Г.С.Сиговцев

Численные методы

Конспект лекций

Петрозаводск 2002

Содержание

Введение

Часто возникает необходимость, как в самой математике, так и ее приложениях в разнообразных областях получать решения математических задач в числовой форме. (Для представления решения в графическом виде также требуется предварительно вычислять его значения.) При этом для многих задач известно только о существовании решения, но не существует конечной формулы, представляющей ее решение. Даже при наличии такой формулы ее использование для получения отдельных значений решения может оказаться неэффективным. Наконец, всегда существует необходимость решать и такие математические задачи, для которых строгие доказательства существования решения на данный момент отсутствуют. Во всех этих случаях используются методы приближенного, в первую очередь численного решения.

Математика как наука возникла в связи с необходимостью решения практических задач: счета, измерений на местности, навигации и т.д. - ее целью являлось получение решения в виде числа. И на протяжении всей ее истории методы численного решения математических задач всегда составляли неотъемлемую часть математики. Многие математики сочетали в своих исследованиях составление математического описания явлений природы (построение математических моделей) и его исследование математическими средствами. При этом создавались методы приближенного решения различных математических задач. Многие выдающиеся математики занимались разработкой этих методов, о чем говорят их названия: методы Ньютона, Эйлера, Лобачевского, Гаусса, Чебышева и многие другие.

И если свою историю численные методы ведут от Архимеда. от древних индийских и арабских математиков, то в отдельную математическую дисциплину «вычислительная математика» методы вычислений выделились сравнительно поздно, на рубеже 19-го и 20-го веков. К этому времени в основном были разработаны разнообразные, достаточно эффективные и надежные алгоритмы приближенного решения широкого круга математических задач, включающего стандартный набор задач из алгебры, математического анализа и дифференциальных уравнений. Первый в мировой литературе отдельный курс методов вычислений «Лекции о приближенных вычислениях» был издан академиком А.Н.Крыловым в 1911 г.

Прогресс в развитии численных методов способствовал постоянному расширению сферы применения математики в других научных дисциплинах и прикладных разработках, откуда в свою очередь поступали запросы на решение новых проблем, стимулируя дальнейшее развитие вычислительной математики. Метод математического моделирования, основанный на построении и исследовании математических моделей различных объектов, процессов и явлений и получении информации о них из решения связанных с этими моделями математических задач, стал одним из основных способов исследования в так называемых точных науках.

Развитие численных методов шло параллельно с разработкой инструментальных средств вычислений. Абак столь же древен, как самые первые способы вычислений. В разных вариантах и с разными названиями он был у древних египтян, в Китае и Японии. Абак получил широкое распространение у греков и римлян, которые называли его

calculi (от этого слова происходит современное название калькулятор), а затем, постепенно совершенствуясь, использовался в Европе до 18 века. В России свой вариант абака (русские счеты) был изобретен в 16 веке. Простота и удобство счетов для повседневных вычислений обеспечили их распространение в других странах и использование вплоть до нашего времени.

17 век связан в истории науки и техники с множеством открытий и изобретений. В области средств вычислений это изобретение вычислительных машин, представлявших собой различные механические устройства для выполнения арифметических операций. Наиболее известные – это арифметическая машина Б.Паскаля, и арифмометр Г.Лейбница. Тогда же в 17 веке появились таблицы логарифмов и логарифмическая линейка, которая свыше 300 лет оставалась основным инструментом массовых инженерных расчетов. В 18 и 19 веках происходило усовершенствование конструкций ранее созданных устройств, появлялись новые изобретения. Наиболее значительными и яркими среди них были вычислительные машины Ч.Бэббиджа, работами по созданию которых, (сначала разностной, а затем аналитической машины) он занимался с 1820 по 1871 год.

Подавляющее большинство изобретенных в течение нескольких веков вычислительных устройств и приборов существовало в виде отдельных образцов и не получало заметного распространения. Причинами этого были как отсутствие в то время реальной потребности в массовых вычислительных инструментах, так и недостаточный уровень технической базы, не позволявшей создавать достаточно надежные и точные устройства. Первым промышленно выпускаемым вычислительным устройством стал арифмометр К.Томаса, который сумел в 1820 г. организовать производство своей машины (за первые 50 лет было продано 1500 арифмометров).

Необходимо отметить, что прогресс в области инструментальных средств не оказывал заметного влияния на ход развития методов вычислений. Принципиальным образом ситуация изменилась со середины нашего столетия, когда было осуществлено изобретение электронных вычислительных машин. В результате появления ЭВМ скорость выполнения вычислительных операций выросла в миллионы раз, что позволило решить широкий круг бывших до этого практически не решаемыми математических задач. Широкое внедрение ЭВМ в практику научных и технических расчетов потребовало интенсивного развития методов численного решения самых разных математических задач, причем методов, рассчитанных на реализацию их именно на ЭВМ. Это связано с тем, что часть из ранее использовавшихся алгоритмов численного решения неэффективна при реализации на ЭВМ, а некоторые просто непригодны для такого использования.

Современной формой метода математического моделирования, базирующейся на мощной вычислительной базе в виде ЭВМ и программного обеспечения, реализующего алгоритмы численного решения, является вычислительный эксперимент, рассматриваемый как новый теоретический метод исследования различных явлений и процессов. Этот теоретический метод включает существенные черты методологии экспериментального исследования, но эксперименты выполняются не над реальным объектом, а над его математической моделью, и экспериментальной установкой является ЭВМ.

Технологическая цепочка вычислительного эксперимента включает в себя следующие этапы:

- построение математической модели исследуемого объекта (сюда же относится и анализ модели, выяснение корректности поставленной математической задачи;
- построение вычислительного алгоритма - метода приближенного решения поставленной задачи и его обоснование;
- программирование алгоритма на ЭВМ и его тестирование;
- проведение серии расчетов с варьированием определяющих параметров исходной задачи и алгоритма;
- анализ полученных результатов;

Каждый из этих этапов допускает возврат к любому из предыдущих с целью его уточнения и корректировки.

В данном курсе рассматриваются вопросы, связанные со вторым этапом вычислительного эксперимента. Во многих случаях вычислительный алгоритм решения сложной задачи строится из набора базовых компонент, представляющих собой алгоритмы решения некоторых стандартных математических задач. Изучение численных методов решения этих задач - необходимый элемент овладения современной технологией математического моделирования.

При этом идея модели лежит в основе того, что можно назвать методом вычислительной математики. Как правило, алгоритмы приближенного решения базируются на том, что исходная математическая задача заменяется (аппроксимируется) некоторой более простой или чаще последовательностью более простых задач. Решение этих более простых задач трактуется как приближенное решение задачи исходной. Т.е. фактически используется некоторая модель исходной задачи.

1. Элементы теории погрешностей

1.1. Источники и классификация погрешностей

1.1.1. Погрешности данных, метода и вычислений

В вычислительной математике погрешность является обязательным атрибутом всех рассматриваемых методов и их важнейшей характеристикой в виде погрешности получаемых тем или иным методом приближенных решений соответствующих задач. Здесь задача – это нахождение решения X по исходным данным Y: $X = A(Y)$, где A – оператор задачи.

Можно выделить три компонента, в совокупности обуславливающие погрешность решения задачи

1. Применяемый метод приближенного решения как правило основан на замене исходного оператора задачи A на некоторый другой оператор (другую задачу или последовательность задач) B, решения которых не совпадают с решением исходной задачи и называются ее приближенными решениями: $\overline{X} = B(Y)$. Разность $X - \overline{X} = A(Y) - B(Y)$ называется погрешность метода.

2. При реализации алгоритма приближенного решения, определяемого задачей B в силу использования конечной разрядной сетки для представления чисел возникают погрешности округлений, выполняемые в алгоритме арифметические операции над приближенными величинами дают приближенные результаты. Это эквивалентно тому, что оператор задачи B при реализации метода заменяется некоторым другим оператором $\hat{B}$, который дает другое приближенное решение $\hat{X} = \hat{B}(Y)$. Разность $\overline{X} - \hat{X} = B(Y) - \hat{B}(Y)$ называется вычислительной погрешностью.

3. Математическое описание задачи может быть неточным. Содержащиеся в нем исходные данные, параметры, коэффициенты задачи, являющейся моделью некоторого объекта или процесса, как правило, известны приближенно, т.к. получены в результате экспериментов, наблюдений, измерений. Т.е. вместо исходных данных Y используются данные $\widetilde{Y}$, что также вызывает изменение приближенного решения: $\widetilde{X} = \hat{B}(\widetilde{Y})$. Разность $\hat{X} - \widetilde{X} = \hat{B}(Y) - \hat{B}(\widetilde{Y})$ называется погрешностью данных.

Таким образом, итоговая погрешность получаемого приближенного решения $X - \widetilde{X}$ складывается из трех величин: $X - \widetilde{X} = (X - \overline{X}) + (\overline{X} - \hat{X}) + (\hat{X} - \widetilde{X})$.

Погрешность исходных данных задачи (неустранимая погрешность) является некоторым ориентиром возможной точности решения задачи, так как нет смысла пытаться решить задачу существенно «точнее», чем это определяется погрешностью исходных данных. Кроме того, хотя при выводе оценок погрешности приближенного метода решения обычно полагают, что все вычислительные операции выполняются точно, реальное получение приближенного решения происходит с некоторой погрешностью, учет которой (по крайней мере в некоторых случаях) может быть необходим.

1.1.2. Абсолютная и относительная погрешности

Если a^*-неизвестное точное значение некоторой величины, а a - известное приближение к нему, то абсолютной погрешностью приближения называют обычно некоторую величину $\Delta(a)$, про которую известно, что она удовлетворяет неравенству: $|a^* - a| \le \Delta(a)$. В соответствии со смыслом понятия погрешность в качестве значения $\Delta(a)$ стараются использовать величину, наиболее близкую к $|a^* - a|$.

Пример 1. Определить абсолютную погрешность числа 1.41, взятого в качестве приближенного значения числа $\sqrt{2}$

Известно, что 1.41 < $\sqrt{2}$ < 1.42. Значит | $\sqrt{2}$ -1.41| < 0.01.

Можно принять **Δa** =0.01.

Если учесть, что 1.41 < $\sqrt{2}$ < 1.41421, то получим лучшую оценку **Δa**=0.00421. Заменяя это число большим, но более простым по записи, получим **Δa** =0.005.

Относительной погрешностью приближения называют некоторую величину $\delta(a)$, про которую известно, что она удовлетворяет неравенству:

$\left|\frac{a^* - a}{a}\right| \le \delta(a)$. При этом обычно полагают $\delta(a) = \frac{\Delta(a)}{a}$

Пример 2. Заменим число $\sqrt{2}$ приближенным значением 1.41.Будем полагать при этом Δ(a)=0.005.

Вычислим $\delta(a) = \frac{0.005}{1.41} = 0.0035 = 0.35\,\%$

Относительную погрешность часто выражают в процентах. Она дает более точное представление о величине ошибки, содержащейся в некоторой величине.

Значащая цифра, входящая в запись приближенного значения некоторой величины называется верной (верной в строгом смысле), если абсолютная погрешность значения не превосходит единицы (половины единицы) разряда, соответствующего этой цифре

1.2. Погрешности арифметических операций.

1.2.1. Погрешность вычисления значений функции.

Пусть $y = f(x_1, x_2, \ldots, x_n)$ непрерывно дифференцируемая функция, $\tilde{x}_i$ - приближенные значения ее аргументов, для которых $|x_i - \tilde{x}_i| \le \Delta(x_i)$ - известные абсолютные погрешности.

Для погрешности приближенного значения функции $\tilde{y} = f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ по формуле Лагранжа получаем $y - \tilde{y} = \sum_{i=1}^{n} a_i(\theta)(x_i - \tilde{x}_i)$,

где $a_i(\theta) = \frac{\partial f}{\partial x_i}(\tilde{x}_1 + \theta(x_1 - \tilde{x}_1), ..., \tilde{x}_n + \theta(x_n - \tilde{x}_n))$

Заменяя $a_i(\theta) \approx \tilde{a}_i = \frac{\partial f}{\partial x_i}(\tilde{x}_1, ..., \tilde{x}_n)$, получаем $y - \tilde{y} \approx \sum_{i=1}^{n} \tilde{a}_i (x_i - \tilde{x}_i)$

Оценка погрешности соответственно:

$$|y - \tilde{y}| \le \Delta(y) = \sum_{i=1}^{n} A_i \Delta(x_i), \text{ где } A_i = sup\left|\frac{\partial f}{\partial x_i}(x_1, \dots x_n)\right|$$

или $|y - \tilde{y}| \approx \sum_{i=1}^{n} \tilde{A}_i \Delta(x_i)$, где $\tilde{A}_i = \left|\frac{\partial f}{\partial x_i}(\tilde{x}_1, ... \tilde{x}_n)\right|$.

Относительная погрешность $\delta(\tilde{y}) = \frac{\sum_{i=1}^{n} \tilde{A}_i \Delta(x_i)}{f(\tilde{x}_1, ..., \tilde{x}_n)}$

1.2.2. Погрешность суммы

Пусть задана функция $y = f(x_1, x_2) = x_1 + x_2$

Тогда $\frac{\partial f}{\partial x_i} = 1,\ i = 1,2,\ A_i = 1$

Для абсолютной погрешности получаем

$$\Delta(y^*) = \Delta(x_1^*) + \Delta(x_2^*).$$

Относительная погрешность

$$\delta(y^*) = \frac{\Delta(x_1^*) + \Delta(x_2^*)}{x_1^* + x_2^*} = \frac{\delta(x_1^*) \cdot x_1^* + \delta(x_2^*) \cdot x_2^*}{x_1^* + x_2^*}.$$

Пусть $m \le \delta(x_1^*),\ \delta(x_2^*) \le M$, тогда $m \le \delta(y^*) \le M$, т.е. при сложении приближенных величин относительная погрешность не возрастает.

1.2.3. Погрешность разности

Пусть задана функция $y = f(x_1, x_2) = x_1 - x_2$

Тогда аналогично предыдущему абсолютная погрешность

$$\Delta(y^*) = \Delta(x_1^*) + \Delta(x_2^*).$$

Для относительной погрешности имеем формулу

$$\delta(y^*) = \frac{\Delta(x_1^*) + \Delta(x_2^*)}{x_1^* - x_2^*}.$$

Отсюда следует, что если приближенные значения x_1^* и x_2^* близки друг к другу, то относительная погрешность их разности $\delta(y^*)$ может оказаться намного больше $\delta(x_1^*)$ и $\delta(x_2^*)$.

1.2.4. Погрешность произведения

Пусть задана функция $y = f(x_1, x_2) = x_1 \cdot x_2$

Тогда абсолютная погрешность

$$\Delta(y^*) = \left|x_2^*\right| \cdot \Delta(x_1^*) + \left|x_1^*\right| \cdot \Delta(x_2^*).$$

Относительная погрешность

$$\delta(y^*) = \frac{\Delta(y^*)}{\left|x_1^* \cdot x_2^*\right|} = \frac{\Delta(x_1^*)}{\left|x_1^*\right|} + \frac{\Delta(x_2^*)}{\left|x_2^*\right|} = \delta(x_1^*) + \delta(x_2^*).$$

1.2.5. Погрешность частного

Пусть задана функция $y = f(x_1, x_2) = \dfrac{x_1}{x_2}$

Тогда абсолютная погрешность

$$\Delta(y^*) = \frac{1}{\left|x_2^*\right|} \cdot \Delta(x_1^*) + \frac{\left|x_1^*\right|}{\left|x_2^*\right|^2} \cdot \Delta(x_2^*).$$

Относительная погрешность

$$\delta(y^*) = \frac{\Delta(x_1^*)}{\left|x_1^*\right|} + \frac{\Delta(x_2^*)}{\left|x_2^*\right|} = \delta(x_1^*) + \delta(x_2^*)$$

1.3. Обратная задача оценки погрешности

Иногда возникает задача определения допустимой погрешности аргументов, при которой погрешность значений функции будет не более заданной величины ε.

Используем ранее полученное неравенство

$$\left|y - y^*\right| \le \sum_{i=1}^{n} A_i \Delta(x_i^*), A_i = \sup\left|\frac{\partial f}{\partial x_i}\right|.$$

Должно быть $\sum_{i=1}^{n} A_i \Delta(x_i^*) \le \varepsilon$.

При n=1 вопрос решается однозначно:

$$\Delta(x_1^*) \le \frac{\varepsilon}{A_1}$$

При $n>1$ возможны разные подходы:

1. Считать погрешности всех аргументов одинаковыми

$$\Delta(x_1^*) = \ldots = \Delta(x_n^*) = \Delta$$

Тогда получаем $\Delta\sum_{i=1}^{n} A_i \le \varepsilon$, следовательно $\Delta \le \dfrac{\varepsilon}{\sum A_i}$

2. Считать, что вклад погрешности каждого аргумента в погрешность результата одинаков. $A_1 \cdot \Delta(x_1^*) = \ldots = A_n \cdot \Delta(x_n^*) = \dfrac{\varepsilon}{n}$, тогда

$$\Delta(x_i^*) = \frac{\varepsilon}{A_i \cdot n}$$

Если для разных аргументов достижение определенной точности их задания существенно различается, то можно ввести функцию стоимости $F(\Delta(x_1^*),\ldots,\Delta(x_n^*))$ затрат на задание точки $x_1^*,\ldots,x_n^*$ с заданными абсолютными погрешностями $\Delta(x_1^*),\ldots,\Delta(x_n^*)$ и искать ее минимум в области

$\sum A_i \Delta(x_i^*) \le \varepsilon$, $\Delta(x_i^*) \ge 0$

1.4. Обратный анализ погрешности.

Доказывается, что в некоторых задачах (в частности для СЛАУ В.В.Воеводин Вычислительные основы линейной алгебры, с. 42-45) результат вычислений с округлениями над точными исходными данными можно трактовать как точные вычисления с искаженными исходными данными, т.е. свести учет вычислительной погрешности к учету погрешности исходных данных. Погрешность исходных данных, эквивалентная в указанном смысле вычислительной погрешности, называется эквивалентным возмущением.

Сравнение величин первоначальной ошибки исходных данных и эквивалентного возмущения при решении задачи позволяет правильно соотнести точность исходных данных и точность вычислений.

1.5. Статистический и технический подход к учету погрешности

Описанный выше классический (аналитический) подход к оценке погрешности вычислений требует оценки погрешности каждой арифметической операции или вычисления значения функции. Кроме того, он учитывает наихудший вариант взаимодействия погрешностей операндов и дает обычно завышенные оценки.

При большом объеме операций применимы статистические законы для описания погрешностей. Например, математическое ожидание абсолютной погрешности суммы n слагаемых с одинаковыми абсолютными погрешностями пропорционально $\sqrt{n}$. В частности, согласно правилу Н. Г. Чеботарева, если все слагаемые округлены до m-го разряда, то: $\Delta S \approx \sqrt{3n} \cdot 0.5 \cdot 10^{-m}$. отсюда следует, например, что погрешность среднего арифметического n слагаемых стремится к нулю при $n \to \infty$.

В практике ручных приближенных вычислений наиболее распространен основанный на статистическом подходе набор правил, сформулированный А. Н. Крыловым:

- в записи приближенного числа все значащие цифры, кроме последней, должны быть верными;
- при сложении и вычитании в результате надо сохранять количество десятичных знаков в дробной части, равное наименьшему количеству этих знаков в операндах;
- при умножении и делении в результате надо сохранять количество значащих цифр, равное наименьшему количеству таких цифр в операндах;
- в результатах промежуточных операций надо дополнительно сохранять один или два десятичных знака или значащие цифры, которые в окончательном результате отбрасываются.

Эти правила выражают технический подход к оценке погрешности. И хотя они не гарантируют правильную оценку погрешности в каждом конкретном случае, но обеспечивают получение разумных результатов как правило (в среднем).

1.6. Особенности машинной арифметики

Фундаментальной особенностью вычислений является тот факт, что они используют представление чисел с помощью конечного числа символов. При использовании позиционной системы счисления это означает, что числа записываются конечным числом разрядов (в конечной разрядной сетке). А это в свою очередь означает что в реальных расчетах как правило неизбежно присутствуют погрешности округления. (Д.В.Беклемишев: ошибки округления в численных расчетах столь же неустранимы, как и ошибки измерения в физике. Беклемишев Дмитрий Владимирович, профессор кафедры высшей математики Московского Физико-технического института.)

Вышесказанное полностью относится и к вычислениям на компьютерах, в которых как правило используется кроме обычной записи целых чисел (в двоичной системе счисления) также запись чисел в форме с плавающей точкой, применяемая для представления вещественных чисел. Точнее, такая запись в силу конечной разрядной сетки позволяет представить только некоторое конечное множество рациональных чисел.

Конкретные характеристики арифметики различны для разных стандартов. Для ПЭВМ наиболее распространённым является IEEE-стандарт (IEEE-754-1985) [Institute of Electrical and Electronic Engineers, его изначально разрабатывал один человек, профессор математики университета Беркли Вильям Каган (William Kahan).], согласно которому вещественные числа представляются в двух основных формах.

Данные с плавающей точкой по IEEE-стандарту

Тип	Размер, бит	Диапазон изменения полож.чисел максимум минимум	Точность десятичных цифр	Машинное e
single	32	3.4*10+38 1.2*10-38	Не менее 6	1.192*10-7
double	64	1.8*10+308 2.2*10-308	Не менее 15	2.221*10-16

2. Численные методы алгебры

В данном разделе рассматриваются методы решения следующих алгебраических задач:

- нахождение корней конечного уравнения $f(x)=0$;
- решение системы линейных алгебраических уравнений;
- решение нелинейной системы конечных уравнений;
- нахождение собственных чисел и собственных векторов матрицы.

2.1. Предварительные сведения

2.1.1. Понятие близости в метрическом пространстве.

Определение 1.

Множество X элементов произвольной природы (не обязательно числовое множество) называется *метрическим пространством*, если любой паре элементов $x, y \in X$ поставлено в соответствие число $\rho(x, y)$, (метрика, или расстояние) в соответствии с аксиомами:

А1. $\rho(x, y) \geq 0$, $\rho(x, y) = 0$ тогда и только тогда, когда $x=y$.

А2. $\rho(x, y) = \rho(y, x)$.

А3. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ – неравенство треугольника.

Определение 2.

Говорят, что последовательность элементов $\{x_n\}$ метрического пространства X сходится к элементу $x \in X$, если $\lim\limits_{n\to\infty} \rho(x_n, x) = 0$.

Определение 3.

Последовательность $\{x_n\}$ элементов метрического пространства X называется *фундаментальной*, если

$$\forall \varepsilon > 0 \; \exists N : n, m > N \Rightarrow \rho(x_n, x_m) < \varepsilon .$$

Определение 4.

Метрическое пространство X называется *полным*, если любая фундаментальная последовательность $\{x_n\}$ его элементов сходится к некоторому элементу этого пространства.

Определение 5.

Множество X называется *нормированным линейным пространством*, если

- оно является линейным пространством, т.е. в нем определены операции сложения элементов и умножения элемента на число с известными свойствами.

- любому элементу $x \in X$ поставлено в соответствие число $\|x\|$ (норма x), удовлетворяющее аксиомам:

A1. $\|x\| \geq 0, \|x\| = 0 \Leftrightarrow x = 0$,

A2. $\|\alpha x\| = |\alpha| \|x\|, \forall \alpha \in R$

A3. $\|x + y\| \leq \|x\| + \|y\|$ – неравенство треугольника.

Замечание.

Любое нормированное линейное пространство X можно считать метрическим, введя метрику по формуле

$$\rho(x, y) = \|x - y\|. \tag{1}$$

Если последовательность $\{x_n\}$ нормированного пространства X сходится в смысле метрики (1), то говорят о сходимости по норме пространства X.

Примеры классов функций и соответствующих нормированных пространств.

Пример 1.

Множество всех функций, заданных на отрезке [a, b] и имеющих на нем непрерывные производные до k -го порядка включительно, называется классом $C^k[a,b]$.

Пример 2.

При $k=0$ получаем класс $C^0[a,b]$ - множество непрерывных на отрезке [a, b] функций.

Если на $C^0[a,b]$ ввести норму по формуле

$$\|f\|_C = \max_{[a,b]} |f(x)|, \tag{2}$$

то получим линейное нормированное пространство $C[a,b]$ (операции сложения и умножения на число вводятся обычным образом $f+g=f(x)+g(x)$, $\alpha f = \alpha f(x)$).

Аксиомы A1, A2 – очевидно, выполняются.

В справедливости A3 нетрудно также убедиться с помощью свойств модуля и теоремы Вейерштрасса.

Пример 3.

Множество всех функций, p-я степень модуля которых интегрируема на отрезке [a, b], называется линейным нормированным пространством $L_p[a,b]$, если на нем введена норма по формуле

$$\|f\|_{l_l} = \left(\int_a^b |f(x)|^p \, dx \right)^{\frac{1}{p}}. \tag{4}$$

Сходимость по норме (4) называется сходимостью в среднем (при $p=2$ - среднеквадратичная сходимость).

2.1.2. Принцип сжатых отображений.

Пусть X – полное метрическое пространство, $\rho(x,y)$ - расстояние между элементами x и y. Пусть, кроме того, S – замкнутое ограниченное множество (компакт): $S \subseteq X$ и T – оператор (вообще говоря, – нелинейный), действующий из S в S, то есть отображающий множество S в себя: $Tx \in S, \forall x \in S$.

Назовем точку $x^* \in S$ *неподвижной точкой* оператора T, если

$$x^*=Tx^* \tag{1}$$

Таким образом, неподвижные точки оператора T являются решениями уравнения (1). Наиболее простой способ решения этого уравнения – итерационный, начиная с некоторого значения x^0

$$x^{n+1}=Tx^n, \ x^0 \in S \tag{2}$$

При этом важно, чтобы такая последовательность $\{x^n\}$ сходилась к единственной точке x^*. Следующая теорема формулирует достаточные условия сходимости итерационного процесса (2).

Теорема 1. (Принцип сжатых отображений).

Пусть T – оператор сжатия на S, то есть

$$\forall x, y \in S : Tx \in S \ \text{ и } \ \rho(Tx,Ty) \le \alpha \cdot \rho(x,y), \alpha \in (0,1) \tag{3}$$

Тогда в S существует единственная неподвижная точка оператора T, являющаяся пределом последовательности $\{x^n\}$, определяемой процедурой итераций, начиная с $\forall x^0 \in S$. При этом скорость сходимости оценивается неравенствами:

$$\rho(x^n,x^*) \le \frac{\alpha^n}{1-\alpha} \cdot \rho(x^1,x^0) \tag{4}$$

$$\rho(x^n,x^*) \le \frac{\alpha}{1-\alpha} \cdot \rho(x^n,x^{n-1}) \tag{5}$$

◅Докажем, что последовательность $\{x^n\}$ – фундаментальная. Рассмотрим

$$\rho(x^{k+1},x^k) = \rho(Tx^k,Tx^{k-1}) \le \alpha \cdot \rho(x^k,x^{k-1}) \le ... \le \alpha^k \cdot \rho(x^1,x^0) \tag{6}$$

Далее при p>1 имеем

$\rho(x^{n+p},x^n) \le$ {неравенство треугольника: вставим точку x^{n+p-1}} $\le$

$\le \rho(x^{n+p},x^{n+p-1}) + \rho(x^{n+p-1},x^n) \le$ {продолжая вставлять точки} $\le$

$\le \rho(x^{n+p},x^{n+p-1}) + \rho(x^{n+p-1},x^{n+p-2}) + ... + \rho(x^{n+1},x^n) \le$ {на основании (6)} $\le$

$\le (\alpha^{n+p-1} + \alpha^{n+p-2} + ... + \alpha^n)\rho(x^1,x^0) \le$ {геометр. прогрессия} $\le$

$$\le \alpha^n(1+\alpha+...+\alpha^{p-1})\rho(x^1,x^0) \le \frac{\alpha^n}{1-\alpha}\rho(x^1,x^0). \tag{7}$$

Отсюда следует, что

$$\rho(x^{n+p}, x^n) \xrightarrow{n\to\infty} 0\,,\ \forall p > 1$$

следовательно, последовательность $\{x^n\}$ – фундаментальная, и согласно критерию Коши-Вейерштрасса последовательность $\{x^n\}$ сходится к элементу $x^* \in S$ (так как S - компакт). Таким образом, имеем

$$\lim_{k\to\infty} x^{k+1} = \lim_{k\to\infty} Tx^k = x^*\,.$$

Далее

$$\rho(x^{k+1}, Tx^*) = \rho(Tx^k, Tx^*) \le \alpha\rho(x^k, x^*) \xrightarrow{k\to\infty} 0\,.$$

Следовательно,

$$Tx^k \to Tx^* \Rightarrow x^* = Tx^*\,.$$

Докажем единственность неподвижной точки x^*.

От противного. Пусть $\exists x^* \neq y^*$: $x^*=Tx^*$, $y^*=Ty^*$. Тогда

$$\rho(x^*, y^*) = \rho(Tx^*, Ty^*) \le \alpha \cdot \rho(x^*, y^*) < \rho(x^*, y^*)\,.$$

Но это противоречие.

Формула (4) следует из формулы (7) при $p \to \infty$:

$$\lim_{p\to\infty} \rho(x^{n+p}, x^n) = \rho(x^*, x^n) \le \frac{\alpha^n}{1-\alpha}\rho(x^1, x^0)\,,$$

т.к. правая часть неравенства (7) не зависит от p.

Докажем (5):

$$\rho(x^n, x^*) \le \{\text{неравенство треугольника}\} \le \rho(x^{n+1}, x^n) + \rho(x^{n+1}, x^*) \le$$

$$\le \rho(Tx^n, Tx^{n-1}) + \rho(Tx^n, Tx^*) \le \alpha\rho(x^n, x^{n-1}) + \alpha\rho(x^n, x^*)\,.$$

Отсюда

$$(1-\alpha)\rho(x^n, x^*) \le \alpha\rho(x^n, x^{n-1})\,.$$

Если разделить обе части этого неравенства на (1-α), то получим (5). ⊳

Замечание 1.

Неравенство (4) показывает, что последовательность $\{x^n\}$ сходится к x^* со скоростью *геометрической прогрессии* (такая скорость называется *линейной*: каждый шаг в α раз приближает к x^*). Кроме того, неравенство (4) позволяет определить, сколько итераций (шагов) необходимо сделать для достижения заданной точности ε. Для этого нужно решить неравенство:

$$\frac{\alpha^n}{1-\alpha}\rho(\mathrm{x}^1, \mathrm{x}^0) \le \varepsilon$$

Ясно, что для хорошей оценки числа итераций необходимо точнее оценивать константу сжатия α, что на практике не всегда просто сделать. При реализации алгоритма полезно также использовать неравенство (5), позволяющее контролировать каждый шаг итерации и установить следующий критерий останова:

$$\rho(x^n, x^{n-1}) \le \frac{1-\alpha}{\alpha}\varepsilon \Rightarrow \rho(x^n, x^*) \le \varepsilon \Rightarrow STOP : x^* \approx x^n .$$

Теорема 2.

Пусть X – *банахово пространство*, то есть полное нормированное пространство с нормой элементов $\|x\|, x \in X$. T- оператор, определенный на замкнутом множестве S и отображающий S в себя. Тогда, если выполняется условие

$$\|Tx - Ty\| \le \alpha\|x - y\|, \alpha \in (0,1), x, y \in S \qquad (8)$$

(это условие Липшица с константой $\alpha \in (0,1)$), то справедливо утверждение теоремы 1.

◁ Действительно, положим $\rho(x, y) = \|x - y\| \Rightarrow$ результат. ▷

2.2. Итерационные методы для функциональных уравнений

Пусть дана непрерывная на некотором промежутке функция $f(x)$. Необходимо найти принадлежащие этому промежутку корни уравнения

$$f(x) = 0 \qquad (1)$$

Как правило, алгоритм приближенного метода состоит из двух этапов:

- поиск приближенного значения корня или содержащего его отрезка;

- уточнение приближенного значения до некоторой заданной степени точности.

Иногда ограничиваются только первым этапом. При этом могут использоваться решения близких задач, графические методы, физические соображения и т.д. На втором этапе для уточнения приближенного значения обычно строится последовательность, элементы которой в пределе сходятся к точному значению корня. Сам метод решения при этом называется итерационным или методом последовательных приближений.

2.2.1. Метод итераций для функциональных уравнений.

Утверждение 1.

Пусть $x \in R$ (одномерный случай) и задана функция *f(x)*, удовлетворяющая условию:

$$|f(x) - f(y)| < \alpha|x - y|, \alpha \in (0,1); x, y \in [a, b] \qquad (9)$$

(Условие Липшица с константой α на отрезке [*a*,*b*].)

Тогда оператор $f(x)$ - сжимающий и уравнение $f(x)=x$ имеет единственную неподвижную точку, которую можно найти методом простых итераций:

$$x_{n+1} = f(x_n), x_0 \in [a,b].$$

⊲Действительно, определим $\|x\| = |x|$. Следовательно, выполняется условие (8) теоремы 2, откуда и следует результат. ⊳

Утверждение 2.

Пусть $f(x) \in C^1[a,b]$, причем

$$|f'(x)| \le \alpha < 1, \forall x \in [a,b] \qquad (10)$$

Тогда оператор $f(x)$ является сжимающим.

⊲ Согласно теореме о среднем

$$f(x) - f(y) = f'(\xi)(x-y), x < \xi < y.$$

Оценим это неравенство по модулю:

$$|f(x) - f(y)| = |f'(\xi)| \cdot |x-y| \le \alpha |x-y|, \alpha \in (0,1).$$

Это говорит о том, что выполняется условие (9) утверждения 1, значит, $f(x)$ действительно сжимающий оператор. ⊳

Рассмотрим задачу поиска корней уравнения $F(x) = 0$. Пусть известны границы для корня этого уравнения и мы хотим найти этот корень методом итераций. Если удастся привести уравнение к виду $x=f(x)$, так чтобы выполнялось одно из условий утверждения 1 или утверждения 2, то в этом случае можно будет применить метод итераций. Такое преобразование, вообще говоря, не единственно, причем главная трудность заключается в определении того замкнутого ограниченного множества S (а в одномерном случае – отрезка $[a,b]$), для которого помимо условия сжатости, выполняется условие $f(x) \in S, \forall x \in S$.

2.2.2. Метод простых итераций

Во многих случаях метод последовательных приближений решения уравнения $f(x) = 0$ может быть описан в рамках следующей общей схемы. Уравнение представляется в виде $x = \varphi(x)$, выбирается начальное приближение x_0, а затем следующие приближения вычисляются по формуле:

$$x_{k+1} = \varphi(x_k), k = 0,1,\ldots \qquad (13)$$

Достаточное условие сходимости этого итерационного процесса - выполнение неравенства $|\varphi'(x)| < 1$, так как и $|x_{k+1} - x^*| = |\varphi'(\xi)| \cdot |x_k - x^*|$ при данном условии отображение φ будет сжимающим.

Если взять функцию $\varphi(x)$ в виде $x - \psi(x) \cdot f(x)$, то за счет соответствующего выбора $\psi(x)$ можно обеспечить сходимость метода.

Методы хорд и касательных также можно трактовать как частные случаи итерационного метода, когда соответственно выбирается $\psi(x)=(b-x)/(f(b)-f(x))$ для метода хорд и $\psi(x)=1/f'(x)$ для метода касательных.

В простейшем случае полагают $\psi(x)=\tau$. Тогда итерационный метод (13) имеет вид: $$x_{k+1}=x_k-\tau\cdot f(x_k),\ n=0,1,\ldots \qquad (14)$$

и называется методом простых итераций.

Достаточное условие сходимости в этом случае означает: $\left|\varphi'(x)\right|=\left|1-\tau\cdot f'(x)\right|<1$.

Оно будет выполнено, если параметр τ выбирать удовлетворяющим неравенству $0<\tau<\frac{2}{M_1}$ при $f'(x)>0$, и неравенству $0>\tau>-\frac{2}{M_1}$ при $f'(x)<0$, где $M_1\geqslant max|f'(x)|$.

Рассмотрим вопрос о нахождении такого значения τ^* из указанного в условии сходимости промежутка, при котором количество итераций, необходимых для получения решения с заданной точностью, будет минимальным (или будет минимальной погрешность решения, полученного после фиксированного количества итераций). С учетом того, что $|x_{n+1}-x^*|\leq q\cdot|x_n-x^*|\leq\ldots\leq q^{n+1}\cdot|x_0-x^*|$, эта минимальность будет достигаться при минимальном значении величины

$$q=\max\left|\varphi'(x)\right|=\max\left|1-\tau\cdot f'(x)\right|=\max\begin{cases}1-\tau\cdot f'(x), 1-\tau\cdot f'(x)>0\\ -1+\tau\cdot f'(x), 1-\tau\cdot f'(x)\leq 0\end{cases}$$

Тогда $|q|\leq\max\{1-m\tau,-1+M\tau\}$, где $m=\min|f'(x)|$, $M=\max|f'(x)|$

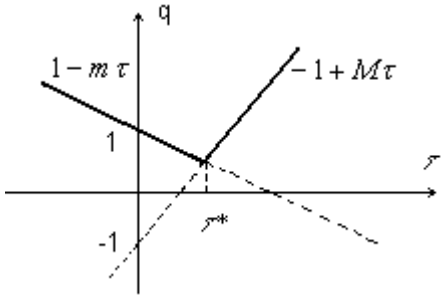


Рис. 4

Правая часть последнего неравенства будет минимальна при значении $\tau=\tau^*$, которое соответствует указанной на рис. 4 точке пересечения прямых. Отсюда получаем $\tau^*=\frac{2}{M+m}$. Тогда $q^*=\frac{M-m}{M+m}$ будет минимальным значением величины q.

2.2.3. Метод деления отрезка пополам

Одним из итерационных методов является метод деления отрезка пополам (дихотомии, бисекции).

На первом этапе должен быть найден отрезка $[a,b]$ такой, что $f(a)\cdot f(b)<0$.

Тогда отрезок $[a,b]$ содержит нечетное число корней уравнения (1) нечетной кратности (x^* - корень кратности p, если $f^{(k)}(x^*)=0, k=0,1,\ldots,p-1,$).

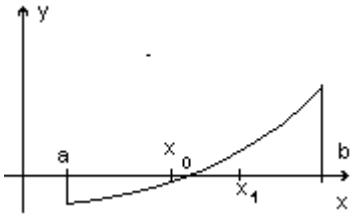


Рис. 1

Начальное приближение $c=\frac{a+b}{2}$, как показано на рис. 1. На втором этапе выбирается тот из двух отрезков $[a,c]$, $[c,b]$, на концах которого функция $f(x)$ имеет значения разных знаков и т. д..

Таким образом, строится последовательность содержащих корень уравнения x^* вложенных промежутков $[x^1_n, x^2_n]$, $n=0,1,\ldots$, где $x^1_0=a$, $x^2_0=b$, $x^1_{n+1}=(x^1_n+x^2_n)/2$ или x^1_n, а $x^2_{n+1}=x^2_n$ или $(x^1_n+x^2_n)/2$ соответственно

После каждой итерации длина отрезка, содержащего корень, уменьшается вдвое

Инерционный процесс продолжается до тех пор, пока длина полученного отрезка не станет меньше заданной величины $2\cdot\varepsilon$. За приближенное решение принимается средняя точка последнего промежутка.

Таким образом погрешности приближенного решения $r^2_n=x^2_n-x^*$ и оценива- $r^1_n=x^*-x^1_n$ ются следующим образом:

$r_n\le|x^2_n-x^1_n|\le\frac{1}{2^n}\cdot|b-a|$ Поскольку мажориру- r_n ется членом убывающей геометрической прогрессии, то говорят, что метод сходится со скоростью геометрической прогрессии.

Другой используемый на практике вариант условия окончания итерационного процесса $|f(x_n)|<\varepsilon$ (по величине невязки).

2.2.4. Метод хорд

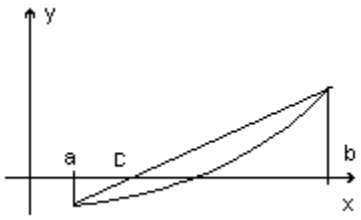


Рис. 2

Геометрический смысл этого метода заключается в замене кривой $y=f(x)$ хордой. Очередное приближение находится как точка пересечения хорды с осью абсцисс. (Рис. 2)

Если $[a,b]$ - отрезок содержащий корень, то уравнение хорды

$$\frac{y-f(a)}{f(b)-f(a)}=\frac{x-a}{b-a}. \qquad (2)$$

Для точки пересечения хорды с осью абсцисс $x=c, y=0$ имеем

$$c=a-\frac{b-a}{f(b)-f(a)}\cdot f(a).$$

$x=c$ принимается за очередное приближение к корню. Далее выбирается тот из промежутков $[a,c],[c,b]$ на концах которого функция имеет значения разных знаков и

т.д. При этом, если $f(x)$ дважды непрерывно дифференцируемая функция и знак $f''(x)$ сохраняется на рассматриваемом промежутке, то полученные приближения будут сходиться к корню монотонно.

Если знаки $f'(x)$ и $f''(x)$ сохраняются на исходном промежутке, содержащем корень, то у всех получаемых промежутков один конец будет общим, а именно тот, на котором совпадают знаки функции и второй производной. Например, если $f(b)\cdot f''(b)>0$, то последовательные приближения к корню вычисляются по формуле

$$x_{n+1}=x_n-\frac{b-x_n}{f(b)-f(x_n)}f(x_n), x=0,1,\ldots,x_0=a \qquad (3)$$

и корень принадлежит последовательности вложенных отрезков $[x_n,b]$

Если оставить неподвижным тот конец промежутка, где знаки $f(x)$ и $f''(x)$ противоположны, то после вычисления x_1 получаем промежуток не содержащий корень уравнения. Дальнейшее развитие событий зависит от поведения конкретной функции и величины промежутка. Возможна как сходимость метода (при этом соседние приближения находятся по разные стороны от корня), так и его расходимость.

Рассмотрим сходимость метода хорд и оценки погрешности приближенных решений.

Пусть на исходном промежутке $[a,b]$ функция $f(x)$ дважды дифференцируема, знаки $f(x)$ и $f''(x)$ сохраняются и $f(b)\cdot f''(b)>0$

Из формулы (3) получаем

$$-f(x_n)=\frac{x_{n+1}-x_n}{b-x_n}\cdot(f(b)-f(x_n)).$$

Прибавляя слева $f(x^*)$ и применяя к разности $f(x^*)-f(x_n)$

формулу конечных приращений (формулу Лагранжа), далее получаем

$$(x^*-x_n)f'(\xi)=(x_{n+1}-x_n)f'(\bar{\xi}) \qquad (4)$$

Из формулы (4), добавляя справа в скобке $\pm x^*$ и группируя члены, получаем

$$(x_{n+1}-x^*)=(x_n-x^*)\cdot\frac{f'(\bar{\xi})-f'(\xi)}{f'(\bar{\xi})}.$$

Так как знаки разностей $(x_{n+1}-x^*)$ и (x_n-x^*) совпадают, то

$$q_n=\frac{f'(\bar{\xi})-f'(\xi)}{f'(\bar{\xi})}>0 ,$$

Причем $f'(\xi)$ и $f'(\bar{\xi})$ одного знака. Тогда $q_n<1$

Следовательно,

$$\left|x_{n+1}-x^*\right| = q_n \cdot \left|x_n - x^*\right| \le q \cdot \left|x_n - x^*\right| \qquad (5)$$

где $q = \sup_n (q_n)$

Из (5) получаем

$$\left|x_n - x^*\right| \le q^n \cdot \left|x_0 - x^*\right|$$

Отсюда следует, что погрешность приближенного решения $(x_n - x^*)$ стремится к нулю при $n \to \infty$. В этом случае говорят, что метод сходится. И когда убывание погрешности приближенного решения характеризуется неравенством вида (5) говорят также, что метод имеет линейную скорость сходимости. (Сходится со скоростью геометрической прогрессии.)

Из формулы (4) также следует неравенство, в котором погрешность приближенного решения оценивается через разность двух последовательных приближений

$$\left|x_n - x^*\right| \le \frac{M_1}{m_1} \cdot \left|x_{n+1} - x_n\right| \qquad (6)$$

Здесь $M_1 = \max\left|f'(x)\right|$ и $m_1 = \min\left|f'(x)\right|$ на рассматриваемом отрезке.

Другой вариант оценки погрешности приближенного решения через невязку приближенного решения $f(x_n)$ дает сама формула конечных приращений

$$f(x_n) - f(x^*) = f'(\xi)(x_n - x^*)$$

Отсюда (с учетом, что $f(x^*) = 0$) получаем

$$\left|x_n - x^*\right| \le \frac{\left|f(x_n)\right|}{m_1} \qquad (7)$$

Фомы (6) и (7) позволяют установить, что для получения приближенного решения с заданной точностью (т.е. такого x_n, для которого $\left|x_n - x^*\right|$ будет меньше заданного числа ε) достаточно выполнить такое количество итераций $(n+1)$, после которого будет выполнено хотя бы одно из условий

$$\frac{M_1}{m_1} \cdot \left|x_{n+1} - x_n\right| < \varepsilon \quad \text{или} \quad \frac{\left|f(x_{n+1})\right|}{m_1} < \varepsilon$$

2.2.5. Метод Ньютона (метод касательных)

Геометрический смысл метода Ньютона (метода касательных) заключается в том, что на отрезке $[a,b]$содержащем корень уравнения (1) график функции $f(x)$ заменяет-

ся отрезком касательной, проведенной к графику $f(x)$ при $x = a$ или $x = b$. (Предполагая что функция $f(x)$ дифференцируема на $[a,b]$.)

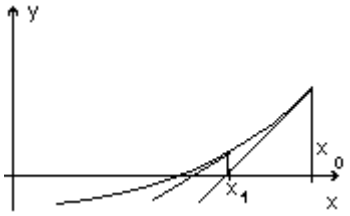


Рис. 3

При этом используется только одна точка, поэтому не обязательно задавать отрезок $[a,b]$, содержащий корень, достаточно задать некоторое приближение x_0.

Уравнение касательной в точке (x_0, y_0)

$$y - f(x_0) = f'(x_0)(x - x_0).$$

Для точки пересечения с осью 0x получаем $x_1 = x_0 - \dfrac{f(x_0)}{f'(x_0)}$, и т. д. (рис. 3).

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (8)$$

Объем вычислений в методе Ньютона на каждом шаге выше, чем в предыдущих методах, т. к. в точке x_n вычисляются значения функции и ее производной, что компенсируется более высокой скоростью сходимости этого метода.

Но в отличие от предыдущих методов метод Ньютона сходится не при всяком выборе начального приближения на отрезке, содержащем корень уравнения.

Легко проверить, что примером достаточных условий сходимости метода будет сохранение знака второй производной $f(x)$ на некотором промежутке, содержащем корень, и выбор начального приближения с той стороны от корня, где знак функции совпадает со знаком второй производной.

При этом последовательные приближения будут сходится к корню монотонно.

Другой вариант достаточного условия сходимости получается при исследовании скорости сходимости вблизи корня

Перепишем формулу (8) в виде

$$x_{n+1} - x^* = x_n - x^* - \frac{f(x_n) - f(x^*)}{f(x_n)}$$

Используя для разности $f(x_n) - f(x^*)$ разложение по формуле Тейлора до членов второго порядка, получим

$$x_{n+1} - x^* = \frac{f''(\xi)}{2 \cdot f'(x_n)} \cdot (x_n - x^*)^2.$$

Отсюда

$$\left| x_{n+1} - x^* \right| \le q \cdot (x_n - x^*)^2, \qquad (9)$$

где $q = \frac{M_2}{2 \cdot m_1}, M_2 = \max \left| f''(x) \right|.$

Из неравенства (9) следует, что

$$\left| x_{n+1} - x^* \right| \cdot q \le (q \cdot \left| x_n - x^* \right|)^2 \le \ldots (q \cdot \left| x - x^* \right|)^{2^{n+1}} \to 0 \text{ при } q \cdot (x_o - x^*) < 1 \quad (10)$$

Условие (10) можно рассматривать как ограничение на выбор начального приближения: выполнение неравенства

$\left| x_0 - x^* \right| < \frac{2m_1}{M_2}$ достаточно для сходимости метод.

Если погрешности приближенных решений, полученных некоторым методом, удовлетворяют неравенству вида (9) (где $q < 1$), то говорят, что метод имеет квадратичную скорость сходимости.

Наличие показателя степени 2 в правой части неравенства (9) определяет большее убывание погрешности приближенных решений, полученных методом Ньютона, по сравню, например, с методом хорд.

Оценки погрешности приближенных решений могут быть получены в аналогичном методу хорд виде. Из формулы (8) следует

$$-f(x_n) = (x_{n+1} - x_n) f'(x_n),$$

отсюда получается формула, аналогичная (4)

$$(x^* - x_n) \cdot f'(\xi) = (x_{n+1} - x_n) f'(x_n).$$

Из этой формулы вытекает оценка (6)

$$\left| x_n - x^* \right| \le \frac{M_1}{m_1} \cdot \left| x_{n+1} - x_n \right|.$$

Для погрешности приближенного решения также справедлива и оценка (7)

2.2.6. Модификации метода Ньютона

1). Модифицированный метод Ньютона.

С целью уменьшить вычислительные затраты для получения приближенных решений можно заменить в формуле (8) значение $f'(x_n)$, вычисляемое для каждого x_n, на постоянное значение $f'(x_0)$. Соответствующая итерационная формула имеет вид:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \quad (9)$$

Модифицированный метод Ньютона обладает линейной скоростью сходимости.

2). Конечно-разностный метод Ньютона.

Значение производной в формуле метода Ньютона можно аппроксимировать конечно-разностным отношением вида $\left[f\ (x_n+h_n)-f\ (x_n)\right]/h_n$, где $h_n \to 0$ при $n \to \infty$. Полагая $h_n = f\ (x_n)$ (что, конечно, имеет смысл делать только если величина $f\ (x_n)$ достаточно мала), получаем итерационную формулу метода Стеффенсена

$$x_{n+1} = x_n - \frac{f^2(x_n).}{f\ (x_n + f\ (x_n))\ - f\ (x_n)} \tag{10}$$

Этот метод сохраняет квадратичную скорость сходимости метода Ньютона.

3). Метод секущих.

Этот метод основан на аппроксимации значения производной $f'(x_n)$ разностным отношением $\left(f\ (x_n)\ - f\ (x_{n-1})\ \right)/(x_n - x_{n-1})$. После задания двух первых приближений для корня последующие приближения вычисляются по формуле:

$$x_{n+1} = x_n - \frac{f(x_n)\cdot(x_n - x_{n-1})}{f\ (x_n)\ - f\ (x_{n-1})} \tag{11}$$

Метод секущих в отличие от предыдущих вариантов итерационных методов требует только одного вычисления функции на каждом шаге. Так как при этом значение k+1 -го приближения определяется значениями k -го и k-1 -го приближений, то метод называется двухшаговым. Анализ его сходимости приводит к неравенству вида:

$\left|x_{n+1} - x^*\right| \le q\cdot(x_{n-1} - x^*)^2$, которое означает, что сходимость метода двухшагово квадратичная. Такая скорость сходимости в сочетании с вычислительной экономичностью делают метод секущим самым эффективным алгоритмом приближенного решения уравнений.

4). Метод Стеффенсена

$$x_0 \in [a,b] \qquad x_{k+1} = x_k - \frac{(f(x_k))^2}{f(x_k + f(x_k)) - f(x_k)} \qquad k = 0, 1, \ldots,$$

2.2.7. Метод Чебышёва

В основе предложенного П.Л.Чебышёвым метода лежит представление функции, обратной к $f(x)$, по формуле Тейлора. Пусть функция $f(x)$ на содержащем корень x^* промежутке $[a,b]$ имеет несколько непрерывных производных и $f'(x) \ne 0$. Тогда существует имеющая соответствующую гладкость обратная функция $g(y)$ и $x^* = g(0)$. Разложение $g(y)$ по формуле Тейлора в окрестности $y_n = f(x_n)$, где x_n некоторое приближение для x^*, имеет вид:

$$g(y) = g(y_n) + g'(y_n)\cdot(y - y_n) + \frac{1}{2}g''(y_n)\cdot(y - y_n)^2 + \ldots$$

Если использовать линейное приближение для $g(y)$ $h(y) = g(y_n) + g'(y_n)\cdot(y - y_n)$, то следующее приближение для x^* определяется формулой: $x_{n+1} = h(0) = g(y_n) - g'(y_n)\cdot y_n$, а с учетом того, что $g(y_n) = x_n$ и $g'(y_n) = 1/f'(x_n)$, получаем формулу метода Ньютона.

Рассмотрим квадратичное приближение для $g(y)$:

$p(y) = g(y_n) + g'(y_n)\cdot(y - y_n) + \frac{1}{2}g''(y_n)\cdot(y - y_n)^2$. Значение

$x_{n+1} = p(0) = g(y_n) - g'(y_n)\cdot y_n + \frac{1}{2}g''(y_n)\cdot y_n^2$ можно рассматривать как новое приближение для x^*.Так как $g''(y_n) = -f''(x_n)/[f'(x_n)]^3$, то

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2}\frac{f''(x_n)}{[f'(x_n)]^3}\cdot f^2(x_n) \qquad (12)$$

Для погрешностей приближений, определяемых по формуле (12), может быть получено неравенство вида $\left|x_{n+1} - x^*\right| \le q\cdot(x_n - x^*)^3$. Это означает, что порядок скорости сходимости равен трем. Метод (12) называют методом Чебышёва третьего порядка.

Метод Вегстейна

Это совместное применение итерационной формулы:

$\bar{x}_{k+1} = x_{k+1} - \frac{(x_{k+1} - x_k)(x_{k+1} - \bar{x}_k)}{(x_{k+1} - x_k) - (\bar{x}_k - \bar{x}_{k-1})}$, где $k = 1, 2, \ldots$ и итерационной формулы простых итераций $x_{k+1} = \bar{x}_k - \tau\cdot f(\bar{x}_k)$, где $k = 0, 1, \ldots$ Начальные значения $\bar{x}_0 = x_0$, $\bar{x}_1 = x_1$

2.3. Системы линейных алгебраических уравнений. Метод Гаусса

Рассмотрим задачу решения системы уравнений вида:

$$\left.\begin{array}{l} a_{11}\cdot x_1 + a_{12}\cdot x_2 + \ldots + a_{1n}\cdot x_n = b_1 \\ a_{21}\cdot x_1 + a_{22}\cdot x_2 + \ldots + a_{2n}\cdot x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}\cdot x_1 + a_{m2}\cdot x_2 + \ldots + a_{mn}\cdot x_n = b_m \end{array}\right\} \qquad (1)$$

$$(\sum_{k=1}^{n} a_{ik}\cdot x_k = b_i, i = 1,2,\ldots,m.)$$

или $Ax = b$, где A - матрица коэффициентов системы, x - вектор неизвестных, b - вектор правых частей.

Известно (теорема Кронекера-Капелли), что система (1) совместна (разрешима) тогда и только тогда, когда ранг матрицы системы (r) равен рангу ее расширенной матрицы (матрицы, полученной добавлением к A столбца свободных членов). Совместная система имеет единственное решение, если $r = n$; иначе система имеет общее решение, в котором число свободных неизвестных равно $n - r$.

Для системы уравнений с квадратной матрицей ($m = n$) это означает, что система (1) имеет единственное решение, если ее матрица невырожденная (т. е. определитель матрицы A отличен от нуля). В этом случае решение системы линейных уравнений можно выразить по формулам Крамера через отношение определителей. Но вычисление определителя требует выполнения такого количества арифметических операций, которое делает практическое использование формул Крамера совершенно не эффективным по сравнению с другими методами решения. (Вычисление определителя порядка n методом разложения по элементам строки или столбца требует выполнения $2n!-1$ операций умножения и сложения.)

2.2.1. Общая схема метода Гаусса

Одним из вычислительно эффективных и универсальных точных методов решения систем линейных алгебраических уравнений является метод Гаусса, известный также в виде различных модификаций (метод Гаусса с выбором главного элемента, метод Гаусса-Жордана).

В основе метода Гаусса лежит идея последовательного исключения неизвестных из уравнений системы. Она реализуется путем приведения расширенной матрицы системы к ступенчатому виду (прямой ход метода Гаусса) и затем нахождения решения, если оно существует (обратный ход метода Гаусса).

Приведение к ступенчатому виду можно выполнить, последовательно заменяя строки расширенной матрицы системы их линейными комбинациями. А именно, пусть первый элемент первой строки отличен от нуля (Иначе поменяем местами эту строку

со строкой, содержащей в первом столбце ненулевой элемент.) Заменим вторую строку ее суммой с первой строкой, умноженной на такое число, чтобы обратился в нуль элемент первого столбца (т.е. коэффициент при x_1). Затем таким же образом преобразуем третью строку, четвертую и т. д. Таким образом обращаются в ноль все коэффициенты первого столбца, лежащие ниже главной диагонали.

Затем аналогичным образом при помощи второй строки преобразуем третью, четвертую и т. д. строки, получая в них нули вместо ненулевых элементов второго столбца. Продолжая этот процесс, будем последовательно получать нули в третьем, четвертом и т.д. столбцах. При этом, если при получении нулей в очередном столбце используемая на этом шаге преобразований строка содержит в этом столбце нулевой элемент, то она меняется местами с одной из ниже расположенных строк, имеющих в этом столбце ненулевой элемент.

Ступенчатый вид матрицы позволяет определить ее ранг, т.к. в этом случае все строки, содержащие ненулевые элементы (и только они) будут линейно независимы. А поскольку преобразования, используемые для получения ступенчатого вида, не могут изменить число линейно независимых строк, то и ранг исходной матрицы будет точно таким же, как ранг полученной из нее матрицы ступенчатого вида. Таким образом, после приведения расширенной матрицы системы к ступенчатому виду становится ясен вопрос о ее разрешимости и о том, имеет ли система единственное или общее решение, и о количестве свободных неизвестных в таком решении.

2.2.2. Метод Гаусса для системы с квадратной невырожденной матрицей

Запишем для этого случая общие формулы вычислительного процесса прямого хода. Пусть проведено исключение неизвестных из *k*-1 столбца. Тогда остались такие уравнения с ненулевыми элементами ниже главной диагонали:

$$\sum_{j=k}^{n} a_{ij}^{(k)} \cdot x_j = b_i^{(k)} \ , k \le i \le n.$$

Умножим *k*-ю строку на число

$$c_{mk} = \frac{a_{mk}^{(k)}}{a_{kk}^{(k)}}, m > k$$

и вычтем из *m*-й строки. Первый ненулевой элемент этой строки обратится в нуль, а остальные изменятся по формулам

$$a_{ml}^{(k+1)} = a_{ml}^{(k)} - c_{mk} \cdot a_{kl}^{(k)},$$

$$b_m^{(k+1)} = b_m^{(k)} - c_{mk} \cdot b_k^{(k)}, k < m, l \le n. \qquad (4)$$

Производя вычисления по этим формулам при всех указанных индексах, исключим элементы *k*-го столбца. После завершения прямого хода система будет иметь вид:

$$\left.\begin{array}{l} c_{11}\cdot x_1 + c_{12}\cdot x_2 + \ldots + c_{1n}\cdot x_n = d_1 \\ 0 \quad + c_{22}\cdot x_2 + \ldots + c_{2n}\cdot x_n = d_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0 + \ldots + c_{n-1,n-1}\cdot x_{n-1} + c_{n-1,n}\cdot x_n = d_{n-1} \\ 0 + \quad \ldots \quad + 0 + c_{nn}\cdot x_n = d_n \end{array}\right\} \qquad (5)$$

Тогда из последнего уравнения сразу определяем

$$x_n = \frac{d_n}{c_{nn}}.$$

Подставляя его в предпоследнее уравнение, находим

$$x_{n-1} = \frac{1}{c_{n-1,\,n-1}} \cdot (d_{n-1} - c_{n-1,n} \cdot x_n)$$

и т. д.

Общие формулы имеют вид

$$x_k = \frac{1}{c_{kk}} \cdot (d_k - \sum_{j=k+1}^{n} c_{ki} \cdot x_j), \quad k=n,\ n\text{-}1,\ldots,\ 1. \qquad (6)$$

Исключение по формулам (3)-(4) нельзя проводить, если в ходе расчета на главной диагонали оказался нулевой элемент $a_{kk}^{(k)} = 0.$ Но в k - ом столбце промежуточной системы (2) все элементы не могут быть нулями: это означало бы, что det A = 0. Перестановкой строк можно переместить ненулевой элемент на главную диагональ и продолжить расчет.

Для уменьшения вычислительной погрешности можно каждое повторение внешнего цикла начинать с выбора максимального по модулю элемента в k-том столбце (главного элемента) и перестановки уравнения с главным элементом так, чтобы он оказался на главной диагонали. Этот вариант называется методом Гаусса с выбором главного элемента.

В качестве одной из характеристик эффективности того или иного алгоритма используют вычислительные затраты, измеряемые количеством элементарных операций, которые необходимо выполнить для получения решения.

Для прямого хода метода Гаусса число арифметических операций, в соответствии с формулами (3), (4), равно

$$Q_1(n) = \sum_{k=1}^{n-1} \sum_{m=k+1}^{n} \left[\textit{деление} + \sum_{p=k}^{n+1} (\textit{умножение} + \textit{вычитание}) \right] =$$

$$\frac{1}{3} \cdot n(n-1)(2 \cdot n + \frac{13}{2}) = \frac{2}{3} \cdot n^3 + \frac{3}{2} \cdot n^2 - \frac{13}{6} \cdot n$$

Для обратного хода по формулам (6) число арифметических операций равно

$$Q_2(n) = \sum_{k=1}^{n} (деление + вычитание \sum_{j=k+1}^{n} умножение) =$$

$$\frac{1}{2} \cdot n(n+3) = \frac{1}{2} \cdot n^2 + \frac{3}{2} \cdot n$$

Общие вычислительные затраты метода Гаусса:

$$Q(n) = \frac{2}{3} \cdot n^3 + 2 \cdot n^2 - \frac{2}{3} \cdot n$$

Таким образом, $Q(n) \approx \frac{2}{3} \cdot n^3 = O(n^3)$

Выполнение прямого хода метода Гаусса позволяет также вычислить значение определителя матрицы системы, который будет равен произведению элементов, стоящих на главной диагонали приведенной к треугольному виду матрицы системы. (Т.к. при замене строк матрицы их линейными комбинациями значение определителя не меняется, а знак меняется при каждой перестановке строк, то знак произведения должен быть скорректирован с учетом выполнявшихся возможно перестановок строк.)

Таким образом, определитель вычисляется по формуле $\det A = \pm \prod_{k=1}^{n} a_{kk}^{(k)}$.

Метод Гаусса может быть использован и для нахождения обратной матрицы. Обозначим ее элементы через α_{jm}. Тогда соотношение $A \cdot A^{-1} = E$ можно записать так:

$$\sum_{k=1}^{n} a_{ik} \cdot \alpha_{kj} = \delta_{ij}, \ 1 \le i, j \le n.$$

Видно, что если рассматривать *j*-й столбец обратной матрицы как вектор, то он является решением линейной системы вида (1) с матрицей A и специальной правой частью (в которой на $j-м$ месте стоит единица, а на остальных нули).

Таким образом, для обращения матрицы надо решить n систем линейных уравнений с одинаковой матрицей A и разными правыми частями. Приведение матрицы A к треугольной делается при этом только один раз, а правые части преобразуются по формулам (3)-(4).

Преобразование матрицы требует порядка $\frac{2}{3}n^3$ операций. Действия по преобразованию правых частей систем и обратный ход метода Гаусса повторяются n раз, а однократное преобразование правых частей и обратный ход требуют порядка $\frac{3}{2}n^2$ опера-

ций. Следовательно, суммарные вычислительные затраты на обращение матрицы составляют: $\frac{2}{3}n^3 + \frac{3}{2}n^3 = \frac{13}{6}n^3$.

2.2.3. Векторные и матричные нормы

Для изучения итерационных методов приближенного решения систем линейных алгебраических уравнений используются различные варианты векторных и матричных норм.

Наиболее часто на практике используются следующие нормы векторов, как элементов n – мерного пространства:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \qquad \|x\|_\infty = \max_i |x_i|$$

Любая векторная норма позволяет определить подчиненную матричную норму:

$\|S\| = \sup_{x \neq 0} \frac{\|S\,x\|}{\|x\|}$ Для указанных выше векторных норм подчиненными нормами матрицы будут соответственно:

$\|S\|_1 = \max_j \sum_{i=1}^{n} |s_{ij}|$, $\|S\|_2 =$ максимальное сингулярное число матрицы S (если S нормальная матрица, т.е. перестановочна со своей сопряженной матрицей, то $\|S\|_2 = \max|\lambda_i|$, где λ_i собственные числа матрицы), $\|S\|_\infty = \max_i \sum_{j=1}^{n} |s_{ij}|$.

Матричная норма согласована с некоторой векторной нормой, если для всех векторов x $\|S\,x\| \le \|S\| \cdot \|x\|$

Подчиненная матричная норма является минимальной согласованной с соответствующей векторной нормой

2.3. Итерационные методы решения систем линейных алгебраических уравнений. Общая схема

Рассмотрим общее описание метода итераций для системы линейных алгебраических уравнений с квадратной невырожденной матрицей

$$Ax = b \qquad (1)$$

В этом случае по некоторому алгоритму, начиная с выбранного вектора $x^{(0)}$ строится последовательность векторов $x^{(k)}$.

При этом вектор $x^{(k+1)}$ выражается через известные предыдущие вектора $x^{(k)}, x^{(k-1)}, \ldots$ Если при вычислении $x^{(k+1)}$ используется только вектор $x^{(k)}$, то итерационный метод называется одношаговым (или двухслойным) методом.

Общий вид линейного одношагового метода

$$x^{(k+1)} = S_k \cdot x^{(k)} + g_k$$

(2)

Важную роль играет запись итерационных методов в единой (канонической) форме.

Она имеет вид:

$$B_k \cdot \frac{x^{(k+1)} - x^{(k)}}{\tau_{k+1}} + A \cdot x^{(k)} = b, k = 0,1,\ldots, \qquad (3)$$

где A - матрица исходной системы уравнений (1), B_k - некоторая последовательность невырожденных матриц, $\tau_1, \tau_2,\ldots, \tau_{k+1},\ldots$ - итерационные параметры, $\tau_{k+1} > 0$.

Связь между записью итерационного метода в виде (2) и в виде (3) выражается формулами:

$$S_k = E - \tau_{k+1} B_k^{-1} A,\ g_k = \tau_{k+1} B_k^{-1} b$$

Выбирая различным образом B_k и τ_k, можно получить разные варианты итерационных методов, которые различаются скоростью сходимости, сложностью реализации.

Если $B_k = E$ - единичная матрица, то метод называют явным: $x^{(k+1)}$ находится по явной формуле

$$x^{(k+1)} = x^{(k)} - \tau_{k+1} \cdot (A \cdot x^{(k)} - b).$$

В общем случае, при $B_k \neq E$, метод называют неявным итерационным методом: для определения $x^{(k+1)}$ надо решать систему уравнений

$$B_k \cdot x^{(k+1)} = B_k \cdot x^{(k)} - \tau_{k+1} \cdot (A \cdot x^{(k)} - b) = F_k, k = 0,1,\ldots \qquad (5)$$

Точность метода характеризуется величиной погрешности $z^{(k)} = x^{(k)} - x^*$, т.е. разностью между решением $x^{(k)}$ уравнения (5) и точным решением x^* исходной системы линейных алгебраических уравнений.

Говорят, что итерационный метод сходится, если $\left\| z^{(k)} \right\| \to 0$ при $k \to \infty$.

В случае, когда S_k и g_k (B_k и τ_k - соответственно) не зависят от номера итерации *k,* итерационный метод называется стационарным (иначе - нестационарным).

Критерий сходимости стационарного линейного одношагового итерационного метода

$$x^{(k+1)} = S \cdot x^{(k)} + g \qquad (6)$$

формулируется в теореме 1.

Теорема 1. Метод (6) сходится для любого начального приближения $x^{(0)}$ тогда и только тогда, когда все собственные числа матрицы перехода *S* по модулю меньше единицы.

Доказательство.

Необходимость. Пусть некоторое собственное число λ матрицы перехода по модулю больше единицы. Возьмем в качестве начального приближения вектор $x^{(0)} = x^* + s$, где s - собственный вектор, соответствующий собственному числу λ.

Тогда $z^{(0)} = s$,

$$z^{(k+1)} = S \cdot z^{(k)} = S(S \cdot z^{(k-1)}) = \ldots = S^k \cdot z^{(0)} = \lambda^{(k)} s$$

$\left\|z^{(k+1)}\right\| = |\lambda|^k \|s\| \to \infty$ при $k \to \infty$. Если $|\lambda| = 1$, то $\lim\limits_{k \to \infty} \left\|z^{(k+1)}\right\| = \|s\| \neq 0$

Достаточность. Пусть матрица S имеет n линейно независимых собственных векторов (т.е. является матрицей простой структуры): $s_1, \ldots, s_n$, соответствующих собственным числам $\lambda_1, \ldots, \lambda_n$, каждое из которых по модулю меньше единицы. Разложим погрешность начального приближения $z^{(0)} = x^{(0)} - x^*$ по базису из собственных векторов. $z^{(0)} = \sum\limits_{j=1}^{n} d_j s_j$ Тогда $z^{(k)} = S^k z^{(0)} = \sum\limits_{j=1}^{n} d_j \lambda_j^k s_j$

$\left\|z^{(k)}\right\| \le \sum\limits_{j=1}^{n} \left|d_j\right| \left|\lambda_j\right|^k \left\|s_j\right\| \le \rho^k \sum\limits_{j=1}^{n} \left|d_j\right| \left\|s_j\right\|$ где $\rho = \max\limits_j \left|\lambda_j\right|$ (спектральный радиус). Так как $\rho < 1$, то $\left\|z^{(k)}\right\| \to 0$ при $k \to \infty$, т.е. метод сходится.

Замечание. В общем случае, когда система собственных векторов матрицы S не является полной, доказательство достаточности проводится с использованием жордановой формы матрицы.

В качестве следствия из теоремы 1 можно получить легко проверяемые на практике достаточные условия. Так как для максимального по модулю собственного числа матрицы S и соответствующего собственного вектора y выполняется равенство $\|Sy\| = \rho \cdot \|y\|$, то для любой согласованной нормы матрицы $\|S\| \ge \rho$.

Таким образом, например, выполнение любого из неравенств $\max\limits_j \sum\limits_{i=1}^{n} \left|s_{ij}\right| < 1$, $\max\limits_i \sum\limits_{j=1}^{n} \left|s_{ij}\right| < 1$ достаточно для сходимости итерационного метода.

Достаточные условия сходимости для итерационного метода, записанного в канонической форме, содержатся в следующей теореме, которая приводится без доказательства.

Теорема 2. Пусть A - симметричная положительная матрица и выполнено условие

$$B > \frac{\tau}{2} \cdot A. \qquad (7)$$

Тогда метод итераций

$$B \cdot \frac{x^{(k+1)} - x^{(k)}}{\tau_{k+1}} + A \cdot x^{(k)} = b, k = 0,1,\ldots$$

сходится.

Напомним, что матрица положительная, если для любого ненулевого вектора x $(Ax,x)>0$.

Неравенство (7) означает, что для любого ненулевого вектора x матрица $B - \frac{\tau}{2}A$ положительна.

2.4. Варианты итерационных методов

2.4.1. Метод простых итераций

В качестве первого примера рассмотрим явный стационарный итерационный метод, каноническая форма которого:

$$\frac{x^{(k+1)} - x^{(k)}}{\tau} + Ax^{(k)} = b \qquad (1)$$

Выясним достаточные условия сходимости этого метода. В соответствии с теоремой 2 для этого достаточно, чтобы матрица системы A была симметричной и положительной и выполнялось неравенство $E > \frac{\tau}{2}A$

Учитывая, что $E \geq \frac{1}{\|A\|} \cdot A$, имеем $E - \frac{\tau}{2} \cdot A \geq (\frac{1}{\|A\|} - \frac{\tau}{2}) \cdot A > 0$.

Это неравенство выполнено при $\frac{1}{\|A\|} - \frac{\tau}{2} > 0$. Следовательно, метод простых итераций сходится при всех значениях τ, удовлетворяющих неравенству $\tau < \frac{2}{\|A\|}$.

С учетом неравенства $\|A\| \geq \max|\lambda_i|$, где λ_i- собственные числа матрицы A, достаточное условие сходимости можно записать в виде

$$\tau < \frac{2}{\max|\lambda_i|} \qquad (2)$$

Условие (2) является также и необходимым для сходимости метода простых итераций. Пусть λ_1 - максимальное по модулю собственное число, y_1 - соответствующий собственный вектор. При начальном приближении $x^{(0)} = x^* + y_1$ для погрешности k-го приближения имеем:

$$z^{(k)} = (E - \tau A)^k \cdot y_m = (1 - \tau\lambda_1)^k \cdot y_1.$$

Тогда

$$\left\|z^{(k)}\right\| = \left|1-\lambda_1\tau\right|^k \left\|y_1\right\|.$$

Если $\tau > \dfrac{2}{\lambda_1}$, то $\left|1-\lambda_1\tau\right|^k > 1$ и $\left\|z^{(k)}\right\| \to \infty$ при $k \to \infty$.

Если $\tau = \dfrac{2}{\lambda_1}$, то $\left\|z^{(k)}\right\| = \left\|y_1\right\|$ и не стремится к нулю при $k \to \infty$.

2.4.2. Метод Якоби

Координатная форма записи этого варианта итерационного метода имеет вид:

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \cdot \left(\sum_{j=1, j\neq i}^{n} a_{ij} \cdot x_j^{(k)} - b_i \right), i = 1, \ldots, n. \qquad (27)$$

Формулы (27) получаются непосредственно из исходной системы, если i - ое уравнение системы разрешить относительно неизвестного x_i.

Подставляя сюда

$$\sum_{j=1, j\neq i}^{n} a_{ij} \cdot x_j^{(k)} = \sum_{j=1}^{n} a_{ij} \cdot x_j^{(k)} - a_{ii} \cdot x_i^{(k)},$$

получаем

$$x_i^{(k+1)} = x_i^{(k)} - \frac{1}{a_{ii}} \cdot \left(\sum_{j=1}^{n} a_{ij} \cdot x_j^{(k)} - b_i \right),$$

или, в каноническом виде,

$$D \cdot \frac{x^{(k+1)} - x^{(k)}}{1} + A \cdot x^{(k)} = b, \quad k = 0,1,\ldots, \tau = 1,$$

где $D = (a_{ij}\delta_i^j)$ - диагональная матрица.

В соответствии с теоремой 2 сходимость этого метода гарантирована, если положительны матрица A и матрица $D - \dfrac{1}{2} \cdot A = \begin{pmatrix} \dfrac{a_{11}}{2} & \dfrac{-a_{12}}{2} & \cdots & \dfrac{-a_{1n}}{2} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{-a_{n1}}{2} & \dfrac{-a_{n2}}{2} & \cdots & \dfrac{-a_{nn}}{2} \end{pmatrix}$

Из положительности матрицы A вытекает, что ее диагональные элементы, а значит и диагональные элементы матрицы

$D-\frac{1}{2}\cdot A$ больше нуля. При этом условии для положительности матрицы достаточно, чтобы она имела свойства диагонального преобладания $\frac{a_{ii}}{2} > \sum_{j\neq i}\frac{|a_{ij}|}{2}, i=1,2,\ldots,n.$

Последнее равносильно тому, чтобы этим свойством обладала сама матрица A. Свойство диагонального преобладания матрицы A как достаточное условие сходимости метода Якоби возникает и в качестве следствия из теоремы 1.

2.4.3. Метод Зейделя

Весьма широко на практике применяется итерационный метод Зейделя:

$$\sum_{j=1}^{i} a_{ij}\cdot x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}\cdot x_j^{(k)} = f_i, a_{ii}\neq 0, i=1,\ldots,n.$$

Компоненты $x^{(k+1)}$ находятся последовательно по формулам:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\cdot(b_i - \sum_{j=i+1}^{n} a_{ij}\cdot x_j^{(k)} - \sum_{j=1}^{i-1} a_{ij}\cdot x_j^{(k+1)}), i=2,...,n.$$

Запишем этот метод в матричной форме. Для этого представим матрицу A в виде суммы

$$A = A^{-} + D + A^{+},$$

где

$A^{-}=(a_{ij}^{-}),\ a_{ij}^{-}=\begin{cases} a_{ij}, i>j \\ 0, i\le j \end{cases}$ - нижняя треугольная матрица,

$A^{+}=(a_{ij}^{+}),\ a_{ij}^{+}=\begin{cases} a_{ij}, i<j \\ 0, i\ge j \end{cases}$ - верхняя треугольная матрица.

В этих обозначениях метод Зейделя записывается следующим образом:

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

Применим теорему 2 для исследования сходимости метода Зейделя.

$B = D + A^{-}, \tau = 1$. В этом случае

$$B-\frac{1}{2}\cdot A = D + A^{-} - \frac{1}{2}\cdot(A^{-}+A^{+}+D) = \frac{D}{2} + \frac{1}{2}\cdot(A^{-}-A^{+}),$$

$$((B-\frac{1}{2}\cdot A)y, y) = \frac{1}{2}\cdot(Dy, y) + \frac{1}{2}\cdot((A^{+}-A^{-})y, y) = \frac{1}{2}\cdot(Dy, y) > 0,$$

если $D > 0$. Следовательно, метод Зейделя сходится, если $D > 0$.

Неравенство $D > 0$ следует из условия $A > 0$.

Таким образом, метод Зейделя всегда сходится, если A - положительная матрица.

2.4.4. Метод релаксации

Можно ускорить сходимость метода Зейделя, если в схему (28) ввести итерационный параметр (параметр релаксации) ω.Получим

$$(D + \omega \cdot A^{-}) \cdot \frac{x^{(k+1)} - x^{(k)}}{\omega} + A \cdot x^{(k)} = b\,.$$

Значение $\omega = 1$ соответствует методу Зейделя.

Применим теорему 2 для исследования сходимости метода релаксации.

$B = D + \omega \cdot A^{-}$, $\tau = \omega$. Найдем разность

$$B - \frac{\tau}{2} \cdot A = D + \omega \cdot A^{-} - \frac{\omega}{2} \cdot (A^{-} + A^{+} + D) = (1 - \frac{\omega}{2}) \cdot D + \frac{\omega}{2} \cdot (A^{-} - A^{+}),$$

$$((B - \frac{\tau}{2} \cdot A)y, y) = (1 - \frac{\omega}{2}) \cdot (Dy, y) > 0 \text{ при } 0 < \omega < 2\,.$$

Таким образом, метод релаксации сходится при любых значениях $\omega \in (0,2)$, если A - положительная матрица.

2.5. Вариационно-итерационные методы

К этой группе относятся итерационные методы приближенного решения систем линейных алгебраических уравнений, в которых приближенное решение на каждом шаге определяется на основе решения некоторой экстремальной задачи.

2.5.1. Метод минимальных невязок

Метод минимальных невязок – явный нестационарный метод, каноническая форма которого имеет вид:

$$\frac{x^{(k+1)} - x^{(k)}}{\tau_{k+1}} + A \cdot x^{(k)} = b \qquad (1)$$

Значение параметра τ_{k+1} определяется из условия получения минимальной по норме невязки $r^{(k+1)} = A \cdot x^{(k+1)} - b$ на k+1 – ом шаге итерационного процесса.

Из (1) следует, что невязки удовлетворяют соотношению $\frac{r^{(k+1)} - r^{(k)}}{\tau_{k+1}} + A \cdot r^{(k)} = 0$,

откуда получаем:

$$\|r^{k+1}\|^2 = \|r^k - \tau_{k+1} A r^k\|^2 = \|r^k\|^2 - 2\tau_{k+1}(r^k, A r^k) + \tau_{k+1}\|A r^k\|^2 = \varphi(\tau_{k+1})$$

Дифференцируя функцию $\varphi(\tau_{k+1})$ и приравнивая нулю производную, получаем:

$$\tau_{k+1} = \frac{(r^k, A r^k)}{(A r^k, A r^k)} \qquad (2)$$

Если $A = A^* > 0$, то для невязок приближенных решений, получаемых по формулам (1),(2), справедливо неравенство:

$\|r^{k+1}\| \le \rho \|r^k\| \le \ldots \le \rho^k \|A x^0 - b\|$, где $\rho = \frac{\mu - \nu}{\mu + \nu}$ (μ ,ν - границы промежутка, содержащего собственные числа матрицы A). Из этого неравенства следует линейная скорость сходимости метода минимальных невязок.

2.5.2. Связь с задачей о минимуме квадратичной формы

Теорема. Решение системы $Ax = b$, где $A = A^* > 0$ является точкой минимума квадратичной формы $F(x) = (Ax, x) - 2(b, x)$ и наоборот.

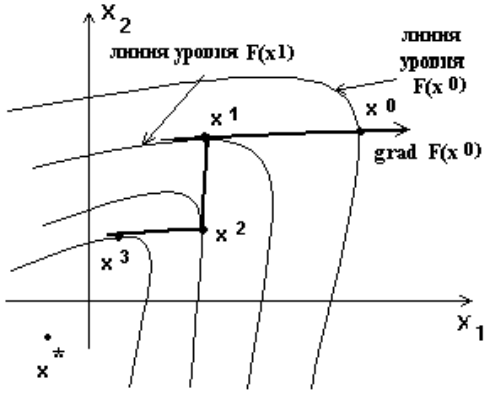
Доказательство Пусть x^* - решение системы уравнений. Произвольный вектор *x* можно представить в виде: $x = x^* + \lambda u$, где λ - скалярный параметр, а u - вектор. Тогда $F(x) = f(x^* + \lambda u) = F(x^*) + \lambda^2 (Au, u) + 2\lambda (Ax^* - b, u) = F(x^*) + \lambda^2 (Au, u) \ge F(x^*)$

Пусть x^* точка минимума квадратичной формы. Для произвольного фиксированного вектора *u* $F(x^* + \lambda u) = \varphi(\lambda)$ - функция скалярного аргумента λ , которая имеет минимум при $\lambda = 0$. Тогда $\varphi'|_{\lambda = 0} = 0$. Отсюда получаем $(Ax^* - b, u) = 0$, а это в силу произвольности вектора *u* означает, что $Ax^* = b$.

На основании доказанной теоремы для получения решения системы линейных алгебраических уравнений с симметричной положительной матрицей можно использовать любой из методов поиска минимума функции *n* переменных, частным случаем которой является данная квадратичная форма.

2.5.3. Метод градиентного спуска

Рассмотрим один из методов поиска точки минимума дифференцируемой функции *n* переменных *F(x)* .



Пусть x^0 точка в некоторой окрестности точки минимума x^* функции *F(x),* которую мы принимаем за начальное приближение. .Так как производная в точке x^0 по направлению, противоположному вектору $grad\ F(x^0)$, равна $-grad\ F(x^0)$, то в этом направлении функция убывает (и скорость убывания функции в этом

направлении наибольшая). Среди точек этого направления $x = x^0 - t\ gradF(x^0)$ найдем ту, в которой значение *F(x)* будет минимально, т.е. найдем точку минимума функции скалярного аргумента $\varphi(t) = F(x^0 - t \cdot gradF(x^0))$. Необходимое условие экстремума $\varphi'(t) = 0$ позволяет определить это значение t_1, которому соответствует следующее приближение к точке минимума функции *F(x)* $x^1 = x^0 - t_1\ gradF(x^0)$ и т.д. Геометрическая иллюстрация этого процесса представлена на рис. 5, причем точка x^k является точкой касания прямой $x = x^{k-1} - t\ gradF(x^{k-1})$ некоторой линии уровня функции *F(x)*.

Рис. 5

Для квадратичной формы $F(x) = (A\,x, x) - 2(b, x)$, о которой идет речь в предыдущей теореме, $gradF(x) = 2(Ax - b) = 2r$ и функция $\varphi(x)$ будет иметь вид: $\varphi(x) = 4t^2 \cdot (Ar^k, r^k) - 4t \cdot (r^k, r^k) - 2 \cdot (x^k, b)$. Вычисляя производную $\varphi'(t)$ и приравнивая ее к нулю, получаем: $t_{k+1} = \dfrac{(r^k, r^k)}{2 \cdot (Ar^k, r^k)}$. Затем находим $x^{k+1} = x^k - 2t_{k+1} \cdot r^k$.

Сравнивая алгоритмы метода градиентного спуска и метода минимальных невязок, можно отметить, что в обоих случаях очередное приближенное решение определяется по формуле $x^{k+1} = x^k - \alpha \cdot r_k$, но значения коэффициента α, а значит и сами получаемые этими двумя методами приближенные решения будут различны.

2.5. Оценка погрешности и мера обусловленности

Предположим, что матрица имеющей единственное решение системы линейных уравнений и вектор правых частей заданы неточно и вместо предъявленной к решению системы

$$A \cdot x = b \tag{30}$$

в действительности решается некоторая система

$$A_1 \cdot x = b_1, \text{ где } A_1 = A + \Delta A,\ b_1 = b + \Delta b$$

Обозначим решения (30) и (31) через x и x_1

Оценим погрешность решения $z = x_1 - x$.

Подставим выражения A_1, b_1 и x_1 в (31)

$$(A + \Delta A) \cdot (x + z) = b + \Delta b$$

Вычитая (30), получим

$$A \cdot z + \Delta A \cdot x + \Delta A \cdot z = \Delta b$$

$$z = A^{-1} \cdot (\Delta b - \Delta A \cdot x - \Delta A \cdot z)$$

Отсюда:

$$\|z\| \le \|A^{-1}\| \cdot (\|\Delta b\| + \|\Delta A\| \cdot \|x\| + \|\Delta A\| \cdot \|z\|) \tag{32}$$

Если $\|A^{-1}\| \cdot \|\Delta A\| < 1$, то из (32) следует оценка погрешности

$$\|z\| \le \frac{\|A^{-1}\| \cdot (\|\Delta b\| + \|\Delta A\| \cdot \|x\|)}{1 - \|A^{-1}\| \cdot \|\Delta A\|}.$$

Если $\|\Delta A\| \to 0$ и $\|\Delta b\| \to 0$, то $\|z\| \to 0$. Но малые погрешности исходных данных СЛАУ (коэффициентов матрицы и правых частей) не всегда гарантируют такую же малость погрешности решения. Чтобы выяснить этот вопрос, получим оценку относительной погрешности решения системы через относительные погрешности ее исходных данных. Обозначим $\delta x = \frac{\|\Delta x\|}{\|x\|}$, $\delta b = \frac{\|\Delta b\|}{\|b\|}$, $\delta A = \frac{\|\Delta A\|}{\|A\|}$. Из (32) следует:

$$\delta x \le \|A^{-1}\| \cdot (\frac{\|\Delta b\|}{\|x\|} + \|\Delta A\| + \|\Delta A\| \cdot \delta x) = \|A^{-1}\| \cdot (\delta b \cdot \frac{\|b\|}{\|x\|} + \delta A \cdot \|A\| + \delta A \cdot \|A\| \cdot \delta x)$$

Т.к. из $A \cdot x = b$ следует $\|b\| \le \|A\| \cdot \|x\|$, то далее получаем:

$$\delta x \le \|A^{-1}\| \cdot \|A\| \cdot (\delta b + \delta A \cdot + \delta A \cdot \delta x)$$

Величину $\|A^{-1}\| \cdot \|A\|$ называют мерой или числом обусловленности матрицы и обозначают $cond(A)$. Число обусловленности зависит от выбора нормы матрицы, но т.к. $A^{-1} \cdot A = E$, то для любой нормы и для любой матрицы $cond(A) \ge 1$.

Если $cond(A) \cdot \|A^{-1}\| < 1$, то из (34) для относительной погрешности решения можно получить оценку, аналогичную (33):

$$\delta x \le \frac{cond(A) \cdot (\delta b + \delta A \cdot)}{1 - cond(A) \cdot \delta A} \qquad (35)$$

Важным частным случаем решения систем с неточными исходными данными является ситуация, когда погрешности элементов матрицы существенно меньше погрешности правой части и их влиянием можно пренебречь. Тогда для абсолютной погрешности решения системы получается оценка: $\|z\| \le \|A^{-1}\| \cdot \|\Delta b\|$, а для относительной погрешности оценка: $\delta x \le cond(A) \cdot \delta b$

Величина относительной погрешности решения при фиксированной величине относительной погрешности правой части может стать сколь угодно большой при достаточно большом числе обусловленности матрицы системы.

Системы уравнений и матрицы с большими значениями мер обусловленности принято называть плохо обусловленными, а с малыми - хорошо обусловленными.

Пример.

2.6. Алгебраическая проблема собственных значений

Большое число научно-технических, экономических и других задач, а также потребности самой вычислительной математики приводят к вопросу нахождения собственных чисел и собственных векторов матриц, т. е. отыскание таких значений λ, для которых существуют нетривиальные решения системы алгебраических уравнений

$$A \cdot x = \lambda \cdot x \qquad (37)$$

Различают две постановки задачи и самих этих решений:

- нахождение одного или нескольких собственных чисел (и соответственно - векторов) - частичная проблема собственных значений;

- нахождение всех собственных чисел (и соответственно - векторов) - полная проблема.

Среди большого числа алгоритмов, предназначенных для решения этих задач, нет такого, который можно было бы рассматривать как универсальный и эффективный во всех случаях. Все существующие методы делятся на две группы: прямые и итерационные.

В прямых методах для нахождения собственных чисел используется характеристический многочлен матрицы $\det(A-\lambda\cdot E)$, корнями которого собственные числа и являются. Здесь первый важный этап – это нахождение коэффициентов характеристического многочлена, так как их вычисление по определению требует большого числа арифметических операций, связанных с раскрытием определителя. Существует ряд достаточно эффективных методов (Крылова, Данилевского и другие), в которых нахождение коэффициентов характеристического многочлена связано с некоторыми преобразованиями исходной матрицы (например, приведение ее к характеристической форме Фробениуса в методе Данилевского). Второй этап прямых методов – нахождение каким-либо способом корней характеристического многочлена.

В итерационных методах, начиная с исходной матрицы строится ряд матриц, собственные числа которых равны, такой что собственные числа (точнее их приближенные значения) последней в этом ряду матрицы определяются легко. Например, полученная матрица имеет диагональный или другой специальный вид. Еще одним вариантом итерационных методов для решения частичной проблемы собственных значений является степенной метод.

2.6.1. Степенной метод

Степенной метод позволяет найти наибольшее по модулю собственное значение и собственный вектор.

Пусть $\lambda_1,\lambda_2,\dots\lambda_n$ - собственные числа матрицы А. Для определенности предположим, что

$$|\lambda_1|>|\lambda_2|\geq\dots\geq|\lambda_n|$$

Берем произвольный ненулевой вектор $y^{(0)}$. Строим последовательность векторов

$$y^{(1)}=A\cdot y^{(0)},\ y^{(2)}=A\cdot y^{(1)}=A^2\cdot y^{(0)},\dots,\ y^{(m)}=A\cdot y^{(m-1)}=A^m\cdot y^{(0)}$$

Тогда

$$\lambda_1\approx\frac{y_i^{m+1}}{y_i^m} \qquad (38)$$

для любого номера *i=1,2,...,n.*

Точнее

$$\lambda_1 = \frac{y_i^{m+1}}{y_i^m} + O((\frac{\lambda_2}{\lambda_1})^m) \qquad (*)$$

Докажем это в предположении, что матрица *A* имеет *n* линейных независимых собственных векторов $x_1,\ldots,x_n$. Запишем разложение вектора $y^{(0)}$ по базису из собственных векторов

$$y^{(0)} = \sum_{j=1}^{n} c_j \cdot x_j .$$

Тогда

$$y^{(1)} = A \cdot y^{(0)} = \sum_{j=1}^{n} c_j \cdot A \cdot x_j = \sum_{j=1}^{n} c_j \cdot \lambda_j \cdot x_j$$

$$y^{(m)} = A \cdot y^{(m-1)} = \sum_{j=1}^{n} c_j \lambda_j^m \cdot x_j$$

$$\frac{y_i^{(m+1)}}{y_i^{(m)}} = \frac{c_1 \cdot \lambda_1^{m+1} \cdot x_{1i} + c_2 \cdot \lambda_2^{m+1} \cdot x_{2i} + \ldots}{c_1 \cdot \lambda_1^m \cdot x_{1i} + c_2 \cdot \lambda_2^m \cdot x_{2i} + \ldots} =$$

$$= \lambda_1 \frac{c_1 \cdot x_{1i} + c_2 \cdot \left(\frac{\lambda_2}{\lambda_1}\right)^{m+1} \cdot x_{2i} + \ldots}{c_1 \cdot x_{1i} + c_2 \cdot \left(\frac{\lambda_2}{\lambda_1}\right)^{m} \cdot x_{2i} + \ldots} =$$

$$= \lambda_1 \left(1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m \frac{c_2 \cdot x_{2i} \cdot \left(\frac{\lambda_2}{\lambda_1} - 1\right) + c_3 \cdot x_{3i} \cdot \left(\frac{\lambda_3}{\lambda_1} - 1\right) + \left(\frac{\lambda_3}{\lambda_2}\right)^m + \ldots}{c_1 \cdot x_{1i} + c_2 \cdot \left(\frac{\lambda_2}{\lambda_1}\right)^m \cdot x_{2i} + \ldots} \right)$$

Так как $\left|\frac{\lambda_k}{\lambda_1}\right| < 1$ для *k=2,...,n* и $\left|\frac{\lambda_k}{\lambda_2}\right| \le 1$ для *k=3,...,n* , то отсюда следует, что при $m \to \infty$ выполняется соотношение (*)

Взяв достаточно большой номер итерации *m*, мы сможем с любой степенью точности определить по формуле (38) наибольшее по модулю собственное число λ_1 матрицы А. Для этого может быть использована любая координата $y_i^{(m)}$ вектора $y^{(m)}$, или их среднее арифметическое значение.

Так как $\frac{y^{(m)}}{c_1 \cdot \lambda_1^m} = x_1 + \frac{c_2}{c_1} \cdot \left(\frac{\lambda_2}{\lambda_1}\right)^m \cdot x_2 + ... + \frac{c_n}{c_1} \cdot \left(\frac{\lambda_n}{\lambda_1}\right)^m \cdot x_n$, то при $m \to \infty$

$$\frac{y^{(m)}}{c_1 \cdot \lambda_1^m} = x_1 + \mathrm{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^m\right).$$

Поскольку собственный вектор определяется с точностью до скалярного множителя, то сам вектор $y^{(m)}$ приближенно представляет собой собственный вектор матрицы A, соответствующий собственному значению λ_1. При этом $\lim_{m\to\infty}\|y^{(m)}\| = \lim_{m\to\infty} c_1\lambda_1\|x_1\|$ Если $|\lambda_1| > 1$, то $\|y^{(m)}\|$ неограниченно растет, если $|\lambda_1| < 1$, то $\|y^{(m)}\|$ стремится к нулю с ростом номера итерации. Для устранения этого может применяться нормирование вектора $y^{(m)}$. $\bar{y}^{(m)} = \frac{y^{(m)}}{\|y^{(m)}\|}$. Тогда $\tilde{y}^{(m+1)} = A\bar{y}^{(m)} = \frac{y^{(m+1)}}{\|y^{(m)}\|}$, и $\frac{\tilde{y}_i^{(m+1)}}{\bar{y}_i^{(m)}} = \frac{y_i^{(m+1)}}{y_i^{(m)}}$, то есть, будет получен тот же результат.

2.6.2. Метод вращений

Метод вращения позволяет для симметрических матриц решать полную проблему собственных значений.

Этот метод основан на следующих фактах из теории матриц:

Лемма 1. . Известно, что для симметрической матрицы A существует ортогональная матрица U такая, что $U'AU = \Lambda$ где U' - транспонированная к U матрица, а Λ - диагональная матрица. Так как $U' = U^{-1}$, то матрица Λ подобна матрице A

Лемма 2. если матрицы A и B подобны (т.е. существует матрица C такая, что $B = C^{-1} A C$), то их собственные числа одинаковы, а собственные вектора связаны соотношением: $y = C x$.

Действительно, пусть λ - собственное число, а x - соответствующий собственный вектор матрицы B. Тогда $(C^{-1} A C) x = B x = \lambda x$, отсюда $(A C) x = C\lambda x$ или $A(C x) = \lambda (C x)$

Лемма 3. Собственные пары диагональной матрицы Λ имеют вид: (λ_i, e^i), где λ_i диагональный элемент матрицы, а e^i - единичный вектор, *i*-ая компонента которого равна 1.

Таким образом, решение полной проблемы собственных значений сводится для симметрической матрицы A к нахождению ортогональной матрицы U, с помощью которой матрица A приводится к диагональному виду.

Суть метода вращений состоит в следующем.

Пусть $\overline{U}'A\overline{U} = \overline{\Lambda}$, где $\overline{\Lambda}$ мало отличается от диагональной матрицы, т. е. элементы вне главной диагонали малы. Тогда можно ожидать, что собственные числа матри-

цы $\overline{\Lambda}$ будут близки к диагональным элементам $\overline{\lambda_i}$ матрицы $\overline{\Lambda}$, и $\overline{\lambda_i}$ можно принять за приближенные значения λ_i.

В методе вращения матрица U строится как предел последовательности произведений матриц простых поворотов, при которых все оси координат кроме двух остаются неподвижными. При этом матрицы простых поворотов подбираются так, чтобы при преобразовании матрицы с помощью матрицы простого поворота на каждом шаге уничтожался максимальный по модулю недиагональный элемент.

Итерационный процесс осуществляется следующим образом.

Пусть A^k - матрица, полученная после k - го преобразования поворота.

В матрице A^k находится максимальный по модулю элемент $a_{ij}^k (i < j)$. Строится ортогональная матрица простого поворота вида

$$U^k(\varphi_k) = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\varphi_k & & & -\sin\varphi_k & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin\varphi_k & & & \cos\varphi_k & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}$$

Угол φ_k подбирается так, чтобы у матрицы

$$A^{k+1} = (U^k)' A U^k$$

элемент a_{ij}^k обратился бы в нуль. Найдем выражение для этого элемента.

Обозначим $B^k = A^k U^k$.

Матрица B^k отличается от матрицы A^k только столбцами с номерами i и j, причем последние имеют такой вид:

$$b_{li}^k = a_{li}^k \cdot \cos\varphi_k + a_{lj}^k \cdot \sin\varphi_k \qquad (39)$$

$$b_{lj}^k = -a_{li}^k \cdot \sin\varphi_k + a_{lj}^k \cdot \cos\varphi_k \qquad (40)$$

Матрица $A^{k+1} = (U^k)' B^k$ отличается от матрицы B^k только строками с номерами i и j, причем эти строки имеют такой вид:

$$a_{il}^{k+1} = b_{il}^{k} \cdot \cos\varphi_k + b_{jl}^{k} \cdot \sin\varphi_k \qquad (41)$$

$$a_{jl}^{k+1} = -b_{il}^{k} \cdot \sin\varphi_k + b_{jl}^{k} \cdot \cos\varphi_k \qquad (42)$$

Таким образом, из формул (41) и (40) получаем:

$$a_{ij}^{k+1} = (-a_{ii}^{k} \cdot \sin\varphi_k + a_{ij}^{k} \cdot \cos\varphi_k) \cdot \cos\varphi_k + (-a_{ji}^{k} \cdot \sin\varphi_k + a_{jj}^{k} \cdot \cos\varphi_k) \cdot \sin\varphi_k =$$

$$= \frac{1}{2} \cdot (a_{jj} - a_{ii}) \cdot \sin(2\varphi_k) + a_{ij} \cdot \cos(2\varphi_k)$$

Из требования, что $a_{ij}^{k+1} = 0$, получаем

$$tg2\varphi_k = \frac{2a_{ij}^{k}}{a_{ii}^{k} - a_{jj}^{k}}, \qquad \text{т. е. } \varphi_k = \frac{1}{2} \cdot arctg \frac{2a_{ij}^{k}}{a_{ii}^{k} - a_{jj}^{k}}$$

Процесс заканчивается, когда все внедиагональные элементы полученной на очередном шаге матрицы A^{k+1} будут достаточно малы. Диагональные элементы этой матрицы являются приближениями для собственных чисел матрицы *A*, а столбцы матрицы $U = U^1 \cdot U^2 \cdot \ldots \cdot U^k$ приближениями для соответствующих собственных векторов.

Сходимость метода вращений.

В качестве меры отличия матрицы от диагональной используем сумму квадратов ее внедиагональных элементов: $t(A) = \sum_{i,j,i \neq j} a_{ij}^2$. Можно показать, что

$$t(A^{k+1}) = t(A^k) - 2 \cdot (a_{ij}^{k})^2 \qquad (43)$$

Из (41) и (42) следует, что $(a_{il}^{k+1})^2 + (a_{jl}^{k+1})^2 = (b_{il}^{k})^2 + (b_{jl}^{k})^2$ для всех $l \neq i, j$.

Аналогично, из (39) и (40) следует, что $(b_{li}^{k})^2 + (b_{lj}^{k})^2 = (a_{li}^{k})^2 + (a_{lj}^{k})^2$ для всех $l \neq i, j$. На основе этих равенств получаем:

Добавляя к первому выражению в этой цепочке равенств равные нулю элементы a_{ij}^{k+1} и a_{ji}^{k+1}, получаем слева $t(A^{k+1})$, а прибавляя и вычитая справа a_{ij}^{k} и a_{ji}^{k}, получаем правую часть формулы (43). Т.к. a_{ij}^{k} максимальный по модулю элемент матрицы A^k ,то $t(A^k) \le n(n-1)(a_{ij}^{k})^2$. Отсюда $(a_{ij}^{k})^2 \ge \frac{t(A^k)}{n(n-1)}$ и, следовательно,

$t(A^{k+1}) \le (1 - \frac{2}{n(n-1)}) \cdot t(A^k)$ Таким образом, метод вращений сходится в том смысле, что $t(A^k) \to 0$ при $k \to \infty$.

На основании леммы 2 приближениями для собственных векторов матрицы *A* будут столбцы матрицы, образованной последовательным перемножением всех использованных матриц поворота.

Метод вращения является одним из самых удобных итерационных методов для определения собственных значений и собственных векторов симметрических матриц. Он прост по вычислительной схеме, быстро сходится, кратные и близкие собственные значения не вызывают никаких трудностей.

2.7. Итерационные методы решения систем нелинейных уравнений.

Рассматриваются методы приближенного решения системы уравнений вида:

$$\left.\begin{array}{l} f_1(x_1,...,x_n)=0 \\ \quad\cdots\cdots \\ f_n(x_1,...,x_n)=0 \end{array}\right\} \text{ или в векторной форме } F(x)=0 \qquad (1)$$

Общий вид одношаговых итерационных методов решения системы (1) в канонической форме:

$$B_{k+1}\cdot\frac{x^{k+1}-x^k}{\tau_{k+1}}+F(x_k)=0 \qquad (2)$$

где B_{k+1} невырожденная квадратная матрица порядка n, τ_{k+1} скалярный параметр. Применение методов вида (2) требует решения на каждом шаге итерационного процесса линейной системы уравнений вида $B_{k+1}\cdot x^{k+1}=G(x^k)$. Если B_{k+1} и τ_{k+1} не зависят от номера итерации k, то метод называется стационарным, иначе - нестационарным.

2.7.1. Метод Ньютона

2.7.2. Нелинейные методы Якоби и Зейделя

Литература

1. Березин И.С., Жидков Н.П., Методы вычислений т.1, М, Наука, 1966, т.2, М, Наука, 1962.
2. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. М, Наука, 1987. – 600 с.
3. Калиткин Н.Н. Численные методы. М, Наука, 1978. – 512 с.
4. Крылов В.И., Бобков В.В., Монастырный П.. Вычислительные методы. Т. 1, М, Наука, 1976. – 432 с., Т. 2, М, Наука, 1977. – 400 с.
5. Самарский А.А. Введение в численные методы. М, Наука, 1987. – 235 с.
6. Самарский А.А., Гулин А.В. Численные методы. М, Наука, 1989. – 432 с.
7. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. М, Мир, 1998. – 575 с.
8. Марчук Г.И. Методы вычислительной математики. М., Наука, 1980. – 536 с.
9. Турчак Л.И. Основы численных методов М., Наука, 1987. – 318 с.
10. Вержбицкий В.М. Численные методы. Линейная алгебра и нелинейные уравнения. М., Высшая школа, 2000. – 266 с.
11. Косарев В.И. 12 лекций по вычислительной математике. М.: Издательство МФТИ Физматкнига, 2000. – 224 с.

Остатки

Подстановка $x^{(k)} = z^{(k)} + x^*$ в (19) приводит к системе уравнений для погрешности:

$$B_k \cdot \frac{z^{(k+1)} - z^{(k)}}{\tau_{k+1}} + A \cdot z^{(k)} = 0, k = 0,1,\ldots \quad (22)$$

$\left\|z^{(k)}\right\|_A = \left\|x^{(k)} - x\right\|_A$ не стремится к нулю при $k \to \infty$.