}, \ldots, p^{(j)}\rangle \;=\; \langle r^{(0)}, \ldots, r^{(j-1)}\rangle \;=\; K_j(A, r^{(0)}) \;=\; \langle r^{(0)}, Ar^{(0)}, \ldots, A^{j-1}r^{(0)}\rangle\,.$$

Тогда $r^{(j)} = r^{(j-1)} - \alpha_k A p^{(j)} \in K_{j+1}(A, r^{(0)})$, поскольку $r^{(j-1)}, p^{(j)} \in K_j(A, r^{(0)})$. Вместе с тем $p^{(j+1)} = r^{(j)} - AP_j z^{(j)} \in K_{j+1}(A, r^{(0)})$, так как $r^{(j)} \in K_{j+1}(A, r^{(0)})$ и $AP_j\mathbb{C}^j \subseteq AK_{j-1}(A, r^{(0)}) \subseteq K_{j+1}(A, r^{(0)})$. Таким образом, $\langle r^{(0)}, \ldots, r^{(j)}\rangle$ и $\langle p^{(1)}, \ldots, p^{(j+1)}\rangle$ входят в крыловское подпространство $K_{j+1}(A, r^{(0)})$ размерности не выше $j+1$. Так как $\dim\langle p^{(1)}, \ldots, p^{(j+1)}\rangle = j+1$, отсюда сразу следует, что $\langle p^{(1)}, \ldots, p^{(j+1)}\rangle = K_{j+1}(A, r^{(0)})$. Остаётся заметить, что по доказанному $r^{(j)} \perp \langle p^{(1)}, \ldots, p^{(j)}\rangle = \langle r^{(0)}, \ldots, r^{(j-1)}\rangle$, а потому $\dim\langle r^{(0)}, \ldots, r^{(j)}\rangle = j+1$ и $\langle r^{(0)}, \ldots, r^{(j)}\rangle = K_{j+1}(A, r^{(0)})$. Тем самым шаг индукции полностью доказан. Заметим, что по ходу этого рассуждения мы доказали и взаимную ортогональность невязок, $r^{(i)} \perp r^{(j)}$, $i \neq j$. □

Теперь мы в состоянии вывести описание вычислительной процедуры метода сопряжённых градиентов.

В первую очередь необходимо показать, что k-ый сопряжённый градиент $p^{(k)}$ выражается через $k-1$-ые сопряжённый градиент $p^{(k-1)}$ и невязку $r^{(k-1)}$. Случай $k = 1$ очевиден ($p^{(1)} = r^{(0)}$) и можно считать, что $k > 1$. Напомним, что $p^{(k)} = r^{(k-1)} - AP_{k-1}z^{(k-1)} \perp_A P_{k-1}\mathbb{k}^{k-1}$ (для большей общности будем считать $\mathbb{k} = \mathbb{C}$) и $r^{(k-1)} = r^{(k-2)} - \alpha_{k-1}Ap^{(k-1)}$. Поэтому, полагая $z^{(k-1)} = (\nu, \mu)^t$, $\nu \in \mathbb{C}^{k-2}$, $\mu \in \mathbb{C}$, мы можем записать

$$p^{(k)} \;=\; r^{(k-1)} - AP_{k-2}\nu - \mu Ap^{(k-1)} \;=\; r^{(k-1)} - AP_{k-2}\nu - \frac{\mu}{\alpha_{k-1}}(r^{(k-2)} - r^{(k-1)}) \;=$$
$$\Big(1 + \frac{\mu}{\alpha_{k-1}}\Big)r^{(k-1)} + \Big(-AP_{k-2}\nu - \frac{\mu}{\alpha_{k-1}}r^{(k-2)}\Big)\,,$$

что возможно, поскольку $\alpha_{k-1} \neq 0$ (иначе $r^{(k-2)} \perp p^{(k-1)}$). При $k = 2$ составляющая ν и связанные с ней слагаемые опускаются. Так как

$$AP_{k-2}\nu \in A\langle r^{(0)}, \ldots, r^{(k-3)}\rangle \;=\; AK_{k-2}(A, r^{(0)}) \subseteq$$
$$K_{k-1}(A, r^{(0)}) \;=\; \langle r^{(0)}, \ldots, r^{(k-2)}\rangle \subseteq \langle r^{(k-1)}\rangle^{\perp}\,,$$

полученная нами запись $p^{(k)}$ есть в точности его разложение в сумму ортогональных проекций на подпростанства $\langle r^{(k-1)}\rangle$ и $\langle r^{(k-1)}\rangle^{\perp}$. Кроме того,

$$\|p^{(k)}\|_2 \;=\; \Big(\Big|1 + \frac{\mu}{\alpha_{k-1}}\Big|^2\|r^{(k-1)}\|_2^2 + \Big\|AP_{k-2}\nu + \frac{\mu}{\alpha_{k-1}}r^{(k-2)}\Big\|_2^2\Big)^{1/2} \;=$$
$$\min_{z\in\mathbb{C}^{k-1}}\|r^{(k-1)} - AP_{k-1}z\|_2 \;=$$
$$\min_{\nu'\in\mathbb{C}^{k-2},\ \mu'\in\mathbb{C}}\Big(\Big|1 + \frac{\mu'}{\alpha_{k-1}}\Big|^2\|r^{(k-1)}\|_2^2 + \Big\|AP_{k-2}\nu' + \frac{\mu'}{\alpha_{k-1}}r^{(k-2)}\Big\|_2^2\Big)^{1/2}\,.$$

Следовательно, если бы выполнялось равенство $\mu = 0$, то

$$\|p^{(k)}\|_2 \;=\; \min_{\nu' \in \mathbb{C}^{(k-2)}} (\|r^{(k-1)}\|_2^2 + \|AP_{k-2}\nu'\|_2)^{1/2} \;=\; \|r^{(k-1)}\|_2\,,$$

что соответствует случаю $\nu = 0$, т.е. $p^{(k)} = r^{(k-1)}$. В таком случае $r^{(k-1)} = p^{(k)} = r^{(k-2)} - \alpha_{k-1}Ap^{(k-1)} \perp r^{(k-2)}, Ap^{(k-1)}$ и значит, $\|r^{(k-1)}\|_2^2 = 0$, а эта ситуация нами изначально была исключена из рассмотрения (мы предполагаем осуществимость всех шагов с первого по k-ый). Поэтому можно считать $\mu \neq 0$ и записать

$$p^{(k)} \;=\; \Big(1 + \frac{\mu}{\alpha_{k-1}}\Big) r^{(k-1)} - \frac{\mu}{\alpha_{k-1}}(r^{(k-2)} + AP_{k-2}\xi)\,,$$

$$\|p^{(k)}\|_2 \;=\; \min_{\xi' \in \mathbb{C}^{k-2},\ 0 \neq \mu' \in \mathbb{C}} \Big(\Big|1 + \frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-1)}\|_2^2 + \Big|\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-2)} + AP_{k-2}\xi'\|_2^2 \Big)^{1/2}.$$

Так как $\min_{\xi' \in \mathbb{C}^{k-2}} \|r^{(k-2)} + AP_{k-2}\xi'\|_2 = \|p^{(k-1)}\|_2$ и достигается при одном единственном $\xi' = -z^{(k-2)}$, при всех $\xi' \in \mathbb{C}^{k-2}$ и $0 \neq \mu' \in \mathbb{C}$

$$\Big|1+\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-1)}\|_2^2 + \Big|\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-2)} + AP_{k-2}\xi'\|_2^2 \geq \Big|1+\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-1)}\|_2^2 + \Big|\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|p^{(k-1)}\|_2^2$$

и, следовательно, $\xi = -z^{(k-2)}$,

$$p^{(k)} \;=\; \Big(1 + \frac{\mu}{\alpha_{k-1}}\Big) r^{(k-1)} - \frac{\mu}{\alpha_{k-1}} p^{(k-1)}\,,$$

где выбор μ определяется из условия

$$\|p^{(k)}\|_2^2 \;=\; \Big(\Big|1 + \frac{\mu}{\alpha_{k-1}}\Big|^2 \|r^{(k-1)}\|_2^2 + \Big|\frac{\mu}{\alpha_{k-1}}\Big|^2 \|p^{(k-1)}\|_2^2 \Big)^{1/2} =$$

$$\min_{0 \neq \mu' \in \mathbb{C}} \Big(\Big|1 + \frac{\mu'}{\alpha_{k-1}}\Big|^2 \|r^{(k-1)}\|_2^2 + \Big|\frac{\mu'}{\alpha_{k-1}}\Big|^2 \|p^{(k-1)}\|_2^2 \Big)^{1/2} =$$

$$\min_{0 \neq \mu' \in \mathbb{C}} \Bigg((\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2) \Big(Re\frac{\mu'}{\alpha_{k-1}} \Big)^2 + 2\|r^{(k-1)}\| Re\frac{\mu'}{\alpha_{k-1}} + \|r^{(k-1)}\|_2^2 +$$

$$(\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2) \Big(Im\frac{\mu'}{\alpha_{k-1}} \Big)^2 \Bigg)^{1/2}.$$

Поэтому

$$Im\frac{\mu}{\alpha_{k-1}} \;=\; 0\,, \quad \frac{\mu}{\alpha_{k-1}} \;=\; Re\frac{\mu}{\alpha_{k-1}} \;=\; -\frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2}\,.$$

В итоге мы получаем, что

$$p^{(k)} \;=\; \frac{\|p^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2} r^{(k-1)} + \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2} p^{(k-1)} \;=\; \gamma_k (r^{(k-1)} + \beta_k p^{(k-1)})\,,$$

где $\gamma_k = \frac{\|p^{(k-1)}\|_2^2}{\|r^{(k-1)}\|_2^2 + \|p^{(k-1)}\|_2^2}$ и $\beta_k = \frac{\|r^{(k-1)}\|_2^2}{\|p^{(k-1)}\|_2^2}$. Кроме того, поскольку $p^{(k)} \perp_A p^{(k-1)}$,

$$0 \;=\; (r^{(k-1)} + \beta_k p^{(k-1)}, p^{(k-1)})_A \;=\; (r^{(k-1)}, p^{(k-1)})_A + \beta_k \|p^{(k-1)}\|_A^2\,,$$

и потому

$$\beta_k \;=\; \frac{\|r^{(k-1)}\|_2^2}{\|p^{(k-1)}\|_2^2} \;=\; -\frac{(r^{(k-1)},p^{(k-1)})_A}{\|p^{(k-1)}\|_A^2}\,.$$

Поскольку для нас значимым является выбор направления (сопряжённого градиента) с точностью до умножения на положительное действительное число, которое может быть включено в коэффициент сдвига, имеет смысл переписать выведенную нами здесь вычислительную схему метода сопряжённых градиентов в форме свободной от коэффициентов $\{\gamma_k\}$. Точнее, полагая $\tilde p^{(1)} = p^{(1)} = r^{(0)}$ и $\tilde p^{(k)} = \gamma_k^{-1} p^{(k)}$ для всех $k \ge 2$, мы можем записать

$$\begin{aligned}
\tilde p^{(k)} \;&=\; r^{(k-1)} + (\beta_k\gamma_{k-1})\tilde p^{(k-1)} \;=\; r^{(k-1)} + \tilde\beta_k \tilde p^{(k-1)}\,,\\
x^{(k)} \;&=\; x^{(k-1)} + (\alpha_k\gamma_k)\tilde p^{(k)} \;=\; x^{(k-1)} + \tilde\alpha_k \tilde p^{(k)}\,,\\
r^{(k)} \;&=\; r^{(k-1)} - (\alpha_k\gamma_k) A\tilde p^{(k)} \;=\; r^{(k-1)} - \tilde\alpha_k A\tilde p^{(k)}\,,
\end{aligned}$$

где, с учётом того, что $\gamma_k > 0$,

$$\begin{aligned}
\tilde\beta_k \;&=\; \beta_k\gamma_{k-1} \;=\; -\frac{(r^{(k-1)},\gamma_{k-1}p^{(k-1)})_A}{\|p^{(k-1)}\|_A^2} \;=\; -\frac{(r^{(k-1)},\tilde p^{(k-1)})_A}{\|\tilde p^{(k-1)}\|_A^2}\,,\\
\tilde\alpha_k \;&=\; \alpha_k\gamma_k \;=\; \frac{(r^{(k-1)},\gamma_k p^{(k)})}{\|p^{(k)}\|_A^2} \;=\; \frac{(r^{(k-1)},\tilde p^{(k)})}{\|\tilde p^{(k)}\|_A^2}\,.
\end{aligned}$$

Таким образом, после указанной замены вычислительная схема остаётся прежней: выбираем начальное приближение $x^{(0)}$ и далее вычисляем для всех $k \ge 1$

$$\tilde p^{(k)} \;=\; r^{(k-1)} + \tilde\beta_k \tilde p^{(k-1)}\,,\quad x^{(k)} \;=\; x^{(k-1)} + \tilde\alpha_k \tilde p^{(k)}\,,\quad r^{(k)} \;=\; r^{(k-1)} - \tilde\alpha_k A\tilde p^{(k)}\,,$$

где $\tilde\beta_1 = 0$, $\tilde\beta_k = -\frac{(r^{(k-1)},\tilde p^{(k-1)})_A}{\|\tilde p^{(k-1)}\|_A^2}$ при $k > 1$ и $\tilde\alpha_k = \frac{(r^{(k-1)},\tilde p^{(k)})}{\|\tilde p^{(k)}\|_A^2}$ при $k \ge 1$ (в частности, $\tilde p^{(1)} = r^{(0)} = b - Ax^{(0)}$ и $\tilde\alpha_1 = \alpha_1$).

Заметим, что перевычисление коэффициента β_k связано именно с необходимостью сохранения формы вычислительной процедуры.

Конечно на этом месте можно было бы и остановиться, но можно и продолжить преобразования с целью получения более компактной записи вычислительной процедуры. Поступим следующим образом: используя взаимную ортогональность векторов $r^{(k-1)}$ и $\tilde p^{(k-1)}$, $\tilde p^{(k-1)} \in \langle r^{(0)},\dots,r^{(k-2)}\rangle$, мы можем записать

$$\tilde\alpha_k \;=\; \frac{(r^{(k-1)},\tilde p^{(k)})}{\|\tilde p^{(k)}\|_A^2} \;=\; \frac{(r^{(k-1)},r^{(k-1)} + \tilde\beta_k\tilde p^{(k-1)})}{\|\tilde p^{(k)}\|_A^2} \;=\; \frac{\|r^{(k-1)}\|_2^2}{\|\tilde p^{(k)}\|_A^2}\,.$$

Далее, с учётом ортогональности невязок и A–сопряжённости направлений (а также того, что $\tilde\alpha_k > 0$) мы имеем

$$\begin{aligned}
\|r^{(k-1)}\|_2^2 \;&=\; (r^{(k-2)} - \tilde\alpha_{k-1}A\tilde p^{(k-1)}, r^{(k-1)}) \;=\\
&\qquad - \tilde\alpha_{k-1}(\tilde p^{(k-1)}, r^{(k-1)})_A \;=\; -\tilde\alpha_{k-1}(r^{(k-1)},\tilde p^{(k-1)})_A\,,\\
\|r^{(k-2)}\|_2^2 \;&=\; (r^{(k-1)} + \tilde\alpha_{k-1}A\tilde p^{(k-1)}, r^{(k-2)}) \;=\; \tilde\alpha_{k-1}(\tilde p^{(k-1)}, r^{(k-2)})_A \;=\\
&\qquad \tilde\alpha_{k-1}(\tilde p^{(k-1)}, \tilde p^{(k-1)} - \tilde\beta_{k-1}\tilde p^{(k-2)})_A \;=\; \tilde\alpha_{k-1}\|\tilde p^{(k-1)}\|_A^2\,,
\end{aligned}$$

а потому, как следствие, при $k > 1$

$$\tilde{\beta}_k \;=\; -\frac{(r^{(k-1)},\tilde{p}^{(k-1)})_A}{\|\tilde{p}^{(k-1)}\|_A^2} \;=\; \frac{\|r^{(k-1)}\|_2^2}{\|r^{(k-2)}\|_2^2}\,.$$

В итоге мы приходим к следующей классической процедуре метода сопряжённых градиентов из работы Хестенса и Штифеля 1952 г.: выбираем начальное приближение $x^{(0)}$, вычисляем первое направление $\tilde{p}^{(1)} = r^{(0)} = b - Ax^{(0)}$, а затем для всех $k \geq 1$ находим

1. $z = A\tilde{p}^{(k)}$;
2. $\tilde{\alpha}_k = \frac{\|r^{(k-1)}\|_2^2}{(z,\tilde{p}^{(k)})}$;
3. $x^{(k)} = x^{(k-1)} + \tilde{\alpha}_k\tilde{p}^{(k)}$;
4. $r^{(k)} = r^{(k-1)} - \tilde{\alpha}_k z$;
5. $\tilde{\beta}_{k+1} = \frac{\|r^{(k)}\|_2^2}{\|r^{(k-1)}\|_2^2}$;
6. $\tilde{p}^{(k+1)} = r^{(k)} + \tilde{\beta}_{k+1}\tilde{p}^{(k)}$.

Условие остановки процесса состоит в проверке условия $\|r^{(k)}\|_2 < \varepsilon$ для заданного $\varepsilon > 0$. Подчеркнём, что экономность данной процедуры в сравнении с предыдущей связана с тем, что в ней при реализации очередного шага следует сохранять значение евклидовой нормы невязки, полученной на предыдущем шаге, и не требуется перевычислять дополнительно скалярные произведения невязки и направления. При этом вычисление нормы невязки само по себе необходимо для проверки условия выхода из процесса.

Уместно будет привести и другие способы организации вычислений в методе сопряжённых градиентов. Начнём с *трёхчленной рекуррентной формы* их организации, которая базируется на следующем простом соображении: при $k > 1$

$$\tilde{p}^{(k)} \;=\; r^{(k-1)} + \tilde{\beta}_k\tilde{p}^{(k-1)} \;=\; b - Ax^{(k-1)} + \frac{\tilde{\beta}_k}{\tilde{\alpha}_{k-1}}(x^{(k-1)} - x^{(k-2)})$$

и, следовательно,

$$x^{(k)} \;=\; x^{(k-1)} + \tilde{\alpha}_k\tilde{p}^{(k)} \;=\; x^{(k-1)} + \tilde{\alpha}_k\Big(b - Ax^{(k-1)} + \frac{\tilde{\beta}_k}{\tilde{\alpha}_{k-1}}(x^{k-1} - x^{k-2})\Big) \;=$$
$$\Big(\Big(1 + \frac{\tilde{\alpha}_k\tilde{\beta}_k}{\tilde{\alpha}_{k-1}} - \tilde{\alpha}_k\Big)E + \tilde{\alpha}_k R\Big)x^{(k-1)} - \frac{\tilde{\alpha}_k\tilde{\beta}_k}{\tilde{\alpha}_{k-1}}x^{(k-2)} + \tilde{\alpha}_k b\,,$$

где $R = E - A$. Поэтому, полагая $\rho_k = 1 + \frac{\tilde{\alpha}_k\tilde{\beta}_k}{\tilde{\alpha}_{k-1}}$ и $\delta_k = \frac{\tilde{\alpha}_k}{\rho_k}$ (заметим, что последнее возможно в силу установленных ранее равенств для $\tilde{\alpha}_k$ и $\tilde{\beta}_k$), мы получаем $x^{(1)} = x^{(0)} + \tilde{\alpha}_1(b - Ax^{(0)}) = \tilde{\alpha}_1(Rx^{(0)} + b) + (1 - \tilde{\alpha}_1)x^{(0)}$ и

$$x^{(k)} \;=\; \rho_k(\delta_k(Rx^{(k-1)} + b) + (1 - \delta_k)x^{(k-1)}) + (1 - \rho_k)x^{(k-2)} \quad (k > 1)\,.$$

Другими словами, в такой записи метод сопряжённых градиентов может быть интерпретирован как процесс полиномиального ускорения стационарного итерационного

процесса $y^{(k)} = Ry^{(k-1)} + b$, использующий систему полиномов $\{p_i\}_{i\geq 0}$, в которой

$$\begin{aligned} p_0 &= 1\,, \quad p_1(t) = \tilde{\alpha}_1 t + 1 - \tilde{\alpha}_1\,, \\ p_{i+1}(t) &= \rho_{i+1}(\delta_{i+1}t + 1 - \delta_{i+1})p_i(t) + (1-\rho_{i+1})p_{i-1}(t) \quad (i \geq 1)\,. \end{aligned}$$

При этом $x^{(k)} - x = p_k(R)(x^{(0)} - x)$ и $\|x^{(k)} - x\|_2 \leq \rho(p_k(R))\|x^{(0)} - x\|_2$, $k \geq 0$ (см. ранее). По этой причине метод сопряжённых градиентов в ряде источников называют *ускорением по методу сопряжённых элементов* процесса $y^{(k)} = Ry^{(k-1)} + b$ с симметризуемой матрицей R.

Ещё одной стандартной формой организации вычислений в методе сопряжённых градиентов является так называемая *полиномиальная* форма. Её идея состоит в следующем. Поскольку $\tilde{p}^{(k)} \in K_k(A, r^{(0)}) = \langle r^{(0)}, Ar^{(0)}, \ldots, A^{k-1}r^{(0)}\rangle = \langle r^{(0)}, \ldots, r^{(k-1)}\rangle$, $k \geq 1$, мы можем записать $\tilde{p}^{(k)} = \psi_k(A)r^{(0)}$ и $r^{(k)} = \phi_k(A)r^{(0)}$ для некоторых многочленов ψ_k, $\deg\psi_k = k-1$, и ϕ_k, $\deg\phi_k = k$. При этом должны выполняться соотношения: $\psi_1 = \phi_0 = 1$ и

$$\begin{aligned} x^{(k)} &= x^{(k-1)} + \tilde{\alpha}_k\psi_k(A)r^{(0)}\,, \\ \tilde{p}^{(k)} &= (\phi_{k-1}(A) + \tilde{\beta}_k\psi_{k-1}(A))r^{(0)} = \psi_k(A)r^{(0)}\,, \\ r^{(k)} &= (\phi_{k-1}(A) - \tilde{\alpha}_k A\psi_k(A))r^{(0)} = \phi_k(A)r^{(0)}\,, \end{aligned}$$

причём, полагая $(\phi,\psi) = (\phi(A)r^{(0)}, \psi(A)r^{(0)}) = (r^{(0)})^*\psi(A)^*\phi(A)r^{(0)}$ для любых полиномов ϕ и ψ, можно также записать

$$\tilde{\alpha}_k = \frac{\|\phi_{k-1}\|^2}{(\psi_{k-1}, t\psi_{k-1})}\,, \quad \tilde{\beta}_k = \frac{\|\phi_{k-1}\|^2}{\|\phi_{k-2}\|^2}\,.$$

Так как коэффициенты $\tilde{\alpha}_k$ и $\tilde{\beta}_k$ являются вещественными, мы можем считать, что полиномы ψ_k, $k \geq 1$, и ϕ_k, $k \geq 0$, связанные между собой равенствами $\psi_1 = \phi_0 = 1$ и

$$\psi_k = \phi_{k-1} + \tilde{\beta}_k\psi_{k-1}\,, \quad \phi_k = \phi_{k-1} - \tilde{\alpha}_k t\psi_k \quad (k \geq 1)$$

являются вещественными. Кроме того, сказанное позволяет реализовать метод сопряжённых градиентов как процесс последовательного построения полиномов ψ_k и ϕ_k с условием выхода $\|\phi_k\| < \varepsilon$.

Связь с алгоритмами Ланцоша и Арнольди

Для матрицы $A = A^*$ и ненулевого вектора q метод Ланцоша строит матрицу Q размера $n \times m$ с ортонормированными столбцами, составляющими базис простроанства $K(A,q)$, $\dim K(A,q) = m$, для которой $AQ = QT$, где T — вещественная симметрическая трёхдиагональная матрица размера $m \times m$ вида

$$T = \begin{pmatrix} a_1 & b_1 & 0 & \ldots & 0 & 0 \\ b_1 & a_2 & b_2 & \ldots & 0 & 0 \\ 0 & b_2 & a_3 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_{m-1} & b_{m-1} \\ 0 & 0 & 0 & \ldots & b_{m-1} & a_m \end{pmatrix}.$$

Процесс построения $Q = (q_1 \dots q_m)$ организуется следующим образом: полагаем $q_0 = 0$, $b_0 = 0$ и $q_1 = q/\|q\|_2$, затем для всех $j \geq 1$ вычисляем

1. $z = Aq_j$, $a_j = q_j^* z$;
2. $z := z - a_j q_j - b_{j-1} q_{j-1}$, $b_j = \|z\|_2$, причём в случае $b_j = 0$ процесс останавливается, в противном случае полагаем $q_{j+1} = z/b_j$.

Заметим, что к описанной вычислительной процедуре можно без труда прийти на основе анализа результата построения. Сходным образом в несимметрическом случае организуется процесс построения матрицы Q для верхней хессенберговой матрицы T. Данная версия алгоритма Ланцоша носит название алгоритма Арнольди.

Обозначим теперь через $R_k = (r^{(0)} \dots r^{(k-1)})$ и $\tilde{P}_k = (\tilde{p}^{(1)} \dots \tilde{p}^{(k)})$ матрицы, столбцами которых служат невязки и сопряжённые градиенты, полученные в результате выполнения первых $k-1$ шагов метода сопряжённых градиентов (при условии их осуществимости), и через B_k — двухдиагональную верхнюю унитреугольную матрицу размера $k \times k$ вида

$$B_k = \begin{pmatrix} 1 & -\tilde{\beta}_2 & 0 & \dots & 0 & 0 \\ 0 & 1 & -\tilde{\beta}_3 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -\tilde{\beta}_k \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Тогда равенства $\tilde{p}^{(1)} = r^{(0)}$ и $\tilde{p}^{(i)} - \tilde{\beta}_i \tilde{p}^{(i-1)} = r^{(i-1)}$, $i = 2, \dots, k$, могут быть записаны в матричном виде как $R_k = \tilde{P}_k B_k$. Вместе с тем из A–сопряжённости направлений $\{\tilde{p}^{(i)}\}$ следует, что матрица

$$R_k^* A R_k = B_k^* \operatorname{diag}(\|\tilde{p}^{(1)}\|_A^2, \dots \|\tilde{p}^{(k)}\|_A^2) B_k$$

является вещественной, симметрической и трёхдиагональной. Полагая $R'_k = R_k \Delta_k^{-1}$, где $\Delta_k = \operatorname{diag}(\|r^{(0)}\|_2, \dots, \|r^{(k-1)}\|_2)$, мы получим матрицу с ортонормированными столбцами, составляющими ортонормированный базис подпространства $K_k(A, r^{(0)})$. При этом по предыдущему матрица $T_k = (R'_k)^* A R_k$ самосопряжена и трёхдиагональна. Следовательно, метод сопряжённых градиентов может быть интерпретирован как процесс последовательного построения ортонормированного базиса пространства $K_k(A, r^{(0)})$, в котором матрица A принимает трёхдиагональный вид. Другими словами, ортонормированные невязки $\|r^{(i)}\|_2^{-1} r^{(i)}$, $i = 0, \dots, k-1$, могут рассматриваться как *вектора Ланцоша* пространства $K_r(A, r^{(0)})$ из алгоритма Ланцоша частичной трёхдиагонализации матрицы A в рамках указанного подпространства.

Сходимость метода сопряжённых градиентов

Обсуждение сходимости метода сопряжённых градиентов мы начнём со следующего простого замечания.

Замечание 0.52. *Пусть* $B \in M_n(\mathbb{C})$, $\operatorname{rk} B = r$. *Тогда* $\dim K(B, x) \leq r + 1$ *для любого* $x \in \mathbb{C}^n$.

Доказательство. Достаточно заметить, что $K(B, x) \subseteq \mathbb{C}x + B\mathbb{C}^n$ и, как следствие,

$\dim K(B,x) \le 1 + \dim\langle b_1,\dots,b_n\rangle = 1+r$, где $B = (b_1 \dots b_n)$. □

Заметим, что данная оценка является точной. В частности, если B — нильпотентная матрица индекса n, тогда для всякого вектора x, $B^{n-1}x \ne 0$, вектора $x, Bx, \dots, B^{n-1}x$ линейно независимы и, как следствие, $\operatorname{rk} B = n-1$.

Применительно к малоранговым возмущениям единичной матрицы это даёт нам

Следствие 0.53. *Если $A = A^* = E + B > 0$, $\operatorname{rk} B = r$, тогда метод сопряжённых градиентов сходится при любом начальном приближении за $r+1$ шаг.*

Доказательство. Достаточно заметить, что

$$\dim K(A,r) \;=\; \dim\langle r, Br, \dots, B^{n-1}r\rangle \le r+1\,.$$

Поэтому метод сопряжённых градиентов гарантировано завершится на $r+1$ шаге (так как построение $p^{(r+2)}$ не может быть осуществлено). □

Теорема 0.54. *На k-ом шаге метода сопряжённых градиентов*

$$\|r^{(k)}\|_{A^{-1}} \le 2\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^k \|r^{(0)}\|_{A^{-1}}\,,$$

$$\|r^{(k)}\|_2 \le 2\sqrt{k_2(A)}\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^k \|r^{(0)}\|_2\,.$$

Доказательство. В методе сопряжённых градиентов $x^{(k)} \in x^{(0)} + K_k(A, r^{(0)})$, причём

$$x^{(k)} \;=\; \operatorname{arginf}_{x' \in x^{(0)}+K_k(A,r^{(0)})} Q(x')\,.$$

Поскольку $\|b - Az\|^2_{A^{-1}} = 2Q(z) + \|b\|^2_{A^{-1}}$, мы можем также записать

$$x^{(k)} \;=\; x^{(0)} + \operatorname{arginf}_{y\in K_k(A,r^{(0)})} \|b - A(x^{(0)}+y)\|^2_{A^{-1}} \;=\; x^{(0)} + \operatorname{arginf}_{y\in K_k(A,r^{(0)})} \|r^{(0)} - Ay\|^2_{A^{-1}}\,.$$

Любой элемент $y \in K_k(A, r^{(0)})$ имеет вид $y = p(A)r^{(0)}$ для некоторого многочлена p, $\deg p \le k-1$, и, следовательно, для $\hat{x} = A^{-1}r^{(0)}$

$$\|r^{(0)} - Ay\|^2_{A^{-1}} \;=\; \|(E - p(A)A)r^{(0)}\|^2_{A^{-1}} \;=\; \|(E - p(A)A)\hat{x}\|^2_A\,.$$

Напомним также, что $x^{(k)} = x^{(0)} + \sum_{i=1}^k \tilde{\alpha}_i \psi_i(A) r^{(0)}$, где $\tilde{\alpha}_i > 0$ и $\psi_i \in \mathbb{R}[t]$, $\deg \psi_i = i-1$, и потому решение нашей минимизационной задачи $y = p_k(A)r^{(0)}$ отвечает вещественному многочлену $p_k = \sum_{i=1}^k \tilde{\alpha}_i \psi_i$, $\deg p_k = k-1$. При этом

$$\|(E - p_k(A)A)\hat{x}\|^2_A \;=\; \|q_k(A)\hat{x}\|^2_A \;=\; \min_{\substack{q\in\mathbb{C}[t],\\ \deg q\le k,\ q(0)=1}} \|q(A)\hat{x}\|^2_A \;=$$

$$\min_{\substack{q\in\mathbb{C}[t],\\ \deg q = k,\ q(0)=1}} \|q(A)\hat{x}\|^2_A \;=\; \min_{\substack{q\in\mathbb{R}[t],\\ \deg q = k,\ q(0)=1}} \|q(A)\hat{x}\|^2_A\,,$$

где $q_k(t) = 1 - tp_k(t)$. В ортонормированном базисе собственных векторов матрицы A, составляющих унитарную матрицу U, матрица A примет диагональный вид, $A = U\Lambda U^*$,

$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, где $\{\lambda_i\} = \operatorname{Spec} A$. Это позволяет записать

$$\|q_k(A)\hat{x}\|_A^2 \;=\; \hat{x}^* q_k(A) A q_k(A) \hat{x} \;=\; \hat{x} U q_k(\lambda) \Lambda q_k(\Lambda) U^* \hat{x} \;=$$

$$\hat{y}^* \operatorname{diag}(\lambda_1 q_k(\lambda_1)^2, \ldots, \lambda_n q_k(\lambda_n)^2) \hat{y} \;=\; \sum_{i=1}^{n} \lambda_i q_k(\lambda_i)^2 |\hat{y}_i|^2 \;=\; \min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \sum_{i=1}^{n} \lambda_i q(\lambda_i)^2 |\hat{y}_i|^2 \leq$$

$$\left(\min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \max_{i=1,\ldots,n} q(\lambda_i)^2 \right) \sum_{i=1}^{n} \lambda_i |\hat{y}_i|^2 \;=\; \left(\min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \left(\max_{\lambda \in \operatorname{Spec} A} |q(\lambda)| \right)^2 \right) \|\hat{x}\|_A^2 ,$$

для $\hat{y} = U\hat{x}$, $\hat{x} = q_0(A)\hat{x}$. Таким образом,

$$\|r^{(k)}\|_{A^{-1}} \;=\; \|q_k(A)\hat{x}\|_A \leq$$

$$\left(\min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \max_{\lambda \in \operatorname{Spec} A} |q(\lambda)| \right) \|r^{(0)}\|_{A^{-1}} \leq \left(\min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \|q\|_{C[m,M]} \right) \|r^{(0)}\|_{A^{-1}} ,$$

где $m = \lambda_{\min} = \min_{\lambda \in \operatorname{Spec} A} \lambda$ и $M = \lambda_{\max} = \max_{\lambda \in \operatorname{Spec} A} \lambda$, $0 < m \leq M$. Заметим, что в случае $m = M$ мы сразу получаем $r^{(1)} = 0$ ($\max_{\lambda \in \operatorname{Spec} A} |q(\lambda)| = 0$ для $q(t) = (m - t)/m$). Поэтому мы можем считать, что $m < M$. Тогда в силу сказанного ранее о многочленах Чебышева

$$\min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(0)=1}} \|q\|_{C[m,M]} \;=\; \min_{\substack{q \in \mathbb{R}[t], \\ \deg q = k,\ q(1)=1}} \|q\|_{C[m+1,M+1]} \;=$$

$$\frac{1}{\left| t_k\left(\frac{m+M}{m-M}\right) \right|} \;=\; \frac{1}{\left| t_k\left(\frac{m+M}{M-m}\right) \right|} \;=\; \frac{1}{\left| t_k\left(\frac{k_2(A)+1}{k_2(A)-1}\right) \right|} ,$$

где $k_2(A) = M/m$ (в данном случае мы воспользовались заменой $q(t)$ на $q(t-1)$).

Остаётся заметить, что

$$\frac{k_2(A)+1}{k_2(A)-1} \pm \sqrt{\left(\frac{k_2(A)+1}{k_2(A)-1}\right)^2 - 1} \;=\; \frac{k_2(A) + 1 \pm 2\sqrt{k_2(A)}}{k_2(A) - 1} \;=\; \frac{\sqrt{k_2(A)} \pm 1}{\sqrt{k_2(A)} \mp 1} ,$$

а потому

$$t_k\left(\frac{k_2(A)+1}{k_2(A)-1}\right) \;=\; \frac{1}{2}\left(\left(\frac{\sqrt{k_2(A)}+1}{\sqrt{k_2(A)}-1}\right)^k + \left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^k \right) \geq \frac{1}{2}\left(\frac{\sqrt{k_2(A)}+1}{\sqrt{k_2(A)}-1}\right)^k$$

и, следовательно,

$$\|r^{(k)}\|_{A^{-1}} \leq 2\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^k \|r^{(0)}\|_{A^{-1}} .$$

Переход к оценке в обычной евклидовой норме $\|\ \|_2$ осуществляется добавлением коэффициента $\sqrt{k_2(A)} = \sqrt{k_2(A^{-1})}$ (см. ранее). □

Заметим, что приведённая здесь оценка скорости сходимости метода сопряжённых градиентов фактически не может быть улучшена. Кроме того, из неё немедленно следует, что для плохо обусловленных матриц эта сходимость будет вообще говоря

весьма медленной. Тем не менее, известны случаи, когда структура спектра матрицы A обеспечивает быструю сходимость метода сопряжённых градиентов.

Замечание 0.55. *Если матрица $A = A^* > 0$ имеет m различных собственных значений, тогда метод сопряжённых градиентов сходится к точному решению за m шагов при любом начальном приближении.*

Доказательство. Действительно, если $\lambda_1, \ldots, \lambda_m$ — все различные собственные значения матрицы A, тогда $\max_{\lambda \in \operatorname{Spec} A} |q(\lambda)| = 0$ для многочлена $q(t) = \prod\limits_{i=1}^{m} \frac{\lambda_i - t}{\lambda_i} \in \mathbb{R}[t]$, $\deg q = m$, $q(0) = 1$, и, следовательно, $r^{(m)} = 0$ (см. выше). □

Стоит подчеркнуть, что используемое нами повсеместно условие выхода из вычислительной процедуры $\|r^{(k)}\|_2 < \varepsilon$ вполне оправдано, поскольку в этом случае $\|x^{(k)} - x\|_2 < \|A^{-1}\|_2 \|r^{(k)}\|_2 = \varepsilon \|A^{-1}\|_2 = \varepsilon / \lambda_{\min}$ и $\|x^{(k)} - x\|_2 = \|r^{(k)}\|_{A^{-2}}$.

Лекция 8. Предобусловленный метод сопряжённых градиентов.

Идея предобуславливания итерационного процесса $x^{(k)} = x^{(k-1)} + \alpha_k p^{(k)}$, $\alpha_k = \frac{(r^{(k-1)}, p^{(k)})}{\|p^{(k)}\|_A^2}$, $k \geq 1$, построенного для решения системы $Ax = b$ с $A = A^* > 0$ одним из описанных выше итерационных методов состоит в выборе невырожденной матрицы S и переходу к системе $\hat{A}\hat{x} = \hat{b}$, где $\hat{A} = SAS^*$, $\hat{b} = Sb$ и $S^*\hat{x} = x$, для которой построенный по аналогичным правилам итерационный процесс $\hat{x}^{(k)} = \hat{x}^{(k-1)} + \hat{\alpha}_k \hat{p}_k$, $\hat{\alpha}_k = \frac{(\hat{r}^{(k-1)}, \hat{p}^{(k)})}{\|\hat{p}^{(k)}\|_{\hat{A}}^2}$, $k \geq 1$, будет иметь более высокую скорость сходимости к $\hat{x} = (S^*)^{-1}x$. Поскольку нас интересует получение приближения к решению $x = A^{-1}b$, имеет реализовавать вычислительную процедуру построения не приближений $\{\hat{x}^{(k)}\}$, а приближений $\{\overline{x}^{(k)} = S^*\hat{x}^{(k)}\}$. Точнее, полагая $\overline{p}^{(k)} = S^*\hat{p}^{(k)}$, $k \geq 1$, и $\overline{r}^{(k)} = S^{-1}\hat{r}^{(k)}$, $k \geq 0$, мы можем записать процесс построения $\{\overline{x}^{(k)}\}$ следующим образом: после выбора начального приближения $\overline{x}^{(0)}$ строим $\overline{x}^{(k)} = \overline{x}^{(k-1)} + \hat{\alpha}_k \overline{p}^{(k)}$, $\overline{r}^{(k)} = \overline{r}^{(k-1)} - \hat{\alpha}_k A\overline{p}^{(k)}$, где

$$\hat{\alpha}_k \;=\; \frac{(S\overline{r}^{(k-1)}, (S^*)^{-1}\overline{p}^{(k)})}{\|(S^*)^{-1}\overline{p}^{(k)}\|_{\hat{A}}^2} \;=\; \frac{(\overline{r}^{(k-1)}, \overline{p}^{(k)})}{\|\overline{p}^{(k)}\|_A^2}\,.$$

При этом, как нетрудно заметить, схема вычислений идентична исходной с точностью до замены системы направлений.

Применительно к методу сопряжённых градиентов, реализуемому процедурой Хестенса — Штифеля, описанный нами процесс дополняется перевычислением коэффициента $\hat{\alpha}_k$

$$\hat{\alpha}_k \;=\; \frac{\|\hat{r}^{(k-1)}\|_2^2}{\|\hat{p}^{(k)}\|_{\hat{A}}^2} \;=\; \frac{\|\overline{r}^{(k-1)}\|_{S^*S}^2}{\|\overline{p}^{(k)}\|_A^2}$$

и сопряжённого градиента $\overline{p}^{(k)}$

$$\overline{p}^{(k)} \;=\; S^*S\overline{r}^{(k-1)} + \hat{\beta}_k \overline{p}^{(k-1)}\,, \quad \hat{\beta}_k \;=\; \frac{\|\hat{r}^{(k-1)}\|_2^2}{\|\hat{r}^{(k-2)}\|_2^2} \;=\; \frac{\|\overline{r}^{(k-1)}\|_{S^*S}^2}{\|\overline{r}^{(k-2)}\|_{S^*S}^2}\,.$$

Полагая $M = (S^*S)^{-1}$, мы приходим к следующей вычислительной процедуре: выбираем начальное приближение $\overline{x}^{(0)}$, вычисляем нулевую невязку $\overline{r}^{(0)} = b - A\overline{x}^{(0)}$ и далее для всех $k \geq 0$

1. решаем систему уравнений $Mz^{(k)} = \overline{r}^{(k)}$ (т.е. находим $z^{(k)} = S^*S\overline{r}^{(k)}$);
2. полагаем $\overline{\beta}_1 = 0$ и $\overline{\beta}_{k+1} = \hat{\beta}_{k+1} = \frac{(z^{(k)}, \overline{r}^{(k)})}{(z^{(k-1)}, \overline{r}^{(k-1)})}$ при $k \geq 1$;
3. находим $\overline{p}^{(k+1)} = z^{(k)} + \overline{\beta}_{k+1}\overline{p}^{(k)}$ ($\overline{p}^{(1)} = z^{(0)}$) и $y = A\overline{p}^{(k+1)}$;
4. вычисляем $\overline{\alpha}_{k+1} = \hat{\alpha}_{k+1} = \frac{(z^{(k)}, \overline{r}^{(k)})}{(y, \overline{p}^{(k+1)})}$;
5. находим $\overline{x}^{(k+1)} = \overline{x}^{(k)} + \overline{\alpha}_{k+1}\overline{p}^{(k+1)}$;
6. находим $\overline{r}^{(k+1)} = \overline{r}^{(k)} - \overline{\alpha}_{k+1} y$;
7. вычисляем $\|\overline{r}^{(k+1)}\|_2$ и проверяем условие продолжения процесса: $\|\overline{r}^{(k+1)}\|_2 \geq \varepsilon$.

Скорость сходимости данного процесса, который называется *предобусловленным методом сопряжённых градиентов с матрицей-предобуславливателем* M, определяется скоростью сходимости к нулю последовательности $\{\hat{r}^{(k)}\}_{k\geq 0}$, которая в силу сказанного ранее определяется величиной числа обусловленности

$$k_2(\hat{A}) \;=\; k_2(SAS^*) \;=\; \frac{\lambda_{\max}(SAS^*)}{\lambda_{\min}(SAS^*)} \;=\; \frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}\,.$$

В соответствии со сказанным выбор предобусловливателя $M = M^* > 0$ должен удовлетворять следующим двум условиям:

1. решение системы $Mz^{(k)} = \overline{r}^{(k)}$ должно иметь как можно более низкую вычислительную стоимость;
2. величина $k_2(\hat{A}) = \frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}$ должна быть как можно более близкой к 1.

Заметим, что условия 1 и 2 являются по сути взаимно исключающими, поскольку оптимальное выполнение первого условия соответствует $M = E$, а второго — $M = A$. Балансирование этих условий привело к появлению различных стратегий предобуславливания, среди которых одной самых естественных является предобуславливание с использованием усечённых рядов.

Предобуславливание при помощи усечённых рядов.

Пусть имеется расщепление $A = P - Q$ с невырожденной матрицей P, удовлетворяющее условию $\rho(H) < 1$, где $H = P^{-1}Q$. В таком случае ряд

$$A^{-1} \;=\; \sum_{k=0}^{\infty} H^k P^{-1}$$

сходится в подходящей операторной норме, а значит, и во всех матричных нормах, причём матрица P^{-1} может быть вынесена за скобки, поскольку это не влияет на значение суммы ряда. В качестве предобуславливателя предлагается использовать матрицу вида

$$M \;=\; P(E + H + \ldots + H^{m-1})^{-1}$$

для некоторого фиксированного $m \geq 1$. При этом обратная к ней матрица

$$M^{-1} \;=\; (E + H + \ldots + H^{m-1})P^{-1}$$

представляет собой частичную сумму ряда для A^{-1} и потому мы вправе расчитывать на уменьшение $\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}$ с ростом m и связанным с этим приближением $M^{-1}A$ к единичной матрице E.

Кроме того, такой выбор M сводит решение линейной системы $Mz = r$ к нахождению

$$z \;=\; M^{-1}r \;=\; (E + H + \ldots + H^{m-1})P^{-1}r \;=\; P^{-1}r + H(E + H + \ldots + H^{m-2})P^{-1}r\,,$$

что в свою очередь равносильно выполнению m шагов вспомогательного итерационного процесса: $r^{(0)} = 0$, $Pr^{(i+1)} = Qr^{(i)} + r$, $i = 0, \ldots, m-1$. Действительно, в указанном процессе $r^{(i+1)} = Hr^{(i)} + P^{-1}r = (E + H + \ldots + H^i)P^{-1}r$, $i = 0, \ldots, m-1$.

Для того, чтобы реализовать описанную нами идею, нам следует обеспечить выбор расщепления матрицы $0 < A = A^* = P - Q$, исходя из требований: $M = M^* > 0$ и $\rho(H) < 1$ для заданного m. Необходимую теоретическую основу для этого содержит в себе следующая серия вспомогательных замечаний.

Замечание 0.56. *Пусть имеются самосопряжённые матрицы $A = A^*$ и $B = B^*$, причём $x^*Bx \neq 0$ при всех $x \neq 0$ (в частности, это имеет место в случае $B > 0$). Тогда* $\operatorname{Spec} AB = \operatorname{Spec} BA \subset \mathbb{R}$.

Доказательство. Для любого собственного значения $\lambda \in \operatorname{Spec} AB$, $ABx_\lambda = \lambda x_\lambda$, $x_\lambda \neq 0$, мы можем записать

$$x_\lambda{}^* BABx_\lambda \;=\; \lambda x_\lambda{}^* Bx_\lambda \in \mathbb{R}.$$

Поскольку $0 \neq x_\lambda{}^* Bx_\lambda \in \mathbb{R}$, отсюда следует, что $\lambda \in \mathbb{R}$. □

Замечание 0.57. *Если $A = A^* > 0$ и $B = B^*$, тогда $B > 0$ в том и только в том случае, если* $\operatorname{Spec} AB = \operatorname{Spec} BA \subset \mathbb{R}_+ = \{r \in \mathbb{R} \mid r > 0\}$.

Доказательство. Если $B > 0$, тогда $BAB > 0$ и потому для всякого $\lambda \in \operatorname{Spec} AB$, $ABx_\lambda = \lambda x_\lambda$, $x_\lambda \neq 0$, мы имеем

$$x_\lambda{}^* BABx_\lambda \;=\; \lambda x_\lambda{}^* Bx_\lambda \in \mathbb{R}_+\,, \quad x_\lambda{}^* Bx_\lambda \in \mathbb{R}_+\,, \quad \lambda \in \mathbb{R}_+\,.$$

Пусть теперь $\operatorname{Spec} AB = \operatorname{Spec} BA \subset \mathbb{R}_+$. Поскольку $A = A^* > 0$, $A = LL^*$ для некоторой невырожденной матрицы L (последнее следует хотя бы из разложения Холесского). Тогда $AB = LL^*B$ и $\operatorname{Spec} AB = \operatorname{Spec} L^{-1}(AB)L = \operatorname{Spec} L^*BL \subset \mathbb{R}_+$, а значит, $L^*BL > 0$, $x^*L^*BLx > 0$ для всех $x \neq 0$, что в свою очередь равносильно $B > 0$. □

Замечание 0.58. *Для всякого расщепления $A = P - Q$ матрицы $A = A^* > 0$ с невырожденной матрицей $P = P^*$ матрица $H = P^{-1}Q$ имеет вещественный спектр,* $\operatorname{Spec} H \subset \mathbb{R}$.

Доказательство. По первому замечанию $\operatorname{Spec} P^{-1}A = \operatorname{Spec}(E - H) = 1 - \operatorname{Spec} H \subset \mathbb{R}$ и, как следствие, $\operatorname{Spec} H \subset \mathbb{R}$. □

Замечание 0.59. *В условиях предыдущего замечания $\rho(H) < 1$, если и только если расщепление $A = P - Q$ является P–регулярным ($P + Q > 0$). Поэтому в данном случае $P > 0$.*

Доказательство. Тот факт, что $\rho(H) < 1$ в случае P–регулярного расщепления матрицы $A = A^* = P - Q > 0$ без дополнительного условия $P = P^*$, был установлен нами ранее в рамках одного из замечаний.

Пусть теперь $\rho(H) < 1$. Поскольку $P = P^*$, из предыдущего замечания следует, что $\operatorname{Spec} H \subset \mathbb{R}$ и, более того, $\operatorname{Spec}(E \pm H) = 1 \pm \operatorname{Spec} H \subset \mathbb{R}_+$. Так как $E \pm H = P^{-1}(P \pm Q)$, $A = A^* = P - Q > 0$ и $\operatorname{Spec} P^{-1}A \subset \mathbb{R}_+$, из одного из предыдущих замечаний следует, что $P^{-1} > 0$ и $P > 0$. Остаётся применить это же замечание и включение $\operatorname{Spec} P^{-1}(P+Q) \subset \mathbb{R}_+$ и получить $P + Q > 0$. □

Теперь мы в состоянии доказать основной результат об интересующих нас предобуславливателях в форме усечённых рядов.

Теорема 0.60. *Пусть* $0 < A = A^* = P - Q$, $P = P^* \in Gl_n(\mathbb{C})$, $m \geq 1$ *и* $R = M^{-1} = (E + H + \ldots + H^{m-1})P^{-1}$, *где* $H = P^{-1}Q$. *Тогда* $R = R^*$ *при всех* m *и* $R > 0$, *если и только если* $P > 0$ *для нечётных* m *и* $P + Q > 0$ *для чётных* m. *Заметим, что в последнем случае также* $P > 0$ *и* $\rho(H) < 1$.

Доказательство. Для начала заметим, что

$$R \;=\; P^{-1} + P^{-1}QP^{-1} + \ldots + (P^{-1}Q)^{m-1}P^{-1}\,,$$

где $P = P^*$, $Q = Q^*$ и $((P^{-1}Q)^s P^{-1})^* = P^{-1}(QP^{-1})^s = ((P^{-1}Q)^s P^{-1})^*$ при любом $s \geq 0$. Поэтому $R = R^*$ и $M = M^*$.

Напомним также, что $\operatorname{Spec} H \subset \mathbb{R}$. Положим $\Delta = E + H + \ldots + H^{m-1}$. Тогда $\operatorname{Spec}\Delta = \{1 + \lambda + \ldots + \lambda^{m-1} \mid \lambda \in \operatorname{Spec} H\}$, где

$$1 + \lambda + \ldots + \lambda^{m-1} \;=\; \begin{cases} \frac{\lambda^m - 1}{\lambda - 1} & \text{при } \lambda \neq 1; \\ m & \text{при } \lambda = 1. \end{cases}$$

Поэтому при $m = 2k + 1$ мы имеем $\operatorname{Spec}\Delta \subset \mathbb{R}_+$. Так как $\Delta = RP$, $R = R^*$, $P = P^*$, из доказанного ранее замечания немедленно следует, что в этом случае $R > 0$, если и только если $P > 0$.

В случае $m = 2k$ мы можем записать

$$\begin{aligned} R \;&=\; P^{-1}(P + PH + \ldots + PH^{m-1})P^{-1} \;= \\ &P^{-1}((P + PH) + (P + PH)H^2 + \ldots + (P + PH)H^{m-2})P^{-1} \;=\; P^{-1}(P + Q)\tilde{\Delta}P^{-1}\,, \end{aligned}$$

где $\tilde{\Delta} = \sum\limits_{i=0}^{k-1} H^{2i}$. Поэтому $R = P^{-1}(P + Q)\tilde{\Delta}P^{-1} = (P^{-1})^*(P + Q)\tilde{\Delta}P^{-1} > 0$ тогда и только тогда, когда $(P + Q)\tilde{\Delta} = PRP = (PRP)^* > 0$. При этом интересующее нас условие $R > 0$ предполагает рассмотрение ситуации $P + Q \in Gl_n(\mathbb{C})$. Поскольку $(P + Q)^{-1}(PRP) = \tilde{\Delta}$ и

$$\operatorname{Spec}\tilde{\Delta} \;=\; \left\{ \sum_{i=0}^{k-1} \lambda^{2i} \,\middle|\, \lambda \in \operatorname{Spec} H \right\} \subset \mathbb{R}_+\,,$$

из сказанного ранее следует, что в данной ситуации $R > 0$ ($PRP > 0$), если и только если $P + Q > 0$. □

Применяя доказанное ранее (перед теоремой) замечание, мы сразу получаем

Следствие 0.61. *Если в условиях теоремы* $\rho(H) < 1$ *(т.е.* $P + Q > 0$*), то* $R > 0$ *при любом* m.

Таким образом, располагая P–регулярным расщеплением $A = A^* = P - Q$, $P = P^*$, мы всегда можем реализовать предобусловленный метод сопряжённых градиентов по описанной выше схеме, с использованием m–шагового вспомогательного итерационного процесса для решения системы $Mz = r$ (такой метод обычно называют методом сопряжённых градиентов с m–шаговым предобуславливанием по соответствующему стационарному итерационному методу). Наиболее популярными вариантами подобных методов являются метод сопряжённых градиентов с m–шаговым предобуславливанием по

методу Якоби и методу $SSOR(\omega)$, $0 < \omega < 2$. При этом в случае метода Якоби без дополнительного требования его сходимости при любом начальном приближении (и, в частности, без требований регулярности диагонального расщепления, которое обеспечивается, к примеру, строгим диагональным преобладанием) рассматриваемый метод может быть реализован вообще говоря только для нечётных m (при наложении указанных условий — при любом m). Для метода $SSOR(\omega)$, $0 < \omega < 2$, реализация возможна при всех m. Впрочем, следует оговориться, что в действительности m–шаговое предобуславливание используется для сравнительно небольших значений m.

Обсудим теперь качество уменьшения величины $\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}$ при использовании m–шаговых предобуславливателей. Записав

$$M^{-1}A \;=\; RA \;=\; (E+H+\ldots+H^{m-1})P^{-1}(P-Q) \;=\; (E+H+\ldots+H^{m-1})(E-H) \;=\; E-H^m\,,$$

мы сразу получаем, что $\operatorname{Spec} M^{-1}A = \{1-\lambda^m \mid \lambda \in \operatorname{Spec} H\}$. Поэтому применительно к интересующей нас ситуации $\operatorname{Spec} H \subset \mathbb{R}$, $\rho(H) < 1$ (см. выше)

$$\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)} \;=\; \begin{cases} \frac{1-\lambda_{\min}(H)^m}{1-\lambda_{\max}(H)^m}\,, & \text{если } \lambda_{\min}(H) \geq 0 \text{ или } \lambda_{\min}(H) < 0,\ m \equiv 1 \quad (2); \\ \\ \frac{1-\delta^m}{1-\lambda_{\max}(H)^m}\,, & \text{если } \lambda_{\min}(H) < 0,\ |\lambda_{\max}(H)| \geq |\lambda_{\min}(H)| \text{ и } m \equiv 0 \quad (2); \\ \\ \frac{1-\delta^m}{1-\lambda_{\min}(H)^m}\,, & \text{если } \lambda_{\min}(H) < 0,\ |\lambda_{\min}(H)| \geq |\lambda_{\max}(H)| \text{ и } m \equiv 0 \quad (2), \end{cases}$$

где $\delta = \min_{\lambda \in \operatorname{Spec} H} |\lambda|$.

Напомним, что в случае $\operatorname{Spec} R_{SSOR(\omega)} \subset (0,1)$ при $0 < \omega < 2$. Поэтому для m–шагового предобуславливания по методу $SSOR(\omega)$ реализуется только первая из указанных здесь возможностей

$$\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)} \;=\; \frac{1-\lambda_{\min}(R_{SSOR(\omega)})^m}{1-\lambda_{\max}(R_{SSOR(\omega)})^m}\,.$$

Аналогичным образом, для нечётных m и m–шагового предобуславливания по методу Якоби

$$\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)} \;=\; \frac{1-\lambda_{\min}(R_J)^m}{1-\lambda_{\max}(R_J)^m}\,.$$

Из этого соотношения сразу следует, что требование $\rho(R_J) < 1$ является в общем случае существенным для стремления указанной здесь величины к единице с ростом m.

Полиномиальное предобуславливание.

Естественным обобщением схемы преобуславливания с использованием усечённых рядов является полиномиальное предобуславливание, т.е. использование в качестве M матрицы обратной к матрице

$$R \;=\; (\alpha_0 E + \alpha_1 H + \ldots + \alpha_{m-1}H^{m-1})P^{-1}$$

где, как и ранее, $A = A^* = P - Q > 0$, $P = P^* \in Gl_n(\mathbb{C})$, $H = P^{-1}Q$, $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$ и, как следствие, $R = R^*$ и $M = R^{-1} = M^*$. Заметим, что

$$RA \;=\; (\alpha_0 E + \alpha_1(E - P^{-1}A) + \ldots + \alpha_{m-1}(E - P^{-1}A)^{m-1})P^{-1}A \;=\; g(P^{-1}A)\,,$$

где $g(x) = \sum\limits_{i=0}^{m-1} \alpha_i (1-x)^i x$. Выбор коэффициентов $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$ следует осуществлять из соображений:

1. $\operatorname{Spec} RA = \operatorname{Spec} g(P^{-1}A) = \operatorname{Spec} g(E-H) \subset \mathbb{R}_+$ (это равносильно согласно одному из приведённых выше замечаний $R = R^* > 0$);

2. элементы $\operatorname{Spec} RA$ должны быть как можно более близки к единице.

Положим $m' = 1 - \lambda_{\max}(H) = \lambda_{\min}(P^{-1}A)$ и $M' = 1 - \lambda_{\min}(H) = \lambda_{\max}(P^{-1}A)$ и будем считать, что $0 < m' < M'$ (как и ранее, мы предполагаем, что $\rho(H) < 1$; случай $m' = M'$ не представляет интерес с точки зрения минимизации отношения $\frac{\lambda_{\max}(RA)}{\lambda_{\min}(RA)}$, так как в данной ситуации оно и так равно своему минимуму, 1). Тогда второе требование можно заменить несколько более грубой минимизационной задачей нахождения

$$\operatorname{arginf}_{\substack{g \in \mathbb{R}[t],\ \deg g = m,\ g(0)=0 \\ g(t)>0,\ m' \le t \le M'}} \frac{\max\limits_{m' \le t \le M'} g(t)}{\min\limits_{m' \le t \le M'} g(x)}$$

или более реалистической задачей в плане нахождения точного решения минимизационной задачей

$$\operatorname{arginf}_{\substack{g \in \mathbb{R}[t],\ \deg g = m,\ g(0)=0 \\ g(t)>0,\ m' \le t \le M'}} \|1 - g\| \ = \ 1 - \operatorname{arginf}_{\substack{f \in \mathbb{R}[t],\ \deg f = m,\ f(0)=1 \\ f(t)<1,\ m' \le t \le M'}} \|f\|$$

для некоторой конкретной нормы $\|\ \|$. Собственно на путях решения последней задачи мы и остановимся, ограничившись рассмотрением двух конкретных норм.

Начнём с обычной чебышевской нормы $\|\ \|_{C[m',M']}$. Поскольку $0 < m' < M'$ из сказанного ранее о многочленах чебышева следует, что одним из решений задачи

$$\operatorname{arginf}_{f \in \mathbb{R}[t],\ \deg f = m,\ f(0)=1} \|f\|_{C[m',M']}$$

является многочлен

$$f_m(t) \ = \ t_m\Big(\frac{2t - m' - M'}{M' - m'}\Big) \Big/ t_m\Big(\frac{M' + m'}{m' - M'}\Big).$$

Для этого достаточно перейти от f к $h(t') = f(t'-1)$ и отрезку $[m'+1, M'+1]$, $h(1) = 0$, где решение рассматриваемой минимизационной задачи нам известно (см. ранее), а затем вернуться обратно подстановкой $t' = t + 1$. Вместе с тем

$$\|f_m\|_{C[m',M']} \ = \ \frac{1}{\Big|t_m\Big(\frac{M'+m'}{m'-M'}\Big)\Big|} \ = \ \frac{2}{\Big(\frac{\sqrt{M'}-\sqrt{m'}}{\sqrt{M'}+\sqrt{m'}}\Big)^m + \Big(\frac{\sqrt{M'}+\sqrt{m'}}{\sqrt{M'}-\sqrt{m'}}\Big)^m}$$

(см. обоснование сходимости метода сопряжённых градиентов). Поскольку $(z+z^{-1}) > 2$ для любого $z > 0$, $z \ne 1$ (это равносильно $(z-1)^2 > 0$), и, в частности, для $z = z' = \frac{\sqrt{M'}+\sqrt{m'}}{\sqrt{M'}-\sqrt{m'}}$, отсюда следует, что $\|f_m\|_{C[m',M']} < 1$. Поэтому многочлен f_m является также решением интересующей нас минимизационной задачи. Покажем, что его выбор является оправданным и с точки зрения минимизации отношения $\frac{\lambda_{\max}(RA)}{\lambda_{\min}(RA)}$. Действительно, полагая $g_m = 1 - f_m$, мы можем записать

$$\frac{\lambda_{\max}(g_m(P^{-1}A))}{\lambda_{\min}(g_m(P^{-1}A))} \le \frac{\max\limits_{m'\le t\le M'}(1-f_m(t))}{\min\limits_{m'\le t\le M'}(1-f_m(t))} = \frac{1+\|f_m\|_{C[m',M']}}{1-\|f_m\|_{C[m',M']}} = \frac{t_m\Big(\frac{M'+m'}{M'-m'}\Big)+1}{t_m\Big(\frac{M'+m'}{M'-m'}\Big)-1} =$$

$$\Big(\frac{z'^{m/2}+z'^{-m/2}}{z'^{m/2}-z'^{-m/2}}\Big)^2 = \Big(\frac{z'^m+1}{z'^m-1}\Big)^2 = \left(\frac{(\sqrt{M'}+\sqrt{m'})^m+(\sqrt{M'}-\sqrt{m'})^m}{(\sqrt{M'}+\sqrt{m'})^m-(\sqrt{M'}-\sqrt{m'})^m}\right)^2 =$$

$$\left(\frac{(\sqrt{1-\lambda_{\min}(H)}+\sqrt{1-\lambda_{\max}(H)}\,)^m+(\sqrt{1-\lambda_{\min}(H)}-\sqrt{1-\lambda_{\max}(H)}\,)^m}{(\sqrt{1-\lambda_{\min}(H)}+\sqrt{1-\lambda_{\max}(H)}\,)^m-(\sqrt{1-\lambda_{\min}(H)}-\sqrt{1-\lambda_{\max}(H)}\,)^m}\right)^2 \to 1$$

при $m\to\infty$.

Другой возможный подход состоит в использовании некоторой интегральной евклидовой нормы $\|\ \|$ на пространстве $C[m',M']$, рассматриваемом как подпространство соответствующего гильбертова пространства $\mathcal{L}^2(m',M')$ (в интересующей нас ситуации речь идёт о вещественных пространствах). В первую очередь следует найти решение задачи

$$\operatorname{arginf}_{f\in\mathbb{R}[t],\ \deg f=m,\ f(0)=1}\|f\|\,.$$

Пусть $\{u_i\}_{i=0}^{\infty}$ система ортононормированных полиномов из $\mathcal{L}^2(m',M')$, полученная в результате переортогонализации системы $\{t^i\}_{i=0}^{\infty}$, где $\deg u_i=i$, $i\ge 0$, и, в частности, u_0 — ненулевая константа. Тогда любой многочлен f степени m может быть записан в виде

$$f(t) = \sum_{i=0}^{m} c_i u_i(t)$$

для подходящих $c_i=(f,u_i)\in\mathbb{R}$, $c_m\ne 0$. Условие $f(0)=1$ означает, что

$$1 = \sum_{i=0}^{m} c_i u_i(0)\,,\quad c_0 = \frac{1}{u_0(0)}\Big(1-\sum_{i=1}^{m} c_i u_i(0)\Big)\,.$$

Поэтому для такого многочлена f

$$\|f\|^2 = \sum_{i=1}^{m} c_i^2 + \frac{1}{u_0(0)^2}\Big(1-\sum_{i=1} c_i u_i(0)\Big)^2 =$$

$$\frac{1}{u_0(0)^2} - \sum_{i=1}^{m} c_i\frac{2u_i(0)}{u_0(0)^2} + \sum_{i=1}^{m} c_i^2\Big(1+\frac{u_i(0)^2}{u_0(0)^2}\Big) + \sum_{1\le i\ne j\le m} c_i c_j \frac{u_i(0)u_j(0)}{u_0(0)^2}\,.$$

Полагая $c=(c_1,\dots,c_m)^t$, $v=\Big(\frac{u_1(0)}{u_0(0)^2},\dots,\frac{u_m(0)}{u_0(0)^2}\Big)^t$ и $T=E+(u_0(0)v)(u_0(0)v)^t$, мы можем переписать последнее равенство в виде

$$\|f\|^2 = 2S(c)\,,\quad S(c) = 1/2c^tTc - v^tc\,.$$

Поскольку матрица $T=T^*>0$ ($w^tTw=\|w\|_2^2+(u_0(0)v,w)^2>0$ при $w\ne 0$), функционал S имеет единственную точку минимума — решение системы уравнений $Tc=v$ (S есть не что иное, как функционал Q для данной системы). Остаётся заметить, что решением $Tc=v$ является вектор

$$c = \frac{1}{1+u_0(0)^2\|v\|_2^2}v = \frac{1}{u_0(0)^2+\ldots+u_m(0)^2}(u_1(0),\dots,u_m(0))^t\,.$$

При этом

$$c_0 \;=\; \frac{1}{u_0(0)}(1-(c,u_0(0)^2v) \;=\; \frac{u_0(0)}{u_0(0)^2+\ldots+u_m(0)^2}\,.$$

Таким образом,

$$f_m \;=\; \mathrm{arginf}_{f\in\mathbb{R}[t],\ \deg f=m,\ f(0)=1}\|f\| \;=\; \frac{\sum\limits_{i=0}^{m}u_i(0)u_i}{\sum\limits_{i=0}^{m}u_i(0)^2}\,,$$

причём, как нетрудно заметить,

$$\|f_m\| \;=\; \left(\sum_{i=0}^{m}u_i(0)^2\right)^{-1/2}\,.$$

Для того, чтобы использовать найденный таким образом многочлен f_m для наших целей, следует убедиться в том, что он принимает значения меньше 1 на отрезке $[m',M']$. В качестве иллюстрации мы рассмотрим реализацию этого условия на примере многочленов Лежандра.

Стандартизированные многочлены Лежандра $\{P_n\}_{n\geq 0}$ определяются следующим образом:

$$P_n(x) \;=\; \frac{1}{2^n n!}\frac{\partial^n}{\partial x^n}(x^2-1)^n \quad (n\geq 0)\,,$$

где, в частности, $P_0(x)=1$, $P_1(x)=x$ и, как легко заметить, каждый многочлен P_n имеет степень n. Многочлены $\{P_n\}_{n\geq 0}$ формируруют ортогональную систему в $\mathcal{L}^2(-1,1)$ относительно стандартного скалярного произведения $(f,g)=\int\limits_{(-1,1)}fgdx$, $f,g\in\mathcal{L}^2(-1,1)$. Более того, полагая $\overline{P_n}=(n+1/2)^{1/2}P_n$, $n\geq 0$, мы получим ортононормированную систему полиномов $\{\overline{P_n}\}_{n\geq 0}$, $\int\limits_{(-1,1)}\overline{P_nP_m}dx=\delta_{n,m}$, $n,m\geq 0$. В действительности, система $\{\overline{P_n}\}_{n\geq 0}$ получается в результате применения процесса переортогонализации к степенным функциям $\{x^n\}_{n\geq 0}$. Стоит также упоминуть следующие свойства полиномов Лежандра:

1. полиномы $\{P_n\}_{n\geq 0}$ связаны рекуррентным соотношением

   $$(n+1)P_{n+1}(x)-(2n+1)xP_n(x)+nP_{n-1}(x) \;=\; 0 \quad (n\geq 1)\,,$$

   одним из следствий которого является тот факт, что в записи P_n участвуют степени x^k, $0\leq k\leq n$, $k\equiv n\ (2)$;

2. корни P_n, $n\geq 1$, лежат в интервале $(-1,1)$ и являются простыми;

3. $P_n(1)=1$, $P_n(-1)=(-1)^n$, $P_n'(1)=\frac{n(n+1)}{2}$ и $P_n'(-1)=(-1)^{n-1}\frac{n(n+1)}{2}$, $n\geq 1$;

4. корни P_n и P_{n+1} перемежаются в том смысле, что они не имеют общих корней и между двумя соседними корнями P_{n+1} расположен один единственный корень P_n;

5. $\|P_n\|_{C[-1,1]}=P_n(1)=|P_n(-1)|=1$;

6. производные полиномов $\{P_n\}_{n\geq 0}$ связаны соотношением

$$P'_{n+1}(x) - P'_{n-1}(x) - (2n+1)P_n(x) \;=\; 0 \quad (n \geq 1)\,,$$

в соответствии с которым (с учётом $P_n(x) > 0$ при $|x| > 1$, $n \equiv 0$ (2), $\operatorname{sign} P_n(x) = \operatorname{sign} x$ при $|x| > 1$, $n \equiv 1$ (2)) многочлен P_n убывает (возрастает) на $(-\infty, -1)$ при $n \equiv 0$ (2) ($n \equiv 1$ (2)) и возрастает на $(1, +\infty)$ при любом n, что гарантирует $|P_n(t)| > 1$ при $|t| > 1$, $n \geq 1$.

Для построения ортонормированной системы полиномов Лежандра на отрезке $[a, b]$ следует взять многочлены

$$\Big(\frac{2n+1}{b-a}\Big)^{1/2} P_n\Big(\frac{2x-a-b}{b-a}\Big) \quad (n \geq 0)\,.$$

Применительно к интересующей нас ситуации $[a, b] = [m', M']$ это будут многочлены

$$u_n(x) \;=\; \Big(\frac{2n+1}{M'-m'}\Big)^{1/2} P_n\Big(\frac{2x-m'-M'}{M'-m'}\Big) \quad (n \geq 0)\,,$$

причём из условия $0 < m' < M'$ следует, что $\left|\frac{m'+M'}{M'-m'}\right| > 1$ и $|u_n(0)| > |u_n(t)|$, $m' \leq t \leq M'$, $n \geq 1$, $u_0 = 1/\sqrt{M'-m'}$. Поэтому в данном случае при всех $m' \leq t \leq M'$, $m \geq 1$

$$|f_m(t)| \;=\; \frac{\Big|\sum\limits_{i=0}^{m} u_i(0)u_i(t)\Big|}{\sum\limits_{i=0}^{m} u_i(0)^2} \leq \frac{\sum\limits_{i=0}^{m} |u_i(0)||u_i(t)|}{\sum\limits_{i=0}^{m} u_i(0)^2} < 1\,.$$

Поэтому такой выбор f_m вполне удовлетворителен. Следует также отметить, что

$$\|f_m\|_{C[m',M']} \leq \frac{\sum\limits_{i=0}^{m} |u_i(0)|\Big(\frac{2i+1}{M'-m'}\Big)^{1/2}}{\sum\limits_{i=0}^{m} u_i(0)^2}\,.$$

Следовательно, интересующее нас отношение можно весьма грубо оценить как

$$\frac{\lambda_{\max}(g_m(P^{-1}A))}{\lambda_{\min}(g_m(P^{-1}A))} \leq \frac{1+\|f_m\|_{C[m',M']}}{1-\|f_m\|_{C[m',M']}} \leq \frac{\sum\limits_{i=0}^{m} |u_i(0)|\Big(|u_i(0)| + \Big(\frac{2i+1}{M'-m'}\Big)^{1/2}\Big)}{\sum\limits_{i=0}^{m} |u_i(0)|\Big(|u_i(0)| - \Big(\frac{2i+1}{M'-m'}\Big)^{1/2}\Big)} \leq$$
$$\frac{\sum\limits_{i=0}^{m} \frac{2i+1}{M'-m'}\Big|P_i\Big(\frac{m'+M'}{m'-M'}\Big)\Big|\Big(\Big|P_i\Big(\frac{m'+M'}{m'-M'}\Big)\Big| + 1\Big)}{\sum\limits_{i=0}^{m} \frac{2i+1}{M'-m'}\Big|P_i\Big(\frac{m'+M'}{m'-M'}\Big)\Big|\Big(\Big|P_i\Big(\frac{m'+M'}{m'-M'}\Big)\Big| - 1\Big)} \to 1$$

при $m \to \infty$.

Скажем теперь несколько слов об организации вычислений при реализации процедуры полиномиального предобуславливания на примере многочленов Чебышева (сходная

идея применима и к решению, которое базируется на многочленах Лежандра). В данном случае

$$g_m(x) \;=\; 1 - f_m(x) \;=\; 1 - t_m\Big(\frac{2x - m' - M'}{M' - m'}\Big)\Big/ t_m\Big(\frac{-m' - M'}{M' - m'}\Big)\,.$$

При этом в действительности нас интересует многочлен

$$h_m(x) \;=\; \alpha_0 + \alpha_1 x + \ldots + \alpha_{m-1}x^{m-1} \;=\; \frac{g_m(1+x)}{1+x}\,,$$

где коэффициенты $\{\alpha_i\}$ те же, что и в начале нашего обсуждения. Заметим, что деление в правой части осуществимо, поскольку $f_m(0) = 1$, $x | g_m(x)$.

Нас интересует нахождение вектора $z^{(m)} = M^{-1}r = h_m(H)P^{-1}r$, построение которого естественно будет осуществить при помощи рекуррентных соотношений. Итак, мы имеем $h_0 = 0$, $h_1(x) = 2x/(m' - M')$. Далее, полагая

$$\xi_m \;=\; t_m\Big(\frac{m' + M'}{m' - M'}\Big) \quad (m \geq 0)\,,$$

мы можем записать

$$h_m(x) \;=\; \frac{\xi_m - t_m\Big(\frac{2x+2-m'-M'}{M'-m'}\Big)}{\xi_m(x+1)}\,, \quad t_m\Big(\frac{2x + 2 - m' - M'}{M' - m'}\Big) \;=\; \xi_m(1 - (x+1)h_m(x))\,.$$

Воспользовавшись рекуррентными соотношениями для многочленов Чебышева, мы получаем

$$\begin{aligned}
&\xi_{m+1}(1 - (x+1)h_{m+1}(x)) \;= \\
&\qquad 2\Big(\frac{2x + 2 - m' - M'}{M' - m'}\Big)\xi_m(1 - (x+1)h_m(x)) - \xi_{m-1}(1 - (x+1)h_{m-1}(x)) \;= \\
&\qquad 4\Big(\frac{x+1}{M' - m'}\Big)\xi_m + (x+1)\Big(-2\Big(\frac{2x + 2 - m' - M'}{M' - m'}\Big)\xi_m h_m(x) + \xi_{m-1}h_{m-1}(x)\Big) + \\
&\qquad\qquad 2\Big(\frac{-m' - M'}{M' - m'}\Big)\xi_m - \xi_{m-1}\,.
\end{aligned}$$

Отсюда следует, что

$$\begin{aligned}
\xi_{m+1}h_{m+1}(x) \;&=\; -\frac{4\xi_m}{M' - m'} + 2\Big(\frac{2x + 2 - m' - M'}{M' - m'}\Big)\xi_m h_m(x) - \xi_{m-1}h_{m-1}(x)\,, \\
h_{m+1}(x) \;&=\; 2\Big(\frac{2x + 2 - m' - M'}{M' - m'}\Big)\frac{\xi_m}{\xi_{m+1}}h_m(x) - \frac{\xi_{m-1}}{\xi_{m+1}}h_{m-1}(x) - \frac{4\xi_m}{\xi_{m+1}(M' - m')}\,,
\end{aligned}$$

где

$$\xi_{m+1} \;=\; 2\Big(\frac{m' + M'}{m' - M'}\Big)\xi_m - \xi_{m-1}\,.$$

Остаётся заметить, что согласно данных равенств мы можем найти $z^{(m)} = h_m(H)P^{-1}r$ на m-ом шаге построения: $z^{(0)} = 0$, $z^{(1)} = \frac{2}{m'-M'}P^{-1}r$ и, далее,

$$z^{(k+1)} \;=\; \frac{2\xi_k}{\xi_{k+1}(M' - m')}(2H + (2 - m' - M')E)z^{(k)} - \frac{\xi_{k-1}}{\xi_{k+1}}z^{(k-1)} - \frac{4\xi_k}{\xi_{k+1}(M' - m')}P^{-1}r\,.$$

Поскольку $m' = 1 - \lambda_{\max}(H)$ и $M' = 1 - \lambda_{\min}(H)$, мы можем переписать последнее соотношение в виде

$$z^{(k+1)} = \frac{2\xi_k}{\xi_{k+1}(\lambda_{\max}(H) - \lambda_{\min}(H))}(2H + (\lambda_{\min}(H) + \lambda_{\max}(H))E)z^{(k)} - \frac{\xi_{k-1}}{\xi_{k+1}}z^{(k-1)} - \frac{4\xi_k}{\xi_{k+1}(\lambda_{\max}(H) - \lambda_{\min}(H))}P^{-1}r\,.$$

Таким образом, в данном случае мы просто заменяем вспомогательный m-шаговый итерационный процесс, используемый в предобуславливании с усечёнными рядами, на m-шаговый процесс, указанного здесь вида.

Предобуславливание с использованием неполного LU–разложения.

Ещё одной стандартной на сегоднящний день схемой построения предобуславливателей является предобуславливание, использующее неполное LU–разложение или неполное разложение Холесского (в симметрическом случае). Последнее в первую очередь актуально для больших систем с разреженными матрицами, где существенным требованием к организации вычислений является экономия памяти. Скажем несколько слов об условиях осуществимости подобной факторизации и особенностях их реализации.

Матрица $A \in Gl_n(\mathbb{R})$ называется *монотонной*, если обратная к ней матрица A^{-1} имеет неотрицательные коэффициенты. Монотонная матрица A с неположительными внедиагональными коэффициентами называется *M–матрицей*.

Теорема 0.62. *Матрица $A \in Gl_n(\mathbb{R})$ с неположительными внедиагональными элементами является M–матрицей в том и только в том случае, если она удовлетворяет условию обобщённого диагонального преобладания, т.е. имеется вектор $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$, $y > 0$, такой, что $Ay > 0$, где для краткости обозначение $z > 0$ для $z = (z_1, \ldots, z_n)^t \in \mathbb{R}^n$ понимается как $z_i > 0$, $i = 1, \ldots, n$.*

Доказательство. Если A — M–матрица, то для всякого $z \in \mathbb{R}^n$, $z > 0$, мы имеем $y = A^{-1}z > 0$ и, как следствие, $Ay = z > 0$.

Пусть теперь $A = (a_{ij}) \in Gl_n(\mathbb{R})$, $a_{ij} \le 0$ при $1 \le i \ne j \le n$ и существует $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$, $y > 0$, для которого $Ay > 0$. Тогда для любого $i = 1, \ldots, n$

$$a_{ii}y_i + \sum_{j \ne i} a_{ij}y_j > 0\,, \quad a_{ii} > -\frac{1}{y_i}\sum_{j \ne i} a_{ij}y_j \ge 0\,.$$

Запишем матрицу A в виде $A = D + A'$, где $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ (по предыдущему $D > 0$, A' — матрица с неположительными элементами). Нам понадобится также векторная норма $\|x\| = \|Y^{-1}x\|_\infty$, где $Y = \operatorname{diag}(y_1, \ldots, y_n)$, и подчинённая ей матричная норма

$$\|B\| = \max_{0 \ne x \in \mathbb{R}^n} \frac{\|Y^{-1}Bx\|_\infty}{\|Y^{-1}x\|_\infty} = \|Y^{-1}BY\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |b_{ij}|\frac{y_j}{y_i} \quad (B \in M_n(\mathbb{R}))\,.$$

Используя эту норму, мы сразу получаем, что

$$\|D^{-1}A'\| = \max_{i=1,\ldots,n} -\sum_{j \ne i} \frac{a_{ij}y_j}{a_{ii}y_i} < 1\,.$$

Поэтому матрица $A^{-1} = (D(E + D^{-1}A'))^{-1}$ представима в виде сходящегося ряда

$$A^{-1} \;=\; \sum_{k=0}^{\infty}(-1)^k(D^{-1}A')^kD^{-1} \;=\; \sum_{k=0}^{\infty}(D^{-1}(-A'))^kD^{-1}\,,$$

состоящего из матриц с неотрицательными элементами. Таким образом, все элементы матрицы A^{-1} неотрицательны и значит, матрица A является монотонной и, более того, M–матрицей. □

Заметим, что применительно к матрице $A \in Gl_n(\mathbb{R})$ с положительными диагональными и неположительными внедиагональными элементами, имеющей строгое диагональное преобладание по строкам, в качестве вектора y можно взять $y = (1, \ldots, 1)^t$.

Симметрические M–матрицы называются также *матрицами Стилтьеса.*

Следствие 0.63. *Любая матрица Стилтьеса A является положительно определённой и имеет регулярное диагональное расщепление.*

Доказательство. Действительно, поскольку $\|D^{-1}A'\| < 1$ (см. доказательство теоремы), $\rho(D^{-1}A') < 1$. При этом $D = D^* > 0$, $A' = A'^*$ и, как следствие, $\operatorname{Spec} D^{-1}A' \subset \mathbb{R}$, $\operatorname{Spec} D^{-1}A \subset \mathbb{R}_+$, что в свою очередь гарантирует $A > 0$ (см. доказанные ранее замечания). Более того, по сходным причинам мы имеем $\operatorname{Spec}(E - D^{-1}A') \subset \mathbb{R}_+$ и $D - A' = D(E - D^{-1}A') > 0$. □

Непосредственным следствием доказательства приведённой нами теоремы является

Следствие 0.64. *Любая матрица $A \in M_n(\mathbb{R})$ с неположительными внедиагональными элементами, удовлетворяющая условию обобщённого диагонального преобладания, является M–матрицей и, в частности, невырождена.*

Следствие 0.65. *Все главные квадратные подматрицы $A_k = (a_{ij})_{i,j=1}^k$, $k = 1, \ldots, n$, M–матрицы A являются M–матрицами и, следовательно, для матрицы A осуществимо LU–разложение.*

Доказательство. Поскольку $Ay > 0$ для некоторого $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$, $y > 0$, для любого $k = 1, \ldots, n$

$$\sum_{j=1}^{k} a_{ij}y_j > -\sum_{j=k+1}^{n} a_{ij}y_j \geq 0 \quad (1 \leq i \leq k)\,,$$

т.е. $A_k y^{(k)} > 0$, где $y^{(k)} = (y_1, \ldots, y_k)$. Согласно доказанной теореме это равносильно тому, что A_k является M–матрицей, $k = 1, \ldots, n$. □

Матрица $A = (a_{ij}) \in M_n(\mathbb{C})$ называется *H–матрицей*, если M–матрицей является матрица $M(A) = (m_{ij}(A))$, в которой

$$m_{ij}(A) \;=\; \begin{cases} |a_{ii}| & \text{при } i = j; \\ -|a_{ij}| & \text{при } i \neq j. \end{cases}$$

Другими словами, это равносильно тому, что матрица $M(A)$ удовлетворяет условию обобщённого диагонального преобладания.

Замечание 0.66. *Любая H–матрица $A \in M_n(\mathbb{C})$ является невырожденной.*

Доказательство. Для начала заметим, что на диагонали такой матрицы A нет нулевых элементов ($|D|$ — диагональ $M(A)$). Из доказательства теоремы также следует, что $\|D^{-1}A'\| = \||D|^{-1}M(A)'\| < 1$ и потому A^{-1} представима в виде сходящегося ряда, указанного в доказательстве. □

Более того, по аналогии с последним следствием мы сразу получаем

Следствие 0.67. *Все главные квадратные подматрицы $A_k = (a_{ij})_{i,j=1}^k$, $k = 1, \ldots, n$, H–матрицы A являются H–матрицами и, следовательно, для матрицы A осуществимо LU–разложение.*

Доказательство. Достаточно заметить, что $M(A_k) = M(A)_k$, $k = 1, \ldots, n$, а затем воспользоваться уже доказанным следствием для M–матриц. □

Пусть теперь Ω — некоторое множество пар индексов (i, j), $1 \le i, j \le n$, включающее в себя пары (i, i), $i = 1, \ldots, n$. Для произвольной матрицы $A \in Gl_n(\mathbb{C})$ рассмотрим следующий процесс построения матриц $A^{(0)}, A^{(1)}, \ldots, A^{(n-1)}$ (в предположении его осуществимости), реализуемый по правилу: $A^{(0)} = A$ и далее $A^{(i)} = L^{(i)}\tilde{A}^{(i)}$, $\tilde{A}^{(i)} = A^{(i-1)} + R^{(i)}$, где $R^{(i)} = (r_{pq}^{(i)})$,

$$r_{pq}^{(i)} = \begin{cases} 0, & \text{если или } p \ne i, \text{ или } q \ne i, \text{ или } (p,q) = (i,q) \in \Omega, \\ & \text{или } (p,q) = (p,i) \in \Omega; \\ -a_{pq}^{(i-1)}, & \text{если } (p,q) = (i,q) \notin \Omega \text{ или } (p,q) = (p,i) \notin \Omega, \end{cases}$$

и $L^{(i)}$ — нижняя унитреугольная матрица, отличающаяся от единичной матрицы лишь компонентами i-го столбца ниже главной диагонали,

$$L^{(i)} = \begin{pmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & -\frac{\tilde{a}_{i+1\,i}^{(i)}}{\tilde{a}_{ii}^{(i)}} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & -\frac{\tilde{a}_{ni}^{(i)}}{\tilde{a}_{ii}^{(i)}} & 0 & \ldots & 1 \end{pmatrix}.$$

Другими словами, переход от $A^{(i-1)}$ к $A^{(i)}$, $i = 2, \ldots, n-1$, сводится к двум действиям:

1. $\tilde{A}^{(i)} = A^{(i-1)} + R^{(i)}$ — аннулирование компонент i-ых строки и столбца матрицы $A^{(i-1)}$, отвечающих парам $(i, q), (p, i) \notin \Omega$;
2. $A^{(i)} = L^{(i)}\tilde{A}^{(i)}$ — один шаг метода Гаусса в реализации равносильной построению LR–разложения, который сводится к обнулению поддиагональных компонент i-го столбца матрицы $\tilde{A}^{(i)}$.

Следует заметить, что при всех $p = i + 1, \ldots, n$

$$\frac{\tilde{a}_{pi}^{(i)}}{\tilde{a}_{ii}^{(i)}} \;=\; \begin{cases} 0\,, & \text{если } (p,i)\notin\Omega; \\ \frac{a_{pi}^{(i-1)}}{a_{ii}^{(i-1)}}\,, & \text{если } (p,i)\in\Omega. \end{cases}$$

Отметим также, что в рассматриваемом процессе структура матриц $A^{(i)}$ идентична структуре матриц в методе Гаусса, причём на этапе перехода от $A^{(i-1)}$ к $A^{(i)}$ обрабатывается только ведущая квадратная подматрица $\overline{A}^{(i-1)} = (a_{pq}^{(i-1)})_{p,q=i}^{n}$, применение к которой первого действия сводится к обнулению компонент её первых строки и столбца с индексами $(i,q),(p,i)\notin\Omega$. Поэтому матрица $R^{(i)}$ имеет ненулевые компоненты только в наддиагональной и поддиагональной частях i-ой строки и i-го столбца, соответственно. Отсюда сразу следует, что $L^{(j)}R^{(i)} = R^{(i)}$, $j<i$.

Скажем несколько слов об осуществимости и свойствах описанного нами процесса для M–матриц и H–матриц.

Теорема 0.68. *Пусть A — M–матрица, $A\in Gl_n(\mathbb{R})$. Тогда для матрицы A и любого индексного множества Ω процесс построения матриц $A^{(i)}$, $i=0,\ldots,n-1$, осуществим и при этом все ведущие квадратные подматрицы $\overline{A}^{(i)}$, $i=0,\ldots,n-1$, на этапах построения являются M–матрицами.*

Доказательство. Доказательство проводится индукцией по i и сводится к проверке осуществимости первого шага построения $A^{(1)}$ и проверке того, что матрица $\overline{A}^{(1)}$ является M–матрицей. Действительно, диагональные компонентны M–матрицы $A^{(0)}=A$ положительны и, в частности, $a_{11}^{(0)} = a_{11} = \tilde{a}_{11}^{(1)} > 0$, что гарантирует осуществимость данного шага. Вместе с тем для всех $p,q=2,\ldots,n$ мы имеем

$$a_{pq}^{(1)} \;=\; \tilde{a}_{pq}^{(1)} - \frac{\tilde{a}_{p1}^{(1)}\tilde{a}_{1q}^{(1)}}{\tilde{a}_{11}^{(1)}} \;=\; \begin{cases} a_{pq}\,, & \text{если } (p,1)\notin\Omega \text{ или (и) } (1,q)\notin\Omega; \\ a_{pq} - \frac{a_{p1}a_{1q}}{a_{11}}\,, & \text{если } (p,1),(1,q)\in\Omega. \end{cases}$$

Поскольку внедиагональные элементы матрицы $A^{(0)}=A$ неположительны, а её диагональные элементы строго положительны, мы сразу получаем, что $a_{pq}^{(1)} \le a_{pq} = a_{pq}^{(0)}$ при всех $p,q=2,\ldots,n$. Последнее сразу гарантирует нам неположительность внедиагональных элементов матрицы $\overline{A}^{(1)} = (a_{pq}^{(1)})_{p,q=2}^{n}$. Вместе с тем в соответствии с доказанной ранее теоремой имеется $y=(y_1,\ldots,y_n)^t\in\mathbb{R}^n$, $y>0$, такой, что $Ay>0$. Полагая $\tilde{A}^{(1)}y = z = (z_1,\ldots,z_n)^t$, мы имеем

$$z_1 \;=\; \sum_{j,\ (1,j)\in\Omega} a_{1j}y_j \;=\; (Ay)_1 + \Big(-\sum_{j,\ (1,j)\notin\Omega} a_{1j}y_j\Big) > 0$$

и для всех $i=2,\ldots,n$

$$z_i \;=\; \tilde{a}_{i1}^{(1)}y_1 + \sum_{j=2}^{n} a_{ij}y_j \;=\; (Ay)_i + (\tilde{a}_{i1}^{(1)} - a_{i1})y_1 > 0\,,$$

поскольку $\tilde{a}_{i1}^{(1)} - a_{i1}$ либо равно нулю при $(i,1)\in\Omega$, либо $-a_{i1}\ge 0$ при $(i,1)\notin\Omega$. Поскольку коэффициенты матрицы $L^{(1)}$ неотрицательны, отсюда следует, что

$$A^{(1)}y \;=\; L^{(1)}\tilde{A}^{(1)}y \;=\; L^{(1)}z \;=\; \begin{pmatrix} z_1 \\ \overline{A}^{(1)}y' \end{pmatrix} > 0\,,$$

где $y' = (y_2, \ldots, y_n)^t$. Таким образом, $\overline{A}^{(1)} y' > 0$ для подходящего $y' > 0$, а это согласно доказанной теореме равносильно тому, что матрица с неположительными внедиагональными элементами $\overline{A}^{(1)}$ является M–матрицей. $\square$

Теорема 0.69. *Пусть A — H–матрица, $A \in Gl_n(\mathbb{C})$. Тогда для матрицы A и любого индексного множества Ω процесс построения матриц $A^{(i)}$, $i = 0, \ldots, n-1$, осуществим и при этом все ведущие квадратные подматрицы $\overline{A}^{(i)}$, $i = 0, \ldots, n-1$, на этапах построения являются H–матрицами.*

Доказательство. Как и в предыдущем утверждении, достаточно рассмотреть первый шаг алгоритма, осуществимость которого не вызывает сомнений (M–матрица $M(A)$ имеет ненулевые коэффициенты на диагонали). Напомним, что при всех $p, q = 2, \ldots, n$

$$a_{pq}^{(1)} = \begin{cases} a_{pq}\,, & \text{если } (p,1) \notin \Omega \text{ или (и) } (1,q) \notin \Omega; \\ a_{pq} - \frac{a_{p1}a_{1q}}{a_{11}}\,, & \text{если } (p,1),(1,q) \in \Omega. \end{cases}$$

Рассмотрим матрицу $M(\overline{A}^{(1)}) = (m_{pq}(A^{(1)}))_{p,q=2}^n$,

$$m_{pq}(A^{(1)}) = \begin{cases} |a_{pp}^{(1)}| & \text{при } p = q; \\ -|a_{pq}^{(1)}| & \text{при } p \neq q, \end{cases}$$

и матрицу $\overline{M(A)}^{(1)} = (m_{pq}(A)^{(1)})_{p,q=2}^n$, полученную в результате выполнения первого шага нашего построения для M–матрицы $M(A)$,

$$m_{pq}(A)^{(1)} = \begin{cases} m_{pq}(A)\,, & \text{если } (p,1) \notin \Omega \text{ или (и) } (1,q) \notin \Omega; \\ m_{pq}(A) - \frac{m_{p1}(A)m_{1q}(A)}{m_{11}(A)}\,, & \text{если } (p,1),(1,q) \in \Omega. \end{cases}$$

По предыдущей теореме матрица $\overline{M(A)}^{(1)}$ является M–матрицей. Вместе с тем, поскольку при всех $p, q = 2, \ldots, n$

$$m_{pq}(A^{(1)}) = \begin{cases} |a_{pp}|\,, & \text{если } p = q,\ (p,1) \notin \Omega \text{ или (и) } (1,p) \notin \Omega; \\ -|a_{pq}|\,, & \text{если } p \neq q,\ (p,1) \notin \Omega \text{ или (и) } (1,q) \notin \Omega; \\ \left|a_{pp} - \frac{a_{p1}a_{1p}}{a_{11}}\right|, & \text{если } p = q,\ (p,1),(1,p) \in \Omega; \\ -\left|a_{pq} - \frac{a_{p1}a_{1q}}{a_{11}}\right|, & \text{если } p \neq q,\ (p,1),(1,q) \in \Omega, \end{cases}$$

и

$$m_{pq}(A)^{(1)} = \begin{cases} |a_{pp}| = m_{pp}(A)\,, & \text{если } p = q,\ (p,1) \notin \Omega \\ & \text{или (и) } (1,p) \notin \Omega; \\ -|a_{pq}| = m_{pq}(A)\,, & \text{если } p \neq q,\ (p,1) \notin \Omega \\ & \text{или (и) } (1,q) \notin \Omega; \\ |a_{pp}| - \frac{|a_{p1}||a_{1p}|}{|a_{11}|} = m_{pp}(A) - \frac{m_{p1}(A)m_{1p}(A)}{m_{11}(A)}\,, & \text{если } p = q,\ (p,1),(1,p) \in \Omega; \\ -|a_{pq}| - \frac{|a_{p1}||a_{1q}|}{|a_{11}|} = m_{pq}(A) - \frac{m_{p1}(A)m_{1q}(A)}{m_{11}(A)}\,, & \text{если } p \neq q,\ (p,1),(1,q) \in \Omega, \end{cases}$$

$m_{pq}(A^{(1)}) \geq m_{pq}(A)^{(1)}$, $p, q = 2, \ldots, n$. Кроме того, имеется $z \in \mathbb{R}^{n-1}$, $z > 0$, для которого

$\overline{M(A)}^{(1)}z > 0$, и, следовательно, $M(\overline{A}^{(1)})z \geq \overline{M(A)}^{(1)}z > 0$. Поэтому в соответствии с доказанной ранее теоремой матрица $M(\overline{A}^{(1)})$ является M–матрицей, а матрица $\overline{A}^{(1)}$ — H–матрицей. $\square$

Замечание 0.70. *Пусть A — M–матрица, $A \in Gl_n(\mathbb{R})$, Ω и Ω' — множества индексных пар, $\Omega \subseteq \Omega'$. Тогда элементы ведущих подматриц $\overline{A}^{(i)}$ и $\overline{A'}^{(i)}$, $i = 0, \ldots, n-1$, возникающих при реализации рассматриваемых процессов для множеств Ω и Ω' соответственно, связаны отношением $a'^{(i)}_{pq} \leq a^{(i)}_{pq}$, $p, q = i+1, \ldots, n$.*

Доказательство. Воспользуемся индукцией по i с очевидным основанием $i = 0$, $A^{(0)} = A'^{(0)} = A$. Пусть наше утверждение доказано при $i < k$ для некоторого $1 \leq k \leq n-1$. Сравним коэффициенты матриц $\overline{A}^{(k)}$ и $\overline{A'}^{(k)}$, где, напомним, при всех $p, q = k+1, \ldots, n$

$$a^{(k)}_{pq} = \begin{cases} a^{(k-1)}_{pq}, & \text{если } (p,k) \notin \Omega \text{ или (и) } (k,q) \notin \Omega; \\ a^{(k-1)}_{pq} - \frac{a^{(k-1)}_{pk} a^{(k-1)}_{kq}}{a^{(k-1)}_{kk}}, & \text{если } (p,k), (k,q) \in \Omega, \end{cases}$$

и

$$a'^{(k)}_{pq} = \begin{cases} a'^{(k-1)}_{pq}, & \text{если } (p,k) \notin \Omega' \text{ или (и) } (k,q) \notin \Omega'; \\ a'^{(k-1)}_{pq} - \frac{a'^{(k-1)}_{pk} a'^{(k-1)}_{kq}}{a'^{(k-1)}_{kk}}, & \text{если } (p,k), (k,q) \in \Omega'. \end{cases}$$

Возможны следующие ситуации:

1. если $(p,k) \notin \Omega'$ или (и) $(k,q) \notin \Omega'$, тогда по предположению индукции
$$a^{(k)}_{pq} = a^{(k-1)}_{pq} \geq a'^{(k-1)}_{pq} = a'^{(k)}_{pq};$$
2. если $(p,k), (k,q) \in \Omega'$, но $(p,k) \notin \Omega$ или (и) $(k,q) \notin \Omega$, тогда в силу индуктивного предположения и того, что матрицы $\overline{A}^{(i)}$ и $\overline{A'}^{(i)}$ являются M–матрицами при всех $i = 0, \ldots, n-1$,
$$a^{(k)}_{pq} = a^{(k-1)}_{pq} \geq a'^{(k-1)}_{pq} \geq a'^{(k)}_{pq} = a'^{(k-1)}_{pq} - \frac{a'^{(k-1)}_{pk} a'^{(k-1)}_{kq}}{a'^{(k-1)}_{kk}};$$
3. если $(p,k), (k,q) \in \Omega \subseteq \Omega'$, тогда
$$a^{(k)}_{pq} = a^{(k-1)}_{pq} - \frac{a^{(k-1)}_{pk} a^{(k-1)}_{kq}}{a^{(k-1)}_{kk}} \geq a'^{(k)}_{pq} = a'^{(k-1)}_{pq} - \frac{a'^{(k-1)}_{pk} a'^{(k-1)}_{kq}}{a'^{(k-1)}_{kk}},$$
поскольку по предположению индукции $a^{(k-1)}_{st} \geq a'^{(k-1)}_{st}$, причём $a^{(k-1)}_{ss} \geq a'^{(k-1)}_{ss} > 0$ и $0 \geq a^{(k-1)}_{st} \geq a'^{(k-1)}_{st}$, $k \leq s \neq t \leq n$, и, как следствие,
$$\frac{a'^{(k-1)}_{pk} a'^{(k-1)}_{kq}}{a'^{(k-1)}_{kk}} = \frac{|a'^{(k-1)}_{pk}||a'^{(k-1)}_{kq}|}{a'^{(k-1)}_{kk}} \geq \frac{a^{(k-1)}_{pk} a^{(k-1)}_{kq}}{a^{(k-1)}_{kk}}.$$

Таким образом, шаг индукции и вместе с ним основное утверждение доказаны. $\square$

Вернёмся теперь к исходному процессу построения матриц $A^{(i)}$, $i = 0, \ldots, n-1$, где $A^{(0)} = A$ и $A^{(i)} = L^{(i)}\tilde{A}^{(i)}$, $\tilde{A}^{(i)} = A^{(i-1)} + R^{(i)}$, $i \geq 1$, предполагая его осуществимость (в частности, можно считать, что исходная матрица A является M–матрицей в вещественном и H–матрицей в комплексном случаях). По построению при $i = 1, \ldots, n-1$

$$A^{(i)} = L^{(i)}(A^{(i-1)} + R^{(i)}) = L^{(i)}(L^{(i-1)}(A^{(i-2)} + R^{(i-1)}) + R^{(i)}) = \ldots =$$
$$L^{(i)}L^{(i-1)}\cdots L^{(1)}A^{(0)} + L^{(i)}L^{(i-1)}\cdots L^{(1)}R^{(1)} + L^{(i)}L^{(i-1)}\cdots L^{(2)}R^{(2)} + \ldots + L^{(i)}R^{(i)}.$$

Поэтому, полагая $L = (L^{(n-1)}\cdots L^{(1)})^{-1}$ и $R = A^{(n-1)}$, мы получаем с учётом отмеченных ранее равенств $L^{(j)}R^{(i)} = R^{(i)}$, $j < i$,

$$R = L^{-1}A + (L^{(n-1)}\cdots L^{(1)}R^{(1)} + L^{(n-1)}\cdots L^{(2)}R^{(2)} + \ldots + L^{(n-1)}R^{(n-1)})$$

и

$$LR = A + R^{(1)} + (L^{(1)})^{-1}R^{(2)} + \ldots + (L^{(n-2)}\cdots L^{(1)})^{-1}R^{(n-1)} =$$
$$A + (R^{(1)} + \ldots + R^{(n-1)}) = A + T.$$

Таким образом, описанный нами процесс доставляет разложение $A = LR - T$, где $L = (L^{(1)})^{-1}\cdots(L^{(n-1)})^{-1}$ — нижняя унитреугольная матрица, в которой поддиагональные компоненты i-го столбца совпадают с поддиагональными компонентами i-го столбца матрицы $L^{(i)}$, записанными с противоположным знаком,

$$L = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ \frac{\tilde{a}^{(1)}_{21}}{\tilde{a}^{(1)}_{11}} & 1 & 0 & \ldots & 0 & 0 \\ \frac{\tilde{a}^{(1)}_{31}}{\tilde{a}^{(1)}_{11}} & \frac{\tilde{a}^{(2)}_{32}}{\tilde{a}^{(2)}_{22}} & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{\tilde{a}^{(1)}_{n-11}}{\tilde{a}^{(1)}_{11}} & \frac{\tilde{a}^{(2)}_{n-12}}{\tilde{a}^{(2)}_{22}} & \frac{\tilde{a}^{(3)}_{n-13}}{\tilde{a}^{(3)}_{33}} & \ldots & 1 & 0 \\ \frac{\tilde{a}^{(1)}_{n1}}{\tilde{a}^{(1)}_{11}} & \frac{\tilde{a}^{(2)}_{n1}}{\tilde{a}^{(2)}_{22}} & \frac{\tilde{a}^{(3)}_{n1}}{\tilde{a}^{(3)}_{33}} & \ldots & \frac{\tilde{a}^{(n-1)}_{nn-1}}{\tilde{a}^{(n-1)}_{n-1n-1}} & 1 \end{pmatrix},$$

R — верхняя треугольная матрица,

$$R = \begin{pmatrix} \tilde{a}^{(1)}_{11} & \tilde{a}^{(1)}_{12} & \ldots & \tilde{a}^{(1)}_{1n-1} & \tilde{a}^{(1)}_{1n} \\ 0 & \tilde{a}^{(2)}_{22} & \ldots & \tilde{a}^{(2)}_{2n-1} & \tilde{a}^{(2)}_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \tilde{a}^{(n-1)}_{n-1n-1} & \tilde{a}^{(n-1)}_{n-1n} \\ 0 & 0 & \ldots & 0 & a^{(n-1)}_{nn} \end{pmatrix} = \begin{pmatrix} a^{(1)}_{11} & a^{(1)}_{12} & \ldots & a^{(1)}_{1n-1} & a^{(1)}_{1n} \\ 0 & a^{(2)}_{22} & \ldots & a^{(2)}_{2n-1} & a^{(2)}_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a^{(n-1)}_{n-1n-1} & a^{(n-1)}_{n-1n} \\ 0 & 0 & \ldots & 0 & a^{(n-1)}_{nn} \end{pmatrix},$$

и $T = R^{(1)} + \ldots + R^{(n-1)}$ — матрица с нулевой диагональю, в которой поддиагональные и наддиагональные компоненты i-ых столбца и строки совпадают с соответствующими компонентами матрицы $R^{(i)}$. При этом, в матрицах L и R ненулевые компонентны могут находится только на позициях с индесами $(p, q) \in \Omega$, а в матрице T — только на позициях с индексами $(p, q) \notin \Omega$.

Представление $A = LR - T$ называется неполным LR–разложением матрицы A относительно множества Ω (перебросив диагональ R в L–множитель, можно говорить о неполном LU–разложении относительно Ω). Как и в обычном (полном) LR–разложении, можно предложить два эквивалентных способа построения неполной LR–факторизации. В терминах матричных умножений её можно организовать следующим образом: на первом шаге полагаем

$$r_{1i} = \begin{cases} 0 & \text{при } (1,i) \notin \Omega; \\ a_{1i} & \text{при } (1,i) \in \Omega \end{cases} \quad (i = 1, \ldots, n),$$

$$l_{j1} = \begin{cases} 0 & \text{при } (j,1) \notin \Omega; \\ \frac{a_{j1}}{a_{11}} & \text{при } (j,1) \in \Omega \end{cases} \quad (j = 2, \ldots, n),$$

а затем перевычисляем

$$a_{pq} := a_{pq} - l_{p1} r_{1q} \quad (p, q = 2, \ldots, n),$$

что соответствует нахождению $\overline{A}^{(1)} = (a_{pq}^{(1)})$; после выполнения $k-1$ шага и нахождения первых $k-1$ строки и столбца мартиц R и L мы находим компоненты их k-ых строки и столбца по формулам

$$r_{ki} = \begin{cases} 0 & \text{при } (k,i) \notin \Omega; \\ a_{ki} & \text{при } (k,i) \in \Omega \end{cases} \quad (i = k, \ldots, n),$$

$$l_{jk} = \begin{cases} 0 & \text{при } (j,k) \notin \Omega; \\ \frac{a_{jk}}{a_{kk}} & \text{при } (j,k) \in \Omega \end{cases} \quad (j = k+1, \ldots, n),$$

после чего полагаем (вычисляем ведущую подматрицу $\overline{A}^{(k)}$)

$$a_{pq} := a_{pq} - l_{pk} r_{kq} \quad (p, q = k+1, \ldots, n).$$

Если же заменить в этих равенствах перевычисляемые компонентны $\{a_{pq}\}$ их точными значениями, то можно получить процедуру вычисления неполной LR–факторизации в форме скалярных произведений: для всех $k = 1, \ldots, n$

$$r_{ki} = \begin{cases} 0 & \text{при } (k,i) \notin \Omega; \\ a_{ki} - \sum\limits_{s=1}^{k-1} l_{ks} r_{si} & \text{при } (k,i) \in \Omega \end{cases} \quad (i = k, \ldots, n),$$

$$l_{jk} = \begin{cases} 0 & \text{при } (j,k) \notin \Omega; \\ \Big(a_{jk} - \sum\limits_{t=1}^{k-1} l_{jt} r_{tk}\Big) / r_{kk} & \text{при } (j,k) \in \Omega \end{cases} \quad (j = k+1, \ldots, n),$$

где в случае $k = 1$ участвующие здесь суммы опускаются и на последнем шаге выисляется только компонента r_{nn}. К тому же результату можно было бы прийти и через решение уравнений, связывающих коэффициенты матриц L и R с коэффициентами исходной матрицы на позициях из множества Ω (см. вывод вычислительной процедуры LU–разложения).

В случае самосопряжённой матрицы A и симметричного относительно диагонали множества Ω (если $(i,j) \in \Omega$, то $(j,i) \in \Omega$) неполное LR–разложение относительно Ω превращается в неполное разложение Холесского $A = LDL^* - T$, где $DL^* = R$, $D = \mathrm{diag}(d_1, \ldots, d_n)$ — диагональная матрица с вещественными числами на диагонали (положительными вещественными числами в случае M–матрицы или H–матрицы A). При этом соответствующая вычислительная процедура в форме скалярных произведений может быть переписана в виде: для всех $k = 1, \ldots, n-1$ вычислить

$$d_k \;=\; a_{kk} - \sum_{s=1}^{k-1} |l_{ks}|^2 d_s$$

и

$$l_{ik} \;=\; \begin{cases} 0 & \text{при } (i,k) \notin \Omega; \\ \left(a_{ik} - \sum\limits_{t=1}^{k-1} l_{it} d_t \overline{l_{kt}}\right)/d_k & \text{при } (i,k) \in \Omega \end{cases} \qquad (i = k+1, \ldots, n)\,.$$

К числу наиболее популярных стратегий построения неполного LR–разложения (LU–разложения) относится неполная факторизация по "портрету матрицы A"

$$\Omega(A) \;=\; \{(i,j) \mid a_{ij} \neq 0\}\,,$$

включающему в себя в случае M–матрицы или H–матрицы A в обязательном порядке все пары (i,i). Разложение такого рода обычно применяется для разреженных матриц и, как правило, обозначается символом $ILU(0)$ (неполная LU–факторизация нулевого уровня). Неполные факторизации более высоких уровней $ILU(k)$, $k \geq 0$, определяются следующим образом: находим разложение $ILU(0)$ по портрету $\Omega_0 = \Omega(A)$, $A = L_0 R_0 - T_0$, полагаем $\Omega_1 = \Omega_0 \cup \Omega(T_0)$ и определяем $ILU(1)$–разложение по Ω_1, $A = L_1 R_1 - T_1$, и т.д. для всех k.

Пользуется популярностью также следующий вариант ILU–разложения относительно множества Ω, который принято называть модифицированным ILU–разложением или $MILU$–разложением. Его идея состоит в том, чтобы отказаться от выполнения равенств $(LR)_{ii} = a_{ii}$, $i = 1, \ldots, n$, обязательных для LU–разложения, заменив их условием сохранения строчных сумм $LRe = Ae$, где $e = (1, \ldots, 1)^t$. При этом остальные требования $(LR)_{ij} = a_{ij}$, $(i,j) \in \Omega$, $i \neq j$, остаются неизменными. Заметим, что первоначально подобные схемы "компенсации ошибок" появились в рамках серии алгоритмов Булеева, предложенных им в конце 50-х годов прошлого века.

Если мы располагаем $MILU$–разложением $A = LR - T$, где $Te = 0$ и $(LR)_{ij} = a_{ij}$, $(i,j) \in \Omega$, $i \neq j$, тогда $a_{ii} = (LR)_{ii} - t_{ii}$, $i = 1, \ldots, n$, и значит,

$$A + \mathrm{diag}(t_{11}, \ldots, t_{nn}) \;=\; LR - T'$$

— ILU–разложение матрицы $A + \mathrm{diag}(t_{11}, \ldots, t_{nn})$, причём $T' = T - \mathrm{diag}(t_{11}, \ldots, t_{nn})$ и $t_{ii} = (-T'e)_i$, $i = 1, \ldots, n$. Поскольку ILU–разложение определено однозначно (расчётные формулы в форме скалярных произведений следуют из уравнений $a_{ij} = (LR)_{ij}$, $(i,j) \in \Omega$), отсюда следует, что $MILU$–разложение является не чем иным как ILU–разложением возмущённой матрицы $A - S = LR - T$, где $S = \mathrm{diag}(s_1, \ldots, s_n)$ — диагональная матрица, такая, что $Se = Te$,

$$s_i \;=\; \sum_{j=1}^{n} t_{ij} \quad (i = 1, \ldots, n)\,.$$

По сравнению с описанным ранее процессом здесь добавляется вспомогательное действие: $A^{(i)} = L^{(i)}\tilde{A}^{(i)}$, $\tilde{A}^{(i)} = A^{(i-1)} + R^{(i)} - E_{ii}s_i$, $i = 1, \ldots, n$, где

$$s_i \;=\; \sum_{j=1}^{i-1} r_{ij}^{(j)} + \sum_{j=i+1}^{n} r_{ij}^{(i)} \;=\; ((R^{(1)} + \ldots + R^{(i)})e)_i \;=\; ((R^{(1)} + \ldots + R^{(n-1)})e)_i\,.$$

Действительно, достаточно заметить, что в результате такого построения мы получим матрицу

$$\begin{aligned} R \;=\; A^{(n)} \;=\; & L^{(n-1)} \cdots L^{(1)} A + L^{(n-1)} \cdots L^{(1)}(R^{(1)} - E_{11}s_1) + \\ & L^{(n-1)} \cdots L^{(2)}(R^{(2)} - E_{22}s_2) + \ldots + L^{(n-1)}(R^{(n-1)} - E_{n-1n-1}s_{n-1}) - E_{nn}s_n\,, \end{aligned}$$

а потому, полагая $L = (L^{(n-1)} \cdots L^{(1)})^{-1}$, мы можем записать

$$\begin{aligned} LR \;=\; & A + (R^{(1)} - E_{11}s_1) + (L^{(1)})^{-1}(R^{(2)} - E_{22}s_2) + \ldots + \\ & (L^{(n-2)} \cdots L^{(1)})^{-1}(R^{(n-1)} - E_{n-1n-1}s_{n-1}) - LE_{nn}s_n \;= \\ & A + (R^{(1)} + \ldots + R^{(n-1)}) - S \;=\; A + T - S\,. \end{aligned}$$

По аналогии с разложениями $ILU(k)$ строятся также разложения $MILU(k)$ для всех $k \geq 0$ (индексное множество k-го уровня получается добавлением к индексному множеству $k-1$-го уровня портрета T–составляющей, полученной при его реализации последнего).

Идея применения неполной факторизации $A = LR - T$ для предобуславливания метода сопряжённых градиентов состоит в использовании LR в качестве матрицы предобуславливателя M, что сводит задачу решения вспомогательной линейной системы $Mz = r$, выполняемой на каждом шаге предобусловленного метода, к решению двух трегольных систем. Вместе с тем нам следует обеспечить выполнение требований $M = M^* > 0$ и добиться по возможности минимизации отношения $\frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}$, где в нашем случае $M^{-1}A = E - M^{-1}T$. Заметим, что применительно к разложениям $ILU(k)$ и $MILU(k)$ мы вправе расчитывать на стремление данного отношения к 1 с ростом k в связи с увеличением индексного множества и приближением T–составляющей к нулю (заметим, что портрет T в разложении $A = LR - T$ по индексному множеству Ω может входить в Ω лишь в случае если он пуст и $T = 0$).

Замечание 0.71. *Пусть A — M–матрица и $A = LR - T$ — её неполное LR–разложение относительного множества Ω. Тогда коэффициенты матрицы T неотрицательны, матрицы L и R имеют неположительные внедиагональные элементы и, как следствие, матрица $M = LR = A + T$ монотонна, а матрица $M^{-1}A = E - M^{-1}T$ является M–матрицей. Кроме того, в этом случае $\rho(M^{-1}T) < 1$.*

Доказательство. Неположительность внедиагональных элементов матриц L и R, а также неотрицательность коэффициентов матрицы T, следует из описания процесса построения неполного LR–разложения и того, что ведущие квадратные подматрицы на этапах построения являются M–матрицами (см. ранее). Поэтому коэффициенты матрицы $M^{-1} = (LR)^{-1}$ неотрицательны и значит, такими же являются коэффициенты матрицы $M^{-1}T$. Отсюда следует, что внедиагональные коэффициенты матрицы

$M^{-1}A = E - M^{-1}T$ неположительны. Для M–матрицы A можно подобрать $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$, $y > 0$, такой, что $Ay > 0$. Тогда $M^{-1}Ay > 0$ и, следовательно, по доказанной ранее теореме матрица $M^{-1}A$ с неположительными внедиагональными элементами является M–матрицей. Полагая, как и в упомянутой теореме, $\|x\| = \|Y^{-1}x\|_\infty$, $x \in \mathbb{R}^n$, $Y = \operatorname{diag}(y_1, \ldots, y_n)$, мы получаем, что

$$\|M^{-1}T\| \;=\; \|Y^{-1}M^{-1}TY\|_\infty \;=\; \max_{i=1,\ldots,n} \sum_{j=1}^{n} (M^{-1}T)_{ij} \frac{y_j}{y_i} < 1\,,$$

поскольку $M^{-1}Ay = (E - M^{-1}T)y > 0$,

$$\sum_{j=1}^{n} (M^{-1}T)_{ij} y_j < y_i \quad (i = 1, \ldots, n)\,.$$

Следовательно, $\rho(M^{-1}T) < 1$. □

Следствие 0.72. *Если A — матрица Стилтьеса, тогда её неполная LR–факторизация $A = LR - T$ относительно симметрического множества Ω является LR–регулярным расщеплением A.*

Доказательство. В данном случае матрица A является симметрической M–матрицей и в её разложении $A = LR - T = M - T$ матрицы M и T являются симметрическими. Так как $\rho(M^{-1}T) < 1$, мы получаем, что в данном случае расщепление $A = M - T$ является M–регулярным, т.е. $M + T > 0$. В частности, это означает, что $M > 0$, $\operatorname{Spec} M^{-1}A \subset \mathbb{R}$ и, более того, $M^{-1}A \subset \mathbb{R}_+$ (напомним, что $A > 0$). □

Последнее наблюдение примечательно тем, что позволяет использовать неполную факторизацию Холесского не только для непосредственного построения предобуславливателя, но и в качестве базового расщепления для любой из описанных выше схем полиномиального ускорения этого метода.

Лекция 9. Алгебраическая проблема собственных значений. Основы теории возмущений.

Алгебраическая проблема собственных значений или задача поиска собственных значений и соответствующих собственных векторов (в ряде случаев задача нахождения канонической формы оператора) первоначально в наиболее ранних алгоритмах сводилась к построению характеристического многочлена матрицы с последующим поиском корней последнего. На сегодняшний день основные алгоритмы решения этой задачи ставят свой целью построение матрицы близкой к некоторой сопряжённой с исходной матрицей, имеющей верхнюю хессенбергову форму с блочной диагональю, составленной из блоков 1×1 и 2×2, соответствующих действительным и парам комплексно-сопряжённых собственных значений, принимаемых в качестве приближений к искомым собственным значениям. Во всяком случае к построениям такого рода относятся основные алгоритмы полной проблемы собственных значений и её симметрической составляющей, ориентированной на самосопряжённые матрицы.

Прежде всего стоит напомнить общие факты о взаимосвязи между матрицами и многочленами. Каждому многочлену $f = x^n + a_{n-1}x^{n-1} + \ldots + a_0 \in F[x]$ степени $n \geq 1$ с коэффициентами из некоторого поля (или в большей общности ассоциативного коммутативного кольца с единицей) F соответствует матрица

$$A_f \;=\; \begin{pmatrix} 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & \ldots & 0 & -a_1 \\ 0 & 1 & \ldots & 0 & -a_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & -a_{n-1} \end{pmatrix} \in M_n(F)\,,$$

с единичной побочной диагональю ниже главной, последним столбцом $(-a_0, \ldots, -a_{n-1})^t$ и равными нулю остальными коэффициентами. Такая матрица традиционно называется *матрицей Фробениуса многочлена* f. Матрица Фробениуса A_f многочлена f примечательна тем, что её характеристический многочлен совпадает с точностью до знака с многочленом f. Точнее $\chi_{A_f} = (-1)^n f$. Действительно, в этом можно убедиться, воспользовавшись индукцией по n с очевидным основанием $n = 1$ ($-a_0 - x = -f$). Если считать наше утверждение доказанным для всех $n < m$, тогда при $n = m$ мы можем разложить определитель $\chi_{A_f}(x) = \det(A - xE)$ по первой строке,

$$\chi_{A_f}(x) \;=\; (-x)\chi_{A_g}(x) + (-1)^{1+m}(-a_0)\,, \quad g \;=\; (f - a_0)/x\,,$$

что вместе с индуктивным предположением даёт нам

$$\chi_{A_f}(x) \;=\; (-x)(-1)^{m-1}((f(x) - a_0)/x) + (-1)^m a_0 \;=\; (-1)^m f(x)\,.$$

Заметим также, что для всякого корня λ многочлена f вектор $v_\lambda = (1, \lambda, \ldots, \lambda^{n-1})$ является левым собственным вектором матрицы A_f, отвечающим собственному значению λ, $v_\lambda A_f = \lambda v_\lambda$.

Своим возникновением матрица Фробениуса обязана понятию циклического подпространства оператора A, действующего на конечномерном векторном F–пространстве V. Напомним, что под циклическим подпространством, порождённым вектором $v \in V$,

понимается циклический подмодуль $F[A]v$ модуля V операторной алгеброй $F[A]$, который представляет собой F–линейную оболочку образов v при действии степеней A или, что одно и то же, Крыловское подпространство $K(A,v)$. Если элементы $e_i = A^{i-1}v$, $i = 1,\ldots,n$, составляют F–базис пространства $K(A,v)$ (а такой базис всегда можно выбрать), тогда, записав $Ae_n = A^n v = b_{n-1}A^{n-1}v + \ldots + b_0 v$ для подходящих $b_j \in F$, мы получаем, что в указанном базисе матрицей ограничения оператора A на A–инвариантном подпространстве $K(A,v)$ является матрица Фробениуса A_h многочлена $h = x^n - b_{n-1}x^{n-1} - \ldots - b_0 \in F[x]$. При этом, по построению $h(A)v = 0$ и, следовательно, $h(A)K(A,v) = h(A)F[A]v = \{0\}$. Последний вывод ещё раз подтверждает утверждение теоремы Гамильтона — Кэли об аннулировании оператора его характеристическим многочленом. Более того, поскольку в действительности пространство V распадается в прямую сумму циклических подпростанств и в базисе этих подпространств характристический многочлен оператора A, определённый, как известно, независимо от выбора базиса, совпадает с произведением характеристических многочленов ограничений A на эти подпространства, мы получаем в силу сказанного выше, что применение к A его характеристического многочлена даёт нуль, т.е. приходим к ещё одному способу доказательства теоремы Гамильтона — Кэли. Заметим, что используемое здесь представление матрицы оператора A в виде блочной диагональной матрицы, составленной из матриц Фробенуса соответствующих многочленов, является одной из известных канонических форм оператора, на основе которой в ранних алгоритмах проблемы собственных значений наподобие методов Крылова и Данилевского проводилось вычисление характеристического многочлена A.

Стоит также отметить, что рассматриваемая нами здесь взаимосвязь между матрицами и многочленами и известная из теории Галуа неразрешимость в радикалах уравнений степени выше 4 (невозможность построения в общем случае радикального расширения основного поля, которое включало бы в себя поле разложения многочлена степени выше 4) позволяет заключить следующее: не существует прямых методов нахождения собственных значений матриц размерности выше 4 за конечное число шагов. Другими словами, для матриц размерности выше 4 методы нахождения собственных значений могут быть только приближёнными (т.е. строящими приближения к собственным значениям матрицы).

Для того, чтобы понять порядок близости между собственными значениями матрицы и её возмущения (собственные значения возмущённой матрицы мы и вычисляем в ходе любого из описанных ниже алгоритмов), нам придётся привести ряд фактов из соответствующей "теории возмущений". Начнём со следующего весьма важного наблюдения.

Замечание 0.73. *Корни комплексного многочлена $f = f_n x^n + \ldots + f_0 \in \mathbb{C}[x]$ непрерывно зависят от его коэффициентов в том смысле, что при малых изменениях коэффициентов максимальное расстояние между корнями многочлена f и изменённого многочлена будет также мало. Следовательно, собственные значения комплексной матрицы $A \in M_n(\mathbb{C})$ непрерывно зависят от её коэффициентов.*

Доказательство. Одним из известных следствий теоремы о вычетах является тот факт, что число корней многочлена f, лежащих внутри области, ограниченной непрерывным контуром Γ, в точках которого f не обращается в нуль, совпадает со значением интеграла

$$\frac{1}{2\pi i}\int\limits_{\Gamma}\frac{f'}{f}dz\,.$$

Пусть $\alpha_1,\ldots,\alpha_s$ — корни многочлена f кратностей $k_1,\ldots,k_s$, $1\le s\le n$. Окружим каждый из корней α_i окружностью Γ_i столь малого радиуса r, что указанные окружности не пересекаются. Выберем ε, $0<\varepsilon<r$, и многочлен g, полученный в результате столь малых изменений коэффициентов многочлена f, что он не имеет корней на окружности $\Gamma_i(\varepsilon)$ радиуса ε с центром в точке α_i и

$$\max_{z\in\Gamma_i(\varepsilon)}\left|\frac{f'(z)}{f(z)}-\frac{g'(z)}{g(z)}\right|<\frac{1}{r}\quad(i=1,\ldots,s)\,.$$

Тогда при всех $i=1,\ldots,s$

$$\left|\frac{1}{2\pi i}\int\limits_{\Gamma_i(\varepsilon)}\frac{f'}{f}dz-\frac{1}{2\pi i}\int\limits_{\Gamma_i(\varepsilon)}\frac{g'}{g}dz\right|<1\,,\quad k_i\;=\;\frac{1}{2\pi i}\int\limits_{\Gamma_i(\varepsilon)}\frac{f'}{f}dz\;=\;\frac{1}{2\pi i}\int\limits_{\Gamma_i(\varepsilon)}\frac{g'}{g}dz\,.$$

Другими словами, многочлен g имеет внутри каждого круга $\{z\mid|z-\alpha_i|<\varepsilon\}$ ровно k_i корней. □

Основу теории возмущений алгебраической проблемы собственных значений (в данном случае речь идёт о зависимости между по грешностями решения и возмущениями в данных) составляет следующая серия результатов, сконцентрированных вокруг теоремы о кругах Гершгорина.

Замечание 0.74. *Для любых матричной нормы* $\|\ \|$, *матриц* A *и* B *и* $\mu\in\operatorname{Spec}(A+B)\setminus\operatorname{Spec}(A)$ *справедливо неравенство*

$$\frac{1}{\|(A-\mu E)^{-1}\|}\le\|B\|\,.$$

Доказательство. Поскольку матрица $A-\mu E$ обратима, а матрица $A+B-\mu E$ вырождена, матрица $E+(A-\mu E)^{-1}B$ также вырождена и, как следствие, $\|(A-\mu E)^{-1}B\|\ge 1$, $\|B\|\|(A-\mu E)^{-1}\|\ge 1$. □

Замечание 0.75. *Пусть матрица* A *диагонализуема,* $C^{-1}AC=\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)$, $\|\ \|$ *— любая матричная норма, для которой* $\|\operatorname{diag}(d_1,\ldots,d_n)\|=\max_i|d_i|$ *(данным свойством обладают, в частности, все матричные нормы* $\|\ \|_{p\ge1,\infty}$*). Тогда для любых матрицы* B *и* $\mu\in\operatorname{Spec}(A+B)$ *найдётся* $\lambda\in\operatorname{Spec}(A)$, *такой, что* $|\mu-\lambda|\le k(C)\|B\|$, *где* $k(C)$ *— число обусловленности матрицы* C *в рассматриваемой норме (иначе говоря,* $\rho(\operatorname{Spec}(A),\operatorname{Spec}(A+B))\le k(C)\|B\|$*).*

Доказательство. В рассмотрении нуждается случай $\mu\in\operatorname{Spec}(A+B)\setminus\operatorname{Spec}(A)=\{\lambda_i\}$. Тогда в духе прежних аргументов матрицы $A+B-\mu E$ и $\Lambda+C^{-1}BC-\mu E$ вырождены, а матрицы $A-\mu E$ и $\Lambda-\mu E$ обратимы, и потому по предыдущему замечанию

$$\frac{1}{\|(\Lambda-\mu E)^{-1}\|}\;=\;\frac{1}{\max_i|(\lambda_i-\mu)^{-1}|}\;=\;\min_i|\lambda_i-\mu|\le\|C^{-1}BC\|\le k(C)\|B\|\,.$$

□

Замечание 0.76. *Пусть $C^{-1}AC = J$ — жорданова нормальная форма матрицы A. Тогда для всякого $\mu \in \mathrm{Spec}(A+B)$ найдётся $\lambda \in \mathrm{Spec}(A)$, для которого*

$$\frac{|\mu-\lambda|^m}{1+|\mu-\lambda|+\ldots+|\mu-\lambda|^{m-1}} \le k_p(C)\|B\|_p \quad (p \ge 1, \infty)\,,$$

где $k_p(C)$ — число обусловленности матрицы C в норме $\|\ \|_p$, а m — максимальная размерность жордановой клетки в составе J, отвечающей λ.

Доказательство. По аналогии с предыдущим рассмотрим $\mu \in \mathrm{Spec}(A+B) \setminus \mathrm{Spec}(A)$. Тогда из вырожденности матрицы $J+C^{-1}BC-\mu E$ и невырожденности матрицы $J-\mu E$ следует, что

$$\frac{1}{\|(J-\mu E)^{-1}\|_p} \le k_p(C)\|B\|\,.$$

Представим матрицу J в виде $J = \mathrm{diag}(J_{m_1}(\lambda_1), \ldots, J_{m_s}(\lambda_s))$, где $J_k(\lambda)$ — жорданова клетка размера $k \times k$ с собственным значением λ. Тогда

$$\|(J-\mu E)^{-1}\|_p \ = \ \max_i \|J_{m_i}(\lambda_i-\mu)^{-1}\|_p\,,$$

где вследствие равенства $\|J_k(0)\|_p = 1$, $k \ge 1$,

$$J_{m_i}(\lambda_i-\mu)^{-1} \ = \ ((\lambda_i-\mu)E_{m_i} + J_{m_i}(0))^{-1} \ = \ (\lambda_i-\mu)^{-1}\sum_{j=0}^{m_s-1}(-1)^j(\lambda_i-\mu)^{-j}J_{m_i}(0)^j\,,$$

$$\|J_{m_i}(\lambda_i-\mu)^{-1}\|_p \le |\lambda_i-\mu|^{-m_i}\Big(\sum_{j=0}^{m_i-1}|\lambda_i-\mu|^j\Big)\,.$$

Если теперь $\hat{i}$ — индекс, такой, что

$$\|J_{m_{\hat{i}}}(\lambda_{\hat{i}}-\mu)^{-1}\|_p \ = \ \max_i \|J_{m_i}(\lambda_i-\mu)^{-1}\|_p\,,$$

$k_{\hat{i}} = \max\{m_i \mid \lambda_i = \lambda_{\hat{i}}\}$, тогда

$$k_p(C)\|B\|_p \ge \frac{1}{\|J_{m_{\hat{i}}}(\lambda_{\hat{i}}-\mu)^{-1}\|} \ge \frac{|\lambda_{\hat{i}}-\mu|^{m_{\hat{i}}}}{\sum\limits_{j=0}^{m_{\hat{i}}-1}|\lambda_{\hat{i}}-\mu|^j} \ge \frac{|\lambda_{\hat{i}}-\mu|^{k_{\hat{i}}}}{\sum\limits_{j=0}^{k_{\hat{i}}-1}|\lambda_{\hat{i}}-\mu|^j}\,.$$

В данном случае мы воспользовались также очевидным неравенством

$$\frac{x^m}{1+x+\ldots+x^{m-1}} \ge \frac{x^n}{1+x+\ldots+x^{n-1}} \quad (x \ge 0,\ 0 < m < n)\,.$$

□

В заключение этого раздела приведём следующую теорему Гершгорина.

Теорема 0.77. *Для всякой матрицы $A = (a_{ij})$*

$$\mathrm{Spec}(A) \subseteq \bigcup_i \{z \mid |z-a_{ii}| \le r_i(A)\}\,,$$

где $r_i(A) = \sum\limits_{j\ne i}|a_{ij}|$, $i = 1, \ldots, n$.

Доказательство. Достаточно рассмотреть ненулевой собственный вектор x, $Ax = \lambda x$, и заметить, что для индекса p, $|x_p| = \max_i\{|x_i|\}$,

$$|a_{pp} - \lambda||x_p| \;=\; \left|\sum_{i\neq p} a_{pi}x_i\right| \leq r_p(A)|x_p|\,, \quad |a_{pp} - \lambda| \leq r_p(A)\,.$$

Можно рассуждать и несколько иначе: $A - \lambda E = \mathrm{diag}(A) - \lambda E + A' \notin GL_n(\mathbb{C})$ и потому $\|(\mathrm{diag}(A) - \lambda E)^{-1}A'\| \geq 1$ для любых $\lambda \in \mathrm{Spec}(A) \setminus \{a_{ii}\}$ и матричной нормы $\|\ \|$. В частности, $\|(\mathrm{diag}(A) - \lambda E)^{-1}A'\|_\infty = \max_i |a_{ii} - \lambda|^{-1} r_i(A) = |a_{\hat{i}\hat{i}} - \lambda|^{-1} r_{\hat{i}}(A) \geq 1$. □

Круги с центрами в диагональных элементах матрицы A, участвующие в формулировке данного утверждения традиционно называются её кругами Гершгорина.

Можно также показать, что в случае если объединение k, $1 \leq k \leq n$, кругов Гершгорина представляет собой связную область, имеющую пустое пересечение с остальными $n-k$ кругами, внутри этой области находится ровно k собственных значений рассматриваемой матрицы с учётом кратности (непрерывная зависимость собственных значений от коэффициентов и $A_\varepsilon = \mathrm{diag}(A) + \varepsilon A' \to A$ при $\varepsilon \to 1$, причём $A_0 = \mathrm{diag}(A)$).

Несколько замечаний о канонических формах.

Заметим для начала, что известная нам из курса линейной алгебры каноническая форма Жордана фактически не применима в вычислительной практике по причине численной неустойчивости процесса её нахождения. Действительно, достаточно заметить, что при любых сколь угодно малых попарно различных ε_i, $i = 1, \ldots, n$, жорданова форма матрицы $J_n(0) + \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_n)$ совпадает с $\mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_n)$, а потому малые изменения в данных могут привести при вычислении жордановой формы к сравнительно большим изменениям в решении вне зависимости от используемой для оценки относительной погрешности матричной нормы.

Значительно более ценной с точки зрения матричных вычислений является каноническая форма Шура, описание которой приводится в следующей теореме.

Теорема 0.78. *Для любой матрицы $A \in M_n(\mathbb{C})$ можно подобрать унитарную матрицу $Q \in U_n(\mathbb{C})$, такую, что $Q^*AQ = T$ — верхняя треугольная матрица (унитарная форма Шура матрицы A), на диагонали которой стоят собственные значения матрицы A.*

Доказательство. Воспользуемся индукцией по $n \geq 1$ с очевидным основанием $n = 1$ ($A = a_{11} = 1 \cdot a_{11} \cdot 1$). Предположим, что наше утверждение доказано при всех $n < m$, $A \in M_n(\mathbb{C})$, для некоторого $m > 1$. Рассмотрим случай $n = m$ и произвольную матрицу $A \in M_n(\mathbb{C})$. Выберем собственное значение $\lambda \in \mathrm{Spec}(A)$ и отвечающий ему собственный вектор $u \in \mathbb{C}^n$ единичной нормы $\|u\|_2 = 1$, $Au = \lambda u$. Дополним этот вектор произвольным образом до унитарной матрицы $U = (u, U')$, где U' — блок из $n-1$ ортонормированных и ортогональных u столбцов. Тогда

$$U^*AU \;=\; \begin{pmatrix} u^*Au & u^*AU' \\ U'^*Au & U'^*AU' \end{pmatrix} \;=\; \begin{pmatrix} \lambda & u^*AU' \\ 0 & U'^*AU' \end{pmatrix}.$$

Используя индуктивное предположение, мы можем выбрать унитарную матрицу $V \in U_{n-1}(\mathbb{C})$, для которой $V^*(U'^*AU')V = T'$ — верхняя треугольная матрица, и записать

$$\begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix} U^*AU \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \;=\; \begin{pmatrix} \lambda & u^*AU'V \\ 0 & T' \end{pmatrix} \;=\; T\,.$$

Остаётся положить $Q = U\,\mathrm{diag}(1, V)$. □

Теорема 0.79. *Для любой матрицы $A \in M_n(\mathbb{R})$ имеется такая ортогональная матрица $Q \in O_n(\mathbb{R})$, что $Q^t A Q = T$ — верхняя хессенбергова матрица матрица с диагональными блоками 1×1 и 2×2 (ортогональная форма Шура матрицы A), собственные значения которых являются собственными значениями матрицы A.*

Доказательство. Схема доказательства вполне аналогична унитарному случаю и проводится индукцией по $n \geq 1$. Рассмотрим шаг индукции для произвольной матрицы $A \in M_n(\mathbb{R})$. Если $\mathrm{Spec}(A) \subset \mathbb{R}$, тогда нам достаточно повторить этапы предыдущего рассуждения с заменой унитарных матриц на ортогональные. Пусть имеется $\lambda \in \mathrm{Spec}(A) \setminus \mathbb{R}$, u — ненулевой комплексный собственный вектор матрицы A, отвечающий её собственному значению λ. В таком случае $Au = \lambda u$, $A\overline{u} = \overline{\lambda}\overline{u}$ и, как следствие, $W_\lambda = {}_{\mathbb{R}}\langle u_R = 1/2(u + \overline{u}), u_I = 1/2i(u - \overline{u})\rangle$ — двумерное вещественное A–инвариантное подпространство, ограничение на которое действия A имеет в базисе $\{u_R, u_I\}$ матрицу

$$\begin{pmatrix} 1/2(\lambda + \overline{\lambda}) & 1/2i(\lambda - \overline{\lambda}) \\ -1/2i(\lambda - \overline{\lambda}) & 1/2(\lambda + \overline{\lambda}) \end{pmatrix}.$$

Выпишем QR–разложение $(u_R, u_I) = VR$ матрицы (u_R, u_I) для подходящих $V = (q_1, q_2)$ — матрицы $n \times 2$ с двумя ортононормированными столбцами и верхней треугольной матрицы R размера 2×2. Дополнив матрицу V произвольным образом до ортонормированной матрицы $U = (V, V')$, мы можем с учётом того, что

$${}_{\mathbb{R}}\langle AV\rangle \;=\; {}_{\mathbb{R}}\langle Au_R, Au_I\rangle \;=\; AW_\lambda \subseteq W_\lambda \;=\; {}_{\mathbb{R}}\langle q_1, q_2\rangle\,,$$

записать

$$U^t A U \;=\; \begin{pmatrix} V^t A V & V^t A V' \\ 0 & V'^t A V' \end{pmatrix}.$$

Остаётся применить индуктивное предположение и подобрать матрицу $Q' \in O_{n-2}(\mathbb{R})$, для которой $T' = Q'^t(V'^t A V')Q'$ — верхняя хессенбергова матрица требуемой структуры,

$$\begin{pmatrix} E_2 & 0 \\ 0 & Q'^t \end{pmatrix} V^t A V \begin{pmatrix} E_2 & 0 \\ 0 & Q' \end{pmatrix} \;=\; \begin{pmatrix} V^t A V & V^t A V' Q' \\ 0 & T' \end{pmatrix}.$$

□

Отметим, что канонические формы Шура определены неоднозначным образом.

Позднее мы покажем, что приводимый ниже QR–алгоритм может быть интерпретирован как процесс построения приближений к ортогональной (унитарной) форме Шура исходной матрицы.

Лекция 10. GR–алгоритм и неявный GR–алгоритм. Сдвиги в GR–алгоритме и стратегия "исчерпания матрицы".

Мы рассмотрим несколько более общие схемы построения трёх взаимосвязанных между собой итерационных процессов, которые естественным образом обобщают степеннные QR–итерацию (LR–итерацию), QR–итерацию (LR–итерацию) и QR–алгоритм (LR–алгоритм), а затем обсудим практические и теоретические аспекты их осуществимости на примере QR–алгоритма, который является на сегодняшний день основным алгоритмом полной алгебраической проблемы собственных значений.

Итак, пусть мы располагаем GR–разложением произвольной обратимой матрицы $A = G(A)R(A)$ (или предполагаем осуществимость подобных разложений в приводимых далее построениях), где $G(A)$ и $R(A)$ — матрицы из двух подгрупп G и R группы $GL_n(\mathbb{F})$ над основным полем $\mathbb{F} = \mathbb{R}, \mathbb{C}$, имеющих единичное пересечение (последнее обеспечивает однозначность указанного разложения). Зафиксируем постоянные обратимые матрицы M и K и рассмотрим следующие итерационные процессы:

1. *Степенная GR–итерация*: $KA^kM = G_kR_k$, $G_k = G(KA^kM)$, $R_k = R(KA^kM)$, $k \geq 1$;
2. *GR–итерация*: для $K = E$, $AM = \tilde{G}_1\tilde{R}_1$, где $\tilde{G}_1 = G(AM)$, $\tilde{R}_1 = R(AM)$, и далее $A\tilde{G}_k = \tilde{G}_{k+1}\tilde{R}_{k+1}$, $\tilde{G}_{k+1} = G(A\tilde{G}_k)$, $\tilde{R}_{k+1} = R(A\tilde{G}_k)$, для всех $k \geq 1$;
3. *GR–алгоритм*: для $K = M = E$, $A = \hat{G}_1\hat{R}_1 = A_{(1)}$, где $\hat{G}_1 = G(A)$, $\hat{R}_1 = R(A)$, и далее $A_{(k)} = \hat{G}_k\hat{R}_k$, $\hat{G}_k = G(A_{(k)})$, $\hat{R}_k = R(A_{(k)})$, $A_{(k+1)} = \hat{R}_k\hat{G}_k$ для всех $k \geq 1$.

Взаимосвязь между этим процессами описывается следующим образом:

1. Для $K = E$ мы начальных $k = 1, 2$ можем записать

$$\begin{aligned} AM &= \tilde{G}_1\tilde{R}_1 = G_1R_1\,, \quad \tilde{G}_1 = G_1\,, \quad \tilde{R}_1 = R_1\,, \\ A^2M &= A\tilde{G}_1\tilde{R}_1 = \tilde{G}_2\tilde{R}_2\tilde{R}_1 = G_2R_2\,, \quad \tilde{G}_2 = G_2\,, \quad \tilde{R}_2\tilde{R}_1 = R_2\,. \end{aligned}$$

Предположим, что для некоторого $k \geq 1$ выполнение равенств $\tilde{G}_k = G_k$ и $\tilde{R}_k \cdots \tilde{R}_1 = R_k$ уже доказано. Тогда $A^kM = G_kR_k$ и

$$A^{k+1}M = AG_kR_k = A\tilde{G}_kR_k = \tilde{G}_{k+1}\tilde{R}_{k+1}R_k = G_{k+1}R_{k+1}\,,$$

что приводит нас к равенствам $\tilde{G}_{k+1} = G_{k+1}$ и $\tilde{R}_{k+1} \cdots \tilde{R}_1 = R_{k+1}$ и тем самым завершает доказательство индуктивного шага.

2. Для $K = M = E$, $A = G_1R_1 = \tilde{G}_1\tilde{R}_1 = \hat{G}_1\hat{R}_1$, $G_1 = \tilde{G}_1 = \hat{G}_1$, $R_1 = \tilde{R}_1 = \hat{R}_1$. Затем

$$\begin{aligned} A_{(2)} &= \hat{G}_2\hat{R}_2 = \hat{R}_1\hat{G}_1 = \hat{G}_1^{-1}A_{(1)}\hat{G}_1 = \hat{R}_1A_{(1)}\hat{R}_1^{-1}\,, \\ A\hat{G}_1 &= \hat{G}_1A_{(2)} = \hat{G}_1\hat{G}_2\hat{R}_2\,, \quad \hat{G}_1\hat{G}_2 = \tilde{G}_2\,, \quad \hat{R}_2 = \tilde{R}_2\,. \end{aligned}$$

Это наводит нас на мысль о соотношениях $\tilde{G}_k = \hat{G}_1 \cdots \hat{G}_k$ и $\tilde{R}_k = \hat{R}_k$, справедливость которых доказывается по индукции. Достаточно заметить, что в предположении об их выполнимости для данного k

$$\begin{aligned} A\tilde{G}_k &= A\hat{G}_1 \cdots \hat{G}_k = \hat{G}_1(\hat{R}_1\hat{G}_1)\hat{G}_2 \cdots \hat{G}_k = \hat{G}_1(\hat{G}_2\hat{R}_2)\hat{G}_2 \cdots \hat{G}_k = \ldots = \\ &\hat{G}_1 \cdots \hat{G}_k\hat{R}_k\hat{G}_k = \hat{G}_1 \cdots \hat{G}_{k+1}\hat{R}_{k+1}\,, \quad \tilde{G}_{k+1} = \hat{G}_1 \cdots \hat{G}_{k+1}\,, \quad \tilde{R}_{k+1} = \hat{R}_{k+1}\,. \end{aligned}$$

Таким образом, при $K = M = E$ для любого $k \geq 1$

$$\tilde{G}_k \;=\; G_k \;=\; \hat{G}_1 \cdots \hat{G}_k\,, \quad \tilde{R}_k \;= \hat{R}_k\,, \quad R_k \;=\; \tilde{R}_k \cdots \tilde{R}_1 \;=\; \hat{R}_k \cdots \hat{R}_1\,.$$

Вместе с тем по построению

$$A_{(k)} \;=\; \hat{G}_k\hat{R}_k \;=\; \hat{R}_{k-1}\hat{G}_{k-1} \;=\; \hat{G}_{k-1}^{-1}A_{(k-1)}\hat{G}_{k-1} \;=\; \hat{R}_{k-1}A_{(k-1)}\hat{R}_{k-1}^{-1} \;=$$
$$\hat{G}_{k-1}^{-1}\cdots\hat{G}_1^{-1}A\hat{G}_1\cdots\hat{G}_{k-1} \;=\; G_{k-1}^{-1}AG_{k-1} \;=\; \hat{R}_{k-1}\cdots\hat{R}_1A\hat{R}_1^{-1}\cdots\hat{R}_{k-1}^{-1} \;=\; R_{k-1}AR_{k-1}^{-1}\,.$$

Помимо перечисленных нами итерационных процессов можно столкнуться с их модернизациями, которые известны как "процессы со сдвигами" (или "мультисдвигами" при одновременном выполнении нескольких из них в рамках одного шага). На уровне степенной итерации это соответствует переходу от матрицы A^k к матрице $f_k(A)$, где f_k — нормализованный многочлен над основным полем степени k, $f_k(x) = f_{k-1}(x)(x - t_k)$, $k \geq 1$ (t_k — сдвиг на k-ом шаге). При этом для фиксированных обратимых матриц M и K мы по-прежнему осуществляем процесс последовательного построения GR–разложений $Kf_k(A)M = G_kR_k$, $k \geq 1$. Для процесса GR–итерации с $K = E$ процесс со сдвигами строится следующим образом: $(A - t_1E)M = \tilde{G}_1\tilde{R}_1$ и далее $(A - t_kE)\tilde{G}_{k-1} = \tilde{G}_k\tilde{R}_k$, $k \geq 2$. Последнее означает, что

$$f_k(A)M \;=\; (A - t_kE)\cdots(A - t_1E)M \;=\; G_kR_k \;=\; (A - t_kE)\cdots(A - t_1E)\tilde{G}_1\tilde{R}_1 \;=$$
$$(A - t_kE)\cdots(A - t_2E)\tilde{G}_2\tilde{R}_2\tilde{R}_1 \;=\; \tilde{G}_k\tilde{R}_k\cdots\tilde{R}_1\,, \quad \tilde{G}_k \;=\; G_k\,, \quad \tilde{R}_k\cdots\tilde{R}_1 \;=\; R_k\,.$$

В свою очередь GR–алгоритм трансфирмуется в следующую свою версию со сдвигами: для $K = M = E$, $A_{(1)} = A$, $A_{(1)} - t_1E = \hat{G}_1\hat{R}_1$, $A_{(2)} = \hat{R}_1\hat{G}_1 + t_1E$ и $A_{(k)} - t_kE = \hat{G}_k\hat{R}_k$, $A_{(k+1)} = \hat{R}_k\hat{G}_k + t_kE$ для всех $k \geq 2$. Заметим, что при такой огранизации процесса

$$A_{(k+1)} \;=\; \hat{G}_{k+1}^{-1}A_{(k)}\hat{G}_{k+1} \;=\; t_kE + \hat{R}_k\hat{G}_k \;=$$
$$\hat{G}_k^{-1}\cdots\hat{G}_1^{-1}A\hat{G}_1\cdots\hat{G}_k \;=\; \hat{R}_kA_{(k)}\hat{R}_k^{-1} \;=\; \hat{R}_k\cdots\hat{R}_1A\hat{R}_1^{-1}\cdots\hat{R}_k^{-1}$$

причём $A_{(1)} = \hat{G}_1\hat{R}_1 + t_1E = A$, $\hat{G}_1 = \tilde{G}_1 = G_1$, $\hat{R}_1 = \tilde{R}_1 = R_1$,

$$(A - t_2E)\hat{G}_1 \;=\; \tilde{G}_2\tilde{R}_2 \;=\; \hat{G}_1(\hat{G}_1^{-1}(A - t_2E)\hat{G}_1) \;=\; \hat{G}_1(A_{(2)} - t_2E) \;=\; \hat{G}_1\hat{G}_2\hat{R}_2\,,$$
$$\tilde{G}_2 \;=\; \hat{G}_1\hat{G}_2\,, \quad \tilde{R}_2 \;=\; \hat{R}_2$$

и, более того, $\tilde{G}_k = \hat{G}_1\cdots\hat{G}_k$, $\tilde{R}_k = \hat{R}_k$. Последнее сразу следует из того, что

$$f_k(A) \;=\; (A - t_kE)\cdots(A - t_1E) \;=$$
$$(A - t_kE)\cdots(A - t_2E)\hat{G}_1\hat{R}_1 \;=\; (A - t_kE)\cdots(A - t_3E)\hat{G}_1\hat{G}_2\hat{R}_2\hat{R}_1 \;=$$
$$(A - t_kE)\cdots(A - t_4E)\hat{G}_1\hat{G}_2(\hat{G}_2^{-1}\hat{G}_1^{-1}(A - t_3E)\hat{G}_1\hat{G}_2)\hat{R}_2\hat{R}_1 \;=$$
$$(A - t_kE)\cdots(A - t_4E)\hat{G}_1\hat{G}_2(A_{(3)} - t_3E)\hat{R}_2\hat{R}_1 \;=$$
$$(A - t_kE)\cdots(A - t_4E)\hat{G}_1\hat{G}_2\hat{G}_3\hat{R}_3\hat{R}_2\hat{R}_1 \;=\; \ldots \;=$$
$$\hat{G}_1\cdots\hat{G}_k\hat{R}_k\cdots\hat{R}_1 \;=\; G_kR_k \;=\; \tilde{G}_k\tilde{R}_k\cdots\tilde{R}_1\,.$$

Таким образом, переход к алгоритмам со сдвигами не меняет соотношений между ними.

В практическом плане наиболее приемлимым является реализация описанных построений для верхних хессенберговых матриц, приведение к которым осуществляется

при помощи любого из известных алогритмов, использующем унитарные или ортогональные подобия (см. методы вращений и отражений).

Замечание 0.80. *Если в используемом GR–разложении G–составляющие и R–составляющие разложения верхних хессенберговых матриц являются верними хессенберговыми и верхними треугольными матрицами, тогда каждая матрица $A_{(k)}$, $k \geq 2$, в GR–алгоритме со сдвигами, реализованном для верхней хессенберговой матрицы $A = A_{(1)}$, имеет верхнюю хессенбергову форму.*

Доказательство. Достаточно рассмотреть первый шаг алгоритма: $A_{(1)} - t_1 E = \hat{G}_1 \hat{R}_1$, $A_{(2)} = \hat{R}_1 \hat{G}_1 + t_1 E$. Поскольку матрица $A_{(1)} - t_1 E$ верхняя хессенбергова, матрицы $\hat{G}_1$ и $\hat{R}_1$ являются верхней хессенберговой и верхней треугольной соответственно, а потому верхней хессенберговой является матрица $\hat{R}_1 \hat{G}_1$ (её i-ый столбец является линейной комбинацией первых $i+1$ столбцов матрицы $\hat{R}_1$ для всех i) и матрица $A_{(2)} = \hat{R}_1 \hat{G}_1 + t_1 E$. □

В качестве иллюстрации приведём некоторые примеры выполнения этого замечания.

1. В LR–разложении верхней хессенберговой матрицы $A = LR$ матрицы L и R при условии его осуществимости имеют двухдиагональную структуру из одной главной и одной побочной диагонали ниже главной и верхнюю треугольную структуры. Действительно, процесс построения LR–разложения такой матрицы реализуется в матричном виде следующим образом:

$$t_{nn-1}(-a_{n-1}) \cdots t_{21}(-a_1) A \;=\; R, \quad A \;=\; LR,$$
$$L \;=\; t_{21}(a_1) \cdots t_{nn-1}(a_{n-1}) \;=\; E - \sum_{i=1}^{n-1} E_{ii+1} a_i,$$

где $t_{ij}(a) = E + E_{ij} a$. Соответствующая вычислительная процедура имеет вид: $r_{1k} = a_{1k}$, $k = 1, \ldots, n$, и далее $r_{ik} = a_{ik} - l_{ii-1} r_{i-1k}$, $k = i, \ldots, n$, и $l_{i+1i} = a_{i+1i} / r_{ii}$ для всех $i \geq 2$.

2. В QR–разложении верхней хессенберговой матрицы $A = QR$ матрицы Q и R верхнюю хессенбергову и верхнюю треугольную форму. Это непосредственно следует из единственности QR–разложения и возможности получения последнего путём последовательной переортогонализации столбцов.

3. В LR–разложении трёхдиагональной матрицы $A = LR$ при осуществимости последнего компонентны разложения L и R имеют верхнюю и нижнюю двухдиагональную структуру.

4. Применительно к самосопряжённой трехдиагональной матрице $A = A^*$ и QR–алгоритму матрицы $A_{(k)}$, $k \geq 2$, подобны исходной матрице, имеют верхнюю хессенбергову форму, а потому самосопряжены и трёхдиагональны.

Стратегии сдвига и исчерпания в GR–алгоритме

Начиная с этого момента, мы будем предполагать согласованность GR–разложения с верхней хессенберговой структурой в принятом выше смысле. В результате выполнения GR–алгоритма со сдвигами или без применямого к верхней хессенберговой матрице A мы получим последовательность сопряжённых друг с другом верхних хессенберговых матриц $A_{(k)} = (a_{ij}(k))$, $k \geq 1$, с $A_{(1)} = A$, имеющих одинаковый спектр.

Идея исчерпания состоит в сведении GR–алгоритма в рекурсивную форму за счёт обнуления "сравнительно малых" поддиагональных компонент $a_{ii-1}(k)$ матрицы $A_{(k)}$ на

k-ом шаге и перехода к матрицам меньшей размерности. Точнее речь идёт о переходе к матрице с тем же спектром, имеющей неразложимые верхние хессенберговы блоки (без нулей на побочной нижней диагонали), что позволяет применить к этим блокам меньшей размерности тот же алгоритм. Итоговый шаг состоит в вычислении собственных значений блоков 1×1 и 2×2. В основе данного соображения лежит теория возмущений собственных значений, позволяющая считать множества собственных значений близких матриц близкими друг другу в смысле стандартного хаусдорфова расстояния.

Наиболее употребимыми критериями малости элемента $a_{ii-1}(k)$ являются

1. $|a_{ii-1}(k)| < \varepsilon \|A_{(k)}\|_\infty$ или $|a_{ii-1}(k)| < \varepsilon \|A_{(k)}\|_1$;
2. $|a_{ii-1}(k)| < \varepsilon(|a_{i-1i-2}(k)|+|a_{i+1i}(k)|)$ или $m(a_{ii-1}(k)) < \varepsilon(m(a_{i-1i-2}(k))+m(a_{i+1i}(k)))$, где $m(x)$ — сумма модулей действительных и мнимых частей x.

для некоторого фиксированного $\varepsilon > 0$, определяющего порядок близости собственных значений найденных в процессе реализации алгоритма к собственным значениям исходной матрицы.

Естественным дополнением к стратегии исчерпания является выбор сдвига, обеспечивающего минимизацию одного или нескольких поддиагональных элементов на шаге: $A_{(k)} - t_k E = \hat{G}_k \hat{R}_k$, $A_{(k+1)} = \hat{R}_k \hat{G}_k + t_k E$. При этом в случае LR–алгоритма выбор сдвига призван обеспечить также осуществимость разложения на выполняемом шаге.

Перечислим ряд наиболее популярных стратегий сдвига. Применительно к LR–алгоритму в качестве t_k используется, как правило, или $a_{nn}(k)$, или $1/2a_{nn-1}(k)+a_{nn}(k)$, или одно из вещественных собственых значений матрицы

$$\begin{pmatrix} a_{n-1n-1}(k) & a_{n-1n}(k) \\ a_{nn-1}(k) & a_{nn}(k) \end{pmatrix},$$

ближайших к $a_{nn}(k)$. В QR–алгоритме в роли t_k используется либо $a_{nn}(k)$, либо в случае описанного ниже двойного сдвига — пара комплексно сопряжённых собственных значенний указанной матрицы (правило Фрэнсиса (сходная идея может быть использована и для LR–алгоритма). При этом любая из перечисленных стратегий в общем случае может оказаться бесполезной. К примеру, применение того же правила Фрэнсиса для нижнего углового блока 2×2 матрицы

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

является абсолютно бесполезным.

Неявная GR–теорема и её применение

Теорема 0.81. *Предположим, что группа G пересекается с группой верхних треугольных матриц по подгруппе группы диагональных матриц, диагональные компоненты которых равны ± 1 (или в комплексном случае равны по модулю 1). Тогда в представлении $B^{-1}AB = C$ с матрицами B и C из групп G и R, в котором матрица C является неразложимой верхней хессенберговой матрицей, первый столбец матрицы B определяет выбор её остальных столбцов однозначно с точностью до умножения на ± 1 (на комплексные числа равные по модулю 1).*

Доказательство. Пусть имеются два разложения $B^{-1}AB = C$ и $B'^{-1}AB' = C'$ с неразложимыми верхними хессенберговыми матрицами C и C' из группы R и матрицами B и C из группы G с одинаковыми первыми столбцами. Требуется показать, что $D = B'^{-1}B = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_n)$, $\varepsilon_i \in \{\pm 1\}$, причём $\varepsilon_1 = 1$. По условию

$$C'D \;=\; C'B'^{-1}B \;=\; B'^{-1}AB \;=\; B'^{-1}BC \;=\; DC\,,$$

причём $D = (d_1, \ldots, d_n)$, $d_1 = (1, 0, \ldots, 0)^t$. Остаётся заметить, что

$$(C'D)_i \;=\; (DC)_i \;=\; C'd_i \;=\; \sum_{j=1}^{i+1} d_j c_{ji}\,, \quad d_{i+1}c_{i+1i} \;=\; C'd_i - \sum_{j=1}^{i} d_j c_{ji}\,,$$

где $c_{i+1i} \neq 0$ в силу неразложимости матрицы C, а потому

$$d_2 \;=\; c_{21}^{-1}(c'_1 - d_1 c_{11}) \;=\; (d_{12}, d_{22}, 0, \ldots, 0)^t$$

и в предположении уже установленных равенств $d_j = (d_{1j}, \ldots, d_{jj}, 0, \ldots, 0)^t$, $j = 1, \ldots, i$, мы имеем

$$d_{i+1} \;=\; c_{i+1i}^{-1}\left(C'd_i - \sum_{j=1}^{i} d_j c_{ji}\right) \;=$$

$$c_{i+1i}^{-1}\left(\sum_{k=1}^{i} c'_k d_{ki} - \sum_{j=1}^{i} d_j c_{ji}\right) \;=\; (d_{1i+1}, \ldots, d_{i+1i+1}, 0, \ldots, 0)^t\,.$$

Итак, матрица D верхняя треугольная и при этом входит в группу G, а значит, по условию имеет вид $\mathrm{diag}(1, \varepsilon_2, \ldots, \varepsilon_n)$, $\varepsilon_i \in \{\pm 1\}$ ($|\varepsilon_i| = 1$). □

Заметим, что в случае LR–алгоритма в роли группы G выступает группа нижних унитреугольных матриц и выбор матрицы B определяется её первым столбцом однозначным образом, а в случае QR–алгоритма и группы $G = O_n(\mathbb{R})$ ($G = U_n(\mathbb{C})$) — с точностью до умножения на ± 1 (комплексные числа равные по модулю 1) всех столбцов, начиная со второго.

Основным примером применения неявной теоремы являются неявные GR–алгоритмы с ординарным и двойным сдвигом, на базе QR– и LR–разложений. Начнём с первого из этих алгоритмов. В данном случае вместно непосредственного построения GR–разложений на k-ом шаге алгоритма, применяемого к верхним хессенберговым матрицам, $A_{(k)} - t_k E = \hat{G}_k \hat{R}_k$, $A_{(k+1)} = \hat{R}_k \hat{G}_k + t_k E = \hat{G}_k^{-1} A_{(k)} \hat{G}_k$ реализуется следующее построение:

1. выбирается первый столбец матрицы $\hat{G}_k$ параллельным первому столбцу матрицы $A_{(k)} - t_k E$ (равным его нормализации в случае QR–алгоритма и результату масштабирования на первый коэффициент в случае LR–алгоритма);

2. данный столбец достраивается до матрицы $\hat{G}_k$ из группы G таким образом, что матрица $\hat{G}_k^{-1} A_{(k)} \hat{G}_k = A_{(k+1)}$ имеет верхнюю хессенбергову форму.

При этом в случае неразложимой матрицы $A_{(k+1)}$ выбор матрицы $\hat{G}_k$ определён однозначно с известной из неявной теоремы степенью точности, а в случае разложимой матрицы $A_{(k+1)}$ мы можем перейти к матрицам меньшей размерности по схеме ис-

черпания без качественного ухудшения итогового результата. Отметим также, что использование условия 1 продиктовано первым этапом: $A_{(k)} - t_k E = \hat{G}_k \hat{R}_k$.

В практическом плане это реализуется для вещественного сдвига t_k в соответствии с описанной ниже стратегией вытеснения одномерного выступа. Для определённости мы будем иметь дело в основном с QR–алгоритмом, а комментарии относительно LR–алгоритма будут делаться в предположении осуществимости описанного процесса.

Итак, пусть $(a_{11}(k) - t_k, a_{21}(k), 0, \ldots, 0)^t$ — первый столбец матрицы $A_{(k)} - t_k E$. Положим $\hat{G}_k(1) = \mathrm{diag}(B_k(1), 1, \ldots, 1)$, где $B_k(1)$ — матрица 2×2, которая для QR–алгоритма определяется как

$$B_k(1) \;=\; \frac{1}{\sqrt{|a_{11}(k) - t_k|^2 + |a_{21}(k)|^2}} \begin{pmatrix} a_{11}(k) - t_k & -\overline{a_{21}(k)} \\ a_{21}(k) & a_{11}(k) - t_k \end{pmatrix},$$

а для LR–алгоритма как

$$B_k(1) \;=\; \begin{pmatrix} 1 & 0 \\ \frac{a_{21}(k)}{a_{11}(k) - t_k} & 1 \end{pmatrix}.$$

Перейдём к матрице

$$A_{(k)}(1) \;=\; \hat{G}_k(1)^{-1} A_{(k)} \hat{G}_k(1) \;=\; \begin{pmatrix} * & * & * & * & \ldots & * & * \\ * & * & * & * & \ldots & * & * \\ + & * & * & * & \ldots & * & * \\ 0 & 0 & * & * & \ldots & * & * \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & * & * \end{pmatrix}.$$

Затем подберём матрицу $\hat{G}_k(2) = \mathrm{diag}(1, B_k(2), 1, \ldots, 1) \in G$ с блоком $B_k(2)$ размера 2×2, такую, что $\hat{G}_k(2)^{-1} A_{(k)}(1)$ — верхняя хессенбергова матрица, а потому

$$A_{(k)}(2) \;=\; \hat{G}_k(2)^{-1} A_{(k)}(1) \hat{G}_k(2) \;=\; \begin{pmatrix} * & * & * & * & \ldots & * & * \\ * & * & * & * & \ldots & * & * \\ 0 & * & * & * & \ldots & * & * \\ 0 & + & * & * & \ldots & * & * \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & * & * \end{pmatrix}.$$

Продолжая подобный процесс вытеснения выступа, мы построим $n-1$ диагональную матрицу $\hat{G}_k(i) \in G$, $i = 1, \ldots, n-1$, с единственным неединичным блоком 2×2 и перейдём к верхней хессенберговой матрице

$$A_{(k+1)} \;=\; \hat{G}_k(n-1)^{-1} \cdots \hat{G}_k(1)^{-1} A_{(k)} \hat{G}_k(1) \cdots \hat{G}_k(n-1) \;=\; \hat{G}_k^{-1} A_{(k)} \hat{G}_k\,,$$

где матрица $\hat{G}_k = \hat{G}_k(1) \cdots \hat{G}_k(n-1)$ имеет первый столбец равный первому столбцу матрицы $\hat{G}_k(1)$.

Перейдём теперь к обсуждению неявного GR–алгоритма с двойным сдвигом. Возможность существования у вещественной матрицы пары комплексно сопряжённых собственных значений заставляет задуматься о выполнении двух последовательных сдвигов t_k и $\overline{t_k}$ на шагах k и $k+1$, реализацию которых следует выполнить в рамках одного

шага с сохранением вещественной арифметики. При этом в качестве t_k, как правило, используется комплексное собственное значение матрицы из правила Фрэнсиса

$$\begin{pmatrix} a_{n-1n-1}(k) & a_{n-1n}(k) \\ a_{nn-1}(k) & a_{nn}(k) \end{pmatrix}.$$

Другими словами речь идёт о следующем переходе:

$$A_{(k)} - t_k E = \hat{G}_k \hat{R}_k\,,\ A_{(k+1)} = \hat{R}_k \hat{G}_k + t_k E = \hat{G}_k^{-1} A_{(k)} \hat{G}_k\,,\ A_{(k+1)} - \overline{t_k} E = \hat{G}_{k+1} \hat{R}_{k+1}\,,$$
$$A_{(k+2)} = \hat{R}_{k+1} \hat{G}_{k+1} + \overline{t_k} E = \hat{G}_{k+1}^{-1} A_{(k+1)} \hat{G}_{k+1} = \hat{G}_{k+1}^{-1} \hat{G}_k^{-1} A_{(k)} \hat{G}_k \hat{G}_{k+1}\,,$$

реализуемом в рамках одного перехода от матрицы $A_{(k)}$ к матрице

$$A_{(k+2)} = (\hat{G}_k \hat{G}_{k+1})^{-1} A_{(k)} (\hat{G}_k \hat{G}_{k+1})\,,$$

являющихся вещественными верхними хессенберговыми матрицами. Осуществимость подобного действия базируется на неявной GR–теореме и возможности выбора матрицы $\hat{G}_k \hat{G}_{k+1}$ среди вещественных матриц. Отметим, что в данном случае мы предполагаем выполнимость GR–разложения над полем $\mathbb{C}$, составляющие которого в вещественном случае могут быть выбраны среди вещественных матриц из групп $G \cap M_n(\mathbb{R})$ и $R \cap M_n(\mathbb{R})$, а также то, что группа R входит в группу верхних треугольных матриц.

Замечание 0.82. *Выбор матриц $\hat{G}_k$ и $\hat{G}_{k+1}$ может быть выполнен таким образом:*

1. $\hat{G}_k \hat{G}_{k+1} \in M_n(\mathbb{R})$;
2. *первый столбец матрицы $\hat{G}_k \hat{G}_{k+1}$ параллелен первому столбцу матрицы $A_{(k)}^2 - (t_k + \overline{t_k}) A_{(k)} + |t_k|^2 E$ и определяет в условиях неявной GR–теоремы выбор остальных её столбцов в случае неразложимой матрицы $A_{(k+1)}$ однозначно с точностью до умножения на ± 1.*

Доказательство. Достаточно заметить, что

$$\hat{G}_{k+1} \hat{R}_{k+1} = A_{(k+1)} - \overline{t_k} E = \hat{R}_k \hat{G}_k + (t_k - \overline{t_k}) E\,,$$
$$\hat{G}_k \hat{G}_{k+1} \hat{R}_{k+1} \hat{R}_k = (\hat{G}_k \hat{R}_k)^2 + (t_k - \overline{t_k})(\hat{G}_k \hat{R}_k) =$$
$$(A_{(k)} - t_k E)^2 + (t_k - \overline{t_k})(A_{(k)} - t_k E) = A_{(k)}^2 - (t_k + \overline{t_k}) A_{(k)} + |t_k|^2 E \in M_n(\mathbb{R})\,,$$

т.е. $(\hat{G}_k \hat{G}_{k+1})(\hat{R}_{k+1} \hat{R}_k)$ — GR–разложение вещественной матрицы $A_{(k)}^2 - (t_k + \overline{t_k}) A_{(k)} + |t_k|^2 E$, первый столбец которой параллелен первому столбцу матрицы $\hat{G}_k \hat{G}_{k+1}$ ввиду того, что матрица $\hat{R}_{k+1} \hat{R}_k$ верхняя треугольная. Поэтому составляющие этого разложения могут быть выбраны среди вещественных матриц. Для завершения доказательства остаётся воспользоваться утверждением неявной GR–теоремы. □

Опираясь на это соображение мы можем осуществить выбор $\hat{G}_k \hat{G}_{k+1}$ и переход от $A_{(k)}$ к $A_{(k+1)}$ по правилу:

1. первый столбец матрицы $\hat{G}_k \hat{G}_{k+1}$ выбирается параллельно первому столбцу матрицы $A_{(k)}^2 - (t_k + \overline{t_k}) A_{(k)} + |t_k|^2 E$, что в случае QR–алгоритма соответствует его нормировке евклидовой нормой, а в случае LR–алгоритма — его масштабированию на первый коэффициент (при осуществимости последнего действия);
2. полученный столбец достраивается до матрицы $\hat{G}_k \hat{G}_{k+1} \in G \cap M_n(\mathbb{R})$ таким обра-

зом, что $(\hat{G}_k\hat{G}_{k+1})^{-1}A_{(k)}(\hat{G}_k\hat{G}_{k+1}) = A_{(k+2)}$ — верхняя хессенбергова матрица.

В ситуации, когда матрица $A_{(k+2)}$ является неразложима, подобный переход отвечает с известной степенью точности двум шагам GR–алгоритма. Если же полученная в результате матрица $A_{(k+2)}$ разложима, тогда мы можем вновь воспользоваться схемой исчерпания. Впрочем, на практике схема исчерпания применяется в любом случае и термин разложимости следует понимать с точностью до выбранного порядка малости поддиагональных элементов.

В практическом плане реализация этого процесса для QR– и LR–алгоритмов выглядит следующим образом. Первый столбец матрицы $A^2_{(k)} - (t_k + \overline{t}_k)A_{(k)} + |t_k|^2E$ имеет вид $(a_1, a_2, a_3, 0, \ldots, 0)^t$, где

$$\begin{aligned} a_1 &= a_{11}(k)^2 + a_{12}(k)a_{21}(k) - (t_k + \overline{t_k})a_{11}(k) + |t_k|^2\,, \\ a_2 &= a_{21}(k)(a_{11}(k) + a_{22}(k) - (t_k + \overline{t_k}))\,, \quad a_3 = a_{32}(k)a_{21}(k)\,. \end{aligned}$$

Для QR–алгоритма столбец $(a'_1, a'_2, a'_3, 0, \ldots, 0)^t = 1/\sqrt{a_1^2 + a_2^2 + a_3^2}(a_1, a_2, a_3, 0, \ldots, 0)^t$ достраивается до ортогональной матрицы $\hat{G}_k(1) = \mathrm{diag}(B_k(1), 1, \ldots, 1)$ с ортогональным блоком

$$B_k(1) = \begin{pmatrix} a'_1 & * & * \\ a'_2 & * & * \\ a'_3 & * & * \end{pmatrix}.$$

Для LR–алгоритма столбец $(a''_1, a''_2, a''_3, 0, \ldots, 0)^t = 1/a_1(a_1, a_2, a_3, 0, \ldots, 0)^t$ (в предположении $a_1 \neq 0$) достаивается до матрицы $\hat{G}_k(1) = \mathrm{diag}(B_k(1), 1, \ldots, 1)$,

$$B_k(1) = \begin{pmatrix} 1 & 0 & 0 \\ a''_2 & 1 & 0 \\ a''_3 & 0 & 1 \end{pmatrix}.$$

Затем осуществляется переход к матрице

$$A_{(k)}(1) = \hat{G}_k(1)^{-1}A_{(k)}\hat{G}_k(1) = \begin{pmatrix} * & * & * & * & * & \ldots & * & * \\ * & * & * & * & * & \ldots & * & * \\ + & * & * & * & * & \ldots & * & * \\ + & + & * & * & * & \ldots & * & * \\ 0 & 0 & 0 & * & * & \ldots & * & * \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & * & * \end{pmatrix},$$

отличающаюшейся от верхней хессенберговой выступом 2×2. Подберём матрицу $\hat{G}_k(2) = \mathrm{diag}(1, B_k(2), 1, \ldots, 1)$ с блоком $B_k(2)$ размера 3×3 таким образом, что

$$\hat{G}_k(2)^{-1}A_{(k)}(1) = \begin{pmatrix} * & * & * & * & * & \ldots & * & * \\ * & * & * & * & * & \ldots & * & * \\ 0 & * & * & * & * & \ldots & * & * \\ 0 & + & * & * & * & \ldots & * & * \\ 0 & 0 & 0 & * & * & \ldots & * & * \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & * & * \end{pmatrix}.$$

Выбор блока $B_k(2)$ осуществляется с целью обнуления компонент первого столбца соответствующего блока, начиная со второй (в QR–алгоритме — с использованием матриц вращений или матрицы отражения, а LR–алгоритме — с использованием нижней унитреугольной матрицы отличной от едичной компонентами первого столбца). Тогда

$$A_{(k)}(2) \;=\; \hat{G}_k(2)^{-1}A_{(k)}(1)\hat{G}_k(2) \;=\; \begin{pmatrix} * & * & * & * & * & * & \dots & * & * \\ * & * & * & * & * & * & \dots & * & * \\ 0 & * & * & * & * & * & \dots & * & * \\ 0 & + & * & * & * & * & \dots & * & * \\ 0 & + & + & * & * & * & \dots & * & * \\ 0 & 0 & 0 & 0 & * & * & \dots & * & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & * & * \end{pmatrix}.$$

Продолжая подобный процесс вытеснения блока 2×2, мы построим $n-3$ матрицы $\hat{G}_k(i) = \mathrm{diag}(\underbrace{1,\dots,1}_{i-1}, B_k(i), 1, \dots, 1) \in G$, $i = 1, \dots, n-3$, с блоками $B_k(i)$ размера 3×3 и матрицы $A_{(k)}(i) = \hat{G}_k(i)^{-1}A_{(k)}(i-1)\hat{G}_k(i)$, $A_{(k)}(0) = A_{(k)}$, где

$$A_{(k)}(n-3) \;=\; \hat{G}_k(n-3)^{-1}\cdots\hat{G}_k(1)^{-1}A_{(k)}\hat{G}_k(1)\cdots\hat{G}_k(n-3) \;=\; \begin{pmatrix} * & * & \dots & * & * & * & * \\ * & * & \dots & * & * & * & * \\ & & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & * & * & * & * \\ 0 & 0 & \dots & * & * & * & * \\ 0 & 0 & \dots & 0 & * & * & * \\ 0 & 0 & \dots & 0 & + & * & * \end{pmatrix}.$$

Остаётся подобрать матрицу $\hat{G}_k(n-2) = \mathrm{diag}(1, \dots, 1, B_k(n-2)) \in G$ с блоком $B_k(n-2)$ размера 2×2 таким образом, что матрица $\hat{G}_k(n-2)^{-1}A_{(k)}(n-3)$ и значит, матрица $A_{(k+2)} = \hat{G}_k(n-2)^{-1}A_{(k)}(n-3)\hat{G}_k(n-2)$ имеет верхнюю хессенбергову форму. Отметим, что построению матрица

$$\hat{G}_k(1)\cdots\hat{G}_k(n-2)$$

имеет первый столбец параллельный первому столбцу матрицы $A_{(k)}^2 - (t_k + \overline{t_k})A_{(k)} + |t_k|^2 E$, что делает её выбор однозначным в случае неразложимой матрицы $A_{(k+2)}$ с точностью до умножения на диагональную матрицу из ± 1.

Сходимость QR– и LR–алгоритмов

Сразу оговоримся, что представленные здесь выводы относятся к частным, весьма идеализированным ситуациям, но, тем не менее, они проясняют природу численной сходимости этих алгоритмов в общем случае, в основе которой лежит ограниченность используемой разрядной сетки. Отметим и то, что результаты о сходимости QR–алгоритма появились сравнительно недавно в работах Е. Е. Тыртышникова, а приведённый ниже

результат о сходимости LR–алгоритма входит в число хорошо известных классических результатов вычислительных методов.

Теорема 0.83. *Пусть $A = D\Lambda D^{-1}$, Λ — диагональная матрица, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $|\lambda_1| \geq \ldots \geq |\lambda_m| > |\lambda_{m+1}| \geq \ldots \geq |\lambda_n|$ для некоторого $1 \leq m \leq n-1$, причём в матрице D^{-1} m-ая главная квадратная подматрица $D_m^{-1} = (d'_{ij})_{i,j=1}^m$ невырождена. Тогда матрицы $A_{(k)}$, $k \geq 1$, в QR–алгоритме, записанные в блочной форме*

$$A_{(k)} \;=\; \begin{pmatrix} A_{11}(k) & A_{12}(k) \\ A_{21}(k) & A_{22}(k) \end{pmatrix}$$

с блоками $A_{11}(k)$, $A_{12}(k)$, $A_{21}(k)$, $A_{22}(k)$ размеров $m \times m$, $m \times (n-m)$, $(n-m) \times m$ и $(n-m) \times (n-m)$, обладают следующим свойством: $A_{21}(k) = O(|\lambda_{m+1}/\lambda_m|^{k-1}) \to 0$ при $k \to +\infty$. При этом в случае если матрица A обладает собственными значениями с попарно различными модулями, матрицы $A_{(k)}$ становятся с ростом k близкими к верхним треугольным матрицам с диагональю Λ, а значит, в данной ситуации QR–алгоритм может быть интерпретирован как процесс итерационного построения формы Шура матрицы A.

Доказательство. Напомним, что $A^k = \hat{Q}_1 \cdots \hat{Q}_k \hat{R}_k \cdots \hat{R}_1 = Q_k R_k$ (см. взаимосвязь между степенной QR–итерацией и QR–алгоритмом),

$$A_{(k+1)} \;=\; \hat{Q}_k^{-1} \cdots \hat{Q}_1^{-1} A \hat{Q}_1 \cdots \hat{Q}_k \;=\; (A^k R_k^{-1})^{-1} A (A^k R_k^{-1}) \,.$$

По условию матрица D^{-1} обладает блочным LR–разложением

$$D^{-1} \;=\; LU \;=\; \begin{pmatrix} E_m & 0 \\ L_{21} & E_{n-m} \end{pmatrix} \begin{pmatrix} U_{11} & U_{21} \\ 0 & U_{22} \end{pmatrix},$$

где E_m и E_{n-m} — единичные матрицы размеров $m \times m$ и $(n-m) \times (n-m)$. Поэтому мы можем записать

$$A^k R_k^{-1} \;=\; D\Lambda^k D^{-1} R_k^{-1} \;=\; D\Lambda^k L \Lambda^{-k} (\Lambda^k U R_k^{-1}) \;=\; D\Lambda^k L \Lambda^{-k} Z_k \,,$$

выделив в отдельный множитель блочную верхнюю треугольную матрицу $Z_k = \Lambda^k U R_k^{-1}$. Тогда

$$A_{(k+1)} \;=\; Z_k^{-1} \Lambda^k L^{-1} \Lambda^{-k} D^{-1} A D \Lambda^k L \lambda^{-k} Z_k \;=\; Z_k^{-1} \Lambda^k L^{-1} \Lambda L \Lambda^{-k} Z_k \;=\; Z_k^{-1} H_k Z_k$$

и

$$\begin{aligned} H_k \;=\; \Lambda^k L^{-1} \Lambda L \Lambda^{-k} \;=\; & \\ \begin{pmatrix} \Lambda_1^k & 0 \\ 0 & \Lambda_2^k \end{pmatrix} \begin{pmatrix} E_m & 0 \\ -L_{21} & E_{n-m} \end{pmatrix} \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} E_m & 0 \\ L_{21} & E_{n-m} \end{pmatrix} \begin{pmatrix} \Lambda_1^{-k} & 0 \\ 0 & \Lambda_2^{-k} \end{pmatrix} \;=\; & \\ \begin{pmatrix} \Lambda_1^k & 0 \\ 0 & \Lambda_2^k \end{pmatrix} \begin{pmatrix} \Lambda_1 & 0 \\ -L_{21}\Lambda_1 + \Lambda_2 L_{21} & \Lambda_2 \end{pmatrix} \begin{pmatrix} \Lambda_1^{-k} & 0 \\ 0 & \Lambda_2^{-k} \end{pmatrix} \;=\; & \\ \begin{pmatrix} \Lambda_1 & 0 \\ -\Lambda_2^k L_{21} \Lambda_1^{1-k} + \Lambda_2^{1+k} L_{21} \Lambda_1^{-k} & \Lambda_2 \end{pmatrix} \;=\; \begin{pmatrix} \Lambda_1 & 0 \\ H_{21}(k) & \Lambda_2 \end{pmatrix}, & \end{aligned}$$

где $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$, $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ и $\Lambda_2 = \mathrm{diag}(\lambda_{m+1}, \ldots, \lambda_n)$ и

$$\|H_{21}(k)\|_2 \;=\; \|-\Lambda_2^k L_{21}\Lambda_1^{1-k} + \Lambda_2^{1+k} L_{21}\Lambda_1^{-k}\|_2 \le$$
$$\|-L_{21}\Lambda_1 + \Lambda_2 L_{21}\|_2 \|\Lambda_2^k\|_2 \|\Lambda_1^{-k}\|_2 \;=\; \|-L_{21}\Lambda_1 + \Lambda_1 L_{21}\|_2 |\lambda_{m+1}/\lambda_m|^k \;=\; c|\lambda_{m+1}/\lambda_m|^k .$$

Ввиду ортогональности (унитарности) матрицы $A^k R_k^{-1}$, $\|A^k R_k^{-1}\|_2 = 1$, и потому

$$Z_k \;=\; (D\Lambda^k L \Lambda^{-k})^{-1}(A^k R_k^{-1}) \;=\; \begin{pmatrix} E_m & 0 \\ -\Lambda_2^k L_{21}\Lambda_1^{-k} & E_{n-m} \end{pmatrix} D^{-1}(A^k R_k^{-1}) ,$$
$$\|Z_k^{\pm 1}\|_2 \le \|\Lambda^k L^{\mp 1}\Lambda^{-k}\|_2 \|D^{\mp 1}\|_2 \to \|D^{\mp 1}\|_2 , \quad k \to +\infty .$$

Отсюда следует также, что $\|Z_k^{\pm 1}\|_2 \le \hat{c}$ для некоторого $\hat{c} > 0$. Таким образом,

$$A_{(k+1)} \;=\; \begin{pmatrix} A_{11}(k+1) & A_{21}(k+1) \\ A_{21}(k+1) & A_{22}(k+1) \end{pmatrix} \;=$$
$$Z_k^{-1} \begin{pmatrix} \Lambda_1 & 0 \\ H_{21}(k) & \Lambda_2 \end{pmatrix} Z_k \;=\; Z_k^{-1}\Lambda Z_k + Z_k^{-1} \begin{pmatrix} 0 & 0 \\ H_{21}(k) & 0 \end{pmatrix} Z_k ,$$

где вследствие блочной верхней трегольной структуры матриц $Z_k^{\pm 1}$ (см. ранее) первое слагаемое является блочной верхней треугольной матрицей и значит, блок $A_{21}(k+1)$ равен в точности тому же блоку второго слагаемого,

$$A_{21}(k+1) \;=\; \left(Z_k^{-1} \begin{pmatrix} 0 & 0 \\ H_{21}(k) & 0 \end{pmatrix} Z_k \right)_{21} .$$

Остаётся заметить, что

$$\left\| Z_k^{-1} \begin{pmatrix} 0 & 0 \\ H_{21}(k) & 0 \end{pmatrix} Z_k \right\|_2 \le k_2(Z_k)\|H_{21}(k)\|_2 \le c\hat{c}^2 |\lambda_{m+1}/\lambda_m|^k \to 0 , \quad k \to +\infty .$$

С учётом эквивалетности матричных норм последнее гарантирует нам, что все коэффициентов указанной здесь матрицы имеют как функции от k вид $O(|\lambda_{m+1}/\lambda_m|^k)$ при $k \to +\infty$. □

Из приведённого доказательства сразу следует, что данное утверждение остаётся справедливым для любого GR–алгоритма, в котором R–факторы GR–разложения являются верхними треугольными матрицами, при условии ограниченности сверху всех норм $\|(A^k R_k^{-1})^{\pm 1}\|_2$, $k \ge 1$ (позднее соображения такого рода будут использоваться в обосновании сходимости LR–алгоритма).

Теорема 0.84. *Пусть $A = D\Lambda D^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, $D^{-1} = LPU$ — модифицированное разложение Брюа матрицы D^{-1}. Тогда матрицы $A_{(k)}$ в QR–алгоритме становятся при $k \to +\infty$ близкими к верхним треугольным матрицам с диагональю $P^{-1}\Lambda P$.*

Доказательство. Как и в предыдущем доказательстве, мы можем записать

$$A_{(k+1)} \;=\; (A^k R_k^{-1})^{-1} A (A^k R_k^{-1}) \;=\; R_k D\Lambda D^{-1} R_k^{-1} \;=$$
$$R_k U^{-1} P^{-1} L^{-1} \Lambda L P U R_k^{-1} \;=\; T_k^{-1}(P^{-1}\Lambda^k L^{-1}\Lambda L \Lambda^{-k} P) T_k ,$$

полагая $T_k = P^{-1}\Lambda^k PUR_k^{-1}$. Нетрудно заметить, что указанная здесь матрица T_k является верхней треугольной. Кроме того, $A^k R_k^{-1} = D\Lambda^k L\Lambda^{-k} PT_k$, $\|A^k R_k^{-1}\|_2 = 1$ и потому

$$\|T_k^{\pm 1}\|_2 \;=\; \|(P^{-1}\Lambda^k L^{-1}\Lambda^{-k} D^{-1} A^k R_k)^{\pm 1}\|_2 \le \|D^{\mp 1}\|_2 \|\Lambda^k L^{\mp 1}\Lambda^{-k}\|_2\,,$$

где матрица $L = (l_{ij})$ нижняя треугольная, $\Lambda^k L\Lambda^{-k} = (l_{ij}(\lambda_i/\lambda_j)^k) \to \operatorname{diag}(l_{11},\ldots,l_{nn})$ при $k \to +\infty$. Вследствие этого найдётся $c > 0$, для которого $\|T_k^{\pm 1}\|_2 \le c$ при всех $k \ge 1$. Запишем матрицу $L^{-1}\Lambda L$ в виде $L^{-1}\Lambda L = \Lambda + \Delta$, где Δ — нижняя треугольная матрица с нулевой диагональю. Тогда

$$P^{-1}\Lambda^k L^{-1}\Lambda L\Lambda^{-k} P \;=\; P^{-1}\Lambda P + P^{-1}\Lambda^k \Delta\Lambda^{-k} P \;=\; P^{-1}\Lambda P + O((\max_i |\lambda_{i+1}/\lambda_i|)^k)\,,$$

а значит,

$$A_{(k+1)} \;=\; T_k^{-1}(P^{-1}\Lambda^k L^{-1}\Lambda L\Lambda^{-k} P)T_k \;=\; T_k^{-1} P^{-1}\Lambda P T_k + O((\max_i |\lambda_{i+1}/\lambda_i|)^k)\,.$$

Остаётся лишь заметить, что матрица T_k является по построению верхней треугольной. □

Скажем теперь несколько слов о сходимости LR–алгоритма. Напомним, что компонентны LR–разложения матрицы A (при условии его осуществимости) могут быть вычислены следующим образом: определив матрицы

$$A_{ik} \;=\; (a_1(i)\ldots a_{i-1}(i)a_k(i)) \;=\; \begin{pmatrix} a_{11} & \ldots & a_{1i-1} & a_{1k} \\ a_{21} & \ldots & a_{2i-1} & a_{2k} \\ \ldots\ldots & \ldots & \ldots\ldots & \ldots\ldots \\ a_{i1} & \ldots & a_{ii-1} & a_{ik} \end{pmatrix},$$

$$A^{ki} \;=\; \begin{pmatrix} a^1(i) \\ \vdots \\ a^{i-1}(i) \\ a^k(i) \end{pmatrix} \;=\; \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1i} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{i-11} & a_{i-12} & \ldots & a_{i-1i} \\ a_{k1} & a_{k2} & \ldots & a_{ki} \end{pmatrix},$$

где $1 \le i \le k \le n$ и $A_{1k} = a_{1k}$, $A^{k1} = a_{k1}$ при $i = 1$, мы можем найти компоненты нижней унитреугольной матрицы L и верхней треугольной матрицы R, составляющих LR–разложение $A = LR$, при помощи равенств $A_{ik} = L_i R_{ik}$, $\det A_{ik} = r_{11}\cdots r_{i-1i-1}r_{ik}$, $1 < i \le k \le n$, и $\det A_{1k} = r_{1k} = a_{1k}$, $k = 1,\ldots,n$,

$$r_{ik} \;=\; \frac{\det A_{ik}}{\det A_{i-1i-1}} \quad (2 \le i \le k \le n),$$

и равенств $A^{ki} = L^{ki}R_i$, $\det A^{ki} = l_{ki}r_{11}\cdots r_{ii}$, $1 < i \le k \le n$, и $\det A^{k1} = l_{k1}r_{11} = a_{k1}$, $k = 1,\ldots,n$,

$$l_{ki} \;=\; \frac{\det A^{ki}}{\det A^{ii}} \quad (1 \le i \le k \le n)\,.$$

При этом $\det A^{ii} = \det A_{ii} = \det A_i$, $i = 1,\ldots,n$.

Замечание 0.85. *Пусть $B = (b_{ij})$ и $C = (c_{ij})$ — матрицы размера $n \times n$. Тогда для любого $k = 1,\ldots,n$*

$$\det(BC)_k \;=\; \left|\begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{k1} & \dots & b_{kn} \end{pmatrix}\begin{pmatrix} c_{11} & \dots & c_{1k} \\ \vdots & \ddots & \vdots \\ c_{n1} & \dots & c_{nk} \end{pmatrix}\right| =$$

$$\sum_{1\le i_1\ne\ldots\ne i_k\le n}\begin{vmatrix} b_{1i_1} & \dots & b_{1i_k} \\ \vdots & \ddots & \vdots \\ b_{ki_1} & \dots & b_{ki_k} \end{vmatrix} c_{i_1 1}\cdots c_{i_k k} \;=\; \sum_{1\le i_1<\ldots<i_k\le n}\begin{vmatrix} b_{1i_1} & \dots & b_{1i_k} \\ \vdots & \ddots & \vdots \\ b_{ki_1} & \dots & b_{ki_k} \end{vmatrix}\begin{vmatrix} c_{i_1 1} & \dots & c_{i_1 k} \\ \vdots & \ddots & \vdots \\ c_{i_k 1} & \dots & c_{i_k k} \end{vmatrix}.$$

Доказательство. Достаточно напомнить, что определитель n-линеен и кососимметричен. □

Теорема 0.86. *Пусть $A = D\Lambda D^{-1}$, $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$, $|\lambda_1| > \ldots > |\lambda_n|$, и при этом для всех матриц степенной LR-итерации KA^kM, $k \ge 1$, (K и M — фиксированные невырожденные матрицы) осуществимо LU-разложение (LR-разложение), $KA^kM = L_kR_k$, $L_k = (l_{ij}(k))$, $R_k = (r_{ij}(k))$. Тогда*

$$\lambda_i \;=\; \lim_{k\to+\infty}\frac{r_{ii}(k)}{r_{ii}(k-1)} \quad (i = 1,\ldots,n),$$

$$L_k \to L, \quad \overline{R}_k = \operatorname{diag}(r_{11}(k),\ldots,r_{nn}(k))^{-1}R_k \to R, \quad (k\to+\infty),$$

где L и R — некоторые нижняя и верхняя унитреугольные матрицы.

Доказательство. Начнём с диагональных элементов. Согласно приведённых ранее формул

$$\det(KA^kM)_i \;=\; \det(B\Lambda^kC)_i \;=\; \sum_{1\le j_1<\ldots<j_i\le n}\lambda_{j_1}^k\cdots\lambda_{j_i}^k\begin{vmatrix} b_{1j_1} & \dots & b_{1j_i} \\ \vdots & \ddots & \vdots \\ b_{ij_1} & \dots & b_{ij_i} \end{vmatrix}\begin{vmatrix} c_{j_1 1} & \dots & c_{j_1 i} \\ \vdots & \ddots & \vdots \\ c_{j_i 1} & \dots & c_{j_i i} \end{vmatrix} =$$
$$\lambda_1^k\cdots\lambda_i^k(\det B_i\det C_i + O((\lambda_{i+1}/\lambda_i)^k)),$$

где $B = KD$ и $C = D^{-1}M$. Поэтому

$$r_{11}(k) \;=\; \det(KA^kM)_1 \;=\; \lambda_1^k(b_{11}c_{11} + O((\lambda_2/\lambda_1)^k)$$

и при $i > 1$

$$r_{ii}(k) \;=\; \frac{\det(KA^kM)_i}{\det(KA^kM)_{i-1}} \;=\; \lambda_i^k\left(\frac{\det B_i\det C_i}{\det B_{i-1}\det C_{i-1}} + O((\lambda_{i+1}/\lambda_i)^k, (\lambda_i/\lambda_{i-1})^k)\right),$$

а, следовательно, $r_{ii}(k)/r_{ii}(k-1) \to \lambda_i$ при $k\to+\infty$. Кроме того,

$$\det(KA^kM)^{ji} \;=\; \left|\begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ \dots & \dots & \dots & \dots \\ b_{i-11} & b_{i-12} & \dots & b_{i-1n} \\ b_{j1} & b_{j2} & \dots & b_{jn} \end{pmatrix}\Lambda^k\begin{pmatrix} c_{11} & c_{12} & \dots & c_{1i} \\ \dots & \dots & \dots & \dots \\ c_{n-11} & c_{n-12} & \dots & c_{n-1i} \\ c_{n1} & c_{n2} & \dots & c_{ni} \end{pmatrix}\right| =$$

$$\sum_{1\le l_1<\ldots<l_i\le n}\lambda_{l_1}^k\cdots\lambda_{l_i}^k\begin{vmatrix} b_{1l_1} & b_{1l_2} & \dots & b_{1l_i} \\ \dots & \dots & \dots & \dots \\ b_{i-1l_1} & b_{i-1l_2} & \dots & b_{i-1l_i} \\ b_{jl_1} & b_{jl_2} & \dots & b_{jl_i} \end{vmatrix}\begin{vmatrix} c_{l_1 1} & c_{l_1 2} & \dots & c_{l_1 i} \\ \dots & \dots & \dots & \dots \\ c_{l_{i-1}1} & c_{l_{i-1}2} & \dots & c_{l_{i-1}i} \\ c_{l_i 1} & c_{l_i 2} & \dots & c_{l_i i} \end{vmatrix} =$$
$$\lambda_1^k\cdots\lambda_i^k(\det B^{ji}\det C_i + O((\lambda_{i+1}/\lambda_i)^k))$$

и значит, при всех $1 \le i < j \le n$

$$l_{ji}(k) \;=\; \frac{\det(KA^kM)^{ji}}{\det(KA^kM)_i} \;=\; \frac{\det B^{ji}}{\det B_i} + O((\lambda_{i+1}/\lambda_i)^k) \to l_{ji} \;=\; \frac{\det B^{ji}}{\det B_i} \quad (k \to +\infty)\,,$$

т.е. $L_k \to L = (l_{pq})$, $l_{pp} = 1$, $l_{pq} = 0$, $p > q$. Аналогичным образом,

$$\det(KA^kM)_{ij} \;=\; \lambda_1^k \cdots \lambda_i^k (\det B_i \det C_{ij} + O((\lambda_{i+1}/\lambda_i)^k))\,,$$

откуда следует, что $r_{1j}(k) = \lambda_1^k(b_{11}c_{1j} + O((\lambda_2/\lambda_1)^k))$, $j = 1, \ldots, n$,

$$r_{ij}(k) \;=\; \frac{\det(KA^kM)_{ij}}{\det(KA^kM)_{i-1}} \;=\; \lambda_i^k \left(\frac{\det B_i \det C_{ij}}{\det B_{i-1} \det C_{i-1}} + O((\lambda_{i+1}/\lambda_i)^k, (\lambda_i/\lambda_{i-1})^k) \right),$$

где $2 \le i \le j \le n$. Таким образом, $\Lambda^{-k} R_k \to R' = (r'_{ij})$, $r'_{1j} = b_{11}c_{1j}$ и

$$r'_{ij} \;=\; \frac{\det B_i \det C_{ij}}{\det B_{i-1} \det C_{i-1}} \quad (2 \le i \le j \le n)\,,$$

$r'_{pq} = 0$, $p > q$. Поэтому $\overline{R}_k = \operatorname{diag}(R_k)^{-1} R_k \to R = \operatorname{diag}(R')^{-1} R'$. □

Следствие 0.87. *При выполнении условий теоремы для $K = M = E$ в LR-алгоритме, реализованном для матрицы A (в предположении его осуществимости),*

$$\hat{L}_k \;=\; L_{k-1}^{-1} L_k \to E\,, \quad \hat{R}_k \;=\; R_k R_{k-1}^{-1} \to L^{-1} A L\,, \quad (k \to +\infty)\,,$$

где $L^{-1}AL$ — верхняя треугольная матрица с диагональю равной диагонали матрицы Λ.

Доказательство. Достаточно заметить, что $L_k^{-1} A L_k = A_{(k+1)} = \hat{R}_k \hat{L}_k$, $k \ge 1$ (после этого остаётся лишь перейти к пределу по k). □

Лекция 11. Некоторые алгоритмы симметрической проблемы собственных значений.

Симметрическая часть проблемы собственных значений в значительной мере проще полной проблемы собственных значений уже хотя бы по той причине в данном случае собственные значения являются вещественными (нахождение вещественных корней многочлена является заведомо более простой задачей нежели задача поиска всех его комплексных корней).

Метод Якоби

Метод вращений Якоби нахождения всех собственных значений симметрической вещественной матрицы $A = A^t \in M_n(\mathbb{R})$ является не самым быстрым, но довольно точным методом полной симметрической проблемы собственных значений. Идея метода состоит в построении последовательности ортогонально подобных матриц исходной матрице A, сходящейся к диагональной матрице.

Положим $\Sigma(B) = \sum_{i \neq j} b_{ij}^2$, $B \in M_n(\mathbb{R})$. Применительно к $A = A^t = U\Lambda U^t$ для подходящих $U \in O_n(\mathbb{R})$ и $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\{\lambda_i\} = \operatorname{Spec}(A)$, мы имеем: $\Sigma(U^t A U) = \Sigma(\Lambda) = 0$ и, как следствие, U является одним из тех элементов $O_n(\mathbb{R})$, на которых достигается минимум функционала $V \longmapsto \Sigma(V^t A V)$, $V \in O_n(\mathbb{R})$. При этом каждому решению $U' \in O_n(\mathbb{R})$ данной минимизационной задачи соответствует диагональная матрица $\Lambda' = {U'}^t A U'$, на диагонали которой находятся собственные значения матрицы A. Предположим, что мы располагаем последовательностью симметрических матриц $\{A_i\}_{i \geq 0}$, в которой

1. $A_0 = A$, $A_k = U_k A_{k-1} U_k^t$ для некоторой $U_k \in O_n(\mathbb{R})$;
2. $\Sigma(A_{k-1}) < \Sigma(A_k)$ при всех $k \geq 1$;
3. $\lim\limits_{k \to \infty} \Sigma(A_k) = 0$.

Замечание 0.88. *Пусть $\varepsilon > 0$ и $A_k = (a_{ij}^{(k)}) \in \{A_i\}$, $\Sigma(A_k) < \varepsilon$. Тогда*

$$\min_i |\lambda - a_{ii}^{(kk)}| \leq \sqrt{(n-1)\varepsilon} \quad (\lambda \in \operatorname{Spec}(A)) .$$

Доказательство. В силу теоремы о кругах Гершгорина для любого $\lambda \in \operatorname{Spec}(A_k) = \operatorname{Spec}(A)$ найдётся i, $1 \leq i \leq n$, такой, что

$$|\lambda - a_{ii}^{(k)}| \leq \sum_{j \neq i} |a_{ij}^{(k)}| \leq \sqrt{n-1} \sqrt{\sum_{j \neq i} |a_{ij}|^2} \leq \sqrt{(n-1)\varepsilon} .$$

□

В методе Якоби последовательность $\{A_i\}$ строится по правилу

$$A_k \;=\; T_{i_k j_k}(\phi_k) A_{k-1} T_{i_j j_k}(-\phi_k) \quad (k \geq 1) ,$$

где выбор угла ϕ_k обеспечивает выполнение условия 2, а пары (i_k, j_k) — условия 3. Начнём с обсуждения выбора угла вращения. Положим $B = T_{ij}(\phi) A T_{ij}(-\phi)$ и вычислим

разность $\Sigma(B) - \Sigma(A)$, пользуясь тем, что B отличается от A только компонентами строк и столбцов с индексами i и j,

$$\Sigma(B)-\Sigma(A) \;=\; \sum_{m\neq i,j} (b_{im}^2+b_{im}^2+b_{jm}^2+b_{mj}^2-a_{im}^2-a_{mi}^2-a_{jm}^2-a_{mj}^2)+(b_{ij}^2+b_{ji}^2-a_{ij}^2-a_{ji}^2) \;=$$
$$2\biggl(\sum_{m\neq i,j} (b_{im}^2+b_{jm}^2-a_{im}^2-a_{jm}^2)+(b_{ij}^2-a_{ij}^2)\biggr) \;=\; 2(b_{ij}^2-a_{ij}^2)\,,$$

поскольку при $m \neq i, j$

$$\begin{pmatrix} b_{im} \\ b_{jm} \end{pmatrix} \;=\; \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} a_{im} \\ a_{jm} \end{pmatrix}, \quad b_{im}^2+b_{jm}^2 \;=\; a_{im}^2+a_{jm}^2\,.$$

Поэтому минимальное значение разности $\Sigma(B) - \Sigma(A)$ при фиксированных A и (i, j) достигается при $b_{ij} = 0$. Так как

$$\begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} \;=\; \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix},$$

$b_{ij} = 1/2 \sin 2\phi(a_{ii} - a_{jj}) + \cos 2\phi a_{ij}$ и равенство $b_{ij} = 0$ выполняется, если и только если $\tan 2\phi = -2a_{ij}/(a_{ii}-a_{jj})$ при $a_{ii} \neq a_{jj}$ или $\phi = \pi/4$ при $a_{ii} = a_{jj}$. Выбирая $\phi \in [-\pi/4, \pi/4]$, мы обеспечим $\cos 2\phi \geq 0$, $\operatorname{sign}\sin\phi = \operatorname{sign}\tan 2\phi$ и $\cos 2\phi = (1+\tan^2 2\phi)^{-1/2}$,

$$\cos\phi \;=\; \biggl(\frac{1}{2}\biggl(1+\frac{1}{(1+\tan^2 2\phi)^{1/2}}\biggr)\biggr)^{1/2},\; \sin\phi \;=\; \operatorname{sign}\tan 2\phi\biggl(\frac{1}{2}\biggl(1-\frac{1}{(1+\tan^2 2\phi)^{1/2}}\biggr)\biggr)^{1/2}.$$

Полагая $x = -2a_{ij}$ и $y = a_{ii} - a_{jj}$, мы можем переписать эти равенства для $y \neq 0$ в виде

$$\cos\phi \;=\; \biggl(\frac{1}{2}\biggl(1+\frac{|y|}{(x^2+y^2)^{1/2}}\biggr)\biggr)^{1/2},\; \sin\phi \;=\; \operatorname{sign}(xy)\biggl(\frac{1}{2}\biggl(1-\frac{|y|}{(x^2+y^2)^{1/2}}\biggr)\biggr)^{1/2}.$$

При этом ввиду возможной близости отношения $\frac{|y|}{(y^2+x^2)^{1/2}}$ к 1 при больших $|y|$ лучше использовать формулу

$$\sin\phi \;=\; \frac{\tan 2\phi\cos 2\phi}{2\cos\phi} \;=\; \frac{x|y|}{y(x^2+y^2)^{1/2}}\frac{1}{2\cos\phi} \;=\; \frac{\operatorname{sign}(xy)|x|}{2\cos\phi(x^2+y^2)^{1/2}}\,.$$

Таким образом, реализация условия $\Sigma(B) < \Sigma(A)$ может быть обеспечена выбором (i, j), $a_{ij} \neq 0$, и угла ϕ, где $\phi = \pi/4$, $\cos\phi = \sin\phi = 1/\sqrt{2}$, при $y = 0$, и

$$\cos\phi \;=\; \biggl(\frac{1}{2}\biggl(1+\frac{|y|}{(x^2+y^2)^{1/2}}\biggr)\biggr)^{1/2},\; \sin\phi \;=\; \frac{\operatorname{sign}(xy)|x|}{2\cos\phi(x^2+y^2)^{1/2}}$$

при $y \neq 0$.

Покажем, что выбором (i, j) при описанном выше правиле подбора угла ϕ можно обеспечить сходимость $\Sigma(A_k) \to 0$. Имеются различные стратегии (правила) такого выбора, определяющие различные скорости сходимости всего алгоритма. Точнее, для практической реализации наиболее значимым является прогнозирование шага выполнения условия $\Sigma(A_k) < \varepsilon$, которое гарантирует близость диагональных элементов ма-

трицы A_k к собственным значениям матрицы A с точностью $\varepsilon\sqrt{n-1}$ (см. выше). Простейшая стратегия такого рода — выбор пары (i_k, j_k), отвечающей максимальному по модулю внедиагональному элементу матрицы A_{k-1} при выполнении k-го шага алгоритма (соответствующая версия алгоритма — метод вращений с выбором максимального элемента), $|a_{i_k j_k}^{(k-1)}| = \max_{i \neq j} |a_{ij}^{(k-1)}|$.

Замечание 0.89. *В методе вращений с выбором максимального элемента*

$$\Sigma(A_k) \leq \left(1 - \frac{2}{n(n-1)}\right)^k \Sigma(A) \quad (k \geq 1)\,.$$

Доказательство. Для любой стратегии выбора $\Sigma(A_k) = \Sigma(A_{k-1}) - 2(a_{i_k j_k}^{(k-1)})^2$, $k \geq 1$. Поскольку в нашем случае

$$\Sigma(A_{k-1}) \leq n(n-1)(a_{i_k j_k}^{(k-1)})^2\,, \quad \Sigma(A_k) \leq \left(1 - \frac{2}{n(n-1)}\right)\Sigma(A_{k-1})\,,$$

мы сразу получаем, что $\Sigma(A_k) \leq q^k \Sigma(A)$ для $q = 1 - \frac{2}{n(n-1)}$. □

Понятно, что указанная оценка является фактически неулучшаемой и гарантирует весьма медленную сходимость алгоритма при больших n. Вычислительную сложность шага алгоритма метода вращений с выбором максимального элемента составляют перевычисления компонент i_k и j_k строк и столбцов матрицы A_k, которые для всех стратегий выбора проводятся за $O(n)$ арифметических операции, и $n(n-1)/2$ сравнений на выбор пары (i_k, j_k), т.е порядка $n^2/2 + O(n)$ операций.

Значительно более оптимальной по сложности шага алгоритма с той же оценкой сходимости является стратегия оптимального выбора в методе вращений с выбором оптимального элемента. В данном случае i_k выбирается как индекс строки матрицы A_{k-1} с максимальной суммой квадратов внедиагональных элеметов, а j_k — как индекс максимального по модулю внедиагонального элемента её i_k-ой строки. В рамках реализации помимо матриц A_k хранится также вектор $(b_1^{(k)}, \ldots, b_n^{(k)})$,

$$b_i^{(k)} \;=\; \sum_{j \neq i} (a_{ij}^{(k)})^2 \quad (i = 1, \ldots, n)\,,$$

в котором на k-ом шаге алгоритма перевычисляются только i_k и j_k координаты. Шаг алгоритма в данном случае обходится в $O(n)$ операций.

Используется также стратегия циклического перебора обнуляемых элементов (последовательно в порядке $(12), \ldots, (1n), (23), \ldots, (2n), \ldots, (n-1\,n)$, а затем снова и снова до тех пор пока $\Sigma(A_k) \geq \varepsilon$), но для неё, естественно, нет оценок сходимости.

Метод бисекции

Метод бисекции имеет различные варианты реализации, которые имеют отношение как к полной, так и к частичной симметрической проблеме собственных значений. В идейном плане он представляет собой вариант метода бисекции поиска вещественных корней вещественного многочлена с использованием последовательности Штурма, применяемой в данном случае к характеристическому многочлену неприводимой трёхдиагональной матрицы. Основу метода составляют следующие два утверждения известные

как теорема о минимаксе (теорема Куранта — Фишера) и её следствие — теорема о перемежаемости собственных значений. Мы приведём их в наболее простом матричном случае (естественные обобщения формулируются для самосопряжённых компактных операторов на гильбертовых пространствах).

Теорема 0.90. *Пусть $A = A^* \in M_n(\mathbb{C})$, $\{\alpha_1, \ldots, \alpha_n\}$ — собственные значения A, записанные в невозрастающем порядке, $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$. Тогда*

$$\alpha_i \;=\; \max_{\substack{{}_{\mathbb{C}}V \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} V = i}} \min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)} \;=\; \min_{\substack{{}_{\mathbb{C}}W \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} W = n-i+1}} \max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)} \quad (i = 1, \ldots, n)\,,$$

где $(x,x) = x^*x$, $Q_A(x) = x^*Ax$, $x = (x_1, \ldots, x_n)^t \in \mathbb{C}^n$.

Доказательство. В первую очередь отметим, что указанные здесь отношения можно заменить на значениях непрерывного функционала Q_A на пересечения соответствующих подпространств с единичной окружностью с центром в нуле, а потому ввиду компактности таких пересечений использования минимума и максимума полностью оправдано.

Выберем ортономированный базис собственных векторов $\{q_i\}$ матрицы A, $\|q_i\|_2 = 1$, $Aq_i = \alpha_i q_i$. Тогда для $V = \langle q_1, \ldots, q_i \rangle$ и $W = \langle q_i, \ldots, q_n \rangle$

$$\min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)} \;=\; \min_{(c_1,\ldots,c_i) \neq (0,\ldots,0)} \frac{\sum_{j=1}^{i} |c_j|^2 \alpha_j}{\sum_{j=1}^{i} |c_j|^2} \;=\; \alpha_i\,,$$

$$\max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)} \;=\; \max_{(d_i,\ldots,d_n) \neq (0,\ldots,0)} \frac{\sum_{j=i}^{n} |d_j|^2 \alpha_j}{\sum_{j=i}^{n} |d_j|^2} \;=\; \alpha_i$$

и, следовательно,

$$\inf_{\substack{{}_{\mathbb{C}}W \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} W = n-i+1}} \max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)} \leq \alpha_i \leq \sup_{\substack{{}_{\mathbb{C}}V \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} V = i}} \min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)}\,.$$

Вместе с тем для любых ${}_{\mathbb{C}}V, {}_{\mathbb{C}}W \subset \mathbb{C}^n$, $\dim_{\mathbb{C}} V = i$, $\dim_{\mathbb{C}} W = n - i + 1$, $\dim_{\mathbb{C}} V \cap W \geq 1$, и потому имеется $0 \neq x_{VW} \in V \cap W$, для которого

$$\min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)} \leq \frac{Q_A(x_{VW})}{(x_{VW}, x_{VW})} \leq \max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)}\,.$$

Поэтому

$$\alpha_i \leq \sup_{\substack{{}_{\mathbb{C}}V \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} V = i}} \min_{0 \neq v \in V} \frac{Q_A(v,v)}{(v,v)} \leq \inf_{\substack{{}_{\mathbb{C}}W \subset \mathbb{C}^n, \\ \dim_{\mathbb{C}} W = n-i+1}} \max_{0 \neq w \in W} \frac{Q_A(w,w)}{(w,w)} \leq \alpha_i\,,$$

а значит, мы можем заменить в этих выражениях sup и inf на max и min и получить требуемое равенство. □

Следствие 0.91. *Пусть $A = A^* \in M_n(\mathbb{C})$, $A_k = (a_{ij})_{i,j=1,k}$, $\{\lambda_1(A_k), \ldots, \lambda_k(A_k)\}$ — собственные значения A_k, записанные в невозрастающем порядке. Тогда*

$$\lambda_1(A_{k+1}) \geq \lambda_1(A_k) \geq \lambda_2(A_{k+1}) \geq \ldots \geq \lambda_k(A_{k+1}) \geq \lambda_k(A_k) \geq \lambda_{k+1}(A_{k+1})$$

при всех $k = 1, \ldots, n-1$.

Доказательство. Обозначим через e_i i-ый столбец единичной матрицы $E \in M_n(\mathbb{C})$ и

положим $Z_k = \langle e_1, \ldots, e_k \rangle$. Тогда в силу теоремы о минимаксе

$$\lambda_i(A_{k+1}) \;=\; \max_{\substack{{}_{\mathbb{C}}V \subset Z_{k+1},\\ \dim_{\mathbb{C}} V = i}} \min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)} \geq \max_{\substack{{}_{\mathbb{C}}V \subset Z_{k},\\ \dim_{\mathbb{C}} V = i}} \min_{0 \neq v \in V} \frac{Q_A(v)}{(v,v)} \;=\; \lambda_i(A_k)$$

и вместе с тем

$$\lambda_i(A_{k+1}) \;=\; \min_{\substack{{}_{\mathbb{C}}W \subset Z_{k+1},\\ \dim_{\mathbb{C}} W = k+2-i}} \max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)} \leq \min_{\substack{{}_{\mathbb{C}}W \subset Z_{k},\\ \dim_{\mathbb{C}} W = k+2-i}} \max_{0 \neq w \in W} \frac{Q_A(w)}{(w,w)} \;=\; \lambda_{i-1}(A_k)\,.$$

□

Пусть теперь A — симметрическая трёхдиагональная матрица, $A = (a_{ij}) \in M_n(\mathbb{R})$, $a_{ii} = a_i$, $i = 1, \ldots, n$, $a_{jj+1} = b_j$, $j = 1, \ldots, n-1$. Используя разложение определителя по последнему столбцу, мы получаем следующее трехчленное рекуррентное соотношение

$$\det(A_k) \;=\; a_k \det(A_{k-1}) - b_{k-1}^2 \det(A_{k-2})$$

и вместе с ним соотношение для характеристических многочленов $\chi_k(x) = \det(A_k - xE_k)$ главных квадратных подматриц A_k,

$$\chi_k(x) \;=\; (a_k - x)\chi_{k-1}(x) - b_{k-1}^2 \chi_{k-2}(x) \quad (k = 2, \ldots, n)\,,$$

где $\chi_0(x) = 1$, $\chi_1(x) = a_1 - x$. Начиная с этого моментна мы будем предполагать дополнительно, что матрица A является неприводимой, т.е. не имеет нулей на побочных диагоналях ($b_i \neq 0$).

Замечание 0.92. *Многочлены χ_k и χ_{k+1}, $k = 1, \ldots, n-1$, не имеют общих корней, а потому*

$$\lambda_1(A_{k+1}) > \lambda_1(A_k) > \lambda_2(A_{k+1}) > \ldots > \lambda_k(A_{k+1}) > \lambda_k(A_k) > \lambda_{k+1}(A_{k+1})$$

и, как следствие, многочлены $\{\chi_k\}_{k \geq 1}$ не имеют кратных корней.

Доказательство. Поскольку $\chi_{k+1}(x) = (a_{k+1} - x)\chi_k(x) - b_k^2 \chi_{k-1}(x)$, где $b_k \neq 0$, наличие α, $\chi_{k+1}(\alpha) = \chi_k(\alpha) = 0$, влечёт за собой равенства $0 = \chi_i(\alpha)$ для всех $i = 0, \ldots, k+1$, но $\chi_0 = 1$?! Остаётся воспользоваться теоремой о перемежаемости. □

Замечание 0.93. *Если $\chi_i(\alpha) = 0$ при i, $1 \leq i \leq n-1$, то* $\operatorname{sign} \chi_{i-1}(\alpha)\chi_{i+1}(\alpha) = -1$.

Доказательство. Так как $\chi_{i+1}(\alpha) = -b_i^2 \chi_{i-1}(\alpha) \neq 0$, $\operatorname{sign} \chi_{i+1}(\alpha) = -\operatorname{sign} \chi_{i-1}(\alpha)$. □

Теорема 0.94. *Пусть $\sigma_k(x) = |\{\mu \in \operatorname{Spec}(A_k) \mid \mu < x\}|$ — количество отрицательных собственных значений матрицы $A_k - xE_k$, $k = 1, \ldots, n$. Тогда $\sigma_k(x)$ совпадает с числом перемен знака $m_k(x)$ в наборе $\{\chi_0(x) = 1, \chi_1(x), \ldots, \chi_k(x)\}$, из которого удалены нулевые элементы, $\sigma_k(x) = m_k(x)$.*

Доказательство. Воспользуемся индукцией по $k \geq 1$. При $k = 1$

$$\sigma_1(x) \;=\; \begin{cases} 0, & \text{если } x \leq a_1; \\ 1, & \text{если } x > a_1 \end{cases}$$

и потому $\sigma_1(x)$ очевидным образом равно числу перемен знака в наборе $\{1, a_1 - x\}$.

Допустим, что равенства $\sigma_i(x) = m_i(x)$ уже установлены при $i < k$ для некоторого $k < n$. Тогда с учётом сказанного ранее $m_k(x) = m_{k-1}(x) + \delta(x) = \sigma_{k-1}(x) + \delta(x)$, где

$$\delta(x) \;=\; \begin{cases} 0, & \text{если } \chi_k(x) = 0 \text{ или } \chi_k(x)\chi_{k-1}(x) > 0; \\ 1, & \text{если } \chi_{k-1}(x) = 0 \text{ или } \chi_k(x)\chi_{k-1}(x) < 0. \end{cases}$$

Рассмотрим перечисленные здесь ситуации.

1. Если $\chi_k(x) = 0$, $\lambda_s(A_k - xE_k) = 0$ для некоторого s, тогда $\chi_{k-1}(x) \neq 0$ и

$$\lambda_{s-1}(A_{k-1} - xE_{k-1}) > \lambda_s(A_k - xE_k) \;=\; 0 > \\ \lambda_s(A_{k-1} - xE_{k-1}) > \ldots > \lambda_{k-1}(A_{k-1} - xE_{k-1}) > \lambda_k(A_k - xE_k)\,,$$

а потому в этом случае $\sigma_k(x) = k - s = \sigma_{k-1}(x) = m_{k-1}(x) = m_k(x)$.

2. Если $\chi_{k-1}(x) = 0$, $\lambda_t(A_{k-1} - xE_{k-1}) = 0$ для некоторого t, тогда

$$\lambda_t(A_k - xE_k) > \lambda_t(A_{k-1} - xE_{k-1}) \;=\; 0 > \\ \lambda_{t+1}(A_k - xE_k) > \ldots > \lambda_{k-1}(A_{k-1} - xE_{k-1}) > \lambda_k(A_k - xE_k)\,,$$

а потому $\sigma_{k-1}(x) = m_{k-1}(x) = k - 1 - t$, $\sigma_k(x) = k - t = m_k(x)$.

3. Пусть теперь $\chi_{k-1}(x)\chi_k(x) \neq 0$, $\operatorname{sign}\chi_k(x) = (-1)^{\sigma_k(x)}$, $\operatorname{sign}\chi_{k-1} = (-1)^{\sigma_{k-1}(x)}$. Предположим, что $\sigma_{k-1}(x) > 0$ и индекс s отвечает первому отрицательному собственному значению матрицы $A_{k-1} - xE_{k-1}$. Тогда

$$\lambda_{s-1}(A_k - xE_k) > \lambda_{s-1}(A_{k-1} - xE_{k-1}) > \lambda_s(A_k - xE_k) > \\ \lambda_s(A_{k-1} - xE_{k-1}) > \lambda_{s+1}(A_k - xE_k) > \ldots > \lambda_k(A_k - xE_k)\,,$$

$\lambda_{s-1}(A_{k-1} - xE_{k-1}) > 0$ и потому $\sigma_k(x) = k - s + \delta = \sigma_{k-1}(x) + \delta$, где

$$\delta \;=\; \begin{cases} 0, & \text{если } \lambda_s(A_k - xE_k) > 0; \\ 1, & \text{если } \lambda_s(A_k - xE_k) < 0. \end{cases}$$

Значит, $\operatorname{sign}\chi_k(x) = (-1)^\delta \operatorname{sign}\chi_{k-1}(x)$ и $m_k(x) = m_{k-1}(x) + \delta = \sigma_{k-1}(x) + \delta = \sigma_k(x)$.

В случае если $\sigma_{k-1}(x) = 0$, т.е. $\lambda_i(A_{k-1} - xE_{k-1}) > 0$ при всех i,

$$\sigma_k(x) \;=\; \begin{cases} 0, & \text{если } \lambda_k(A_k - xE_k) > 0; \\ 1, & \text{если } \lambda_k(A_k - xE_k) < 0 \end{cases}$$

и, следовательно, $\operatorname{sign}\chi_{k-1}(x) = 1$, $\operatorname{sign}\chi_k(x) = \operatorname{sign}\lambda_k(A_k - xE_k)$, $m_k(x) = \sigma_k(x)$. Шаг индукции доказан полностью. □

Таким образом, мы располагаем способом вычисления $\sigma(x) = \sigma_n(x) = m_n(x)$ числа собственных значений матрицы A меньших x, а значит, и числа её собственных значений в заданном интервале $[a, b)$ равного $\sigma(b) - \sigma(a) = |\{\lambda \in \operatorname{Spec}(A) \mid a \leq \lambda < b\}|$. Это позволяет предложить следующий алгоритм вычисления $\lambda_{n-k+1}(A)$ с заданной точностью ε:

1. выберем a_0 и b_0, $a_0 < b_0$, из условий: $\sigma(a_0) < k$ и $\sigma(b_0) \geq k$ (в частности, можно положить $a_0 = -\|A\|$ и $b_0 = \|A\| + \varepsilon$ для любой матричной нормы $\|\ \|$);

2. до тех пор пока $b_i - a_i > \varepsilon$, $i = 0, 1, \ldots$, вычисляем $c_i = \frac{a_i + b_i}{2}$ и полагаем $a_{i+1} = c_{i+1}$, $b_{i+1} = b_i$, если $\sigma(c_i) < k$, и $a_{i+1} = a_i$, $b_{i+1} = b_i$, если $\sigma(c_i) \geq k$;

3. в случае если $b_i - a_i \leq \varepsilon$, считаем $c_i = \frac{a_i + b_i}{2}$ искомым ε–близким к $\lambda_{n-k+1}(A)$ приближением кратности $\sigma(b_i) - \sigma(a_i)$.

В действительности найденное таким образом приближение ε–близко к $\lambda_{n-k+1+i}(A)$, $i = 0, \ldots, \sigma(b_i) - \sigma(a_i) - 1$. Поскольку $b_i - a_i = \frac{b_0 - a_0}{2^i}$ условие выхода выполняется при $i \geq \log_2(b_0 - a_0) - \log_2 \varepsilon$.

Сходная идея используется и в алгоритме нахождения всех собственных значений матрицы A в заданном интервале $[a, b)$, $a < b$, с заданной точностью ε, который может быть реализован рекурсивным и последовательным образом:

R применяем алгоритм к интервалам $[a_{i+1}, b'_{i+1})$ и $[a'_{i+1}, b_{i+1})$ для $a_{i+1} = a_i$, $b_{i+1} = b_i$ и $a'_{i+1} = b'_{i+1} = \frac{a_i + b_i}{2}$ до тех пор пока длина $b' - a'$ интервала $[a', b')$, к которому применяется алгоритм больше ε и $\sigma(b') - \sigma(a') > 0$; при $b' - a' \leq \varepsilon$ число $\frac{b' + a'}{2}$ полагается равным искомым приближением точности ε к $\sigma(b') - \sigma(a')$ собственным значением A в интервале $[a', b')$;

S полагая $\sigma(b) = k_2$ и $\sigma(a) = k_1$, последовательно вычисляем искомые приближения к $\lambda_{n-k_2+i}(A)$, $i = 1, \ldots, k_2 - k_1$, при помощи описанного ранее алгоритма с естественной поправкой на ускорение данного процесса, использующее кратности найденных приближений.

Для $a = -\|A\|$ и $b = \|A\| + \varepsilon$ описанные алгоритмы позволяют найти все приближённые собственные значения A с точностью ε.

Основное требование к практической реализации алгоритма состоит в обеспечении численной устойчивости вычисления числа $\sigma(x) = m_n(x)$ перемен знака в наборе $\{\chi_0(x) = 1, \chi_1(x), \ldots, \chi_n(x)\}$, где $\chi_1(x) = a_1 - x$ и далее $\chi_k(x) = (a_k - x)\chi_{k-1}(x) - b^2_{k-1}\chi_{k-2}(x)$. Для этого можно поступить следующим образом:

1. $\delta_0 = \chi_0(x) = 1$, $\delta_1 = \chi_1(x) = a_1 - x$;

2. полагаем $x := \delta_1$, $y := \delta_0$ и далее для $k = 2, \ldots, n$: $a := a_k - x$, $b := b_{k-1}$, $\gamma := 1/(\varepsilon \max\{|x|, |b^2 y|\})$, $u := \gamma(ax - b^2 y)$, $v := \gamma x$, $\delta_k := u$, а затем $x := u$ и $y := v$, где ε — машинная точность;

3. удаляем нули из набора $\{\delta_0, \ldots, \delta_n\}$ и вычисляем число $\sigma(x)$ перемен знака в этом наборе.

Сказанное здесь относится к неприводимым трёхдиагональным матрицам, а в приводимой ситуации применяется к неприводимым составляющим.

Скажем несколько слов о взаимосвязи системы $\{\chi_k\}$ с системой Штурма многочлена χ_n. Для этого в первую очередь напомним само это понятие.

Пусть $f \in \mathbb{R}[x]$, $(f, f') = 1$. Система многочленов $\{f_0, \ldots, f_s\}$, $s \geq 1$, с $f_0 = f$ называется системой Штурма многочлена f, если

1. f_s не имеет действительных корней;

2. любые два соседних многочлена f_i и f_{i+1} не имеют общих корней;

3. если $f_k(\alpha) = 0$, $\alpha \in \mathbb{R}$, $1 \le 1 \le s-1$, то $\operatorname{sign}(f_{k-1}f_{k+1})(\alpha) < 1$;

4. $f_0 f_1$ возрастает в любом корне $\alpha \in \mathbb{R}$ многочлена f_0 в том смысле, что $(f_0 f_1)(\alpha - 0) < 0$ и $(f_0 f_1)(\alpha + 0) > 0$.

Располагая системой Штурма, мы можем вычислять число перемен знака $\omega(x)$ в наборе $\{f_0(x), \ldots, f_s(x)\}$ с вычеркнутыми нулевыми значениями и число корней многочлена f в интервале $(a, b]$ — $\omega(a) - \omega(b)$ (теорема Штурма). Примером системы Штурма могут служить многочлены $\{f_i\}$, $f_0 = f$, $f_1 = f'$, $f_i = f_{i+1}q_{i+2} - f_{i+2}$, $i = 0, \ldots, s$ (алгоритм Евклида), где $f_s = (f, f') = const$.

В нашей ситуации с точностью до порядка последовательность $\{\chi_k\}$ удовлетворяет всем условиям определения последовательности Штурма для χ_n с той лишь разницей, что $\chi_{n-1}\chi_n$ убывает в корне χ_n. Действительно, при движении от a к b, $a < b$, число перемен знака $\sigma(x)$ меняется только при прохождении корня χ_n (см. отмеченные ранее свойства) и при этом увеличивается на 1, т.е. или $\chi_{n-1}(\alpha - 0), \chi_n(\alpha - 0) < 0$ и $\chi_{n-1}(\alpha - 0) < 0$, $\chi_n(\alpha + 0) > 0$, или $\chi_{n-1}(\alpha - 0), \chi_n(\alpha - 0) > 0$ и $\chi_{n-1}(\alpha + 0) > 0$, $\chi_n(\alpha + 0) < 0$, т.е. $(\chi_{n-1}\chi_n)(\alpha - 0) > 0$, $(\chi_{n-1}\chi_n)(\alpha + 0) < 0$, где $\chi_n(\alpha) = 0$.

Симметрические варианты LR–алгоритма и QR–алгоритма

Естественным вариантом LR–алгоритма для симметрических и самосопряжённых матриц является алгоритм Холесского, применяемый в общем случае к матрице $A = A^* > 0$. В наиболее простой форме (без сдвигов) это процесс вида: $A_0 = A = L_0 L_0^*$, $A_1 = L_0^* L_0 = L_1 L_1^*$, и т.д. $A_k = L_k L_k^*$, $A_{k+1} = L_k^* L_k$, $k \ge 0$, где каждый раз вычисляется разложение Холесского матрицы $A_k = A_k^* > 0$ в форме с корнями. При этом практическая реализация процесса предполагает на начальном этапе приведение матрицы A к трёхдиагональному виду методами унитарного (ортогонального) подобия или алгоритмами Ланцоша — Арнольди, что позволит оценивать сложность шага как $O(n)$. Возможна реализация и неявной версии алгоритма. При использовании сдвигов (ординарных) вычислительный процесс преобразуется в форму: $A_k - \tau_k E = L_k L_k^*$, $A_{k+1} = L_k^* L_k + \tau_k E$. Заметим, что в отличие от LR–алгоритма реализация алгоритма Холесского на всех шагах выполнения обеспечивается на начальном этапе условием $A = A^* > 0$.

Наиболее популярные стратегии выбора сдвига: $\tau_k = a_{nn}^{(k)}$ или τ_k — ближайшее к $a_{nn}^{(k)}$ собственное значение матрицы

$$\begin{pmatrix} a_{n-1n-1}^{(k)} & a_{n-1n}^{(k)} \\ a_{nn-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix},$$

где $A_k = (a_{ij}^{(k)})$ (стратегия Уилкинсона). Алгоритм Холесского эквивалентен в симметрической версии QR–алгоритма, точнее справедливо

Замечание 0.95. *Пусть $A_0 = A = A^* > 0$ и $A_k = L_k L_k^*$, $A_{k+1} = L_k^* L_k$, $k \ge 0$. Тогда $A_{k+2} = Q_k^* A_k Q_k = R_k Q_k$, где $A_k = Q_k R_k$ — QR–разложение матрицы A_k, т.е. два шага алгоритма Холесского равносильны одному шагу QR–алгоритма, применяемого к матрице A.*

Доказательство. Достаточно заметить, что $A_k = L_k L_k^* = Q_k R_k$,

$$A_k^2 \;=\; A_k^* A_k \;=\; R_k^* R_k \;=\; L_k L_k^* L_k L_k^* \;=\; L_k L_{k+1} L_{k+1}^* L_k^*\,,$$

а значит, $R_k^* = L_k L_{k+1}$,

$$R_k Q_k \;=\; R_k (Q_k R_k) R_k^{-1} \;=\; R_k A_k R_k^{-1} \;=$$
$$(L_k L_{k+1})^* L_k L_k^* ((L_k L_{k+1})^*)^{-1} \;=\; L_{k+1}^* L_k^* L_k L_k^* (L_k^*)^{-1} (L_{k+1}^*)^{-1} \;=\; L_{k+1}^* L_{k+1} \;=\; A_{k+2}\,.$$

□

Как и для алгоритма Холесского, для симметрической версии QR–алгоритма также возможна неявная реализация с использованием тех же стратегий ординарного сдвига.

Степенные итерации и RQ–алгоритм

Скажем несколько слов о наиболее простых алгоритмах частичной проблемы собственных значений, позволяющих находить уникальные по модулю собственные значения матриц и приближённые собственные вектора по приближённым собственным значениям.

Замечание 0.96. *Пусть матрица $A \in M_n(\mathbb{C})$ имеет собственные значения $\lambda_1, \ldots, \lambda_n$ и отвечающие им собственные вектора $e_1, \ldots, e_n$, $Ae_i = \lambda_i e_i$, $\|e_i\|_2 = 1$, формируют базис $\mathbb{C}^n$, причём $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$. Выберем $x^{(0)} \in \mathbb{C}^n$, такой, что $x^{(0)} = \sum_i c_i e_i$, $c_1 \neq 0$, и положим $x^{(k+1)} = Ax^{(k)}$, $\lambda_1^{(k)} = \frac{(x^{(k+1)}, x^{(k)})}{(x^{(k)}, x^{(k)})}$, $e_1^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|_2}$ для $k \geq 0$. Тогда при $k \to +\infty$*

$$\lambda_1^{(k)} \;=\; \lambda_1 + O(|\lambda_2/\lambda_1|^k)\,, \quad e_1^{(k)} \;=\; \alpha_k e_1 + O(|\lambda_2/\lambda_1|^k)\,,$$

где $\alpha_k \in \mathbb{C}$, $|\alpha_k| = 1$.

Доказательство. Действительно,

$$x^{(k)} \;=\; A^k x^{(0)} \;=\; \sum_i c_i \lambda_i^k e_i = c_1 \lambda_1^k e_1 + O(|\lambda_2|^k)\,,$$

$$(x^{(k)}, x^{(k)}) \;=\; |c_1|^2 |\lambda_1|^{2k} + O(|\lambda_1 \lambda_2|^k)\,, \quad (x^{(k+1)}, x^{(k)}) \;=\; |c_1|^2 \lambda_1 |\lambda_1|^{2k} + O(|\lambda_1 \lambda_2|^k)$$

и потому

$$\lambda_1^{(k)} \;=\; \lambda_1 (1 + O(|\lambda_2/\lambda_1|^k)\,, \quad e_1^{(k)} \;=\; (c_1/|c_1|)(\lambda_1/|\lambda_1|)^k e_1 + O(|\lambda_2/\lambda_1|^k)\,.$$

□

Процесс описанного здесь вида называется степенной итерацией (или степенным методом). При его практической реализации условие $c_1 \neq 0$ обеспечивается за несколько шагов процесса накоплением погрешности. Вместе с тем при $|\lambda_1| > 1$ или $|\lambda_1| < 1$ норма вектора $x^{(k)}$ может значительно увеличиваться или уменьшаться, что делает целесообразным её нормирование через фиксированное число итераций. Имеются и такие модификации степенного метода, как нахождение наименьшего по модулю собственного значения или собственного значения ближайшего к данному λ (в случае уникальности подобных условий).

Вариантом степенного метода является и его обращённая версия, позволяющая находить приближённый собственный вектор по приближённому собственному значению.

Замечание 0.97. *Пусть матрица A диагонализируема, $A = S\Lambda S^{-1}$ для некоторых $S = (s_1, \ldots, s_n) \in GL_n(\mathbb{C})$ и $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, λ_k — однократное собственное значение A и σ — такое приближение к нему, что*

$$q \;=\; \max_{i \neq k} \frac{|\lambda_k - \sigma|}{|\lambda_i - \sigma|} < 1\,.$$

Положим $x^{(0)} = Sz$, где $z = (z_1, \ldots, z_n)^t$ и $z_k \neq 0$, и далее $y^{(m)} = (A - \sigma E)^{-1} x^{(m)}$, $x^{(m+1)} = \frac{y^{(m)}}{\|y^{(m)}\|_2}$, $m \geq 0$. Тогда $x^{(m)} = \beta_m \frac{s_k}{\|s_k\|_2} + O(q^m)$, $\beta_m \in \mathbb{C}$, $|\beta_m| = 1$.

Доказательство. В данном случае $(\lambda_k - \sigma)^{-1}$ является наибольшим по модулю собственным значением матрицы $(A - \sigma E)^{-1}$ и при том её единственным собственным значением с таким свойством. Кроме того,

$$\begin{aligned}(A - \sigma E)^{-m} x^{(0)} \;&=\; S(\Lambda - \sigma E)^{-m} z \;= \\ &S((\lambda_1 - \sigma)^{-m} z_1, \ldots, (\lambda_n - \sigma)^{-m} z_n)^t \;=\; (\lambda_k - \sigma)^{-m}(z_k s_k + O(q^m))\,,\end{aligned}$$

и

$$\begin{aligned}\|(A - \sigma E)^{-m} x^{(0)}\|_2 \;&=\; |\lambda_k - \sigma|^{-m}(|z_k| \|s_k\|_2 + O(q^m))\,, \\ x^{(m)} \;=\; \frac{(A - \sigma E)^{-m} x^{(0)}}{\|(A - \sigma E)^{-m} x^{(0)}\|_2} \;&=\; (z_k/|z_k|)(|\lambda_k - \sigma|/(\lambda_k - \sigma))^m \frac{s_k}{\|s_k\|_2} + O(q^m)\,,\end{aligned}$$

□

К алгоритмам этой серии относится и RQ–итерация, применямая к самосопряжённой матрице $A = A^*$. В данном случае выбирается начальный вектор $x^{(0)}$, $\|x^{(0)}\|_2 = 1$, и запускается процесс: $\rho_k = \frac{(Ax^{(k)}, x^{(k)})}{(x^{(k)}, x^{(k)})}$, $y^{(k)} = (A - \rho_k E)^{-1} x^{(k)}$, $x^{(k+1)} = \frac{y^{(k)}}{\|y^{(k)}\|_2}$, $k \geq 0$, который роднит со степенным методом использование в качестве приближений к собственным значениям чисел ρ_k (отношений Релея матрицы A на векторе $x^{(k)}$), а с обратным степенным методом — использование в качестве приближённых собственных векторов $x^{(k)}$. Условием выхода из процесса является выполнение для заданного $\varepsilon > 0$ неравенства $\|Ax^{(i)} - \rho_i x^{(i)}\|_2 < \varepsilon$, после чего ρ_i и $x^{(i)}$ подаются на выход в качестве искомых приближений.

Отношение Релея фигурирует также в целом ряде минимизационных алгоритмов поиска наибольшего и наименьшего по модулю собственного значения (см., к примеру, теорему о минимаксе).

Алгоритм "разделяй и властвуй"

Данный рекурсивный алгоритм применяется к трёхдиагональной неприводимой вещественной симметрической матрице $A = (a_{ij})$, $a_{ii} = a_i$, $a_{ii+1} = a_{i-1i} = b_i$, $a_{ij} = 0$, $|i-j| > 1$, где $b_i \neq 0$ при всех i. Приведение к матрице такого вида осуществляется любым известным способом (при помощи ортогонального (унитарного) подобия и алгоритма трёхдиагонализации Ланцоша). Выберем m, $1 \leq m < n$, и запишем матрицу A в виде

$$A \;=\; \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} + b_m v v^t\,,$$

где $T_1 = (a'_{ij})_{i,j=1}^{m}$, $T_2 = (a'_{ij})_{i,j=m+1}^{n}$, $a'_{ij} = a_{ij}$, $(i,j) \neq (m,m), (m+1,m+1)$, $a'_{mm} = a_m - b_m$, $a'_{m+1m+1} = a_{m+1} - b_m$, $v = (\underbrace{0,\ldots,0}_{m-1}, 1, 1, \underbrace{0,\ldots,0}_{n-m-1})^t$. Предположим, что мы располагаем спектральным разложением трёхдиагональных матриц T_1 и T_2, $T_i = Q_i \Lambda_i Q_i^t$, $i = 1,2$, для соответствующих ортогональных матриц собственных векторов Q_i и диагональных матриц собственных значений Λ_i, $i = 1,2$. Тогда $A = Q(D + b_m u u^t)Q^t$, где

$$D \;=\; \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \;=\; \mathrm{diag}(d_1,\ldots,d_n)\,, \quad Q \;=\; \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}\,, \quad u \;=\; Q^t v \;=\; \begin{pmatrix} q \\ q' \end{pmatrix}\,,$$

q и q' — последний и первый столбцы матриц Q_1^t и Q_2^t соответственно, и, как следствие, матрицы A и $D + b_m u u^t$ имеют одинаковый спектр. Без ограничения общности мы можем считать, что $d_1 \geq d_2 \geq \ldots \geq d_n$, поскольку необходимую упорядоченность $\{d_i\}$ можно обеспечить дополнительным сопряжением с подходящей матрицей-перестановкой P:

$$A \;=\; (QP^t)(PDP^t + b_m(Pu)(Pu)^t)(QP^t)^t\,.$$

Замечание 0.98. *Для любых столбцов $x, y \in K^n$ (K — любое поле или коммутативное кольцо с 1) справедливо равенство:* $\det(E + xy^t) = 1 + y^t x$.

Доказательство. Обозначив через e_i i-ый столбец единичной матрицы E, $i = 1,\ldots,n$, мы можем записать

$$E + xy^t \;=\; \begin{pmatrix} e_1^t + x_1 y^t \\ e_2^t + x_2 y^t \\ \ldots \\ e_n^t + x_n y^t \end{pmatrix}\,.$$

Тогда, используя n–линейность и кососимметричность определителя, мы получаем

$$\det(E + xy^t) \;=\; |E| + \sum_i \begin{vmatrix} e_1^t \\ \ldots \\ e_{i-1}^t \\ x_i y^t \\ e_{i+1}^t \\ \ldots \\ e_n^t \end{vmatrix} \;=\; 1 + \sum_i x_i y_i\,,$$

поскольку остальные слагаемые — нулевые определители матриц с не менее чем двумя строками вида $x_i y^t$. □

Вернёмся к матрице $D+\rho uu^t$, $\rho=b_m$, $D=\operatorname{diag}(d_1,\dots,d_n)$, $d_i\geq d_{i+1}$. Для любого $\lambda\neq d_i$, $i=1,\dots,n$, равенство

$$\det(D+\rho uu^t-\lambda E) \;=\; \det((D-\lambda E)(E+\rho(D-\lambda E)^{-1}uu^t) \;=$$
$$\prod_{i=1}^n(d_i-\lambda)\left(1+\rho\sum_{i=1}^n\frac{u_i^2}{d_i-\lambda}\right) \;=\; \prod_{i=1}^n(d_i-\lambda)+\rho\sum_{i=1}^n u_i^2\prod_{j\neq i}(d_j-\lambda) \;=\; 0$$

(точная проверка совпадения $\det(D+\rho uu^t-\lambda E)$ осуществляется аналогично предыдущему замечанию) выполняется тогда и только тогда, когда

$$f(\lambda) \;=\; \det(E+\rho(D-\lambda E)^{-1}uu^t) \;=\; 1+\rho u^t(D-\lambda E)^{-1}u \;=\; 1+\rho\sum_{i=1}^n\frac{u_i^2}{d_i-\lambda} \;=\; 0\,,$$

где $u=(u_1,\dots,u_n)^t$. Уравнение $f(\lambda)=0$ традиционно называется *вековым уравнением*.

Рассмотрим для начала случай "общего положения": $u_i\neq 0$, $d_p\neq d_q$ при всех i, $p\neq q$ ($\rho=b_m\neq 0$ в силу неприводимости матрицы A). В этой ситуации функция f имеет горизонтальную асимптоту $y=1$ и n вертикальных асимптот $x=d_i$, $i=1,\dots,n$. Так как её производная

$$f'(\lambda) \;=\; \rho\sum_{i=1}^n\frac{u_i^2}{(d_i-\lambda)^2}$$

знакопостояна на всей области определения функции f ($\mathbb{R}\setminus\{d_1,\dots,d_n\}$), а точнее имеет одинаковый знак с ρ, при $\rho>0$ она всюду возрастает (от 1 к $+\infty$ на $(-\infty,d_n)$, от $-\infty$ до $+\infty$ на (d_i,d_{i-1}), $i=2,\dots,n$, от $-\infty$ до 1 на $(d_1,+\infty)$), а при $\rho<0$ всюду убывает (от 1 к $-\infty$ на $(-\infty,d_n)$, от $+\infty$ до $-\infty$ на (d_i,d_{i-1}), $i=2,\dots,n$, от $+\infty$ до 1 на $(d_1,+\infty)$ (график f при $\rho<0$ — отражение её графика при $\rho>0$). Поэтому функция f имеет $n-1$ корень в интервалах (d_{i+1},d_i), $i=1,\dots,n-1$, к которым добавляется один корень в интервале $(d_1,+\infty)$ при $\rho>0$ и один корень в интервале $(-\infty,d_n)$ при $\rho<0$. Отсюда следует также, что в рассматриваемой ситуации нули f составляют полный набор собственнных значений матриц A и $D+\rho uu^t$.

Замечание 0.99. *Каждому $\lambda\in\operatorname{Spec}(D+\rho uu^t)\setminus\{d_1,\dots,d_n\}$ отвечает собственный вектор $(D-\lambda E)^{-1}u$.*

Доказательство. Действительно, в данном случае

$$0 \;=\; f(\lambda) \;=\; 1+\rho\sum_i\frac{u_i^2}{d_i-\lambda} \;=\; 1+\rho u^t(D-\lambda E)^{-1}u\,,$$

а потому

$$(D+\rho uu^t)(D-\lambda E)^{-1}u \;=\; (D-\lambda E+\lambda E+\rho uu^t)(D-\lambda E)^{-1}u \;=$$
$$u+\lambda(D-\lambda E)^{-1}u+\rho u(u^t(D-\lambda E)^{-1}u) \;=\; \lambda(D-\lambda E)^{-1}u\,.$$

При этом мы нигде не использовали требуемые в случае "общего положения" условия $d_i\neq d_j$ и $u_i\neq 0$. □

Обсудим теперь "дефляционные" случаи совпадения d_i и d_j, $i \neq j$, и равенства нулю компонент вектора u.

1. Допустим, что $u_i = 0$ и, как следствие, матрица uu^t имеет нулевые i-ую строку и i-ый столбец. В такой ситуации i-ый столбец e_i единичной матрицы E является собственным вектором матрицы $D + \rho uu^t$, отвечающим собственному значению d_i,

$$(D + \rho uu^t)e_i \;=\; De_i + \rho(u^t e_i)u \;=\; De_i \;=\; d_i e_i\,.$$

2. Если $d_i = d_j$ для некоторых $1 \leq i \neq j \leq n$, тогда вектор $u_j e_i - u_i e_j$ является собственным вектором матрицы $D + \rho uu^t$, соответствующим собственному значению d_i, так как

$$(D + \rho uu^t)(u_j e_i - u_i e_j) \;=\; d_i(u_j e_i - u_i e_j) + \rho(u^t(u_j e_i - u_i e_j))u \;=\; d_i(u_j e_i - u_i e_j)\,.$$

Описанные ситуации естественным образом объединяет

Замечание 0.100. *Пусть $d_i = \ldots = d_{i+k} = d$, $d_{i-1}, d_{i+k+1} \neq d$ для некоторого $k = k_d \geq 0$ и n_d — число нулевых элементов среди $\{u_i, \ldots, u_{i+k}\}$. Тогда при $n_d + k > 0$ собственное подпространство V_d матрицы $D + \rho uu^t$, отвечающее собственному значению d, содержит подпространство размерности $m_d = n_d + \max\{k - n_d - 1, 0\}$, базис которого составляют вектора e_j, $i \leq j \leq i + k$, $u_j = 0$, и вектора $u_s e_t - u_t e_s$, $u_s, u_t \neq 0$, $i \leq s < t \leq i + k$, где s выбран из условия $u_j = 0$, $i \leq j < s$.*

Доказательство. По построению выбранные нами m_d векторов являются линейно независимыми (тот факт, что их ровно m_d проверяется непосредственно) и входят в рассматриваемое собственное подпространство, отвечающее d (см. сделанные ранее наблюдения). □

Для реализации нашего алгоритма нам необходимо решить вековое уравнение $f = 0$, а точнее представить приемлимый в плане точности и скорости способ нахождения приближений к его решениям. Для простоты рассмотрим случай "общего положения"

$$f(\lambda) \;=\; 1 + \rho \sum_i \frac{u_i^2}{d_i - \lambda}$$

с $u_i \neq 0$ и $d_p \neq d_q$, $p \neq q$.

Реализуем следующий итерационный процесс. Выбирем начальное приближение $\lambda^{(0)} \in (d_{i+1}, d_i)$ и построим приближения $\{\lambda^{(j)}\}$ при помощи следующего индуктивного правила: если для некоторого $j \geq 0$ приближение $\lambda^{(j)}$ уже найдено, тогда, полагая $f = 1 + \psi_1 + \psi_2$,

$$\psi_1(\lambda) \;=\; \rho \sum_{k=1}^{i} \frac{u_k^2}{d_k - \lambda}\,, \quad \psi_2(\lambda) \;=\; \rho \sum_{k=i+1}^{n} \frac{u_k^2}{d_k - \lambda}$$

(функции ψ_1 и ψ_2 знакопостоянны и противоположны по знаку в (d_{i+1}, d_i)), построим функцию

$$h_1(\lambda) \;=\; \hat{c}_1 + \frac{c_2}{d_i - \lambda}$$

из условий $h_1(\lambda^{(j)}) = \psi_1(\lambda^{(j)})$ и $h_1'(\lambda^{(j)}) = \psi_1'(\lambda^{(j)})$,

$$c_1 \;=\; \psi_1'(\lambda^{(j)})(d_i-\lambda^{(j)})^2\,,\quad \hat c_1 \;=\; \psi_1(\lambda^{(j)})-\psi_1'(\lambda^{(j)})(d_i-\lambda^{(j)})\,,$$

функцию

$$h_2(\lambda) \;=\; \hat c_2+\frac{c_2}{d_{i+1}-\lambda}$$

из условий $h_2(\lambda^{(j)})=\psi_2(\lambda^{(j)})$, $h_2'(\lambda^{(j)})=\psi_2'(\lambda^{(j)})$,

$$c_2 \;=\; \psi_2'(\lambda^{(j)})(d_{i+1}-\lambda^{(j)})^2\,,\quad \hat c_2 \;=\; \psi_2(\lambda^{(j)})-\psi_2'(\lambda^{(j)})(d_{i+1}-\lambda^{(j)})\,,$$

найдём нуль функции

$$h(\lambda) \;=\; 1+h_1(\lambda)+h_2(\lambda) \;=\; (1+\hat c_1+\hat c_2)+\frac{c_1}{d_i-\lambda}+\frac{c_2}{d_{i+1}-\lambda} \;=\; c_3+\frac{c_1}{d_i-\lambda}+\frac{c_2}{d_{i+1}-\lambda}$$

в интервале (d_{i+1},d_i), решив квадратное уравнение

$$c_3(d_i-\lambda)(d_{i+1}-\lambda)+c_1(d_{i+1}-\lambda)+c_2(d_i-\lambda) \;=\; 0\,,$$

и положим $\lambda^{(j+1)}$ равным найденному решению.

В результате выполнения некоторого фиксированного изначально числа итерации будет найдено приближение $\lambda^{(j)}$, которое будет использоваться в дальнейшем в качестве искомого приближения к решению $f=0$ в указанном интервале.

Аналогичный процесс поиска приближённого решения уравнения $f=0$ в интервале $(d_1,+\infty)$ при $\rho>0$ или в интервале $(-\infty,d_n)$ при $\rho<0$ реализуется с использованием функции

$$h(\lambda) \;=\; c_1+\frac{c_2}{d_1-\lambda}$$

или функции

$$h(\lambda) \;=\; c_1+\frac{c_2}{d_n-\lambda}\,,$$

которые на каждом шаге процесса строятся по найденному ранее приближению $\lambda^{(j)}$ из условий $h(\lambda^{(j)})=f(\lambda^{(j)})$, $h'(\lambda^{(j)})=f'(\lambda^{(j)})$.

Имеется довольно простой способ построения по заданному набору попарно различных чисел $\{\alpha_i\}$, разделяемых числами $\{d_i\}$, вектора u, такого, что указанный набор составляет полную систему собственных значений матрицы $D+uu^t$. Этот способ реализуется в рамках следующей теоремы Лёвнера.

Теорема 0.101. *Пусть* $D=\operatorname{diag}(d_1,\ldots,d_n)$, $d_1>\ldots>d_n$, $\{\alpha_i\}$, $\{\alpha_i'\}$, *причём*

$$d_n<\alpha_n<d_{n-1}<\alpha_{n-1}<\ldots<d_1<\alpha_1\,,\quad \alpha_n'<d_n<\alpha_{n-1}'<d_{n-1}<\ldots<\alpha_1'<d_1\,.$$

Тогда найдутся вектора $w=(w_1,\ldots,w_n)^t$ *и* $w'=(w_1',\ldots,w_n')^t$, $w_i,w_i'\ge 0$, *такие, что наборы* $\{\alpha_i\}$ *и* $\{\alpha_i'\}$ *составляют полные системы собственных значений матриц* $D+ww^t$ *и* $D-w'w'^t$ *соответственно.*

Доказательство. Достаточно заметить, что для такого вектора v

$$\det(D+ww^t-\lambda E) \;=\; \prod_{j=1}^n(\alpha_j-\lambda) \;=\; \prod_{j=1}^n(d_j-\lambda)+\sum_{i=1}^n w_i^2\prod_{j\ne i}(d_i-\lambda)\quad (i=1,\ldots,n)\,,$$

а потому при $\lambda = d_i$, $i = 1, \ldots, n$, должно выполняться равенство

$$\prod_{j=1}^{n}(\alpha_j - d_i) \;=\; w_i^2 \prod_{j\neq i}(d_j - d_i)\,, \quad w_i \;=\; \sqrt{\frac{\prod\limits_{j=1}^{n}(\alpha_j - d_i)}{\prod\limits_{j\neq i}(d_j - d_i)}}\,.$$

Заметим, что выбор $\{w_i\}$ определён однозначно, поскольку многочлен степени n с фиксированным старшим коэффициентом однозначно определяется своими значениями в n различных точках (в нашем случае речь идёт о значениях в точках $\{d_i\}$). Впрочем, к тому же выводу можно прийти и непосредственно.

$$\det(D + ww^t - \lambda E) \;=\; \prod_{j=1}^{n}(d_j - \lambda) + \sum_{j=1}^{n} w_j^2 \prod_{k\neq j}(d_k - \lambda) \;=$$

$$\prod_{j=1}^{n}(d_j - \lambda) + \sum_{j=1}^{n}\prod_{k=1}^{n}(\alpha_k - d_j)\prod_{k\neq j}\frac{d_k - \lambda}{d_k - d_j} \;=\; \prod_{j=1}^{n}(d_j - \lambda)\left(1 + \sum_{j=1}^{n}\frac{\alpha_j - d_j}{d_j - \lambda}\prod_{k\neq j}\frac{\alpha_k - d_j}{d_k - d_j}\right),$$

где

$$1 + \sum_{j=1}^{n}\frac{\alpha_j - d_j}{d_j - \alpha_s}\prod_{k\neq j}\frac{\alpha_k - d_j}{d_k - d_j} \;=\; 1 - \prod_{k\neq s}\frac{\alpha_k - d_s}{d_k - d_s} + \sum_{j\neq s}\frac{\alpha_j - d_j}{d_j - \alpha_s}\prod_{k\neq j}\frac{\alpha_k - d_j}{d_k - d_j} \;=$$

$$1 - \prod_{k\neq s}\frac{\alpha_k - d_s}{d_k - d_s} - \sum_{j\neq s}\frac{\prod\limits_{k\neq s}(\alpha_k - d_j)}{\prod\limits_{l\neq j}(d_l - d_j)} \;=\; 1 - \sum_{j=1}^{n}\frac{\prod\limits_{k\neq s}(\alpha_k - d_j)}{\prod\limits_{l\neq j}(d_l - d_j)} \;=\; 1 - \sum_{j=1}^{n}\frac{1}{\alpha_s - d_j}\prod_{k\neq j}\frac{\alpha_k - d_j}{d_k - d_j}\,.$$

Выразив многочлен $(x - d_1)\cdots(x - d_n)$ через интерполяционный многочлен по точкам $\{\alpha_i\}$, мы получим

$$\prod_{i=1}^{n}(x - d_i) \;=\; \sum_{i=1}^{n}\prod_{j=1}^{n}(\alpha_i - d_j)\prod_{j\neq i}\frac{x - d_j}{d_i - d_j}\,, \quad 1 \;=\; \sum_{i=1}^{n}\frac{\alpha_i - d_i}{x - d_i}\prod_{j\neq i}\frac{\alpha_i - d_j}{d_i - d_j}\,,$$

$$1 + \sum_{i=1}^{n}\frac{\alpha_i - d_i}{d_i - x}\prod_{j\neq i}\frac{\alpha_i - d_j}{d_i - d_j} \;=\; 0\,.$$

Остаётся подставив в последнее равенство $x = \alpha_s$ и прийти к соотношениям

$$\det(D + ww^t - \alpha_s E) \;=\; 0 \quad (s = 1, \ldots, n)\,, \quad \operatorname{Spec}(D + ww^t) \;=\; \{\alpha_1, \ldots, \alpha_n\}\,.$$

В свою очередь для набора $\{\alpha'_i\}$ и вычисленного по сходным формулам вектора $w' = (w'_1, \ldots, w'_n)^t$ будет выполняться равенство $\operatorname{Spec}(D - w'w'^t) = \{\alpha'_1, \ldots, \alpha'_n\}$.

В силу сказанного ранее собственные вектора матриц $D + ww^t$ и $D - w'w'^t$, отвечающие собственным значениям α_i и α'_i, можно определить при помощи равенств $(D - \alpha_i E)^{-1}w$ и $(D - \alpha'_i E)^{-1}w'$, $i = 1, \ldots, n$. □

Нам также понадобится следующее уточнение теоремы Лёвнера.

Теорема 0.102. *Пусть в наборе* $\{d_i\}$, $d_1 \geq d_2 \geq \ldots \geq d_n$, *элементы* $\{d_{i_j}\}$ *составляют полную систему попарно различных элементов, причём каждый* d_{i_j} *встречается* $k_{d_{i_j}}$ *раз,* $k_{d_{i_j}} \geq 1$, *а в наборе* $\{\alpha_i\}$, $d_n \leq \alpha_n \leq d_{n-1} \leq \ldots \leq \alpha_2 \leq d_1 \leq \alpha_1$, *каждому* j *отвечает* l_j, $k_{d_{i_j}} + 1 \geq l_j \geq k_{d_{i_j}} - 1 \geq 0$ *элементов равных* d_{i_j}. *Тогда вектор* $w = (w_1, \ldots, w_n)^t$, *для которого* $\mathrm{Spec}(D + ww^t) = \{\alpha_i\}$, *может быть найден из соотношений:*

$$w_i \;=\; \begin{cases} 0 & \text{при } i = i_j,\ l_j \geq k_{d_{i_j}} = 1, \text{ и} \\ & \text{при } d_i = d_{i_j},\ l_j > k_{d_{i_j}} - 1 > 0; \\ \sqrt{\dfrac{\prod\limits_{j=1}^{n}(\alpha_j - d_i)}{\prod\limits_{j\neq i}(d_j - d_i)}} & \text{при } i = i_j,\ l_j = 0,\ k_{d_{i_j}} = 1; \\ \sqrt{\dfrac{\prod\limits_{i,\ \alpha_i \neq d_{i_j}}(\alpha_i - d_{i_j})}{k_{d_{i_j}} \prod\limits_{s,\ d_s \neq d_{i_j}}(d_s - d_{i_j})}} & \text{при } d_i = d_{i_j},\ l_j = k_{d_{i_j}} - 1 > 0. \end{cases}$$

При этом каждому $\alpha_i \notin \{d_i\}$ *отвечает одномерное собственное подпространство, порождённое вектором* $(D - \alpha_i E)^{-1} w$. *Если* $\alpha_i = d_{i_j}$, $l_j = k_{d_{i_j}} - 1$, *то* $w_i \neq 0$ *для всех* i, $d_i = d_{i_j}$, *соответствующее собственное подпространство* l_j*–мерно и в качестве его базиса можно взять вектора* $w_s e_t - w_t e_s$, $s \neq t$, $d_s = d_t = d_{i_j}$, *где* s *— некоторый фиксированный индекс с таким свойством. Если* $\alpha_i = d_{i_j}$, $l_j = k_{d_{i_j}}$, *то* $w_i = 0$ *для всех* i, $d_i = d_{i_j}$, *и отвечающее* α_i *собственное подпространство порождается* l_j *линейно независимыми векторами* e_i, $d_i = d_{i_j}$. *Наконец, если* $\alpha_i = d_{i_j}$, $l_j = k_{d_{i_j}} + 1$, *то* $w_i = 0$ *для всех* i, $d_i = d_{i_j}$, *и в качестве базиса отвечающего* α_i *собственного подпространства можно использовать вектора* e_i, $d_i = d_{i_j}$, *и вектор* $z = (z_1, \ldots, z_n)^t$, $z_i = 0$, $d_i = d_{i_j}$, *и* $z_i = (d_i - d_{i_j})^{-1} w_i$ *при* i, $d_i \neq d_{i_j}$.

Доказательство. Начнём с соотношения, определяющего выбор вектора w,

$$\det(D + ww^t - \lambda E) \;=\; \prod_{i=1}^{n}(\alpha_i - \lambda) \;=\; \prod_{i=1}^{n}(d_i - \lambda) + \sum_{i=1}^{n} w_i^2 \prod_{j\neq i}(d_j - \lambda)\,.$$

Тогда для $i = i_j$ с $k_{d_{i_j}} = 1$

$$w_i \;=\; \sqrt{\frac{\prod\limits_{j=1}^{n}(\alpha_j - d_i)}{\prod\limits_{j\neq i}(d_j - d_i)}}\,,$$

при $l_j \geq 1$ в последнем равенстве $w_i = 0$, для $i = i_j$ с $k_{d_{i_j}} > 1$ мы можем разделить первое равенство на $(d_{i_j} - \lambda)^{k_{d_{i_j}} - 1}$ и прийти к соотношению

$$(d_{i_j} - \lambda)^{l_j - k_{d_{i_j}} + 1} \prod_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - \lambda) \;=$$

$$(d_{i_j} - \lambda)\left(\prod_{i,\ d_i \neq d_{i_j}} (d_i - \lambda) + \sum_{i,\ d_i \neq d_{i_j}} w_i^2 \prod_{j\neq i,\ d_j \neq d_{i_j}} (d_j - \lambda)\right) + \left(\sum_{i,\ d_i = d_{i_j}} w_i^2\right) \prod_{j,\ d_j \neq d_{i_j}} (d_j - \lambda)\,,$$

в соответствии с которым $w_i = 0$ для всех i, $d_i = d_{i_j}$, в случае $l_j > k_{d_{i_j}} - 1$ и

$$\prod_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - d_{i_j}) \ = \ \left(\sum_{i,\ d_i = d_{i_j}} w_i^2 \right) \prod_{s,\ d_s \neq d_{i_j}} (d_s - d_{i_j})$$

в случае $l_j = k_{d_{i_j}} - 1$. В последней ситуации w_i можно положить равным

$$w_i \ = \ \sqrt{\frac{\prod\limits_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - d_{i_j})}{k_{d_{i_j}} \prod\limits_{s,\ d_s \neq d_{i_j}} (d_s - d_{i_j})}} \, .$$

Отметим, что в данном случае однозначным является выбор не самих $\{w_i\}$, а коэффициентов

$$\sum_{i,\ d_i = d_{i_j}} w_i^2 \, .$$

Для этого достаточно заметить, что исходное равенство может быть равносильным образом переписано в виде

$$\left(\prod_{i,\ \alpha_i \notin \{d_i\}} (\alpha_i - \lambda) \right) \left(\prod_{j,\ l_j > 0} (d_{i_j} - \lambda)^{l_j - k_{d_{i_j}} + 1} \right) \ =$$

$$\prod_j (d_{i_j} - \lambda) + \sum_j \left(\sum_{i,\ d_i = d_{i_j}} w_i^2 \right) \prod_{l,\ l \neq j} (d_{i_l} - \lambda)$$

после сокращения на общий множитель

$$\prod_j (d_{i_j} - \lambda)^{k_{d_{i_j}} - 1} \ = \ \prod_{j,\ k_{d_{i_j}} > 1} (d_{i_j} - \lambda)^{k_{d_{i_j}} - 1} .$$

Остаётся заметить, что стоящие в левой и правой частях предпоследнего соотношения многочлены имеют равные старшие коэффициенты и одинаковую степень $|\{d_{i_j}\}|$, а потому однозначно определяются своими значениями в точках $\{d_{i_j}\}$.

Обозначим через V_{α_i} собственное подпространство матрицы $D + ww^t$, отвечающее её собственному значению α_i. В случае $\alpha_i \notin \{d_i\}$ значение α_i имеет единичную кратность и согласно одному из предыдущих замечаний подпространство V_{α_i} порождается вектором $(D - \alpha_i E)^{-1} w$.

Если $\alpha_i = d_{i_j}$ для некоторого j, тогда возможна одна из следующих ситуаций $\dim V_{\alpha_i} = l_j = k_{d_{i_j}} - 1, k_{d_{i_j}}, k_{d_{i_j}} + 1$. При этом в первом случае $w_i \neq 0$ для каждого i, $d_i = d_{i_j}$, в двух остальных $w_i = 0$ при всех таких i (см. выше). Выбор базиса V_{α_i} в первых двух случаях осуществляется согласно проделанному ранее рассмотрению дефляционных ситуаций. Остаётся исследовать случай $l_j = k_{d_{i_j}} + 1$. Здесь мы заведомо имеем $k_{d_{i_j}}$ линейно независимых элементов e_i, $d_i = d_{i_j}$, подпространства V_{α_i} и нам остаётся их дополнить одним единственным линейно независимым с ними вектором до базиса пространства. В качестве такого вектора можно взять вектор $z = (z_1, \ldots, z_n)^t$,

$$z_i \;=\; \begin{cases} 0 & \text{при } i,\ d_i = d_{i_j}; \\ \frac{w_i}{d_i - d_{i_j}} & \text{при } i,\ d_i \neq d_{i_j}. \end{cases}$$

Понятно, что z отличен от нуля и линейно независим с векторами $\{e_i \mid d_i = d_{i_j}\}$. Кроме того, $z \in V_{\alpha_i}$, поскольку

$$(D + ww^t)z \;=\; (D - d_{i_j}E)z + d_{i_j}z + (w^t z)w \;=$$

$$w + d_{i_j}z + (w^t z)w \;=\; d_{i_j}z + \left(1 + \sum_{i,\ d_i \neq d_{i_j}} \frac{w_i^2}{d_i - d_{i_j}}\right)w \;=\; d_{i_j}z\,.$$

В данном случае мы используем тот факт, что в рассматриваемой ситуации

$$(d_{i_j} - \lambda)^{l_j - k_{d_{i_j}} + 1} \prod_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - \lambda) \;=\; (d_{i_j} - \lambda)^2 \prod_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - \lambda) \;=$$

$$(d_{i_j} - \lambda)\left(\prod_{i,\ d_i \neq d_{i_j}} (d_i - \lambda) + \sum_{i,\ d_i \neq d_{i_j}} w_i^2 \prod_{j \neq i,\ d_j \neq d_{i_j}} (d_j - \lambda)\right) + \left(\sum_{i,\ d_i = d_{i_j}} w_i^2\right) \prod_{j,\ d_j \neq d_{i_j}} (d_j - \lambda) \;=$$

$$(d_{i_j} - \lambda)\left(\prod_{i,\ d_i \neq d_{i_j}} (d_i - \lambda) + \sum_{i,\ d_i \neq d_{i_j}} w_i^2 \prod_{j \neq i,\ d_j \neq d_{i_j}} (d_j - \lambda)\right) \;=$$

$$(d_{i_j} - \lambda) \prod_{i,\ d_i \neq d_{i_j}} (d_i - \lambda)\left(1 + \sum_{i,\ d_i \neq d_{i_j}} \frac{w_i^2}{d_i - \lambda}\right)$$

и, как следствие,

$$1 + \sum_{i,\ d_i \neq d_{i_j}} \frac{w_i^2}{d_i - d_{i_j}} \;=\; 0\,.$$

Таким образом, вектора z и $\{e_i \mid d_i = d_{i_j}\}$ составляют базис пространства V_{α_i}.

Аналогичные выводы могут быть получены и для набора $\{\alpha'_i\}$ и соответствующего вектора w', $\mathrm{Spec}(D - w'w'^t) = \mathrm{Spec}(\{\alpha'_i\})$. □

Замечание 0.103. *Представленный в теореме способ нахождения компонент вектора w и базиса подпространства V_{α_i} в случае $\alpha_i = d_{i_j}$, $l_j = k_{d_{i_j}} - 1$, можно упростить следующим образом: положим*

$$w_{i_j} \;=\; \sqrt{\frac{\prod\limits_{i,\ \alpha_i \neq d_{i_j}} (\alpha_i - d_{i_j})}{\prod\limits_{s,\ d_s \neq d_{i_j}} (d_s - d_{i_j})}}\,.$$

и $w_i = 0$ для всех $i \neq i_j$, $d_i = d_{i_j}$, а в качестве базиса V_{α_i} возьмём вектора e_i, $i \neq i_j$, $d_i = d_{i_j}$.

Заметим, что такой выбор допустим (см. доказательство теоремы) и не потребует дополнительной переортогонализации векторов для нахождения ортонормированного базиса V_{α_i}.

Перейдём к описанию самого алгоритма. Выберем параметр $\varepsilon > 0$, позволяющий считать величины меньшие ε пренебрежительно малыми, и число итерации k для вычисления приближений к корням "векового уравнения".

Предположим, что мы располагаем представлением исходной матрицы в виде

$$A \;=\; \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} + b_m v v^t$$

и вычисленными по нашему алгоритму приближёнными спектральными разложениями $Q_i \Lambda_i Q_i^t \sim T_i$. Перейдём по описанной ранее схеме к приближённой к матрице A матрице $Q(D + \rho u u^t)Q^t$ и домножим матрицу Q при необходимости на подходящую матрицу-перестановку, обоспечив требуемую упорядоченность набора $\{d_i\}$, $D = \mathrm{diag}(d_1, \ldots, d_n)$. Объявим ε–близкие элементы набора $\{d_i\}$ равными, а ε–близкие к нулю элементы вектора u равными нулю. Найдём приближённые решения векового уравнения, которое имеет вид

$$f(\lambda) \;=\; 1 + \rho \sum_i \frac{u_i^2}{d_i - \lambda} \;=\; 1 + \rho \sum_j \frac{k_j}{d_{i_j} - \lambda} \;=\; 0\,,$$

где $\{d_{i_j}\}$ — полный набор попарно различных элементов в наборе $\{d_i\}$,

$$k_j \;=\; \sum_{i,\ d_i = d_{i_j}} u_i^2 \geq 0\,,$$

Общее число решений такого уравнения равно $|\{j \mid k_j \neq 0\}|$, из которых при необходимости следует исключить решения равные d_{i_j} для $k_j = 0$. Полученные приближения вне множества $\{d_i\}$ являются однократными собственными значениями приближённой матрицы, стоящейся на следующем этапе алгоритма. Остаётся разобраться с кратностями l_j тех d_{i_j}, для которых $k_j = 0$. Для удобства мы можем записать матрицу $D + \rho u u^t$ в виде $D \pm u' u'^t$, полагая $u' = \sqrt{|\rho|}u$. Воспользовавшись уточнением к теореме Лёвнера, применённым к точному набору собственных значений матрицы $D \pm u' u'^t$ в качестве $\{\alpha_i\}$ или $\{\alpha_i'\}$, мы получаем следующий набор возможностей для кратности собственного значения d_{i_j} (при реализации одной из дефляционных ситуаций):

1. $l_j = k_{d_{i_j}} - 1 > 0$, если и только если $k_j \neq 0$;
2. $l_j = k_{d_{i_j}} > 0$, если и только если $k_j = 0$, $f(d_{i_j}) \neq 0$;
3. $l_j = k_{d_{i_j}} + 1$, если и только если $k_j = 0$, $f(d_{i_j}) = 0$.

Следует отметить, что в случае $k_j = 0$ вековое уравнение $f = 0$ является вековым уравнением матрицы, полученной из матрицы $D + \rho u u^t$ вычёркиванием строк и столбцов с номерами i, $d_i = d_{i_j}$.

Сформируем теперь из полученных приближений к собственным значениям матрицы $D + \rho u u^t$ (с учётом их кратностей в дефляционных случаях) набор $\{\alpha_i\}$ или $\{\alpha_i'\}$ и построим матрицу $\hat{D} = D + ww^t$, $\mathrm{Spec}(D + ww^t) = \{\alpha_i\}$, или $\hat{D}' = D - w'w'^t$, $\mathrm{Spec}(D - w'w'^t) = \{\alpha_i'\}$, а вместе с ней и наборы её ортонормированных собственных векторов (ортонормированные базисы собственных подпространств). Используя матрицу $\hat{D}$ или $\hat{D}'$ в качестве приближения к матрице $D + \rho u u^t$, построим на основе её спектрального

разложения при помощи найденной ранее матрицы Q приближённое спектральное разложение матрицы A.

Отметим, что рассмотренный здесь переход от матрицы $D + \rho uu^t$ к матрице $\hat{D} = D + ww^t$ при $\rho > 0$ или $\hat{D'} = D - w'w'^t$ при $\rho < 0$ вполне оправдан с точки зрения качества получаемых приближений. Действительно, $D + \rho uu^t = D \pm u'u'^t$ для подходящего вектора $u' = (u'_1, \ldots, u'_n)^t$, $u'_i \geq 0$, (знак совпадает со знаком ρ), где в случае "общего положения" вектор u' однозначно определяется через собственные значения матрицы $D + \rho uu^t$ и $\{d_i\}$ по приведённым выше формулам, а в общей ситуации однозначно определёнными являются некоторые суммы квадратов координат этого вектора, на которые разбивается его скалярный квадрат. В любом случае порядок близости норм векторов u' и w или u' и w' (для "общего положения" самих этих векторов) определяется качеством приближений $\{\alpha_i\}$ к точным собственным значениям матрицы $D + \rho uu^t$, а значит, аналогичный вывод верен и для порядка близости матриц $D + \rho uu^t$ и $\hat{D}$ или $D + \rho uu^t$ и $\hat{D'}$ в любой матричной норме.

Необходимо подчеркнуть одну немаловажную особенность представленного нами алгоритма: компоненты вектора w вычисляются нами без учёта неоднозначности определения столбцов матрицы Q, которые вычисляются с точностью до умножения на ± 1, по одним и тем же формулам, а потому целесообразно их видоизменить, потребовав дополнительно совпадение знаков u_i и w_i для всех i.