Российская академия наук
Федеральное государственное бюджетное учреждение науки
Институт электрофизики
Уральского отделения Российской академии наук
(ИЭФ УрО РАН)

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ
«КОМПЬЮТЕРНЫЕ МЕТОДЫ В ФИЗИКЕ»

Б1.В.ОД.3

Специальность 03.06.01 – «Физика и астрономия»

Численные методы в физике

Материалы разработали:

к.ф.-м.н., доцент Болтачев Г.Ш.

ЧИСЛЕННЫЕ МЕТОДЫ В ФИЗИКЕ

Введение

Математические модели явлений, которые изучаются в физике (а если быть более конкретным, в молекулярной физике, в физике сплошной среды, в теплофизике), очень часто представляют собой сложные системы нелинейных дифференциальных уравнений (ДУ) в ЧП. Математическая сложность обусловлена не только нелинейностью уравнений, но и непростыми граничными условиями (ГУ), зависимостью теплофизических свойств от температуры и т.п. Аналитически такие задачи, как правило, не решаются. В связи с этим основная тема спецкурса: Численные методы решения уравнений и систем уравнений в частных производных (ЧП). Но решение нелинейных ДУ в ЧП — это высший пилотаж численных методов. А чтобы исполнять высший пилотаж, для начала, хорошо бы научиться заводить самолет и взлетать. Поэтому, чтобы подойти к задаче численного решения систем ДУ и плодотворно о ней побеседовать, мы обязаны вкратце остановиться на многих других, возможно, более простых задачах и проблемах численных методов.

Задачи, которые наиболее часто решаются численными методами:

1. Решение алгебраических уравнений (систем). Здесь же работа с матрицами: определитель, ранг, собственные числа и вектора.
2. Интерполяция и оптимизация (МНК)
3. Интегрирование. Здесь же преобразования Фурье (прямое и обратное).
4. Решение обыкновенных ДУ (систем)
5. Решение ДУ в ЧП (систем)
6. Описание системы многих частиц (МД и МК методы)

Чем обусловлена необходимость использования численных методов? Дело в том, что аналитически решить даже алгебраическое уравнение можно далеко не всегда. Еще хуже дело обстоит с интегралами, и еще хуже с ДУ. Откуда берутся такие нерешаемые алгебраические уравнения, интегралы и ДУ? Перед физикой всегда ставится задача описать некоторое явление, которое мы наблюдаем в жизни. А жизнь, как вы уже должно быть догадываетесь, сложная штука. Но(!) физика никогда не описывает нашу жизнь и явления в ней происходящие. Вместо этого явление подменяется моделью, математической моделью, которая отражает основные, главные черты изучаемого явления. В результате такой, по сути, подмены, а по-научному говоря, моделирования появляется алгебраическое уравнение, или интеграл, или ДУ, или система каких-то уравнений. Конечно, любая модель физического явления строится с помощью упрощения, идеализации реального явления. И все начинается, как правило, с очень простых моделей, которые можно решить аналитически. Однако, для более полного описания модель постепенно усложняется, и очень быстро выходит на уровень, когда проанализировать такую модель аналитическими методами уже не удается. Более того, на сегодня нет даже надежды на то, что в обозримом будущем аналитические методы будут развиваться в соответствии с растущими потребностями физики. Именно в этой связи резко возросла роль численных методов анализа модели и такое положение, образованное, в конечном счете, потребностями нашей жизни, привело к появлению вычислительного эксперимента (ВЭ).

ВЭ — это не эксперимент в классическом смысле этого слова, а технология анализа модели и ее совершенствования. Что включает в себя ВЭ. Я бы выделил в ВЭ следующие этапы.

1. Построение модели.

2. Разработка алгоритма.
   Как правило, для решения одной и той же математической задачи можно предложить множество алгоритмов, неодинаковых по своим свойствам. Свойства алгоритма: достоверность (точность) и скорость (время счета). Физик не должен заниматься разработкой алгоритмов. Физик должен уметь выбирать, для физика, этот этап должен называться — выбор алгоритма. Поэтому вы должны иметь представление о свойствах различных алгоритмов.
3. Создание программы.
   Вот программу вы выбрать не сможете, ее придется написать самостоятельно.
4. Расчеты.
   Этот этап имеет максимальное сходство с реальным экспериментом. Экспериментатор задает вопросы математической модели. Точность и достоверность "ответов" зависят от качества модели, алгоритма и программы. Поэтому сразу сложные расчеты не делают. Обязательный этап — тестирование программы, т.е. расчет каких-то упрощенных ситуаций (возможно имеющих аналитическое решение, или имеющих надежные экспериментальные данные). Здесь полная аналогия с предварительными измерениями в реальном эксперименте.
   Цель тестирования: отыскать и исправить ошибки модели (надежные экспериментальные данные), алгоритма (аналитическое решение) и программы.
5. Обработка результатов.
   Этот этап смыкается с четвертым. Выводы могут быть такими:
   • Есть необходимость уточнения модели или алгоритма
   • Возможно упрощение модели или алгоритма
   • Результаты идут в "дело"

В ходе нашего спецкурса вам не придется проводить ВЭ-ов (этот курс пока, к сожалению, — чисто теоретический), но иметь понять о ВЭ вы должны. Мы же с вами познакомимся с различными алгоритмами (или методами), применяемыми для решения различных задач. Методы решения любых уравнений бывают двух типов (алгебраические уравнения, ДУ, системы линейных уравнений):

- прямые (точные, конечные),
- итерационные (приближенные, бесконечные).

Прямые — решение находится за заранее известное число арифметических действий, решение строгое. Примеры: корни квадратного уравнения, кубического и четвертой степени.

Историческая справка. Алгебраические уравнения третьей и четвертой степени не поддавались усилиям математиков около 2000 лет. Эту задачу решили итальянские математики эпохи Ренессанса: решение кубического уравнения, опубликованное Кардано в 1545г., связывают с именами Спицион дель Ферро (1456–1526), Никколо Тарталья (1500–1557), Джироламо Кардано (1501–1576); решение уравнений 4-ой степени найдено Людовико Феррари (1522–1565). Для уравнений 5-й и более высоких степеней аналогичных формул не существует. Этот факт известен как теорема Абеля (Нильс Хенрик Абель, норвежский математик, 1802–1829), доказанная им в возрасте около 22 лет (1824).

Итерационные — (методы последовательных приближений) это такие методы, в которых нельзя заранее предсказать число арифметических действий, которое потребуется для решения уравнения (системы) с заданной точностью. Конечно, в случае одного уравнения прямое решение, если оно возможно, предпочтительнее. Мы же остановимся на случае, когда прямого решения нет, или его очень сложно найти. Подчастую даже кубическое уравнение проще решить итерационным способом. Причем, точность решения при этом будет ничуть не меньше, а иногда даже выше прямого (казалось бы строгого) решения.

Причина этому — ошибки округления, которые лавинообразно накапливаются в прямых методах.

Итак, итерационный метод. Он состоит из двух этапов:
1. Задание начального (приближенного) значения корня x_0.
2. Последовательное уточнение приближенного значения до некоторой заданной степени точности (расчет более точных значений $x_1, x_2, \ldots, x_k$, пока $|x_k - \bar{x}| \le \varepsilon$).

Первый этап нас сейчас не интересует. Эта задача высокотворческая и решается индивидуально в каждом конкретном случае. Обычно начальное значение можно задать из каких-либо физических соображений.

Второй этап состоит из ряда последовательных итераций. В ходе итераций получают значения x, которые все ближе и ближе приближаются к истинному значению корня (метод сходится), или не приближаются (метод не сходится, бывает и такое). Важнейшей характеристикой итерационного метода является скорость сходимости (или порядок сходимости). Чем выше скорость сходимости, тем меньше итераций придется сделать для достижения требуемой точности. Пусть итерационный метод обладает следующим свойством: существует некоторая δ-окрестность корня $\bar{x}$ такая, что если приближение x_k принадлежит этой окрестности, то справедлива оценка

$$|x_{k+1} - \bar{x}| \le q|x_k - \bar{x}|^p \ ,$$

где $q > 0$ и $p \ge 1$ — постоянные. Тогда число p называют порядком сходимости метода. Если $p = 1$ и $q < 1$, то говорят, что метод обладает линейной скоростью сходимости в указанной δ-окрестности корня. Другое название линейной сходимости — скорость геометрической прогрессии, знаменателем которой является число q. Действительно, при $p = 1$ можем записать

$$|x_k - \bar{x}| \le q^k|x_0 - \bar{x}| \ ,$$

откуда и пошло такое название. Если $p > 1$, то говорят, что метод обладает сверхлинейной скоростью сходимости. При $p = 2$ — квадратичная скорость, при $p = 3$ — кубическая.

В методах со сверлинейной скоростью сходимости для прекращения итераций можно использовать простые критерии

$$|x_k - x_{k-1}| < \varepsilon \ , \quad \text{или} \quad \frac{|x_k - x_{k-1}|}{|x_k|} < \varepsilon \ .$$

Эти же критерии пригодны и для метода с линейной скоростью сходимости, если $q \le 0.5$. При $q > 0.5$ условие приходится ужесточать

$$|x_k - x_{k-1}| < \varepsilon\frac{1-q}{q} \ , \quad \text{или} \quad \frac{|x_k - x_{k-1}|}{|x_k|} < \varepsilon\frac{1-q}{q} \ .$$

Мы познакомимся с решением алгебраических уравнений и систем, причем, если теоретики в этом пункте основной упор делают на проблему собственных значений и векторов, то мы эти вопросы вообще рассматривать не будем; вопросы интерполяции и оптимизации мы также рассматривать не будем — они достаточно хорошо автоматизированы; мы остановимся на вопросах численного интегрирования, причем, без обсуждения тех же Фурье преобразований — они, на мой взгляд, не составят для вас проблемы, если вы будете уметь численно рассчитывать интегралы; затем мы порешаем обыкновенные ДУ, и наконец перейдем к ДУ в ЧП.

Ну а начнем мы с того, что попроще.

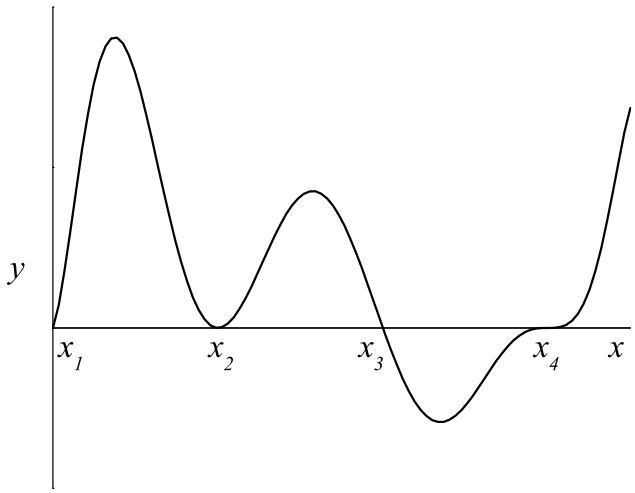
I. Алгебраические уравнения и системы.

I.1. Решение алгебраических уравнений

Пусть дано уравнение

$$f(x) = 0\ , \quad \text{например,} \quad x^2 - 9 = 0\ .$$

Мы должны найти корень (корни) этого уравнения. Корни бывают разные. Геометрически корень $\bar{x}$ соответствует точке пересечения графика функции $y = f(x)$ с осью Ox. Корень



$\bar{x}$ называется простым, если $f'(\bar{x}) \neq 0$. В противном случае корень называется кратным. Целое число m называется кратностью корня $\bar{x}$, если $f^{(k)} = 0\ \forall\ k < m$ и $f^{(m)}(\bar{x}) \neq 0$. Так на рис. корни x_1 и x_3 — простые, x_2 — как минимум, второй кратности, x_4 — как минимум, третьей кратности. Мы рассмотрим только методы нахождения простых корней:

- Метод половинного деления;
- Метод простой итерации;
- Метод Ньютона–Рафсона.

Теперь поговорим о том, какие же собственно методы используются для решения нелинейного уравнения.

Деление отрезка пополам.

Допустим вы установили, что функция $f(x)$ обращается в ноль на интервале (x_n, x_k), причем левее корня $f(x) < 0$, а правее $f(x) > 0$. Тогда найти искомый корень не представляет труда. Вы просто делите отрезок пополам и смотрите знак функции в полученной точке x_i. Если $f > 0$, то сдвигаем верхнюю границу $x_k = x_i$, в противном случае сдвигаем нижнюю границу $x_n = x_i$. Метод сходится медленно, но зато очень прост, и если вам некуда спешить, пользуйтесь этим методом. Порядок сходимости этого метода равен 1, т.е. метод обладает линейной скоростью сходимости, или сходится со скоростью геометрической прогрессии, знаменатель которой равен 1/2.

Однако зачастую поведение функции оказывается достаточно сложным и трудно заранее определить необходимый интервал (x_n, x_k), или же требуется более быстрый метод.

Вдобавок к этому метод деления отрезка пополам не обобщается на системы из нескольких уравнений. По этим причинам мы с вами вынуждены познакомиться с другими, более хитрыми методами.

Метод простой итерации.

Для этого метода необходимо из функции $f(x)$ выделить x, и записать исходное уравнение в виде

$$x = s(x) \ .$$

Конечно, сделать это можно по-разному (кто на что горазд). Например, для нашего уравнения возможны следующие варианты

$$x = x^2 + x - 9 \ , \quad x = \frac{9}{x} \ , \quad x = \frac{1}{2}\left(x + \frac{9}{x}\right) \ .$$

Задаем начальное приближение $x_0 = 2.5$. Последующие приближения находятся подстановкой предыдущего в функцию $s(x)$, т.е.

$$x_1 = s(x_0) \ , \quad x_2 = s(x_1) \ , \quad \dots \ , \quad x_{n+1} = s(x_n) \ , \quad \dots \ .$$

Для записанных нами функций $s(x)$ получаем (см. табл.). Сходимость наблюдается только

x_i	$s = x^2 + x - 9$	$s = 9/x$	$s = (x + 9/x)/2$	$(-2.0; 4.0)/2$
0	2.5	2.5	2.5	2.5
1	-0.25	3.6	3.05	3.75
2	-9.2	2.5	3.0004	3.125
3	66	3.6	$3.0^{7}3$	2.8125
4	4443	2.5	$3.0^{15}1$	2.96875

в одном случае. Почему? На этот вопрос достаточно просто ответить. Пусть $x = a$ — корень исходного уравнения. Приближение получаемое на $n+1$-ом шаге вычисляется, как

$$x_{n+1} = s(x_n) \ .$$

Тогда отклонение x_{n+1} от a можно представить в следующем виде:

$$x_{n+1} - a = s(x_n) - s(a) = s'(a)\,(x_n - a) + O((x_n - a)^2) \ .$$

Теперь если максимальное значение производной на всем интервале изменения x обозначить как q, то получаем

$$|x_n - a| \le q^n |x_0 - a| \ .$$

Таким образом, если во всей области изменения x производная записанной вами функции $|s'(x)| < 1$, то итерации сходятся. Причем сходятся в геометрической прогрессии, т.е. достаточно быстро.

По сути, мы (почти) доказали известную теорему для метода простой итерации:
Пусть в некоторой окрестности корня $\bar{x}$ функция s дифференцируема и удовлетворяет неравенству

$$|s'(x)| \le q \ , \text{ где } \ 0 \le q < 1 - const.$$

Тогда независимо от выбора начального приближения x_0 из указанной δ-окрестности корня итерационная последовательность не выходит за пределы этой окрестности и метод сходится со скоростью геометрической прогрессии со знаменателем q.

К сожалению, если вы возьмете начальное приближение за пределами указанной окрестности, то сходимость вам никто не сможет гарантировать, и, как правило, ее и не будет. Эту окрестность называют еще областью сходимости.

В нашем примере при $x = a = 3$

$$s_1'(x) = 2x + 1 = 7 \ , \quad s_2'(x) = -\frac{9}{x^2} = -1 \ , \quad s_3'(x) = 0 \ .$$

Условию сходимости удовлетворяет лишь третья функция. Нетрудно догадаться, что метод простой итерации, если он сходится, то обладает линейной скоростью сходимости, или скоростью геометрической прогрессии со знаменателем равным q. Если q меньше 0.5, то данный метод сходится быстрее половинного деления. Случай же с $s'(a) = 0$ вообще-то соответствует сверхлинейной скорости сходимости. Это, конечно, большая редкость для метода простой итерации, предложенная формула (третья) настолько уникальна, что именно она используется для извлечения квадратных корней во многих численных пакетах и в калькуляторах.

Теперь, казалось бы, вы знаете какую из функций $s(x)$ нужно выбирать. Однако, метод простой итерации применяют обычно как раз тогда, когда трудно что-либо сказать о значении производной вблизи искомого корня. Когда заранее невозможно выбрать нужный вид функции $s(x)$, возникает необходимость заставить сходиться метод итераций с любой функцией. Этого можно достичь благодаря незначительной модификации.

$$x_{n+1} = x_n + \alpha \Delta x_n \ ,$$

где Δx_n — это поправка, которая соответствует простому методу итераций

$$\Delta x_n = s(x_n) - x_n \ .$$

Для функций $s_1(x)$ и $s_2(x)$ достаточно взять $\alpha = -1/7$ и 0.99, соответственно, чтобы получить сходимость. Однако однозначных рекомендаций по выбору параметра α не существует. Каждый раз это приходится делать эмпирически.

Из исходного нелинейного уравнения вовсе не обязательно всегда выделять переменную x в чистом виде. В более общем случае метод простой итерации можно представить в виде

$$\phi(x_k) = s(x_{k-1}) \ , \qquad *$$

где, конечно, желательно, функцию $\phi(x)$ подбирать так, чтобы уравнение

$$\phi(x) = const$$

решалось сравнительно просто (возможно, вообще без итераций), например, $\phi(x) = \exp(x)$ или $\phi(x) = x^t$.

Нетрудно, догадаться, что в этом, более общем случае, итерации сходятся, при выполнении условия

$$|s'(x)| < |\phi'(x)| \ .$$

Если существует и легко вычисляется функция обратная к ϕ — ϕ^{-1}, то действием ϕ^{-1} на уравнение (*) опять приходим к уже известному варианту изложенного метода.

Однако, в случае, когда не представляет труда продифференцировать заданную уравнением функцию, более предпочтительным оказывается

метод Ньютона – Рафсона.

Запишем функцию $f(x)$ при $\bar{x}$ (корень) в виде ряда

$$f(\bar{x}) = 0 = f(x_n) + f'(x_n)\,(\bar{x} - x_n) + \dots \ .$$

Обрывая ряд на втором члене, получим... нет, не $\bar{x}$, конечно, мы ведь оборвали ряд, поэтому точное значение $\bar{x}$ не получим. А получим $n+1$-ое приближение

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

Это и есть знаменитая формула Ньютона – Рафсона. Поскольку мы отбросили члены, пропорциональные $(x_n - \bar{x})^2$, то в отличие от метода простой итерации

$$x_{n+1} - \bar{x} = O\left[(x_n - \bar{x})^2\right] \ .$$

Строго: "Если существует окрестнось корня O_1, в которой функция f дважды непрерывно дифференцируема, то $\exists$ окрестность $O_2 \in O_1$: итерацинная последовательность не выходит за пределы этой окрестности и метод сходится с квадратичной скоростью". Таким образом, если этот метод сходится, то сходится он, в общем случае, значительно быстрее метода простой итерации. Но в нашем конкретном примере метод НР $\equiv s_3$.

Недостатки метода:

- малая область сходимости; зачастую полезно начинать поиск корня с помощью другого (медленного) метода, а завершать быстрым уточнением полученного приближения по методу Ньютона–Рафсона;
- необходимость расчета первой производной.

Преимущества метода:

- быстрая сходимость;
- возможность обобщения на системы алгебраических уравнений.

I.2. Решение систем нелинейных алгебраических уравнений

Используются, в основном, только итерационные методы. Исходная система, которую надо решить:

$$\begin{cases} f_1(x_1, x_2, \ldots, x_m) &= 0 \\ f_2(x_1, x_2, \ldots, x_m) &= 0 \\ \cdots & \cdots \ \cdots \\ f_m(x_1, x_2, \ldots, x_m) &= 0 \end{cases} \quad \text{или} \quad \vec{f}(X) = 0$$

Методы похожи на методы решения отдельных нелинейных алгебраических уравнений. В ходе итераций получаем последовательность векторов $X^{(k)} = (x_1^{(k)}, \ldots, x_m^{(k)})$. Итерационный метод сходится, если при $k \to \infty$ норма отклонения стремится к нулю:

$$|X^{(k)} - \bar{X}| \to 0 \ .$$

Что такое норма?

В случае чисел под нормой понимают модуль числа. В случае векторов под нормой чего только не понимают, лишь бы она (норма) удовлетворяла известным свойствам:
1) $||X|| \geq 0; \quad ||X|| = 0 \quad \Leftrightarrow \quad X = 0$;
2) $||\alpha X|| = |\alpha| \cdot ||X||$, α — скаляр;
3) $||X + Y|| \leq ||X|| + ||Y||$ (неравенство Минковского).

Например:
$||X|| = \sqrt{\sum x_i^2}$ — евклидова норма;
$||X|| = \max_i |x_i|$ — равномерная норма.

В случае матриц под нормой понимают число $||A||$, удовлетворяющее свойствам:
1) $||A|| \geq 0; \quad ||A|| = 0 \quad \Leftrightarrow \quad A = 0$;
2) $||\alpha A|| = |\alpha| \cdot ||A||$, α — скаляр;
3) $||A + B|| \leq ||A|| + ||B||$;
4) $||AB|| \leq ||A|| \cdot ||B||$.

Более того, норма матрицы $||A||$ согласована с нормой вектора $||X||$ (а только такие нормы мы и будем использовать), если $||AX|| \leq ||A|| \cdot ||X||$.

Например, с евклидовой метрикой векторного пространства согласована норма $||A|| = \sqrt{\rho\left(A^T A\right)}$. Здесь A^T — транспонированная матрица, $\rho(B)$ — спектральный радиус матрицы B, т.е. максимальное собственное значение.

С равномерной метрикой согласована норма $||A|| = \max_i \sum_j |A_{ij}|$.

Метод простой итерации

Систему приводят к виду:

$$\begin{cases} x_1 &= s_1\left(x_1, x_2, \ldots, x_m\right) \\ x_2 &= s_2\left(x_1, x_2, \ldots, x_m\right) \\ \ldots & \ldots \quad \ldots \\ x_m &= s_m\left(x_1, x_2, \ldots, x_m\right) \end{cases} \qquad \text{или} \qquad X = \vec{s}(X)$$

Далее задают исходное приближение:

$$X^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}\right) \ .$$

От близости исходного приближения к решению как быстрота сходимости, так и сам факт сходимости. Подставляя нулевое приближение в правые части записанных выражений, получаем $X^{(1)}$, затем — $X^{(2)}$, и т.д. Формулу итераций можно записать в следующем виде

$$x_i^{n+1} = s_i\left(x_1^n; x_2^n; \ldots; x_m^n\right) \ , \quad i = 1, 2, \ldots, m,$$

или в векторном виде

$$X^{n+1} = \vec{s}\left(X^n\right) \ .$$

Область исходных значений $x_1^{(0)}$, $x_2^{(0)}$, ..., $x_m^{(0)}$, при которых такая процедура сходится к решению (или говорят просто "сходится"), называют областью сходимости. Если исходное приближение лежит за пределами этой области, то решение вы не получите. К сожалению, с увеличением числа неизвестных область сходимости уменьшается. В случае очень больших систем сходимость бывает только при условии, что $X^{(0)}$ очень близко к решению. При решении эволюционных нелинейных ДУ в ЧП это часто не беда, т.к. за нулевое приближение можно принять решение, полученное на предыдущем временном слое.

Аналогично случаю с одним уравнением, если метод сходится, то погрешность расчета убывает по геометрической прогрессии. А именно, может быть доказана следующая
Теорема:
Если $\exists$ D-$O(\bar{X})$ такая, в которой (т.е. $\forall$ $X \in D$-$O(\bar{X})$) функции $s_i(X)$ дифференцируемы и $|| \left(\mathbf{s}'(\mathbf{X})\right) || \leq q$, где $0 \leq q < 1$ ($(\mathbf{s}'(\mathbf{X}))$ — матрица Якоби);
то независимо от выбора начального приближения X_0 из указанной окрестности ($\forall X_0 \in D$-O) выполняется: 1) итерационная последовательность не выходит за пределы этой окрестности ($X_k \in D$-O $\forall k$); 2) метод сходится ($|X_k - \bar{X}| \to_{k\to\infty} 0$) со скоростью геометрической прогрессии со знаменателем q, т.е. $|X_k - \bar{X}| \leq q^k |X_0 - \bar{X}|$.

Обычно расчет идет до выполнения неравенств

$$\left|x_i^{(k)} - x_i^{(k-1)}\right| < \Sigma\frac{1-q}{q} \ , \quad \text{для} \quad i = 1, \ldots, m \ ,$$

где Σ — заданная точность решения.

Обычно систему нелинейных уравнений не решают методом простой итерации, а используют незначительную модификацию этого метода.

Метод Зейделя.

Все точно также, как и в методе простой итерации. Отличие состоит лишь в том, что определив из первого уравнения $x_1^{(1)}$, используют уже его в правой части выражения для $x_2^{(1)}$, а для определения $x_3^{(1)}$ используют уже вычисленные к этому этапу $x_1^{(1)}$ и $x_2^{(1)}$, и т.д.

$$x_i^{n+1} = s_i\left(x_1^{n+1}, x_2^{n+1}, \ldots, x_{i-1}^{n+1}, x_i^n, \ldots, x_m^n\right) \ .$$

В отличие от одного уравнения в случае системы уравнений, гораздо сложнее определить заранее будет ли при выбранных функциях s_i метод сходиться. Поэтому здесь еще более актуальной является задача "заставить" сходиться метод. При одном уравнении мы этого добились введя параметр α

$$x^{(n+1)} = x^{(n)} + \alpha\Delta x_{(n)} \ .$$

С системами поступают точно также, и называют это...

Метод релаксации.

Записывают формулы для релаксации в таком виде:

$$x_i^{n+1} = x_i^{(n)} + \alpha\left[\left(x_i^{(n+1)}\right)_{\text{Зей}} - x_i^{(n)}\right] \ .$$

К сожалению, выбор α — задача чисто эмпирическая. Обычно $\alpha \in (-1; +1)$. При $\alpha > 0$ метод называют методом последовательной верхней релаксации (SOR-метод "successive over relaxation"), при $\alpha < 0$ — последовательная нижняя релаксация.

Если же вы хотите красиво жить, вспомните о методе Ньютона – Рафсона. Обобщение этого метода на систему нелинейных уравнений называется:

Метод Ньютона.

Это наиболее распространенный метод для решения систем нелинейных уравнений.

Решается та же система. Пусть известно какое-то n-е приближение $X^{(n)} = (x_1^{(n)}, \ldots, x_m^{(n)})$. Тогда "более хорошее" приближение $X^{(n)}$ ищем аналогично методу для одного уравнения. Представим все $f_i\left(\bar{X}\right)$ по ф-ле Тейлора вблизи $X^{(n)}$ ($\bar{X}$ — вектор истинного решения):

$$f_i\left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m\right) = 0 = f_i\left(x_1^{(n)}, x_2^{(n)}, \ldots, x_m^{(n)}\right) + \left(\frac{\partial f_i}{\partial x_1}\right)_{X^{(n)}}\left(\bar{x}_1 - x_1^{(n)}\right) +$$
$$\left(\frac{\partial f_i}{\partial x_2}\right)_{X^{(n)}}\left(\bar{x}_2 - x_2^{(n)}\right) + \ldots\left(\frac{\partial f_i}{\partial x_m}\right)_{X^{(n)}}\left(\bar{x}_m - x_m^{(n)}\right) + O\left[\left(\bar{x}_1 - x_1^{(n)}\right)^2, \ldots, \left(\bar{x}_m - x_1^{(n)}\right)^2\right] \ .$$

Опять же обрываем эти ряды на линейных членах. Получили систему линейных уравнений для нахождения .... Нет, опять же не самого истинного решения $\bar{X}$, а лишь $(n+1)$-го приближения.

$$0 = \vec{f}(X_n) + (\mathbf{f}')\,(X_{n+1} - X_n)\ ,$$

где опять $(\mathbf{f}'(\mathbf{X}))$ — матрица Якоби функций $f_i(X)$. Решив эту систему, найдем $X^{(n+1)}$ и т.д. до выполнения условия окончания расчетов (достижение требуемой точности). В методе Ньютона мы отбрасываем члены со второй производной пропорциональные $\left(x_i^{(n)} - \bar{x}_i\right)^2$. Отброшенные члены определяют погрешность нашего расчета, поэтому в отличие от метода простой итерации $x_i^{n+1} - \bar{x}_i = O\left[(x_i^n - \bar{x}_i)^2\right]$ и метод Ньютона достаточно быстро сходится. А именно, может быть доказано, что метод Ньютона имеет квадратичную скорость сходимости, если функции $f_i(X)$ дважды непрерывно дифференцируемы, и матрица Якоби невырождена, т.е. ее определитель отличен от нуля.

Недостатки метода:

- малая область сходимости*; зачастую полезно начинать поиск решения с помощью другого (медленного) метода (например: сведение к задаче минимизации), а завершать быстрым уточнением полученного приближения по методу Ньютона;
- необходимость расчета производных;
- необходимость на каждом шаге итераций решать систему линейных алгебраических уравнений.

Преимущества метода:

- ввиду быстрой сходимости метод Ньютона является одним из наиболее распространенных методов решения систем нелинейных уравнений.

Лишь одна загвоздка: по ходу итераций необходимо каждый раз решать систему линейных алгебраический уравнений. Здесь можно отметить, что неодходимость решения систем линейных алгебраических уравнений возникает в численных методах очень часто. В частности, такая задача возникает, как правило, постоянно при решении ДУ в ЧП. Ввиду исключительной важности систем линейных алгебраических уравнений, мы с вами обязательно остановимся на этом вопросе. А пока, в заключение, разговора о системе нелинейных уравнений рассмотрим еще один способ ее численного решения.

Сведение к задаче минимизации.

Одной из наиболее трудных проблем, возникающих при решении систем нелинейных уравнений, является задача предварительной локализации решения, т.е. задание начального приближения X_0. Иногда бывает полезно свести задачу отыскания решения системы нелинейных уравнений к задаче отыскания минимума функции многих переменных. В простейшем варианте это сведение выглядит следующим образом. Вводится функция

$$\Phi(X) = \sum_{i=1}^{m} \left(f_i(X)\right)^2 .$$

Данная функция неотрицательна и достигает своего минимума (равного нулю) тогда и только тогда, когда $f_i(X) = 0$ для всех $i = 1, 2, \ldots, m$, т.е. X является решением заданной системы.

Далее используется какой-нибудь метод минимизации — метод градиентного спуска, покоординатного спуска, сопряженных градиентов и т.п. Мы на них не будем останавливаться. Выгода от указанного сведения исходной задачи к задаче минимизации состоит в том, что, как правило, методы спуска имеют более широкую область сходимости. Использование методов спуска можно рассматривать как один из способов локализации решения исходной системы. Применение на заключительном этапе методов, специально ориентированных на решение систем нелинейных уравнений, вызвано тем, что вблизи искомого решения (вблизи минимума функции $\Phi(X)$) методы спуска сходятся медленее.

Обычно поступают как раз наоборот: задачу минимизации сводят к решению системы алгебраических уравнений. Например, минимум функции $f(x)$ ищут, решая уравнение $f'(x) = 0$.

I.3. Решение систем линейных алгебраических уравнений

Итак, систему линейных алгебраических уравнений, как вы понимаете, можно представить в виде

$$AX = B \ ,$$

где A — заданная квадратная матрица порядка "m", X — искомый вектор решения, B — заданный вектор (вектор правых частей системы уравнений).

$$A = \left\| \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{array} \right\| , \quad X = \left\| \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_m \end{array} \right\| , \quad B = \left\| \begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_m \end{array} \right\| .$$

Естественно, что предполагается $det||A|| \neq 0$, т.е. решение существует и оно единственное.

I.3.1. Прямые методы решения систем линейных алгебраических уравнений

Систему нелинейных уравнений мы в общем случае не можем решить точно, а вот линейную систему вы умеете решать. Мне на память приходят следующие методы:

1. Доморощенный
2. Метод обратной матрицы
3. Метод Крамера
4. Метод Гаусса

Начнем анализ этих методов.

Доморощенный.

Вы выражаете x_1 из первого уравнений и подставляете полученное выражение для x_1 во все оставшиеся уравнения. Тем самым x_1 исключается из системы. Далее вы выражаете x_2 из второго уравнения и подставляете в оставшиеся, и т.д. Последовательность таких действий называется прямой ход решения системы. На обратном ходе вы последовательно вычисляете все неизвестные, начиная, как вы понимаете, с конца.

Метод работает, я его проверял, и дает на самом деле очень точное решение. Но есть одна загвоздка. Хорошо если вам надо решить систему из 2 или 3 уравнений. А если система содержит 30 уравнений? Вы застрелитесь быстрее. А ведь 30 уравнений — это

далеко не предел. Задача решения системы линейных уравнений в большинстве случаев возникает не сама по себе, а как часть каких-то других алгоритмов. Так вот, при решении ДУ в ЧП возникают линейные системы по 100, 1000 и более уравнений. Но не нужно сразу же ставить крест на "доморощенном" методе. На самом деле рациональное зерно в нем есть. Если вы попробуете запрограммировать его, то хорошенько поломав голову вы придете, скорее всего, к методу Гаусса — самому эффективному методу точного решения системы линейных алгебраических уравнений. Но мы пойдем по порядку.

Метод обратной матрицы.

Вообще-то все оставшиеся методы: обратной матрицы, Крамера и Гаусса — это та или иная работа с матрицами. И, в частности, все они подразумевают, что вы умеете считать определитель матрицы. Пара слов о том, как это делается. Определителем квадратной матрицы A называется число $||A||$, равное сумме $m!$ слагаемых

$$||A|| = \sum_{1}^{m!} (-1)^r a_{1,j_1} a_{2,j_2} \dots a_{m,j_m} \ ,$$

где r — число попарных перестановок в множестве $\{j_1, j_2, \dots, j_m\}$. Чтобы непосредственно, вычислить такую сумму, например для матрицы размером 100×100, при быстродействии компьютера 10^9 операций в секунду, потребовалось бы время намного превышающее возраст Вселенной.

Единственный, известный мне, разумный способ расчета определителя — это приведение исходной матрицы к треугольному виду:

$$A = \left\| \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{array} \right\| , \quad \Longrightarrow \quad \left\| \begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1m} \\ 0 & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{mm} \end{array} \right\| .$$

После этого определитель просто равен произведению диагональных элементов. Что нужно для такой процедуры. Для получения нулей в первом столбце первую строку, домноженную на различные коэффициенты, прибавить к нижележащим. Это $\sim m^2$ арифметических операций. Затем для получения нулей во втором столбце вторую строчку, домноженную на различные коэффициенты, прибавим к нижележащим. Это $\sim (m-1)^2$ арифметических операций. В итоге

$$\text{Число арифметических операций} \sim \sum_{i=1}^{m} i^2 \simeq \int_0^m x^2 dx \sim m^3 \ .$$

Это несложно запрограммировать. Но имеется одна тонкость. Ее называют — метод главного элемента.

Перед тем как получать нули в каком-нибуть i-ом столбце из оставшихся строк выбирают строчку с максимальным коэффициентом a_{ji} и именно эту строчку сдвигают вверх на i-ую позицию. Благодаря этой несложной процедуре вы, во-первых, избежите тривиального деления на ноль, а, во-вторых, существенно повысите точность расчета. Иногда ввиду погрешности округления чисел в компьютере (а такое округление происходит всегда) погрешность полученного результата превосходит само рассчитанное значение. Единственный метод борьбы с таким лавинообразным накоплением ошибки округления — метод

главного элемента. Не забудьте, что при перестановке строк в матрице определитель меняет знак.

Теперь продолжим разговор о методе обратной матрицы. Согласно этому методу решение исходной системы записывается в виде

$$X = A^{-1}B \ ,$$

где A^{-1} — матрица, обратная к матрице A. Для расчета обратной матрицы, как вы несомненно помните, необходимо рассчитать алгебраические дополнения ко всем элементам прямой матрицы, т.е. вычислить m^2 определителей порядка $(m-1)$. Это значит

$$\text{Число арифметических операций} \sim m^2(m-1)^3 \simeq m^5 \ .$$

Столь большое число необходимых арифметических операций делает, в общем случае, метод обратной матрицы неприемлимым для численных расчетов. Единственный случай когда этот метод может оказаться полезен, это когда вам необходимо решать очень много систем, которые отличаются лишь столбцом свободных членов B.

Метод Крамера.

Решение имеет вид

$$x_i = \frac{\Delta_i}{\Delta} \ ,$$

где Δ — определитель матрицы A, а Δ_i — определители матрицы A, в которой i-ый столбец заменен столбцом свободных членов. В этом методе необходимо рассчитать m определителей порядка m, т.е.

$$\text{Число арифметических операций} \sim m^4 \ .$$

Это уже гораздо лучше. Метод хорош, когда вас не интересуют значения всех неизвестных, а вам необходимо определить лишь какое-то одно из них. Но если вас все же интересуют все неизвестные, то нет ничего лучше метода Гаусса.

Метод Гаусса.

Этот метод — есть не что иное, как логическое продолжение самого первого нашего метода, доморощенного. Метод состоит из прямого и обратного хода. На прямом ходе исходную систему линейных уравнений приводят к треугольному виду. Это делается точно также, как и приведение к треугольному виду матрицы, для расчета определителя. На обратном ходе вычисляются неизвестные x_i, начиная естественно с конца. Число арифметических действий, необходимых для решения:

$$\text{Число арифметических операций} \sim m^3 \ , \quad O(m^3) \ .$$

Важную роль играет та же тонкость, что и при расчете определителей. Используйте метод главного элемента! Приведу один показательный пример

$$\begin{aligned} &\frac{1}{7}x + 10^5 y = 10^5 + \frac{1}{7} \ , \\ &3 \cdot 10^5 x + \frac{1}{11}y = 3 \cdot 10^5 + \frac{1}{11} \ . \end{aligned}$$

Решение этой системы $x = 1$, $y = 1$. Применим метод Гаусса без перестановки строк. Если расчеты ведутся с точностью до 20 значащих цифр (это очень высокая точность), то получим, что $x = 1 + 2.13 \cdot 10^{-14}$. Потеряли 6 порядков точности! Системы, подобные приведенной, отмечает, то что на их диагонали стоят малые по абсолютной величине коэффициенты. Если же переставить строки перед расчетом, то все 20 значащих цифр будут верными.

Но и это еще не предел совершенству. При применении многих численных методов возникают системы уравнений, матрицы которых имеют вполне определенный вид. Например, матрица A может содержать много нулевых элементов, расположенных в матрице не беспорядочно, а плотными массивами на заранее известных местах. Тогда расчет по методу Гаусса можно организовать так, чтобы не включать эти элементы. Типичные случаи матрицы A: ленточная, ящичная, блочно-диагональная, квазитреугольная, и т.д. Примечательно то, что во всех таких вариантах нельзя делать выбор ведущего (главного) элемента, так как перестановки разрушают специальную структуру матриц. Но обычно в этом нет и необходимости, так как задачи, приводящие к таким матрицам, обычно таковы, что приводят к хорошо обусловленным матрицам с большими (преобладающими) элементами на главной диагонали. Варианты метода Гаусса, разработанные для решения систем уравнений с такими матрицами (матрицами специального вида), позволяют существенно уменьшить число арифметических действий, необходимых для решения [$O(m^2)$ и даже $O(m)$].

Мы рассмотрим только одну модификацию метода Гаусса, широко применяющуюся при решении ДУ в ЧП. Речь идет о решении систем с трехдиагональной матрицей. Название этого метода —

метод прогонки или алгоритм Томаса (англ.).

Трехдиагональная матрица:

$$\left\| \begin{array}{ccccc} a_{11} & a_{12} & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & 0 \\ 0 & a_{32} & a_{33} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_{m-1,m} \\ 0 & 0 & 0 & a_{m,m-1} & a_{mm} \end{array} \right\| ,$$

т.е. ненулевые элементы расположены только на главной диагонали и двух, примыкающих к ней линиях.

Запишем систему уравнений с такой матрицей в каноническом виде:

$$\left\| \begin{array}{ccccc} -b_1 & c_1 & 0 & \dots & 0 \\ a_2 & -b_2 & c_2 & \ddots & 0 \\ 0 & a_3 & -b_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{m-1} \\ 0 & 0 & 0 & a_m & -b_m \end{array} \right\| , \qquad \begin{array}{l} a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i \ , \\ 1 \le i \le m \ , \quad a_1 = c_m = 0 \ . \end{array}$$

Если подвергнуть прямому ходу метода Гаусса систему уравнений с трехдиагональной матрицей, то получим систему, содержащую в каждом уравнении только два неизвестных — x_i и x_{i+1}. Поэтому формулы обратного хода можно представить в виде

$$x_i = \delta_i + \gamma_i x_{i+1} \ , \qquad i = m, \dots, 1 \ . \qquad (*)$$

γ_i и δ_i — некоторые пока неизвестные коэффициенты. Перепишем эту формулу для x_{i-1}:

$$x_{i-1} = \delta_{i-1} + \gamma_{i-1} x_i \ .$$

Подставим в каноническую запись:

$$a_i \left(\delta_{i-1} + \gamma_{i-1} x_i\right) - b_i x_i + c_i x_{i+1} = d_i \ , \qquad 1 \le i \le m \ ,$$

и выразим отсюда x_i

$$x_i = \frac{a_i \delta_{i-1} - d_i}{b_i - a_i \gamma_{i-1}} + \frac{c_i}{b_i - a_i \gamma_{i-1}} x_{i+1} \ , \qquad i = m, \ldots, 1 \ , \qquad c_m = 0 \ . \qquad (**)$$

Для того, чтобы найти неизвестные пока коэффициенты γ_i, δ_i, сравним это с формулами обратного хода (*). Получаем:

$$\delta_i = \frac{a_i \delta_{i-1} - d_i}{b_i - a_i \gamma_{i-1}} \ , \qquad \gamma_i = \frac{c_i}{b_i - a_i \gamma_{i-1}} \ , \qquad i = 1, \ldots, m \ , \qquad a_1 = 0$$

— это формулы прямого хода. γ_0, δ_0 не нужны (см. ур. (**), так как умножаются на $a_1 = 0$. $\gamma_m = 0$, что очевидно, поскольку (см. (*)) никакого x_{m+1} у нас нет.

Число арифметических действий, необходимых для решения системы с треугольной матрицей методом прогонки $\sim m$, т.е. $O(m)$.

Преимущества метода:

- экономичность — максимально использована специальная структура исходной системы.

Недостатки метода:

- Возможность лавинообразного накопления ошибок округления.
  Данный недостаток присущ всем прямым методам.
  Особенностью метода прогонки является невозможность использования метода главного элемента для снижения ошибок округления: нарушится специальная (трехдиагональная) структура исходной системы. В связи с этим высокую актуальность имеет анализ устойчивости метода прогонки.

В знаменателях формул стоит выражение:

$$b_i - a_i \gamma_{i-1} \ .$$

Не получится ли при вычислениях, что оно близко к 0?

Можно доказать, что достаточными условиями устойчивости вычислений по этому алгоритму являются:

$$a_i \neq 0 \quad (2 \le i \le m) \ ; \qquad c_i \neq 0 \quad (1 \le i \le m-1) \ ;$$

$$|b_i| \ge |a_i| + |c_i| \ . \qquad (***)$$

Причем строгое неравенство в (***) должно быть хотя бы для одного уравнения. Докажем, что данных условий достаточно для устойчивости метода прогонки.
Предположим, что условие (***) выполнено. Тогда

$$|\gamma_1| = \frac{|c_1|}{|b_1|} \le 1 \ ;$$

$$|\gamma_i| = \frac{|c_i|}{|b_i - a_i\gamma_i|} \le \frac{|c_i|}{|b_i| - |a_i||\gamma_i|} \le \frac{|c_i|}{|b_i| - |a_i|} \le 1 \ .$$

Это уже показывает устойчивость метода прогонки, поскольку исключена возможность лавинообразного нарастания ошибок округления

$$|\Delta x_i| = |\gamma_i||\Delta x_{i+1}| \le |\Delta x_{i+1}| \ .$$

Зачем нужно хотя бы одно строгое неравенство в (***)? Чтобы в самом начале обратного хода (расчет $x_m = \delta_m$) знаменатель δ_m не обратился в ∞. Этого не произойдет, если

$$|b_m| > |a_m||\gamma_{m-1}| \ ,$$

т.е.

$$\text{либо} \quad |b_m| > |a_m| \ , \qquad \text{либо} \quad |\gamma_{m-1}| < 1 \ .$$

Условие (***) — условие преобладания диагональных элементов. Оно является достаточным, но не необходимым, т.е. в практических расчетах прогонка оказывается устойчивой даже при нарушении этого условия.

I.3.2. Итерационные методы решения систем линейных алгебраических уравнений

Они те же, что и для решения систем нелинейных алгебраических уравнений:
— метод простой итерации (метод Якоби),
— метод Зейделя (метод Гаусса – Зейделя).

Аналога методу Ньютона (или Ньютона – Рафсона) для линейных систем не существует.

Исходную систему

$$AX = B$$

для метода Якоби записывают в виде уравнений

$$x_i^{(n+1)} = -\sum_{j=1, j\neq i}^{m} \frac{a_{ij}}{a_{ii}} x_j^{(n)} + \frac{b_i}{a_{ii}} \ ,$$

а для метода Гаусса–Зейделя в виде уравнений

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

Преимущества метода Гаусса–Зейделя можно продемонстрировать на следующем примере. Решим систему:

$$\begin{cases} 2x + y = 2 \\ x - 2y = -2 \end{cases}$$

Решение тривиально: $x = 0.4$, $y = 1.2$. Заметьте, если переписать эту систему в виде

$$\begin{cases} x = 2y - 2 \\ y = -2x + 2 \end{cases}$$

"пригодном" для итераций, то вас уже ничто не спасет. Итерации не сходятся. Вы уже должны понимать почему это происходит — норма матрицы Якоби записанной системы

> 1. А именно, как для равномерной так и для евклидовой метрики $||s'(x,y)|| = 2$. Мы запишем так:

$$x = \frac{2-y}{2} \ , \quad y = \frac{2+x}{2} \ .$$

Теперь $||s_i'|| = 1/2 < 1$, т.е. итерации будут сходиться. Причем, чем замечательны линейные уравнения, если итерации сходятся, то они сходятся при любом выборе начального приближения. Сравним методы Якоби и Гаусса–Зейделя (начальные значения: $x, y = 0$). Окончание вычислений: $\left|\dfrac{x_i^{n+1} - x_i^n}{x_i^n}\right| < \Sigma$, для всех i, Σ — заданная.

	Якоби		Зейдель	
0	0	0	0	0
1	0.5	1.5	0.5	1.25
2	0.25	1.25	0.375	1.1875
3	0.375	1.125	0.40625	1.203175
4	0.4375	1.1875	$\simeq 0.3984$	$\simeq 1.1992$

Сходимость методов Якоби и Гаусса–Зейделя. Используя критерий сходимости метода простой итерации для алгебраических уравнений с равномерной метрикой, нетрудно получить его прообраз для линейных систем:

$$|a_{ii}| > |a_{i1}| + \ldots + |a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}| \ , \qquad \forall i \ .$$

Однако для линейных систем этот критерий можно смягчить. Покажем это на примере метода Гаусса–Зейделя. Запишем систему из двух уравнений в общем виде:

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases}$$

Вычисления в методе Гаусса–Зейделя мы будем проводить по формулам (на k-ой итерации)

$$x^{(k)} = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} y^{(k-1)} \ ,$$

$$y^{(k)} = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x^{(k)} \ .$$

Теперь представим, что (x, y) — истинное решение системы. Тогда

$$x^{(k)} = x + \Delta x^{(k)} \ , \qquad y^{(k)} = y + \Delta y^{(k)} \ .$$

Подставим это в наши итерационные формулы. Получаем

$$\Delta x^{(k)} = -\frac{a_{12}}{a_{11}} \Delta y^{(k-1)} \ ,$$

$$\Delta y^{(k)} = -\frac{a_{21}}{a_{22}} \Delta x^{(k)} = \frac{a_{21}a_{12}}{a_{11}a_{22}} \Delta y^{(k-1)} \ .$$

Поскольку это справедливо для каждой итерации, то в итоге после k итераций получим

$$\Delta y^{(k)} = \left(\frac{a_{21}a_{12}}{a_{11}a_{22}}\right)^k \Delta y^{(0)} \ .$$

Т.е. метод Гаусса–Зейделя сходится, если выполняется условие

$$\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| < 1 \ .$$

Заметим, что такому условию система будет удовлетворять, в частности, если в каждой строчке диагональный член по абсолютной величине превышает сумму абсолютных значений других членов, т.е.

$$\begin{cases} |a_{11}| > |a_{12}| \\ |a_{22}| > |a_{21}| \end{cases}$$

Причем знак строгого неравенства необходимо потребовать лишь в одном из неравенств, в остальных можно допустить и равенство левых и правых частей. Мы получили не что иное, как условием преобладания диагональных элементов. При этом условии, как вы помните, наиболее устойчив метод прогонки (т.е. в ходе прогонки не возникнет больших ошибок округления); аналогичное условие при расчете определителей сводится к методу главного элемента; и это же условие мы получили для сходимости теперь уже итерационного метода Гаусса–Зейделя. Оно не столь строгое, как изначальное условие. А именно, первое условие было необходимым и достаточным, условие же преобладания диагональных членов — лишь достаточное, но уже не необходимое. Однако условие преобладания диагональных элементов легко обобщается на случай большего числа уравнений. А именно, в общем случае линейной системы из n уравнений с n неизвестными метод Гаусса–Зейделя сходится, если для каждого из уравнений выполняется неравенство

$$|a_{ii}| \geq |a_{i1}| + \ldots + |a_{i,i-1}| + |a_{i,i+1}| + \ldots + |a_{in}| \ ,$$

и если по крайней мере для одного из уравнений мы имеем строгое неравенство.

Таким образом, если вы имеете систему с преобладающими диагональными элементами, используйте для ее решения итерационный метод Гаусса–Зейделя.

Преимущества метода:

- надежность — сходимость метода не зависит от выбора начального приближения;
- отсутствие лавинообразного нарастания ошибок округления — возможность получения результата с любой (разумной) точностью.
- высокая скорость сходимости — метод Гаусса–Зейделя по сравнению с методом Якоби обычно требует вдвое меньше итераций для достижения требуемой точности.

II. Численное интегрирование.

Требуется вычислить определенный интеграл

$$J = \int_a^b y(x)dx \ .$$

Методы решения такой задачи.

1. Метод прямоугольников.
2. Аппроксимация рядом Тейлора.
3. Построение квадратурных формул:
   (a) Формулы Ньютона – Котеса.
   (b) Метод Чебышева.
   (c) Метод Лежандра – Гаусса (или просто Гаусса).
4. Метод Монте-Карло.

В пунктах 1 – 3 общий подход к решению задачи следующий. Определенный интеграл представляет собой площадь под кривой $y(x)$ между прямыми $x = a$ и $x = b$. Для вычисления интеграла J интервал (a, b) разбивается на n маленьких подинтервалов размером

$$h = \frac{b-a}{n} \ .$$

Как правило, n — достаточно велико (сотни, тысячи и т.п.). Длина интервалов h называется шагом интегрирования. Приближенно находится площадь каждой полоски s_i и найденные площади суммируются.

Метод МК стоит несколько особняком. В его основе лежат статистические понятия.

В нулевом приближении аппроксимация рядом Тейлора и все квадратурные формулы тождественны и представляют собой тривиальный...

Метод прямоугольников.

Разобьем интервал (a, b) на n равных отрезков длиной

$$h = \frac{b-a}{n} \ .$$

В качестве приближенного значения площади каждой полоски примем площадь прямоугольника, ширина которого равна h, а высота — значению функции $y(x)$ на левом краю интервала. Площадь i-ой полоски

$$s_i \simeq y_i h \ , \qquad \text{где} \quad y_i = y(x_i) \ ,$$

— это локальная формула. Тогда весь интеграл

$$J = \sum_{i=0}^{n-1} y_i h = (y_0 + y_1 + \ldots + y_{n-1})\, h \ .$$

Удивительно простая формула — формула левых прямоугольников. Аналогично строятся формулы правых и центральных прямоугольников:

$$s_i \simeq y_{i+1} h \ , \qquad J = \sum_{i=1}^{n} y_i h = (y_1 + y_2 + \ldots + y_n)\, h \ ;$$

$$s_i \simeq y_{i+1/2} h \ , \qquad J = \sum_{i=0}^{n-1} y_{i+1/2} h = (y_1 + y_2 + \ldots + y_n)\, h \ , \qquad \text{где} \qquad y_{i+1/2} = y(x_i + h/2) \ .$$

Вот это как раз и называется квадратурными формулами. Более строго: квадратурной формулой называется формула

$$J \simeq \sum_{k=0}^{m} \sum_{i=0}^{n} c_{ki} y^{(k)}(x_{ki}) \ ,$$

которая определяет приближенное значение интеграла как сумму значений подинтегральной функции (и ее производных) с некоторыми весовыми коэффициентами c_{ki} в некоторых точках x_{ki}, которые называются узлами.

Оценим погрешность метода левых прямоугольников. На каждом шаге погрешность пропорциональна

$$e_i \sim h \cdot (y_{i+1} - y_i) \sim h \cdot (y_i + y_i' h - y_i) \sim h^2 \ .$$

Но это на одном шаге. На нескольких шагах погрешности имеют обыкновение складываться. По крайней мере, мы должны предполагать худшее...

$$E \sim n \cdot e_i \sim \frac{b-a}{h} h^2 \sim h \ .$$

Погрешность пропорциональна величине шага по x. Это очень много. Такова же погрешность в методе правых прямоугольников, а вот в методе центральных прямоугольников уже $E \sim h^2$ (!), что гораздо лучше. Как уменьшить погрешность расчета? Это можно сделать по-разному. Первый способ.

Аппроксимация рядом Тейлора.

Этот метод хорош, когда не составляет труда вычисление производных от интегрируемой функции $y(x)$. Если вы на i-ом шаге знаете и значение функции y_i, и значение производной y_i', то площадь i-ой полоски

$$s_i \simeq \int_{x_i}^{x_{i+1}} \left(y_i + y_i'(x - x_i) + O\left((x - x_i)^2\right) \right) dx = y_i h + \frac{y_i'}{2} h^2 + O\left(h^3\right) \ .$$

Поскольку мы пренебрегли в итоге членами $\sim h^3$, то погрешность расчета в этом случае будет

$$e_i \sim h^3 \ , \qquad E \sim h^2 \ .$$

Это уже лучше. Далее можно подключить вторую производную y_i''. Получим

$$s_i = y_i h + \frac{y_i'}{2} h^2 + \frac{y_i''}{6} h^3 + \ldots \ .$$

Суммарная погрешность расчета при подключении второй производной

$$E \sim h^3 \ .$$

И т.д.

$$s_i = y_i h + \frac{y_i'}{2} h^2 + \frac{y_i''}{6} h^3 + \ldots + \frac{y_i^{(k)}}{(k+1)!} h^{k+1} + O(h^{k+2}) \ .$$

Чем больше производных вы примете во внимание, тем большей точности вы сможете достичь. Если вы сможете определить весь ряд Тейлора для площади s_i, вы получите точное значение интеграла. Если же вы точно учли первые несколько слагаемых, то погрешность вашего расчета определяется первым из отброшенных (неучтенных) членов. Последнее дает замечательный критерий сравнения для любых других методов. Вы можете не использовать тейлоровское разложение для расчета интеграла, а использовать какой-то другой метод, но чтобы оценить погрешность вашего расчета вам нужно сравнить выражение для площади i-ой полоски, которое получилось в вашем методе, с точным тейлоровским разложением, и вы тут же увидите, какова ваша погрешность. Именно так мы и будем поступать при изучении других методов.

Что же касается непосредственного применения метода аппроксимации рядом Тейлора. Главным недостатком такого метода является необходимость расчета производных интегрируемой функции. Иногда функцию $y(x)$ бывает очень трудно продифференцировать. Вы можете получить очень сложные выражения, которые сильно усложнят вашу программу и замедлят скорость расчета. А бывают ситуации, когда производную рассчитать просто невозможно, например, интегрируемая функция задана в виде таблицы, или каждое текущее ее значение получается в ходе каких-то предварительных вычислений.

Оказывается, что достичь тех же результатов можно и не рассчитывая производных от функции $y(x)$.

Формулы Ньютона – Котеса.

Давайте теперь в качестве площади i-ой полоски примем не площадь прямоугольника $y_i h$, а пойдем чуть дальше (чуть точнее) — площадь трапеции

$$s_i \simeq \frac{y_i + y_{i+1}}{2} h \ . \qquad (*)$$

Суммируя площади всех трапеций получим

$$J \simeq \frac{h}{2} \left(y_0 + 2y_1 + 2y_2 + \ldots + 2y_{n-1} + y_n \right) \ .$$

Эту квадратурную формулу называют правилом трапеций для численного интегрирования. Эта же формула является первой из формул Ньютона – Котеса. Формулу прямоугольников можно назвать нулевой формулой Ньютона – Котеса, ввиду их идейного родства.

Оценим погрешность правила трапеций. Для этого аппроксимируем значение y_{i+1} в формуле (*) тейлоровским рядом в i-ой точке

$$y_{i+1} = y_i + y_i' h + \frac{y_i''}{2} h^2 + \ldots \ .$$

Подставляя это, для i-ой полоски получаем

$$s_i = y_i h + \frac{y_i'}{2} h^2 + \frac{y_i''}{4} h^3 \ .$$

Это вплоть до членов $\sim h^2$ в точности совпадает со строгим тейлоровским разложением. Следовательно, погрешность в методе трапеций

$$e_i \sim h^3 \ , \qquad E \sim h^2 \ .$$

Теперь смотрите, что происходит. В методе прямоугольников мы предположили, что в окрестности i-го узла функция $y(x)$ постоянна и просто равна своему значению в этом i-ом узле. В методе трапеций мы предположили, что в окрестности i-го узла функция линейна по x:

$$y = y_i + a(x - x_i) \ ,$$

а неизвестный коэффициент этой линейной аппроксимации мы установили по значению функции в соседней точке. Это сразу же позволило уменьшить погрешность расчета на порядок. Можно пойти еще дальше. Теперь мы предположим, что в окрестности i-го узла функция квадратична по x, т.е. имеет вид

$$y = y_0 + a(x - x_0) + b(x - x_0)^2 \ .$$

Для определения трех неизвестных коэффициентов необходимо использовать три узла: при $x = x_i$, x_{i+1} и еще один дополнительный. Поделим интервал (x_i, x_{i+1}) на два, и возьмем дополнительный узел в появившейся точке $x = x_{i+1/2} = x_i + h/2$. Теперь коэффициенты квадратичной аппроксимации $y(x)$ найдем по значению функции в трех точках. В качестве x_0 удобно (но не обязательно) взять центральную точку, тогда

$$y = y_{i+1/2} + a(x - x_{i+1/2}) + b(x - x_{i+1/2})^2 \ . \qquad (*)$$

По сути мы уменьшили шаг интегрирования в два раза. Теперь $h = x_{i+1/2} - x_i$, и

$$\left\{ \begin{array}{l} y_i = y_{i+1/2} - ah + bh^2 \\ y_{i+1} = y_{i+1/2} + ah + bh^2 \end{array} \right. \quad \Longrightarrow \quad \begin{array}{l} y_{i+1} + y_i = 2y_{i+1/2} + 2bh^2 \\ y_{i+1} - y_i = 2ah \end{array}$$

От сюда коэффициент b:

$$b = \frac{1}{2h^2} \left(y_i - 2y_{i+1/2} + y_{i+1} \right) \ ,$$

а коэффициент a вообще нам не понадобится. Площадь i-ой полоски

$$s_i = \int_{x_i}^{x_{i+1}} y(x)dx = y_{i+1/2} \cdot 2h + b \cdot \frac{2}{3} h^3 = \frac{h}{3} \left(y_i + 4y_{i+1/2} + y_{i+1} \right) \ . \qquad (**)$$

Весь же интеграл дается выражением

$$J = \frac{h}{3} \sum_{i=1,3,5,\ldots}^{n-1} \left(y_i + 4y_{i+1/2} + y_{i+1} \right) \ .$$

Замечу, что количество интервалов для такого метода должно быть четным. Что это означает? По существу, происходит следующее. Сначала вы делите ваш отрезок (a, b) на некоторое число одинаковых интервалов, а потом каждый из этих интервалов делите пополам и используя значения функции в трех заданных точках аппроксимируете на каждом интервале вашу функцию полиномом второй степени. Замечу, что эта же процедура лежит в основе т.н. сплайн-интерполяции. В итоге

$$J = \frac{h}{3} \left(y_0 + 4y_{1/2} + 2y_1 + 4y_{1+1/2} + 2y_2 + \ldots + 2y_{n-1} + 4y_{n-1/2} + y_n \right) \ .$$

Получили знаменитую и самую оптимальную для численного интегрирования формулу Симпсона. Полная погрешность в методе прямоугольников была $E \sim h$, в методе трапеций $E \sim h^2$. Можно ожидать, что в следующем приближении (по формуле Симпсона) мы

получим $E \sim h^3$. Дак вот это не так. Если расписать значения $y_{i+1/2}$ и y_{i+1} в формуле (**) по формуле Тейлора в окрестности i-го узла, то мы получим для формулы Симпсона

$$E \sim h^4 \ .$$

Как так получилось? Вспомните, что коэффициент a квадратичной зависимости (*) нам не понадобился. Аналогично нам не понадобился бы и коэффициент c перед членом x^3, да и вообще любой коэффициент перед нечетной степенью x. Это означает, что учитывая квадратичные члены мы сразу повышаем точность не на один а на два порядка. Эти свойства формулы Симпсона: ее незначительное усложнение по сравнению с методом трапеций или прямоугольников и, в то же время, значительное повышение точности расчета — делают формулу Симпсона самым распространенным методом численного интегрирования.

Формула Симпсона является второй из семейства формул Ньютона – Котеса. Теперь вы уже можете догадаться как выводятся все остальные. Следующая формула (т.е. третья) называется правилом трех восьмых:

$$\begin{aligned} J = \frac{3}{8}h(y_0 + \ & 3y_{1/3} + 3y_{2/3} + 2y_1 + 3y_{1+1/3} + 3y_{1+2/3} + 2y_2 + 3y_{2+1/3} + \dots \\ + \ & 3y_{n-4/3} + 2y_{n-1} + 3y_{n-2/3} + 3y_{n-1/3} + y_n) \ . \end{aligned}$$

Для построения этой формулы функция аппроксимируется полиномом третьей степени. Коэффициенты полинома находятся по значениям функции в четырех соседних узлах: x_i, $x_{i+1/3}$, $x_{i+2/3}$, x_{i+1}. Т.е. весь интервал изначально разбивается на несколько равных интервалов, а затем каждый из них делится на три равные части. И по значению функции в четырех полученных точках строится полином третьей степени. Ввиду подмеченного нами выпадания нечетных степеней x погрешность расчета по этой формуле оказывается такой же, как и по формуле Симпсона $E \sim h^4$. А вот при использовании четвертой формулы Ньютона – Котеса

$$\begin{aligned} J = \frac{2}{45}h(7y_0 + \ & 32y_{1/4} + 12y_{1/2} + 32y_{3/4} + 14y_1 + 32y_{1+1/4} + 12y_{1+1/2} + \dots \\ + \ & 14y_{n-1} + 32y_{n-3/4} + 12y_{n-1/2} + 32y_{n-1/4} + 7y_n) \end{aligned}$$

погрешность уменьшается опять на два порядка $E \sim h^6$. Однако обычно формулы Симпсона вполне достаточно, и более высокие формулы Ньютона – Котеса не нужны. Хотя все можем статься... при необходимости, я думаю, вы вспомните об их существовании и найдете их в соответствующей литературе.

Теперь, что объединяет все формулы Ньютона – Котеса? Изначальный (большой) отрезок разбивается на ряд маленьких интервалов. Затем в рамках каждого полученного маленького интервала искомая функция аппроксимируется полиномом некоторой степени n. Для этого каждый интервал мы делили на n равных частей, после чего легко получали для каждого интервала некоторую квадратурную формулу

$$s_i = \sum_{i=0}^{n} c_i y(x_i) \ ,$$

с узлами, координаты которых нам известны заранее (они расположены на одинаковом расстоянии друг от друга), и единственное что мы делали — определяли необходимые весовые коэффициенты. При этом полученная в итоге квадратурная формула дает в общем случае приближенное значение интеграла. Погрешность можно определить, сопоставляя результат с точным тейлоровским разложением. Но если подинтегральная функция представляет собой полином степени n, то наша квадратурная формула даст точное значение

интеграла. Это позволяет сформулировать следующее условие для определения неизвестных весовых коэффициентов c_i: весовые коэффициенты квадратурной формулы определяются таким образом, что полученная в итоге формула точно интегрирует любой полином степени n.

У нас имеется $n+1$ неизвестное — коэффициенты произвольного полинома степени n и $n+1$ свободный параметр — весовые коэффициенты $n+1$ узла. Поэтому задача однозначно решалась. Но кто сказал, что в качестве свободных параметров необходимо использовать именно весовые коэффициенты? Никто. И нашлись люди, которые пошли другим путем. Мы последуем за ними.

Метод Чебышева.

Итак, мы должны построить квадратурную формулу для определенного интеграла

$$s_i = \int_{x_i}^{x_{i+1}} y(x)dx \ .$$

Для упрощения дальнейших записей мы от интервала (x_i, x_{i+1}) перейдем к интервалу $(-1, +1)$. Это всегда можно сделать простой заменой переменных. Итак, мы ищем квадратурную формулу для интеграла

$$s_i = \int_{-1}^{+1} y(x)dx \simeq \sum_{i=1}^{n} c_i y(x_i) \ .$$

В чем состоит другой путь, по которому пошел т.Чебышев. В качестве свободных параметров теперь будут не весовые коэффициенты, а координаты узлов. Весовые же коэффициенты пусть будут одинаковы и равны некоторой константе c. Мы имеем таким образом $n+1$ свободный параметр — это n координат x_i и общий весовой коэффициент c — и, следовательно, максимум на что мы можем рассчитывать это точно проинтегрировать полином степени n. Этим условием мы и воспользуемся, чтобы определить неизвестные свободные параметры. Мы требуем, чтобы наша квадратура точно интегрировала любой полином степени n, т.е. чтобы выполнялось условие

$$\int_{-1}^{+1} Q_n(x)dx = \sum_{i=1}^{n} c_i Q(x_i) \ , \quad Q_n(x) = q_0 + q_1 x + q_2 x^2 + \ldots + q_n x^n \ , \quad \text{для любых} \quad q_1, q_2, \ldots, q_n \ .$$

Простейшим базисом в пространстве полиномов степени n является набор следующих полиномов

$$\{n\} \ : \quad x^0 \ , \ \ x^1 \ , \ \ x^3 \ , \ \ \ldots \ , \ \ x^n \ .$$

Что означает, когда я произношу слово "базис"? Любой полином степени n можно представить как линейную комбинацию базисных полиномов (мы именно так и делаем, когда записываем полином). Теперь нетрудно доказать следующее утверждение: квадратура точно интегрирует любой полином степени n тогда и только тогда, когда она точно интегрирует любой элемент базиса. Действительно, докажем достаточность: пусть некая квадратурная формула с определенными весовыми коэффициентами c_i и узлами x_i точно интегрирует любой элемент базиса $\{n\}$, тогда эта квадратурная формула точно интегрирует любой полином степени n. Доказательство:
имеем

$$\forall j \ : 0 \le j \le n \quad \Rightarrow \quad \int_{-1}^{+1} x^j dx = \sum_{i=1}^{n} c_i x_i^j \ .$$

Получаем:

$$\int_{-1}^{+1} Q_n(x)dx = \int_{-1}^{+1} \sum_{j=0}^{n} q_j x^j dx = \sum_{j=0}^{n} q_j \int_{-1}^{+1} x^j dx =$$
$$= \sum_{j=0}^{n} q_j \sum_{i=1}^{n} c_i x_i^j = \sum_{i=1}^{n} c_i \sum_{j=0}^{n} q_j x_i^j = \sum_{i=1}^{n} c_i Q_n(x_i) \ .$$

Необходимость же следует из того, что каждый из элементов базиса является частным примером полинома степени n.

Таким образом для определения свободных параметров получаем систему из $n+1$ уравнения (условия того, что наша квадратура точно интегрирует каждый из элементов базиса)

$$\int_{-1}^{+1} x^0 dx = \sum_{i=1}^{n} c_i \ , \quad \to \quad nc = 2 \ , \quad \text{т.е.} \quad c = \frac{2}{n} \ ,$$
$$\int_{-1}^{+1} x^1 dx = \sum_{i=1}^{n} c_i x_i^1 \ , \quad \to \quad x_1 + x_2 + \ldots + x_n = 0 \ ,$$
$$\int_{-1}^{+1} x^2 dx = \sum_{i=1}^{n} c_i x_i^2 \ , \quad \to \quad x_1^2 + x_2^2 + \ldots + x_n^2 = \frac{2}{3c} \ ,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\int_{-1}^{+1} x^n dx = \sum_{i=1}^{n} c_i x_i^n \ , \quad \to \quad x_1^n + x_2^n + \ldots + x_n^n = \frac{1}{c} \frac{1 + (-1)^n}{n+1} \ .$$

Последние n уравнений определяют координаты точек, по которым строится квадратура. Решить эту систему для $n = 2$ просто. Получим

$$n = 2 \quad \to \quad x_2 = -x_1 = \frac{1}{\sqrt{3}} \ .$$

Для $n = 3$ посложнее:

$$n = 3 \quad \to \quad x_3 = -x_1 = \frac{1}{\sqrt{2}} \ , \quad x_2 = 0 \ .$$

Заметьте, что координаты располагаются симметрично относительно нуля. Еще сложнее решить эту систему для $n > 3$, да это и не нужно. Это давно уже сделано. Решения этой системы можно взять из справочника. Но ... вещественные решения этой системы существуют только для $n < 8$ и $n = 9$, при других n система не имеет вещественных решений. В этом основной недостаток метода Чебышева.

Достоинство: если функция вычисляется с некоторой погрешностью, то эта погрешность не умножается на большой весовой коэффициент, как это происходит в старших методах Ньютона–Котеса. Но ... точность интегрирования при этом мы не повысили.

Оценим погрешность 2-х точечного метода Чебышева. Имеем 2 узла, т.е. 3 свободных параметра. Следовательно, должны точно интегрировать любой полином степени 2, т.е. $Q_2(x)$. Погрешность должна быть $e_i \sim \int x^3 dx = h^4$. Однако, в действительности погрешность на порядок меньше, ввиду симметрии расположения узлов (вспомните метод Симпсона). Действительно, можем записать для 2-х точечного метода

$$s_i = y(-1/\sqrt{3}) + y(+1/\sqrt{3}) = 2y_0 + 2\frac{y_0''}{2}\frac{1}{3} + 2\frac{y_0^{(4)}}{4!}\frac{1}{9} + \ldots \ .$$

С другой стороны, используя аппроксимацию рядом Тейлора, относительно точки $x = 0$, имеем

$$s_i = \int_{-1}^{+1} y(x)dx = 2y_0 + 2\frac{y_0''}{2}\frac{1}{3} + 2\frac{y_0^{(4)}}{5!} + O(h^6) \ .$$

Видим, что различие $\sim h^5$. Таким образом, 2-х точечный метод Чебышева — аналог методу Симпсона. Таким образом, m-ая формула Ньютона – Котеса (при $m+1$ узле) точно интегрирует полином m-ой степени (для четных m — $m+1$-ой степени), также как и m-ая формула Чебышева (с $m+1$ своб. параметром). Вот если при том же количестве узлов мы смогли бы точно проинтегрировать полином более высокой степени, это было бы круто. И что самое важное, этого можно достичь и это достигается в методе Лежандра – Гаусса.

Метод Лежандра – Гаусса.

Решаем ту же задачу, что и в методе Чебышева, т.е. пытаемся аппроксимировать интеграл квадратурной формулой

$$J = \int_{-1}^{+1} y(x)dx \simeq \sum_{i=1}^{n} c_i y(x_i) \ .$$

В формулах Ньютона – Котеса мы искали весовые коэффициенты c_i при заданных координатах. В методе Чебышева — искали координаты узлов при одинаковых весовых коэффициентах. Теперь же мы будем считать свободными параметрами и весовые множители и координаты узлов. Это дает при n узлах $2n$ свободных параметра. Т.е. мы можем надеяться, что мы сможем точно проинтегрировать полином степени $(2n-1)$. Это — абсолютный максимум, чего можно достичь, используя информацию об n узлах подинтегральной функции. Поэтому метод Лежандра – Гаусса называют методом наивысшей алгебраической точности.

Теперь посмотрим, как же достичь этой наивысшей алгебраической точности. Как и в методе Чебышева, му требуем, чтобы наша квадратура точно интегрировала любой полином степени теперь уже $(2n-1)$, т.е. чтобы выполнялось условие

$$\int_{-1}^{+1} Q_{(2n-1)}(x)dx = \sum_{i=1}^{n} c_i Q_{(2n-1)}(x_i) \ , \qquad Q_{(2n-1)}(x) = q_0 + q_1 x + \ldots + q_{(2n-1)} x^{2n-1} \ .$$

Теперь нашим базисом в пространстве полиномов степени $(2n-1)$ является набор

$$x^0 \ , \quad x^1 \ , \quad x^3 \ , \quad \ldots \ , \quad x^{2n-1} \ .$$

Опять же я утверждаю: квадратура точно интегрирует любой полином степени $(2n-1)$ тогда и только тогда, когда она точно интегрирует любой элемент базиса. Доказательство этого мы уже провели в методе Чебышева. Таким образом для определения свободных параметров получаем систему из $2n$ уравнений (условия того, что наша квадратура точно интегрирует каждый из элементов базиса)

$$\int_{-1}^{+1} x^0 dx = \sum_{i=1}^{n} c_i = 2 \ ,$$

$$\int_{-1}^{+1} x^1 dx = \sum_{i=1}^{n} c_i x_i^1 = 0 \ ,$$

$$\int_{-1}^{+1} x^2 dx = \sum_{i=1}^{n} c_i x_i^2 = \frac{2}{3} \ ,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\int_{-1}^{+1} x^{2n-2} dx = \sum_{i=1}^{n} c_i x_i^{2n-2} = \frac{2}{2n-1} \ ,$$

$$\int_{-1}^{+1} x^{2n-1}dx = \sum_{i=1}^{n} c_i x_i^{2n-1} = 0 \ .$$

Записанная система $2n$ уравнений, казалось бы, еще сложнее, чем в методе Чебышева, и решить ее еще труднее. Но ... здесь как раз мы подходим к самой изюминке метода Гаусса, или, точнее, метода Лежандра – Гаусса. Имя Лежандра упоминается в этом методе не с проста.

Что мы теперь сделаем? Мы представим наш произвольный полином $Q_{2n-1}(x)$, интеграл от которого мы вычисляем, в виде

$$Q_{2n-1}(x) = Q'_{n-1}(x)P_n(x) + Q''_{n-1}(x) \ ,$$

где $P_n(x)$ — полином Лежандра степени n. Замечу, что такое разложение всегда можно сделать для любого полинома, и надеюсь, что вы знаете как это делается (?). Полином Q' — частное от деления на полином Лежандра, а Q'' — остаток. Не грех напомнить, что такое полиномы Лежандра:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[\left(x^2-1\right)^n\right] \ .$$

Основные свойства полиномов Лежандра:

- $P_n(1) = 1$, $P_n(-1) = (-1)^n$.
- На интервале $(-1;+1)$ полином P_n имеет n действительных корней.
- Ортогональность. $\int_{-1}^{+1} P_n(x)Q_k(x)dx = 0$, для любого полинома $Q_k(x)$ степенью $k < n$.

К чему теперь сводится наше условие, которое должно определять квадратурную формулу (т.е. весовые коэффициенты и координаты узлов).

$$\int_{-1}^{+1}\left(Q'_{n-1}(x)P_n(x) + Q''_{n-1}(x)\right)dx = \sum_{i=1}^{n} c_i\left(Q'_{n-1}(x_i)P_n(x_i) + Q''_{n-1}(x_i)\right) \ .$$

Интеграл от первого слагаемого в левой части тождественно обращается в нуль в силу свойства ортогональности полиномов Лежандра, а сумма от первого слагаемого в правой части обратится в нуль, если в качестве n узлов нашей квадратуры взять n действительных корней полинома Лежандра. При таком выборе узлов записанное условие сводится к соотношению

$$\int_{-1}^{+1} Q''_{n-1}(x)dx = \sum_{i=1}^{n} c_i Q''_{n-1}(x_i) \ ,$$

которое дает n уравнений для определения n оставшихся неизвестных, т.е. весовых коэффициентов. Это как раз первые n уравнений записанной нами ранее системы. Тут можно заметить, что если относительно координат узлов данная система нелинейна, т.е. ее очень трудно решать, а как показывает метод Чебышева во многих случаях и невозможно достичь решения в области действительных чисел, то относительно весовых коэффициетнов при заданных числах x_i мы имеем систему линейных алгебраических уравнений, решение которой не представляет никаких проблем.

Возьмем для примера $n = 2$.

$$P_2(x) = \frac{1}{2}\left(3x^2-1\right) \ .$$

Корни: $x_{1,2} = \pm 1/\sqrt{(3)}$. Весовые коэффициенты находим из системы

$$c_1 + c_2 = 2 \ ,$$
$$c_1 \frac{-1}{\sqrt{3}} + c_2 \frac{+1}{\sqrt{3}} = 0 \ .$$

Получаем $c_1 = c_2 = 1$.

В случае двух узлов методы Гаусса и Чебышева оказываются тождественны. При большем количестве узлов метод Гаусса дает уже разные весовые коэффициенты и, естественно, результаты отличаются уже от метода Чебышева. Приведу без доказательства значения весов для n-точечного метода Гаусса:

$$c_i = \frac{2}{(1 - x_i^2)\left[P_n'(x_i)\right]^2} \ .$$

Достоинства метода:

- Точно проинтегрировать полином степени $(2m - 1)$ — это максимум, чего можно достичь, используя информацию о m узлах подинтегральной функции. Поэтому метод Лежандра – Гаусса называют методом наивысшей алгебраической точности.

Метод Монте-Карло.

Пусть на интервале (a, b) задана последовательность случайных чисел $\{x_i\}$ с законом распределения вероятностей $f_x(x)$. Если подвергнуть эту последовательность функциональной обработке

$$y = y(x) \ ,$$

то математическое ожидание величины y дается соотношениями

$$M_y = \frac{\int_a^b y(x) f_x(x) dx}{\int_a^b f_x(x) dx} = \frac{1}{n} \sum_{i=1}^{n} y_i \ .$$

Чтобы получить нужный нам интеграл, достаточно рассмотреть мат. ожидание величины y/f_x

$$M_{y/f_x} = \frac{\int_a^b y(x) dx}{\int_a^b f_x(x) dx} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{f_x(x_i)} \ .$$

В простейшем случае используют равномерный закон распределения

$$f_x(x) = const \ ,$$

и нормируют функцию распределения, исходя из условия,

$$\int_a^b f_x(x) = 1 \ .$$

Это дает

$$f_x(x) = \frac{1}{b - a} \ .$$

В итоге получаем

$$J = \int_a^b y(x)dx = \frac{b-a}{n}\sum_{i=1}^{n} y(x_i) \ .$$

Погрешность определяется объемом выборки и согласно статистическому анализу

$$\Delta J \sim \frac{1}{\sqrt{n}} \ .$$

Как отмечается в литературе, метод Монте-Карло имеет некоторые преимущества перед другими методами при вычислении многократных интегралов.

Достоинства метода:

- Возможность остановить вычисления при любом значении n.
  Оказывается предпочтительным перед другими методами при вычислении многократных интегралов, особенно при наличии сложной области интегрирования.

Недостатки метода:

- Крайне низкая скорость уменьшения погрешности.

Экстраполяционный переход к пределу.

Рассмотрим такую ситуацию. Вы написали программу заложив правило прямоугольников, провели расчеты и вдруг обнаружили, что уменьшая шаг интегрирования вы не можете достичь требуемой точности. Например, при шаге $h = 0.1$ расчет продолжался 3 часа, а при шаге вдвое меньшем $h = 0.05$ расчет шел уже, естественно, 6 часов. И сопоставив результаты вы обнаружили, что погрешность еще на порядок больше требуемого значения. Можно, конечно, уменьшить шаг на порядок и прождать 30 часов. Можно переписать программу заложив правило Симпсона. Но... можно почти ничего не делая получить весьма точное значение.

Для этого построим зависимость J от шага интегрирования. Поскольку для метода прямоугольников $E \sim h$, то мы получим линейную зависимость $J(h)$. Эту зависимость легко можно проэкстраполировать к значению $h = 0$. Получим

$$J = J_{0.05} + (J_{0.05} - J_{0.1}) \ .$$

Естественно, подобный экстраполяционный переход к пределу можно совершить не только в случае правила прямоугольников, но и в случае правила трапеций или правила Симпсона (но не для метода Монте-Карло). Нужно будет только не забыть, что там погрешность расчета пропорциональна, соответственно, h^2 и h^4.

Иногда ввиду сложности интеграла, например, двойные, тройные интегралы, сложно определить заранее характер зависимости погрешности E от величины шага интегрирования h. Тогда для применения экстраполяционного перехода к пределу предварительно требуется провести 3—4 расчета при разных шагах интегрирования и определить зависимость $E(h)$, или, что то же самое, зависимость $J(h)$.

Интегрирование несобственных интегралов.

Рассмотрим ситуацию, когда одним из пределов интегрирования является $\pm\infty$:

$$J = \int_a^{\infty} y(x)dx \ .$$

Методы интегрирования:

- Аналитическое интегрирование хвоста.
- Замена переменной.
- Экстраполяционный переход к пределу.

Аналитическое интегрирование хвоста. Требуется определить асимптотическое поведение функции $y(x)$ при $x \to \infty$. Часто оказывается, что асимптотика функции $y(x)$ (обозначим ее $\phi(x)$) является легко интегрируемой. Тогда исходный интеграл записывают в виде

$$J = \int_a^b y(x)dx + \int_b^\infty \phi(x)dx \ ,$$

b — значение x, при котором погрешность замены исходной функции $y(x)$ ее асимптотическим пределом пренебрежима мала. Теперь первый интеграл считают уже известными вам методами, а второй вычисляют аналитически.

Рассмотрим следующий пример.

$$J = \int_0^\infty \left(e^{-\frac{4}{\tau}\left(\frac{1}{x^{12}} - \frac{1}{x^6}\right)} - 1 \right) x^2 dx \ .$$

Здесь в степени экспоненты стоит т.н. леннард-джонсовский потенциал межмолекулярного взаимодействия. А вся экспонента представляет собой не что иное, как радиальную функцию распределения молекул в разреженном газе. Различные интегралы от этой функции частенько возникают в физике жидкостей и газов.

При $x \to \infty$ (например, при $x \geq 5$) в показателе экспоненты можно пренебречь первым слагаемым, а всю экспоненту разложить в ряд. Оставляя в скобках первый ненулевой член получим

$$J = \int_0^5 \left(e^{-\frac{4}{\tau}\left(\frac{1}{x^{12}} - \frac{1}{x^6}\right)} - 1 \right) x^2 dx + \int_5^\infty \left(\frac{4}{\tau}\frac{1}{x^6} \right) x^2 dx \ .$$

Вычисляя второй интеграл аналитически имеем

$$J = \int_0^5 \left(e^{-\frac{4}{\tau}\left(\frac{1}{x^{12}} - \frac{1}{x^6}\right)} - 1 \right) x^2 dx + \frac{4}{3\tau \cdot 5^3} + O\left((b=5)^{-9} \right) \ .$$

Недостатки метода:

- Бывают ситуации, когда невозможно представить асимптотическое поведение подинтегрального выражения какой-нибудь простой легко интегрируемой функцией. Например, интеграл $\int_a^\infty \frac{1}{x}\, e^{-x^2}\, dx$ таким методом не рассчитать.

Замена переменной. От ∞ в пределе интегрирования избавляются, переходя к новой переменной $x \to t$, такой чтобы при $x \to \infty$, $t(x) \to \text{const}$. Например,

$$J = \int_a^\infty \frac{e^{-x^2}}{x} dx \ .$$

Вводим $t = 1/x$:

$$J = \int_0^{1/a} \frac{e^{-1/t^2}}{t} dt \ .$$

А этот интеграл уже не представляет для вас проблем.

Недостатки метода:

- Требуемая замена переменной может слишком усложнить поведение подинтегральной функции на одном из пределов интегрирования. Например, замена $t = 1/x$ в интеграле: $\int_a^\infty \frac{dx}{\sqrt{1+x^3}}$ приводит к неаналитичности в точке $t = 0$: $\int_0^{1/a} \frac{dt}{\sqrt{t^4+t}}$.

Экстраполяционный переход к пределу. Вы в любом случае можете оборвать интегрирование на некотором значении $x = b$ и просто пренебречь оставшимся хвостом. Но ... вы должны построить зависимость величины интеграла J от обратного значения b

$$J = J(1/b) \ .$$

Анализируя эту зависимость вы, возможно, сможете установить вид функции $J = J(1/b)$, и значит легко проэкстраполируете ее на значение $1/b = 0$. Если же даже вы не установите вид функции, такой анализ поможет вам оценить погрешность, обусловленную обрезанием интеграла.

Недостатки метода:

- Зависимость $J(1/b)$ может оказаться неаналитичной в точке $1/b = 0$. Например, $J(1/b) = J_0 + \xi\sqrt{1/b} + \dots$.

Вычисление интегралов в нерегулярных случаях.

Нередко приходится вычислять интегралы от функций, имеющих те или иные особенности. Например, функция

$$y(x) = \frac{\exp(-x^2)}{\sqrt{x}}$$

интегрируема на участке $x \in (0, 1)$, но в точке $x = 0$ она расходится. Расходимости плохо аппроксимируются многочленами и поэтому для вычисления подобных интегралов применение стандартных квадратурных формул может оказаться неэффективным.

Методы:

1. Переход к обратной зависимости $x(y)$ — сведение к интегралу с бесконечным пределом.
2. Аналатическое интегрирование хвоста.
3. Выделение особенности: весовая функция, аддитивное выделение особенности.

Введение весовой функции.

Бывает полезно разложить интегрируемую функцию на два сомножителя. Искомый интеграл представляется в виде

$$J = \int_a^b \rho(x) f(x) dx \ ,$$

где $\rho(x)$ — весовая функция. Весовая функция должна выбираться так, чтобы она была интегрируема аналитически и содержала всю особенность исходной функции. В приведенном примере в качестве весовой функции разумно взять $\rho(x) = x^{-1/2}$. Далее на каждом подинтервале весовая функция интегрируется аналитически, а оставшаяся функция $f(x)$ интегрируется численно. Другими словами строится квадратура

$$J \simeq \sum_{i=0}^{n} c_i f(x_i) \ ,$$

где веса c_i определяются различными интегралами от весовой функции.

В нашем примере правило прямоугольников, вводя весовую функцию $\rho(x) = x^{-1/2}$, можно модифицировать следующим образом. По правилу правых прямоугольников мы бы записали (левые прямоугольники дадут деление на ноль)

$$s_i = \int_{x_i-h}^{x_i} \frac{\exp(-x^2)}{\sqrt{x}} dx \simeq \frac{\exp(-x_i^2)}{\sqrt{x_i}} h \ .$$

Теперь же мы запишем

$$s_i = \int_{x_i-h}^{x_i} \frac{\exp(-x^2)}{\sqrt{x}} dx \simeq \exp(-x_i^2) \int_{x_i-h}^{x_i} \frac{dx}{\sqrt{x}} = 2\exp(-x_i^2) \left(\sqrt{x_i} - \sqrt{x_i - h} \right) \ .$$

Вес c_i теперь определяется интегралом от весовой функции. Вдали от особенности $c_i \to h/\sqrt{x_i}$, и мы приходим к обычной формуле прямоугольников.

Приведу погрешность расчета интеграла в нашем примере по формуле прямоугольников с выделением весовой функции и без нее. В таблице: E_1 — правые прямоугольники, E_2 — прав. прям. с весовой функцией, E_3 — центр. прям., E_4 — центр. прям. с весовой функцией.

h	E_1	E_2	E_3	E_4
10^{-1}	0.44	5e-2	0.2	7e-4
10^{-2}	0.14	5e-3	6e-2	5e-6
10^{-3}	5e-2	5e-4	2e-2	4e-8
10^{-4}	1.4e-2	5e-5	6e-3	4e-10
	$\sim \sqrt{h}$	$\sim h$	$\sim \sqrt{h}$	$\sim h^2$

Аддитивное выделение особенности

Здесь особенность выделяется аддитивно:

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

Интеграл от первой функции (с особенностью) берется аналитически, а от второй — численно. Результаты аналогичны введению весовой функции.

III. Решение обыкновенных ДУ.

ОДУ:

$$f\left(x, y, y', y'', \ldots, y^{(k)}\right) = 0 \ ,$$

где k — порядок ОДУ. Так, ОДУ 1-го порядка:

$$f\left(x, y, y'\right) = 0 \ .$$

ОДУ имеет бесконечное множество решений. Для отыскания какого-либо конкретного решения требуются дополнительные условия. Эти условия могут быть двух типов:
1) Дополнительные условия задаются при одном значении независимой переменной, т.е. например при $x = a$ заданы значения функции y_0, и возможно некоторые производные искомой функции y'_0, y''_0 и т.д. В этом случае говорят, что поставлена задача Коши (задача с НУ).
2) Условия задаются при двух (или более) значениях независимой переменной. Такая задача называется краевой, а дополнительные условия — ГУ.

При решении этих задач используются разные методы и алгоритмы, и поэтому мы будем рассматривать эти задачи отдельно.

III.1. Методы решения задачи Коши.

Сформулируем простейшую задачу Коши. Требуется решить уравнение

$$y' = f(x, y) \ , \qquad \text{при НУ} \qquad y(x_0) = y_0 \ .$$

Несмотря на простоту исходной задачи, методы, которые мы рассмотрим, имеют гораздо более широкое приложение.
Поскольку:
1) Разрешить относительно y' мы уже можем (хотя бы численно);
2) Методы, которые мы рассмотрим, легко обобщаются на системы из нескольких ДУ первого порядка;
3) Уравнения высших порядков можно свести к системе уравнений первого порядка. Например, уравнения второго и третьего порядка

$$y'' = f(y', y, x) \ , \qquad y''' = f\left(y'', y', y, x\right)$$

эквивалентны системам уравнений первого порядка

$$\begin{cases} z' = f(z, y, x) \ , \\ y' = z \ . \end{cases} \qquad \begin{cases} \phi' = f(\phi, z, y, x) \ , \\ y' = z \ , \\ z' = \phi \ . \end{cases}$$

Мы рассмотрим следующие методы решения задачи Коши:

1. Аппроксимация рядом Тейлора,
2. Методы Рунге–Кутты,
3. Методы прогноза–коррекции.

Здесь методы Рунге–Кутты и прогноза-коррекции являются неким обобщением квадратурных формул интегрирования. Задача о решении обыкновенного ДУ во многом перекликается с задачей расчета определенного интеграла. Так, если исходная функция f не зависит от y, то решением ДУ просто является соответствующий определенный интеграл

$$y(x) = y(x_0) + \int_{x_0}^{x} f(x)dx \ .$$

Ввиду такого, я бы сказал, родства душ и методы численного решения обыкновенных ДУ во многом перекликаются с методами численного интегрирования.

Начнем мы с того, что по-проще. С того, что опять же является неким общим нулевым приближением для всех методов.

Метод Эйлера.

Подставим НУ в исходное ДУ:

$$y_0' = f(y_0, x_0) \ .$$

Получили тангенс угла наклона функции $y(x)$ в нач. точке. Теперь решение в следующей точке $x_1 = x_0 + h$ можно приближенно, конечно, представить в виде

$$y_1 = y_0 + y_0' \cdot h \ . \qquad (*)$$

Погрешность допускаемая при этом, как нетрудно догадаться,

$$e_0 \sim h^2 \ .$$

Теперь в качестве начальной используем только что полученную точку (y_1, x_1), и с ее помощью находим следующую точку (y_2, x_2), и т.д. Этот самый простой способ численного решения ДУ. Общая формула метода Эйлера

$$y_{i+1} = y_i + h \cdot f(y_i, x_i) \ .$$

Ошибка допускаемая нами на каждом шаге

$$e_i \sim h^2 \ .$$

Суммируя ошибки по всем шагам, получим

$$E \sim h \ .$$

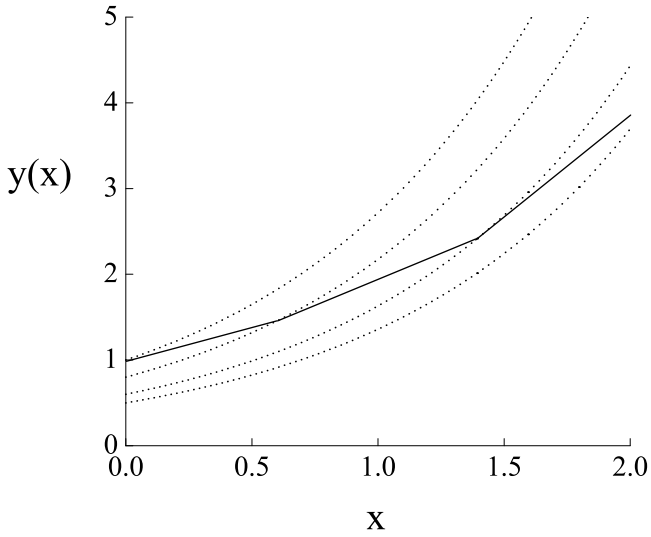
Можно отметить полную аналогию с методом прямоугольников для численного интегрирования. Действительно, если правая часть исходного ДУ не зависит от y, то значение искомой функции в точке $x_1 = x_0 + h$ определяется интегралом

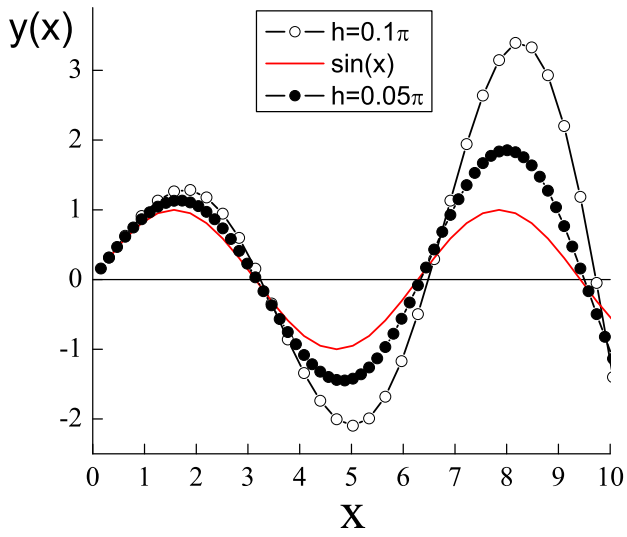
$$y_1 = y_0 + \int_{x_0}^{x_0+h} f(x)dx \ ,$$

и тогда формула (*) представляет собой не что иное, как формулу левых прямоугольников. Однако в отличие от интегрирования, где ошибки просто суммировались, при решении ДУ ошибка на предыдущем шаге порождает еще большую ошибку на следующем шаге, и нередко суммарная ошибка растет экспоненциально с количеством пройденных шагов. Пусть, например, мы решаем уравнение

$$y' = y \ , \quad y(0) = 1 \ .$$

Общее решение — семейство кривых (см. рис.) $y = A_0 e^x$. Нетрудно видеть, что двигаясь по прямым отрезкам, мы перескакиваем на все более низко расположенные кривые, т.е. на каждом шаге мы переходим к новой интегральной кривой все дальше и дальше удаляясь от истинной функции $y(x)$. Подобного эффекта не происходило при интегрировании. Теперь же это "перескакивание" приводит к дополнительному источнику увеличения суммарной погрешности.





Проанализируем, для метода Эйлера зависимость погрешности от координаты $E(x)$. Погрешность в $i+1$ узле

$$E(x_{i+1}) = y_{i+1} - \bar{y}_{i+1} = E(x_i) + h\left[f\left(x_i, y_i\right) - f\left(\bar{y}_i, x_i\right)\right] + \frac{h^2}{2}\ddot{y}_{[i;i+1]} \ .$$

Поскольку приращение функции можно представить, как

$$f\left(x_i, y_i\right) - f\left(\bar{y}_i, x_i\right) = f_{y,[i,i+1]}\left(y_i - \bar{y}_i\right) \ ,$$

то вводя обозначения

$$C_y = \max_{(x,y)} |f_y| \ , \qquad C_2 = \max_{(x,y)} |\ddot{y}| \ ,$$

можем оценить $E(x_{i+1})$ сверху:

$$\begin{aligned} |E(x_{i+1})| \ &\leq \ (1+hC_y)\,|E(x_i)| + \frac{h^2}{2}C_2 \leq (1+hC_y)^2\,|E(x_{i-1})| + (1+hC_y)\,\frac{h^2}{2}C_2 + \frac{h^2}{2}C_2 \\ &\leq \ldots \leq (1+hC_y)^{i+1}\,|E(x_0)| + \frac{h^2}{2}C_2\sum_{k=0}^{i}(1+hC_y)^i \ . \end{aligned}$$

Полагая начальную погрешность $E(x_0)$ равной нулю и используя формулу для суммы геометрической прогрессии

$$S_i = \sum_{k=0}^{i} a^k = \frac{1-a^{i+1}}{1-a} \ ,$$

приходим к

$$|E(x_{i+1})| \leq \frac{h}{2}\frac{C_2}{C_y}\left[(1+hC_y)^{i+1} - 1\right] \ .$$

Теперь заменяя $i+1$ на x_{i+1}/h в пределе $h \to 0$ получаем

$$|E(x_{i+1})| \leq \frac{h}{2}\frac{C_2}{C_y}\left[e^{xC_y} - 1\right] \ .$$

Последнее соотношение показывает, что при $C_y \neq 0$ погрешность E растет экспоненциально с координатой (при $C_y = 0$ получим $E \sim x$).

На рис. приведены расчеты для уравнения $y'' = -y$ (решение: $y = \sin(x)$.)

Единственный способ оценить масштабы набежавшей погрешности — провести несколько пробных расчетов с разными шагами по x.

При таком катастрофическом нарастании ошибок при решении ДУ чрезвычайно актуальной становится проблема повышения точности расчета. И первый из способов повышения точности...

Аппроксимация рядом Тейлора.

В чем суть этого метода? Для тех, кто еще не догадался поясняю. Помимо первой производной в точке i-ой точке мы можем вычислить вторую производную, продифференцировав исходное ДУ,

$$y'' = f_x + f_y \cdot y' \ .$$

Используя найденную вторую производную, мы можем более точно оценить значение функции в следующей точке:

$$y_{i+1} = y_i + y'_i h + \frac{y''_i}{2} h^2 + \dots \ .$$

При этом погрешность уже на порядок ниже

$$e_i \sim h^3 \ .$$

Далее, мы можем учесть третью производную, четвертую и т.д.

Метод хорош до тех пор, пока у вас, как и в случае численного интегрирования нет проблем с расчетом старших производных. Если же проблемы есть, то переходите к другому методу. И опять я должен сообщить, что тех же результатов (в смысле точности решения поставленной задачи) можно достичь не вычисляя никаких старших производных, а вычисляя только значения первой производной в различных точках (x, y).

Методы Рунге–Кутты.

Метод Эйлера является первым из семейства методов Рунге–Кутты. Говорят, он является методом первого порядка, т.к. согласуется с точным тейлоровским разложением вплоть до членов первого порядка малости по h. Чтобы получить методы Рунге–Кутты более высокого порядка точности, необходимо использовать значение производной $y'(x)$, вычисленной в нескольких точках плоскости (x, y). Посмотрим, какой точности можно достичь, если вычислять производную $y'(x)$ в двух точках. Первая точка — это, та где мы стоим в данный момент (x_m, y_m), а координаты второй точки обозначим как $(x_m + b_1 h, y_m + b_2 h y'_m)$. Если параметры b_1, b_2 положить равными 1, то это будет точка, которая является последующей в простейшем методе Эйлера. Сейчас же мы будем вычислять функцию в последующей точке по формуле

$$y_{m+1} = y_m + h \cdot \Phi\left(x_m, y_m, h\right) \ ,$$

где Φ — некоторое среднее производных в первой и во второй точках

$$\Phi\left(x_m, y_m, h\right) = a_1 f(x_m, y_m) + a_2 f(x_m + b_1 h, y_m + b_2 h y'_m) \ .$$

Параметры a_1, a_2 — некие веса наших точек. Наша с вами задача выбрать параметры a_1, a_2, b_1, b_2 такими, чтобы достичь максимального совпадения с точным тейлоровским

разложением. Для этого мы разложим функцию Φ в ряд Тейлора. В итоге для последующей точки получаем

$$y_{m+1} = y_m + h(a_1 + a_2)f + h^2 \left(a_2 b_1 f_x + a_2 b_2 f_y f\right) + O(h^3) \ .$$

Вспоминаем точное тейлоровское разложение для значения y в $(m+1)$-ой точке

$$y_{m+1} = y_m + h \cdot f + \frac{h^2}{2} \left(f_x + f_y f\right) + O(h^3) \ .$$

Сопоставляя коэффициенты при f, f_x и f_y, получаем три уравнения на определение наших четырех неизвестных

$$\begin{cases} a_1 + a_2 = 1 \ , \\ a_2 b_1 = \frac{1}{2} \ , \\ a_2 b_2 = \frac{1}{2} \ . \end{cases}$$

Обозначим параметр a_2 как $\omega \neq 0$, тогда решение системы запишется в виде

$$a_1 = 1 - \omega \ , \quad b_1 = \frac{1}{2\omega} \ , \quad b_2 = \frac{1}{2\omega} \ .$$

А для итераций, которые мы должны заложить в программу, получаем в итоге

$$y_{m+1} = y_m + h \cdot \left[(1-\omega) f(x_m, y_m) + \omega f\left(x_m + \frac{h}{2\omega}, y_m + \frac{h}{2\omega} f(x_m, y_m)\right)\right] + O(h^3) \ .$$

Это наиболее общая форма записи методов Рунге–Кутты второго порядка. Как видим, методов Рунге–Кутты второго порядка бесконечно много. Задавая различные значения параметру ω (от 0 до 1 — разумные пределы), получаем различные методы второго порядка точности. Наиболее известные из них: исправленный метод Эйлера ($\omega = 1/2$) и модифицированный метод Эйлера ($\omega = 1$). Но все это — методы второго порядка. Достичь третьего порядка точности с использованием двух точек невозможно (разве что случайно), поскольку для третьего порядка точности мы должны были бы потребовать равенства коэффициентов перед величинами f — для согласования первой производной y', f_x, $f_y f$ — для согласования y'' (это, то что мы уже сделали), а также f_{xx}, $f_{xy} f$, $f_{yy} f^2$, $f_y f_x$, $f_y^2 f$ — $y''' = f_{xx} + 2 f_{xy} f + f_{yy} f^2 + f_y f_x + f_y^2 f$. Восемь соотношений для четырех параметров — многовато. Если ввести в рассмотрение еще одну точку (третью) и определить среднее от трех производных, как

$$\Phi(x_m, y_m, h) = a_1 f(x_m, y_m) + a_2 f\left(x_m + b_1 h, y_m + b_2 h k_1\right) + a_3 f\left(x + c_1 h, y_m + c_2 h k_2\right) \ ,$$

$$k_1 = f(x_m, y_m) \ , \quad k_2 = f\left(x_m + b_1 h, y_m + b_2 h k_1\right) \ , \quad k_3 = f\left(x + c_1 h, y_m + c_2 h k_2\right) \ ,$$

то переменных становится семь. И система решается! Причем не все из получающихся восьми уравнений независимы, некоторые из них выражаются через другие, благодаря чему система не просто имеет решение, а, как можно показать, опять имеет бесконечный набор решений. Одним из этих решений, например, является следующий набор коэффициентов: $(a_1, a_2, a_3) = (2, 3, 4)/9$, $b_1 = b_2 = 1/2$, $c_1 = c_2 = 3/4$.

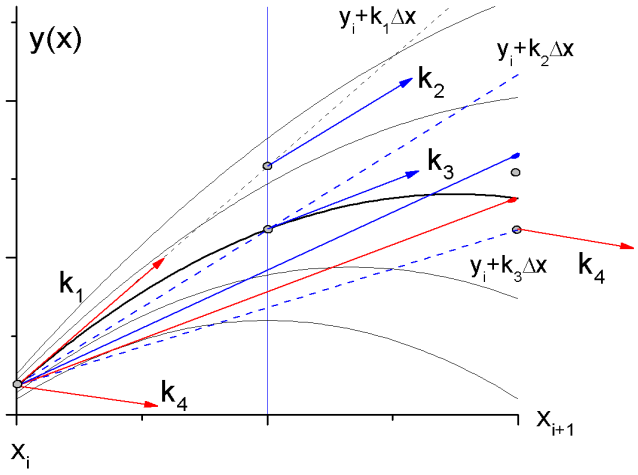
Я не буду приводить общих формул третьего или четвертого порядка, а выпишу лишь одну из самых распространенных формул четвертого порядка. Эта формула применяется настолько широко, что в литературе именно она подчас называется "методом Рунге–Кутты". Этот классический метод Рунге–Кутты описывается системой из пяти соотношений

$$y_{m+1} = y_m + \frac{h}{6} \left(k_1 + 2k_2 + 2k_3 + k_4\right) \ ,$$

где

$$k_1 = f(x_m, y_m) \ ,$$
$$k_2 = f(x_m + \tfrac{h}{2}, y_m + \tfrac{h}{2}k_1) \ ,$$
$$k_3 = f(x_m + \tfrac{h}{2}, y_m + \tfrac{h}{2}k_2) \ ,$$
$$k_4 = f(x_m + h, y_m + hk_3) \ .$$

Погрешность допускаемая на i-ом шаге: $O(h^5)$. Смысл коэффициентов k_i демонстрирует рис.



Еще раз подчеркну родство душ методов численного решения ОДУ и методов численного интегрирования. Если функция f не зависит от y, то решение ДУ — определенный интеграл:

$$y = y_0 + \int_{x_0}^{x} f(x)dx \ .$$

А по формуле Рунге–Кутты получаем

$$y_{m+1} - y_m = \frac{h}{6}\left[f\left(x_m\right) + 4f\left(x_m + \frac{h}{2}\right) + f\left(x_m + h\right)\right] \ .$$

Но ведь это не что иное, как площадь i-ой полоски

$$s_i = \int_{x_m}^{x_m+h} f(x)dx \ ,$$

если использавать формулу Симпсона! По этой причине классическую формулу Рунге–Кутты часто называют формулой Рунге–Кутты–Симпсона.

Методы Рунге–Кутты чрезвычайно широко используются и имеют неоспоримые достоинства, например, перед аппроксимацией рядом Тейлора. Однако же в них содержится одно нерациональное зерно, которое их губит. Для того, чтобы сделать очередной шаг вам необходимо вычислить значение функции $f(x, y)$ в нескольких промежуточных точках. При этом информация о предыдущих, уже пройденных, точках не используется. А ведь вместо расчета новых промежуточных точек можно использовать уже накопленную информацию — значения функции $f(x, y)$ в предыдущих точках. На этой идее построены, так называемые, ...

Методы прогноза и коррекции.

Как получаются методы прогноза второго порядка точности? Представим искомую функцию $y(x)$ в окрестности точки (x_m, y_m) в виде полинома 2-ой степени

$$y(x) = y_m + y'_m(x - x_m) + a(x - x_m)^2 \ ,$$

а коэффициенты a найдем, подключая информацию о предыдущей точке (сейчас мы используем y_{m-1}, а вообще-то можно и y'_{m-1})

$$y_{m-1} = y_m - y'_m h + ah^2 \qquad \rightarrow \qquad a = \frac{y_{m-1} - y_m + y'_m h}{h^2}.$$

В итоге получаем формулу прогноза второго порядка по точности

$$y^{(0)}_{m+1} = y_{m-1} + 2hy'_m \ .$$

Оценим погрешность прогноза, представив значение y_{m-1} рядом Тейлора в окрестности m-го узла,

$$e^{(0)} = \bar{y} - y^{(0)}_{m+1} \ = \ \left(y_m + y'_m h + y''_m \frac{h^2}{2} + y'''_m \frac{h^3}{6} + \dots\right) - \left(y_m + y'_m h + y''_m \frac{h^2}{2} - y'''_m \frac{h^3}{6} + \dots\right) = y'''_m \frac{h^3}{3} + \dots \ .$$

Я снабдил значение функции в последующей точке индексом "0", т.к. методы прогноза обычно используются в связке с методами коррекции, в ходе которых полученное нулевое приближение затем корректируется в ходе нескольких итераций, и получаются более точные значения $y^{(1)}$, $y^{(2)}$ и т.д. в точке $(m+1)$. Естественно, вы можете взять полином более высокой степени, и получите метод прогноза более высокого порядка точности.

Теперь о коррекции полученного значения y_{m+1}.
Методы коррекции тоже бывают разных порядков точности. Для их вывода нужно опять представить искомую функцию в окрестности m-го узла полиномом требуемой степени (смотря какую точность хотите получить)

$$y(x) = y_m + y'_m(x - x_m) + a(x - x_m)^2 + \dots \ ,$$

а коэффициенты этого полинома найти, подключая (если потребуется) информацию о предыдущих точках и (обязательно! чтобы получилась коррекция, а не прогноз) значение y'_{m+1} (именно производная, а не значение функции y_{m+1}). В нашем простейшем случае получаем

$$y^{(i)}_{m+1} = y_m + \frac{h}{2}\left(y'_m + y'^{(i-1)}_{m+1}\right) \ , \qquad y'^{(i-1)}_{m+1} = f\left(x_{m+1}, y^{(i-1)}_{m+1}\right) \ . \qquad (*)$$

Формулы такого типа, когда искомая величина определяется через свое значение на предыдущей итерации, называются неявными. Мы с вами уже имели дело с неявными формулами для итераций, когда решали уравнения и системы уравнений, и еще будем иметь дело, когда займемся ДУ в ЧП. Итерации по формуле (*) продолжаются до выполнения условия

$$\left|y^{i}_{m+1} - y^{i-1}_{m+1}\right| < \varepsilon \ .$$

И сразу же возникает вопрос, удастся ли вообще когда-нибуть удовлетворить записанному условию прекращения итераций, иными словами, сходится ли процесс коррекции вообще? Для ответа на этот вопрос запишем формулу коррекции для $i+1$ итерации и составим разность

$$y^{(i+1)}_{m+1} - y^{(i)}_{m+1} = \frac{h}{2}\left[f(x_{m+1}, y^{(i)}_{m+1}) - f(x_{m+1}, y^{(i-1)}_{m+1})\right] = \frac{h}{2}\left(\frac{\partial f}{\partial y}\right)\left[y^{(i)}_{m+1} - y^{(i-1)}_{m+1}\right] \ .$$

Теперь видно, что коррекция сходится при выполнении условия

$$\frac{h}{2}M < 1 \ , \quad \text{где} \quad M = \max_{x,y}\left|\left(\frac{\partial f}{\partial y}\right)\right| \ .$$

В ходе итераций (как правило достаточно 2–3 итераций) вы получите, некоторую стационарную точку такого (*) отображения. Причем эта стационарная точка совсем не обязательно дает истинное значение y в $(m+1)$-ой точке. Посмотрим с какой погрешностью мы получили решение на этом этапе. Стационарная точка коррекции точно удовлетворяет условию

$$y^{(\infty)}_{m+1} = y_m + \frac{h}{2}\left[f(x_m, y_m) + f(x_{m+1}, y^{(\infty)}_{m+1})\right] \ .$$

Истинное решение

$$\bar{y}_{m+1} = y^{(\infty)}_{m+1} + e_{kor} \ .$$

Выразим от сюда стационарную точку и подставим в формулу коррекции:

$$\bar{y}_{m+1} - e_{kor} = y_m + \frac{h}{2}\left[f(x_m, y_m) + f(x_{m+1}, \bar{y}_{m+1} - e_{kor})\right] \ .$$

Погрешностью e_{kor} в аргументе f можно пренебречь. Аппроксимируя теперь величины $\bar{y}_{m+1}$ и $f(x_{m+1}, \bar{y}_{m+1}) = \dot{\bar{y}}_{m+1}$ рядами Тейлора легко показать, что

$$e_{kor} = -y'''_m \frac{h^3}{12} + \dots \ .$$

Разница между стационарной точкой и истинным решением пропорциональна $O(h^3)$, если вы используете методы прогноза и коррекции второго порядка точности. Зачем же, спросите вы, мы использовали коррекцию, если порядок погрешности мы за счет нее не уменьшили? Действительно, если вы остановитесь на достигнутом, то лучше бы вы вообще не вспоминали о методах коррекции. Вся прелесть методов прогноза-коррекции состоит в том, что если вы используете методы одного порядка точности, то можно точно оценить погрешность следующих (неучтенных нами) членов малости, и тем самым уменьшить погрешность на порядок. Так в нашем случае мы получили

$$\bar{y}_{m+1} = y^{(0)}_{m+1} + y'''_m \frac{h^3}{3} \ ,$$

$$\bar{y}_{m+1} = y^{(\infty)}_{m+1} - y'''_m \frac{h^3}{12} \ ,$$

Вычитая одно из другого мы выразим неизвестную нам третью производную (а значит и погрешность наших расчетов) через известные нам величины

$$e_{kor} = -y'''_m \frac{h^3}{12} = \frac{y^{(0)}_{m+1} - y^{(\infty)}_{m+1}}{5} \ .$$

Таким образом, если вы после проведенных корректирующих итераций возьмете значение

$$y_{m+1} = y^{\infty}_{m+1} + \frac{1}{5}\left(y^{(0)}_m - y^{\infty}_m\right) \ ,$$

то погрешность $m+1$-го шага будет уже

$$e \sim h^4 \ .$$

Главным недостатком методов прогноза-коррекции является то, что с их помощью нельзя запустить процесс решения. Для расчета нескольких первых точек вам обязательно придется использовать методы Рунге–Кутты.

Ну и в заключение скажу, что методы Рунге–Кутты, поскольку они не используют информации о предыдущих точках, называют одноступенчатыми методами. В противоположность этому, методы прогноза-коррекции называют многоступенчатыми методами.

III.2. Методы решения краевой задачи.

Кревую задачу рассмотрим на примере ОДУ второго порядка

$$\frac{\partial^2 y}{\partial x^2} = f(x, y, y') \ ,$$

при ГУ

$$y(x = a) = y_a \ , \qquad y(x = b) = y_b \ .$$

Методы решения краевой задачи можно разделить на три группы:

- Метод стрельбы.
- Проекционные методы.
- Метод конечных разностей (МКР).

Метод стрельбы (пристрелки).

Данный метод сводится к замене решения краевой задачи решением некоторой последовательности задач Коши. Что делается? В начальной точке $x = a$ помимо сторого заданного ГУ-ми значения функции y_a задается значение производной $y'(x = a) = y'_a$. После этого производится выстрел: при заданных НУ решается задача Коши, вплоть до точки $x = b$. А там происходит проверка, удалось ли при заложенной нами производной удовлетворить ГУ в последней точке. Не удалось. Затем задается другое значение производной в точке $x = a$, решается задача Коши, и смотрят, лучше стало или хуже. И так, постепенно, анализируя анализируя характер получаемых решений и их зависимость от закладываемого значения производной в начальной точке, находится решение удовлетворяющее одновременной обоим заданным ГУ.

Если исходное ДУ — нелинейное, то поиск решения по методу стрельбы может оказаться неоднозначным и высокотворческим процессом. Если же исходное ДУ линейное, то жизнь резко облегчается. Действительно, если ДУ линейное, т.е. имеет вид

$$y'' = f_1(x)y' + f_2(x)y + f_3(x) \ ,$$

то для поиска истинного решения достаточно всего двух прицелочных выстрелов. Третий будет смертельным. Пусть найденные вами два решения $y_1(x)$ и $y_2(x)$ дают

$$y_1(x = b) = \beta_1 \ , \qquad y_2(x = b) = \beta_2 \ .$$

Поскольку исходное ДУ — линейно, то любая линейная комбинация этих двух решений вида

$$y(x) = cy_1(x) + (1 - c)y_2(x)$$

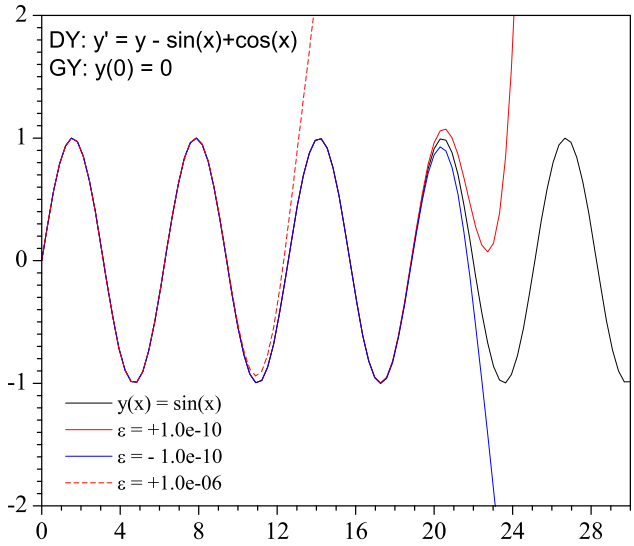
тоже будет являться решением. Причем эта комбинация уже отвечает первому ГУ: $y(a) = y_a$. Остается потребовать выполнение второго ГУ:

$$c\beta_1 + (1-c)\beta_2 = y_b \qquad \rightarrow \qquad c = \frac{y_b - \beta_2}{\beta_1 - \beta_2} \ .$$

Теперь вы сразу можете сконструировать необходимое вам решение

$$y(x) = \frac{y_b - \beta_2}{\beta_1 - \beta_2} y_1(x) + \frac{\beta_1 - y_b}{\beta_1 - \beta_2} y_2(x) \ .$$

Недостатки метода стрельбы. Первый очевиден. Это большой объем вычислений (если ДУ нелинейно). Второй недостаток присущ всем представленным до сих пор методам. Он связан с т.н. жесткостью ДУ. Жестким называется ДУ, если масштаб на котором требуется найти решение много больше некоторого характерного масштаба(ов) данного уравнения. Характерный масштаб называют еще постоянной времени ДУ. Чему равен характерный масштаб (постоянная времени) ДУ? Он определяется поведением решения ДУ. Обычно под характерным масштабом понимают интервал изменения независимой переменной на котором решение убывает в e раз, или возрастает в e раз. Если же решение осциллирует, то это — период осцилляций. На рис. приведено решение ДУ



$$y'(x) = y(x) - \sin(x) + \cos(x) \ , \qquad \text{НУ}: \quad y(0) = 0 \ .$$

Общее решение: $y = \sin(x) + \varepsilon \exp(x)$. Мы же ищем частное решение: $y(x) = \sin(x)$. Видно, что незначительная погрешность в задании НУ приводит к катастрофическому искажению результата.

Если интервал, на котором вы хотите получить решение, много больше постоянной времени вашего ДУ, т.е. вы имеете жеткую задачу, то вам всегда будет очень трудно получить хорошее решение. Численное решение при этом как-бы пытается вырваться из слишком жесткого каркаса, под который вы его загоняете. И иногда при решении таких жестких задач могут помочь ...

Проекционные методы.

Мы рассмотрим с вами классические методы:
1. Метод Ритца.

2. Метод Галеркина.
3. Метод конечных элементов.

Сущность проекционных методов состоит в разложении решения по базису некоторых функций (проекций)

$$y_n(x) = \sum_{i=1}^{n} c_i \phi_i(x) \ . \qquad (*)$$

Базис $\{\phi_i\}$ вы выбираете сами. Это — то творческое (непредсказумое) зерно проекционных методов, которое никто за вас не посеет. В роли базисных функций могут выступать обычные полиномы $\{x^i\}$, полиномы Лагранжа $\{P_i(x)\}$, Фурье-гармоники, наборы синусов $\{\sin(i * x)\}$ и т.д.

Обязательные условия, которым должны удовлетворять базисные функции:
1) разложение (*) должно аппроксимировать ваше решение с любой, сколь угодно малой точностью ($\forall \varepsilon > 0$ существует такое n: ...);
2) функции должны быть линейно независимы.
3) любая их комбинация должна удовлетворять поставленным ГУ.

Для автоматического выполнения третьего условия я могу предложить два способа (может быть, есть еще):
1. От ненулевых ГУ перейти к нулевым — элементарная замена переменных. А базисные функции брать только такие, которые отвечают этим самым, нулевым ГУ. Тогда и любая их комбинация тоже будет отвечать ГУ ($y(a) = 0$, $y(b) = 0$).
2. Никакой замены не надо. Базисные функции берем нулевые на краях, кроме первой ($\phi_1(a) = 1$, $\phi_1(b) = 0$) и последней ($\phi_n(a) = 0$, $\phi_n(b) = 1$). Теперь первый и последний коэффициенты c_i вам известны — они определяются граничными условиями ($c_1 = y_a$, $c_n = y_b$).

После выбора базисных функций вы подставляете ваше разложение в ДУ, и получаете систему для расчета неизвестных коэффициентов c_i (проекций). Система, в общем случае, нелинейная алгебраическая, что уже не должно вас пугать — решать такие системы вы умеете.

Теперь, собственно, как же получить требуемую систему. Сделать это можно по-разному. Первый способ — Метод Ритца.

Метод Ритца применим к решению тех краевых задач, которые позволяют *вариацинную постановку.*

Что это значит? Это означает, что искомая функция $y(x)$ (решение задачи) является стационарной точкой некоторого вариационного функционала

$$J\left[y(x)\right] = \int_a^b F\left(x, y, y'\right) dx \ .$$

Таковым функционалом может быть действие (в теор. механике), энергия (у теоретиков), время (в оптике) и пр. Как известно, функция $y(x)$ является стационарной точкой функционала J, если она удовлетворяет уравнению Эйлера

$$\frac{d}{dx} F'_{y'}(x, y, y') = F'_y(x, y, y') \ .$$

Таким образом, если ваше ДУ совпадает с уравнением Эйлера для какого-нибуть функционала, то ваша задача допускает вариационную постановку. Например, если ДУ имеет вид (одномерное стационарное уравнение теплопроводности)

$$-\frac{d}{dx}\left(k(x)y'(x)\right) + q(x)y(x) = f(x) \ ,$$

то соответствующий функционал

$$J\left[y(x)\right] = \int_a^b \left\{ \frac{k(x)y'(x)}{2} + \frac{q(x)y^2(x)}{2} - f(x)y(x) \right\} dx \ .$$

И, как утверждает Ритц, система для поиска коэффициентов c_i уже построена. Действительно, если в функционал J вместо искомой функции подставить ее разложение по выбранному базису, то мы получим функцию n переменных

$$J\left[y(x)\right] = J\left(c_1, c_2, \ldots, c_n\right) \ .$$

Условие экстремума принимает вид

$$\frac{\partial}{\partial c_i} J = 0 \ , \qquad i = 1, 2, \ldots, n \ .$$

Получили n уравнений для поиска n неизвестных. Алгебраическая и, в общем случае, нелинейная система.

К сожалению, не всегда исходное ДУ позволяет вариационную постановку. Например, незначительная модификация одномерного стационарного уравнения теплопроводности

$$-\frac{d}{dx}\left(k(x)y'(x)\right) + v(x)y'(x) + q(x)y(x) = f(x)$$

уже не допускает вариационной постановки. В этом случае может помочь более общий метод — *Метод Галеркина*.

Если исходная краевая задача не позволила вариационной постановки, тогда переходят к т.н. *проекцинной постановке*.

Домножим исходное ДУ ($f(x, y, y', y'', \ldots)$) на произвольную функцию $\tilde{y}(x)$ и проинтегрируем по всему интервалу x

$$\int_a^b f\left(x, y, y', \ldots\right) \tilde{y}(x) dx = 0 \ .$$

Проекционная постановка формулируется следующим образом: требуется найти функцию $y(x)$, которая удовлетворяет записанному интегральному тождеству для произвольной функции $\tilde{y}(x)$. Далее мы опять вводим некоторый базис $\{\phi_i(x)\}$, и значит ограничиваем наше рассмотрение только такими функциями, которые могут быть аппроксимированы (в пределах известной погрешности) линейной комбинацией базисных функций. Тогда записанное интегральное тождество выполняется для любой пробной функции $\tilde{y}(x)$, если это тождество выполняется для каждой из базисных функций. Получаем

$$\int_a^b f\left(x, y_n, y_n'\right) \phi_i(x) dx = 0 \ , \qquad i = 1, \ldots, n$$

$$y_n(x) = \sum_{j=1}^{n} c_j \phi_j(x)$$

— систему n уравнений для нахождения n неизвестных коэффициентов c_i. Полученная система, в общем случае, — нелинейная.

Мы же с вами сейчас поподробнее остановимся на более простом случае, когда исходное ДУ — линейное, т.е. имеет вид

$$y'' + f_1(x)y' + f_2(x)y = f_3(x) \ .$$

Неизвестными нашей задачи теперь являются n коэффициентов c_j. Для определения этих коэффициентов домножим ДУ на i-ую базисную функцию и проинтегрируем все ДУ по x (под функцией $y(x)$ мы теперь понимаем ряд $y_n(x)$).

$$\int_a^b \left[y_n''(x) + f_1(x)y_n'(x) + f_2(x)y_n(x)\right]\phi_i(x)dx = \int_a^b f_3(x)\phi_i(x)dx \ , \quad i = 1, 2, \ldots, n \ .$$

Индекс i теперь нумерует строки системы из n уравнений. Данные уравнения называют "моментными уравнениями Галеркина". Подставляя $y_n(x)$, получаем систему n линейных алгебраических уравнений для поиска n неизвестных коэффициентов c_j.

$$\sum_{j=1}^{n} c_j \left(a_{ij} + b_{ij} + d_{ij}\right) = F_i \ , \quad i = 1, 2, \ldots, n \ ,$$

где

$$a_{ij} = \int_a^b \phi_j''(x)\phi_i(x)dx \ , \qquad b_{ij} = \int_a^b f_1(x)\phi_j'(x)\phi_i(x)dx \ ,$$
$$d_{ij} = \int_a^b f_2(x)\phi_j(x)\phi_i(x)dx \ , \qquad F_j = \int_a^b f_3(x)\phi_i(x)dx \ .$$

Линейную систему из n уравнений с n неизвестными вы решать умеете.

Посмотрим для наглядности, что дает метод Галеркина для уравнения

$$y'' + y = 0 \ , \quad y(0) = 0 \ , \ y(20\pi) = 0 \ .$$

Имеем

$$f_1(x) = 0 \ , \quad f_2(x) = 1 \ , \quad f_3(x) = 0 \ .$$

В качестве базисных функций возьмем синусы: $\phi_i(x) = \sin(ix)$. Подсчитаем коэффициенты системы моментных уравнений Галеркина.

$$F_i = \int_a^b 0 \cdot \phi_i(x)dx = 0 \ ,$$

т.е. вектор правых частей — нулевой, следовательно, одно из решений — тривиальное. Идем дальше.

$$a_{ij} = \int_0^{20\pi} (-j^2)\sin(jx)\sin(ix)dx \ .$$

Используем тригонометрическое соотношение

$$\cos(B) - \cos(A) = 2\sin\left(\frac{A+B}{2}\right)\sin\left(\frac{A-B}{2}\right) \ .$$

У нас

$$A = (j+i)x \ , \quad B = (j-i)x \ .$$

Получаем

$$a_{ij} = \frac{-j^2}{2}\int_0^{20\pi} \left[\cos\left((i-j)x\right) - \cos\left((i+j)x\right)\right]dx \ .$$

Видим, что все a_{ij} при $i \neq j$ равны нулю. При $i = j$ имеем

$$a_{ii} = \frac{-i^2}{2}\int_0^{20\pi} [1]\,dx = -10\pi i^2 \ .$$

Далее

$$b_{ij} = \int_0^{20\pi} 0 \cdot \phi_i(x)dx = 0 \ .$$

И наконец,

$$d_{ij} = \int_0^{20\pi} \sin(ix)\sin(jx) = 10\pi\delta_{ij} \ .$$

Теперь система моментных уравнений Галеркина сводится к виду

$$\begin{array}{rcl} c_1(-10\pi + 10\pi) + c_2 \cdot 0 + c_3 \cdot 0 + \ldots = 0 \ , & \rightarrow & c_1 = \forall \ , \\ c_1 \cdot 0 + c_2(-10\pi 2^2 + 10\pi) + c_3 \cdot 0 + \ldots = 0 \ , & \rightarrow & c_2 = 0 \ , \\ c_1 \cdot 0 + c_2 \cdot 0 + c_3(-10\pi 3^2 + 10\pi) + \ldots = 0 \ , & \rightarrow & c_2 = 0 \ , \\ \ldots & \ldots & \ldots \ . \end{array}$$

Т.е. решение есть

$$y(x) = c_0 \sin(x) \ .$$

На практике наиболее трудным этапом является выбор системы базисных функций. Значительное упрощение метода Галеркина достигается в случае кусочно-линейных базисных функций, что приводит к методу конечных элементов.

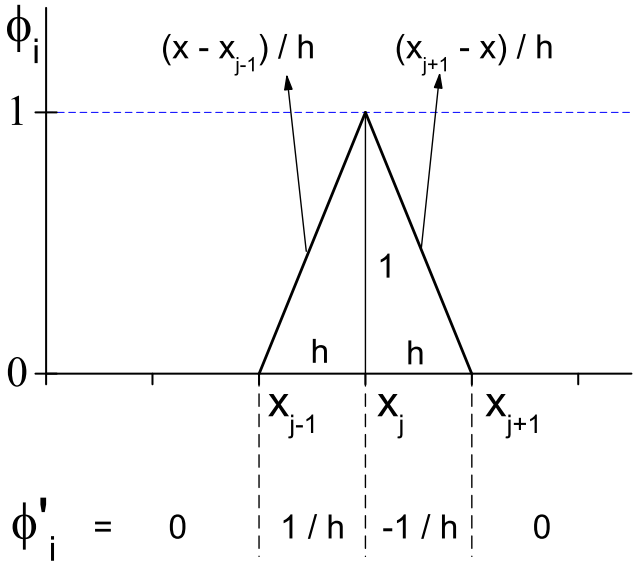
Рассмотрим метод конечных элементов на примере одномерного стационарного уравнения теплопроводности

$$-\frac{d}{dx}\left(q_1(x)y'(x)\right) + q_2(x)y(x) = q_3(x) \ .$$

Исходный отрезок (a, b) разбивается на n частей

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$$

с шагом $h = (b - a)/n$. Говорят: "На интервал (a, b) наложена равномерная сетка с шагом h". Под функциями ϕ_i понимают кусочно-линейные функции вида ... (см. рис.). Искомая



функция аппроксимируется линейной комбинацией функций ϕ_i, причем коэффициенты c_j теперь есть не что иное, как значения функции в точках x_j:

$$y_n(x) = y_0(=0) + y_1\phi_1 + \ldots + y_{n-1}\phi_{n-1} + y_n(=0) \ .$$

Поскольку базисные функции кусочно-линейны, то и результат будет кусочно-линейной функцией, производная которой на участке (x_{j-1}, x_j)

$$y_n'(x) = \frac{y_j - y_{j-1}}{h} \ .$$

Составляем моментное уравнение Галеркина

$$\int_a^b \left[-\frac{d}{dx}\left(q_1 y_n'\right) + q_2 y_n \right] \phi_i(x) dx = \int_a^b q_3(x)\phi_i(x)dx \ , \quad i = 1, 2, \dots \ , n-1 \ .$$

Первое слагаемое в левой части интегрируем по частям

$$\int_{x_{i-1}}^{x_{i+1}} q_1(x) y_n'(x) \phi_i'(x) dx + \int_{x_{i-1}}^{x_{i+1}} q_2(x) y_n(x) \phi_i(x) dx = \int_{x_{i-1}}^{x_{i+1}} q_3(x)\phi_i(x)dx \ , \quad i = 1, 2, \dots \ , n-1 \ ,$$

и подставляем функции $y_n(x)$ и $\phi_i(x)$

$$\begin{aligned}
&\int_{x_{i-1}}^{x_i} q_1(x) \left(\frac{y_i - y_{i-1}}{h} \right) \left(\frac{1}{h} \right) dx + \int_{x_i}^{x_{i+1}} q_1(x) \left(\frac{y_{i+1} - y_i}{h} \right) \left(\frac{-1}{h} \right) dx + \\
&+ \int_{x_{i-1}}^{x_i} q_2(x) \left[y_i + \frac{y_i - y_{i-1}}{h} \left(x - x_i \right) \right] \frac{x - x_{i-1}}{h} dx + \\
&+ \int_{x_i}^{x_{i+1}} q_2(x) \left[y_i + \frac{y_{i+1} - y_i}{h} \left(x - x_i \right) \right] \frac{x_{i+1} - x}{h} dx = \int_{x_{i-1}}^{x_{i+1}} q_3(x)\phi_i(x)dx \ ,
\end{aligned}$$

$i = 1, 2, \dots \ , n-1$.

Видим, что наша система сводится к виду

$$a_i y_{i-1} + b_i y_i + c_i y_{i+1} = F_i \ , \quad i = 1, 2, \dots \ , n-1 \ ,$$

где

$$\begin{aligned}
a_i &= \frac{-1}{h^2} \int_{x_{i-1}}^{x_i} q_1(x)dx + \frac{-1}{h^2} \int_{x_{i-1}}^{x_i} q_2(x) \left(x - x_i \right) \left(x - x_{i-1} \right) dx \ , \\
b_i &= \frac{1}{h^2} \int_{x_{i-1}}^{x_{i+1}} q_1(x)dx + \frac{1}{h^2} \int_{x_{i-1}}^{x_i} q_2(x) \left(x - x_{i-1} \right)^2 dx + \frac{1}{h^2} \int_{x_i}^{x_{i+1}} q_2(x) \left(x_{i+1} - x \right)^2 dx \ , \\
c_i &= \frac{-1}{h^2} \int_{x_i}^{x_{i+1}} q_1(x)dx + \frac{1}{h^2} \int_{x_i}^{x_{i+1}} q_2(x) \left(x - x_i \right) \left(x_{i+1} - x \right) dx \ .
\end{aligned}$$

В итоге мы имеем линейную систему из $n-1$ уравнения с $n-1$ неизвестным. Матрица системы — трехдиагональная, для решения — метод прогонки.

Пример:

$$3\phi'' + \phi = 0 \ , \quad \text{ГУ}: \quad \phi(0) = 0 \ , \ \ \phi\left((20\pi + \pi/2)\sqrt{3} \right) = (20\pi + \pi/2)\sqrt{3} \ .$$

Чтобы перейти к нулевым ГУ, делаем замену функции $y = \phi + x$. Теперь исходная задача принимает вид

$$3y'' + y = -x \ , \quad \text{ГУ}: \quad y(0) = 0 \ , \ \ y((20\pi + \pi/2)\sqrt{3}) = 0 \ .$$

Общее решение:

$$y(x) = -x + k_1 \sin(x/\sqrt{3}) + k_2 \cos(x/\sqrt{3}) \ .$$

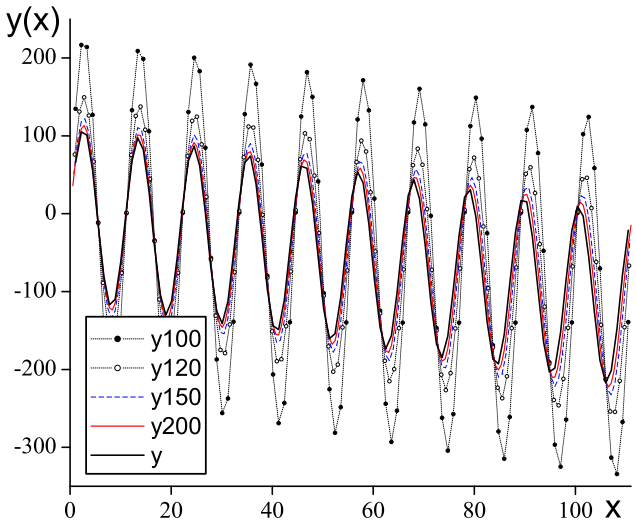
Искомое решение:

$$y(x) = -x + \sqrt{3}\left(20\pi + \frac{\pi}{2} \right) \sin(x/\sqrt{3}) \ .$$

Коэффициенты:

$$\begin{aligned}
a_j &= \frac{-3}{h} - \frac{h}{6} = a \ , \\
c_j &= \frac{-3}{h} - \frac{h}{6} = c = a \ , \\
b_j &= \frac{6}{h} - \frac{2}{3}h = b \ , \\
F_j &= x_j h \ .
\end{aligned}$$

Решения, полученные методом прогонки для разных h, представлены на рис.


y(x)
200
100
0
-100
-200
-300
0
20
40
60
80
100
X
y100
y120
y150
y200
y

IV. Решение ДУ в ЧП.

Дифференциальное уравнение в ЧП:

$$F\left(x_1, x_2, \ldots, x_n, U, \frac{\partial U}{\partial x_1}, \frac{\partial U}{\partial x_2}, \ldots, \frac{\partial U}{\partial x_n}, \frac{\partial^2 U}{\partial x_1^2}, \ldots\right) = 0 \ .$$

Здесь U — зависимая переменная, x_i — независимые переменные. Существенно то, что неизвестная функция $U(x_i)$ зависит более, чем от одной переменной. Мы с вами ограничим наше рассмотрение случаем двух переменных. Все методы, о которых мы будем вести речь, могут быть обобщены на случай большего количества переменных. Порядок старшей частной производной — порядок уравнения. Квазилинейное уравнение — линейное уравнение относительно всех старших производных от неизвестной функции. Линейное уравнение — линейно относительно функции и всех ее частных производных. Мы будем решать квазилинейные уравнения второго порядка. Причем под квазилинейностью мы будем понимать линейность по всем частным производным. Общий вид такого уравнения:

$$A(x,y,U)\frac{\partial^2 U}{\partial x^2} + 2B(\ldots)\frac{\partial^2 U}{\partial x \partial y} + C(\ldots)\frac{\partial^2 U}{\partial y^2} + D(\ldots)\frac{\partial U}{\partial x} + E(\ldots)\frac{\partial U}{\partial y} + f(x,y,U) = 0 \ .$$

Вот на этом мы остановимся. Наш выбор продиктован тем, что именно такие уравнения чрезвычайно часто встречаются в физике. Это и волновое уравнение, и уравнение теплопроводности, и уравнение Лапласа и многие другие.

Классификация ДУ в ЧП второго порядка.

Выделяют три типа ДУ второго порядка:

- Уравнение называется эллиптическим, если $AC - B^2 > 0$.
- Уравнение называется гиперболическим, если $AC - B^2 < 0$.
- Уравнение называется параболическим, если $AC - B^2 = 0$.

Аналогия с алгебраическими уравнениями для эллипса, параболы и гиперболы.

Уравнение может принадлежать к нескольким типам в зависимости от значений коэффициентов. Так, например, уравнение Трикоми

$$y\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0$$

при $y > 0$ эллиптического типа, при $y < 0$ — гиперболического, а линия $y = 0$ называется линией параболичности.

Чем приятен случай двух переменных, дак это тем, что в этом случае ДУ второго порядка с постоянными коэффициентами всегда можно привести к т.н. каноническому виду, т.е. путем замены переменных можно исключить из уравнения смешанную переменную. Тогда классифицировать уравнение особенно просто. Если коэффициенты при вторых производных одного знака, то мы имеем уравнение эллиптического типа. Примеры эллиптических уравнений: уравнение Пуассона

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = -f(x,y) \ ,$$

уравнение Лапласа

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0 \ .$$

Примером параболического уравнения является уравнение теплопроводности или, что то же самое, уравнение диффузии

$$\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2} \ .$$

Уравнением гиперболического типа является т.н. волновое уравнение

$$\frac{\partial^2 U}{\partial t^2} = c^2\frac{\partial^2 U}{\partial x^2} \ .$$

Естественно, для нахождения однозначного решения любого из этих уравнений мы должны задать некоторые дополнительные условия. Если речь идет о некоторых условиях на границе пространственной области, внутри которой ищется решение, то условия называются граничными (ГУ). Если же речь идет о каком-то моменте времени, то условия называются начальными (НУ). Если в уравнении присутствует только первая производная по времени (как в уравнении теплопроводности), то достаточно задать одно начальное условие, например, распределение температуры в начальный момент. Если же в уравнении присутствует вторая производная по времени (волновое уравнение), то необходимы два НУ, например, положение струны в нач. момент и скорость смещения ее точек. Что же касается ГУ, то выделяют три типа ГУ:

1) Задаются значения функции на границе — задача Дирихле,
2) Задаются значения производной функции по нормали к границе — задача Неймана,
3) И смешанная граничная задача (условия Робина) — $\left(\frac{\partial U}{\partial n}\right)_G + kU_G = F$.

Методы решения ДУ в ЧП:

- Проекционные методы;
- Метод конечных разностей (МКР) или метод сеток.

Проекционные методы называют также вариационно – разностными методами, проекционно – разностными, проекционно – сеточными. Сюда относится известный метод Галеркина решения ДУ в ЧП. Основным моментом этих методов является разложение искомой функции по некоторому ортогональному базису известных функций. С данными методами мы познакомились на примере решения краевой задачи для ОДУ. А основной темой наших дальнейших бесед будет МКР.

IV.1. МКР : описание метода.

Это наиболее универсальный численный метод решения ДУ. Возьмем в качестве примера известное уравнение теплопроводности:

$$\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2} \ .$$

Это уравнение параболического типа. Однако МКР одинаково применим к ДУ любого типа. Поэтому все идеи МКР, с которыми мы познакомимся, в одинаковой степени относятся также и к уравнениям гиперболического и эллиптического типов.

Пусть область изменения независимых переменных, в которой нам надо найти решение:

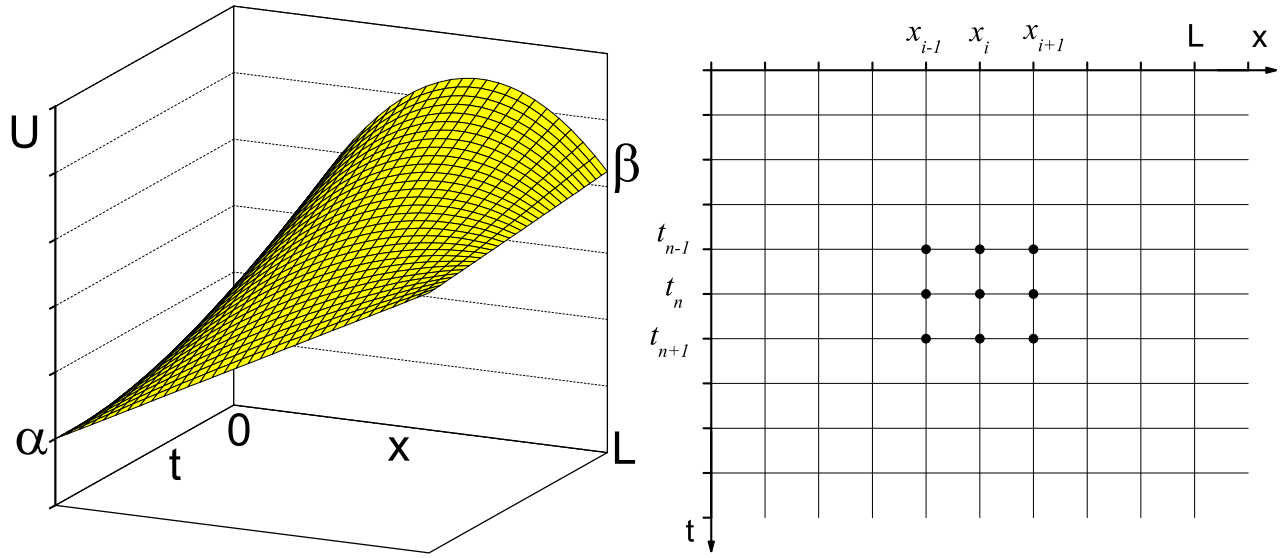
$$G = \{0 \le x \le L \ ; \quad 0 < t < \infty\} \ .$$

Дополнительные условия к уравнению теплопроводности возьмем в самом простом виде (задача Дирихле):

$$\begin{array}{lcl} \text{Начальные условия:} & U(x,0) = \phi(x) \ , & 0 \le x \le L \ . \\ \text{Граничные :} & \begin{array}{c} U(0,t) = \alpha = const_1 \\ U(L,t) = \beta = const_2 \end{array} & t > 0 \ . \end{array}$$

Геометрическая интерпретация задачи

Если независимых переменных всего две, то это решение можно изобразить в виде поверхности в трех измерениях $U(x,t)$. ГУ, приведенные выше, изобразятся в виде двух



прямых параллельных оси t. Начальные условия — кривая в плоскости $t = 0$. *Область изменения независимых переменных — выделить.*

Основная идея МКР:

на область изменения независимых переменных накладывается т.н. сетка и решение ищется не в виде непрерывной функции $U(x,t)$, а в виде дискретного, конечного множества чисел U_i^n, представляющих (заменяющих) функцию $U(x,t)$ на дискретном, конечном множестве значений независимых переменных, т.е. в узлах наложенной сетки. Мы ищем не сплошную поверхность решения (как это делается при аналитическом решении и в проекционных методах, и даже в методе конечных элементов), а конечный дискретный набор чисел U_i^n, лежащих вблизи (на это мы надеемся) этой поверхности, над соответствующими узлами сетки.

Простейшей сеткой является равномерная прямоугольная сетка. Такая сетка образуется при пересечении линий $x = x_i = i \cdot h$ ($i = 0, 1, \ldots, N$) и $t = t_n = n \cdot \tau$ ($n = 0, 1, \ldots$) (см. рис.). h — шаг сетки по координате, τ — по времени. Это какие-то фиксированные числа, причем $N =, L/h$. Если независимых переменных две, то сетка представляет собой множество точек на поверхности (плоскости), если три — в обьеме, четыре — .... Точки пересечения называются — узлами сетки. Положение каждого узла характеризуется двумя целыми числами i и n — задача с двумя независимыми переменными. Если независимых переменных три, то — три числа, и т.д.

Теперь можно еще раз сформулировать идею:
ищем не $U(x,t)$, а набор чисел U_i^n, и надеемся, что $U_i^n \simeq U\left(x = ih, t = n\tau\right)$.
Вообще говоря, если мы найдем какой-то "правильный" метод (способ) вычисления U_i^n, то должно выполняться:

$$\lim_{h\to 0,\tau\to 0} U_i^n = U\left(x = ih, t = n\tau\right) \ .$$

Иными словами, мы, в принципе, сможем сколь угодно близко приблизиться к решению ДУ. Но только в принципе, т.к. уменьшение h и τ требует увеличения числа арифметических действий. Для нас важно то, что мы можем приблизиться к решению с любой заданной наперед степенью точности. Это свойство называют сходимостью численного метода. Именно в этом смысле множество чисел U_i^n представляет решение ДУ.

Множество U_i^n называют численным решением ДУ, если выполняется условие сходимости

$$||U_i^n - U\left(x = ih, t = n\tau\right)||_v \to 0 \ , \quad \text{при} \quad h \to 0 \ , \ \text{и} \ \ \tau \to 0 \ ,$$
$$||\ldots||_v = \max_v |U - U_i^n| \ , \quad \text{так, например, можно ввести норму} \ .$$

Все это было сказано лишь для пояснения основных идей метода.

Как конкретно вычислить множество U_i^n? Как запрограммировать исходное ДУ? Чтобы это сделать нужно запрограммировать частные производные. По определению:

$$\frac{\partial f}{\partial x} = \lim_{\Delta x\to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \ .$$

Но компьютер не понимает слова предел. Максимум, что мы можем использовать — приближенное соотношение

$$\frac{df}{dx} \simeq \frac{f(x+\Delta x) - f(x)}{\Delta x} \ .$$

Т.е. все производные в исходном ДУ мы должны заменить отношением конечных разностей (отсюда название метода). Какими точками мы можем располагать для составления этих конечных разностей? Только узлы сетки.

Таким образом, перед нами ставится задача: научиться определять различные производные (приближенно конечно), используя значения функции в узлах сетки. Данная процедура называется...

Аппроксимация производных.

Рассмотрим какой-то произвольный узел сетки, соответствующий точке x_i и t_n. Значение искомой функции в этом узле, обозначим U_i^n. Саму неизвестную функцию мы можем представить в окрестности выбранного узла в виде разложения Тейлора, например, по переменной x (t пока зафиксируем).

$$U(x) = U_i^n + a(x - x_i) + b(x - x_i)^2 + c(x - x_i)^3 + \ldots \ .$$

Здесь искомые производные с точностью до известных нам численных коэффициентов я обозначил буквами a, b, c, и т.д. Если нас интересует только первая производная, мы,

казалось бы, можем оборвать ряд на втором члене. Тогда

$$U(x) = U_i^n + a(x - x_i) \ ,$$

и для того, чтобы определить одно единственное неизвестное, нам помимо исходного узла достаточно воспользоваться еще одним, любым из соседних, узлом. Воспользуемся правым узлом

$$U_{i+1}^n = U_i^n + ah \ ,$$

от куда

$$a = \left(\frac{\partial U}{\partial x}\right)_{i,n} = \frac{U_{i+1}^n - U_i^n}{h} \ .$$

Это т.н. формула правой аппроксимации первой производной. Оценим погрешность такой аппроксимации. Для этого достаточно разложить значение функции в соседней точке, которую мы использовали, в тейлоровский ряд.

$$U_{i+1}^n = U_i^n + U_x h + \frac{1}{2} U_{xx} h^2 + \dots \ .$$

Подставляя это разложение, видим

$$a = U_{x(i,n)} + \frac{1}{2} U_{xx(i,n)} h + \dots \ , \qquad \text{Погрешность} \sim O(h) \ .$$

Говорят: аппроксимация первого порядка точности. Это означает: если уменьшить шаг вдвое, то и ошибка уменьшится вдвое.

Вместо последующей точки мы могли бы взять предыдущую $(i-1)$. Тогда мы бы получили т.н. левую аппроксимацию первой производной

$$a = \left(\frac{\partial U}{\partial x}\right)_{i,n} = \frac{U_i^n - U_{i-1}^n}{h} \ ,$$

с той же самой погрешностью аппроксимации. Возникает вопрос, как уменьшить погрешность аппроксимации первой производной. Единственный способ — использовать информацию не из одной соседней точки, а из нескольких. Возьмем две соседние точки. Параллельно с увеличением числа точек, которые мы используем для аппроксимации, мы должны увеличивать и число неизвестных параметров. Иначе мы просто не сможем решить задачу. Число соседних точек, т.е. число условий, которые мы накладываем на функцию $U(x)$, должно быть равно числу неизвестных. В случае двух точек мы должны представить искомую функцию в виде полинома второй степени

$$U(x) = U_i^n + a(x - x_i) + b(x - x_i)^2 \ .$$

Составляем систему двух уравнений для определения неизвестных коэффициентов

$$\begin{cases} U_{i-1}^n = U_i^n - ah + bh^2 \ , \\ U_{i+1}^n = U_i^n + ah + bh^2 \ , \end{cases}$$

Из этой системы легко найти

$$a = \frac{U_{i+1}^n - U_{i-1}^n}{2h} \ .$$

Это т.н. аппроксимации первой производной центральной разностью. Нетрудно убедиться, что погрешность центральной аппроксимации уже на порядок меньше $O(h^2)$. Для этого,

как и прежде, достаточно разложить значение функции во всех соседних точках, которые мы использовали в ряды Тейлора. Таким образом, аппроксимация центральной разностью более точная в том смысле, что уменьшение h в два раза приведет к уменьшению погрешности аппроксимации в 4 раза. При необходимости можно построить формулы аппроксимации первой производной еще более высоких порядков точности. Для этого лишь потребуется информация о большем числе соседних узлов.

Аналогично производной по координате строются формулы аппроксимации для производной по времени, только там соответствующие формулы называются, соответственно, аппроксимация вперед по времени, аппроксимация назад по времени и центральная аппроксимация.

Теперь, как аппроксимировать вторую производную. А это мы с вами уже почти сделали. Нам нужно лишь найти из записанной системы коэффициент b

$$b = \frac{1}{2}\frac{\partial^2 U}{\partial x^2} = \frac{U_{i+1}^n - 2U_i^n + U_{i-1}^n}{2h^2} \ .$$

Погрешность аппроксимации второй производной по такой формуле $O(h^2)$. Вместо двух соседних точек вы вольны, конечно, взять, например, две последующие точки. Вы тоже без труда решите систему и получите аппроксимационную формулу для второй производной. Однако погрешность аппроксимации в этом случае будет уже $O(h)$. По этой причине всегда предпочтительнее использовать симметричные точки для построения формул аппроксимации. Хотя иногда по самым разным причинам полезным может оказаться и несимметричный выбор точек, например, когда вы находитесь на краю сетки. Несимметричные разностные аппроксимации приходится строить также для аппроксимации ГУ 2-го и 3-го рода.

На этом мы остановимся. При необходимости вы теперь сможете аппроксимировать любую производную и с любой требуемой степенью точности. Единственное, что необходимо для повышения точности аппроксимации, — это ипользование большего количества соседних точек.

Теперь, когда мы умеем аппроксимировать производные, мы приступим к аппроксимации всего ДУ.

Схемы аппроксимации ДУ в ЧП

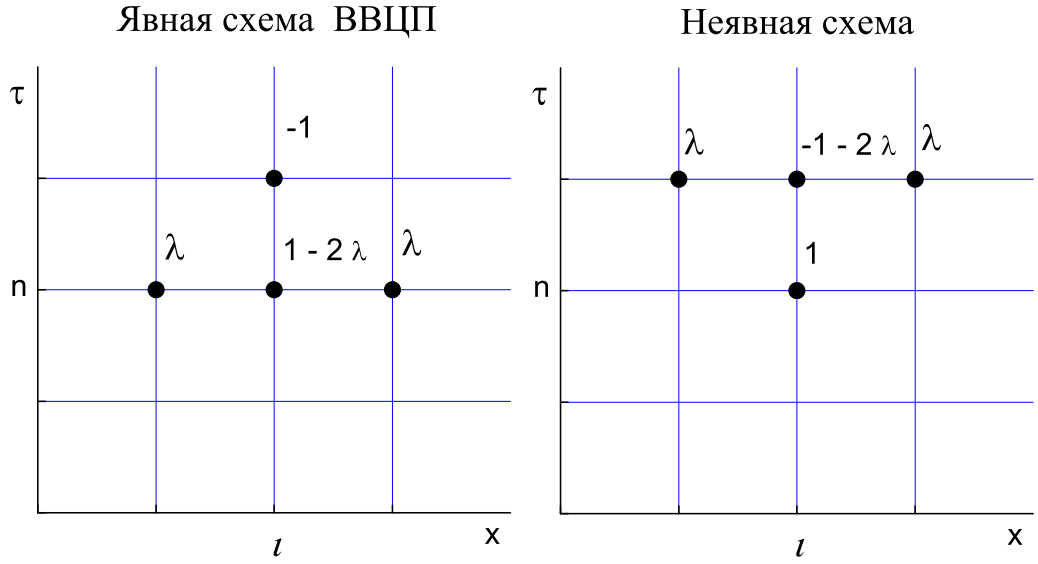
Используя приведенные выше выражения аппроксимации производных, мы можем записать аппроксимацию ДУ теплопроводности. Возьмем для аппроксимации производной по времени формулу вперед по времени, а для второй производной по координате, записанную нами центральную формулу

$$\frac{U_i^{n+1} - U_i^n}{\tau} = \frac{a}{h^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \ , \quad i = 1 \ , \dots \ , N-1 \ . \qquad (*)$$

Это т.н. схема ВВЦП (вперед по времени, центральная по пространству), а англ. лит. — FTCS. Записанную РС можно представить схематически, начертив узлы сетки. Такой рисунок называется трафаретом (stencil в англ.лит.). Для этого сначала домножим РС на τ, обозначим $\lambda = a\tau/h^2$, и перенесем все в правую часть:

$$0 = -U_i^{n+1} + U_i^n + \lambda\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \ , \quad i = 1 \ , \dots \ , N-1 \ .$$

Теперь рисуем сетку вблизи узла (i, n) и возле каждого узла, задействованного в РС, запишем соответствующий ему коэффициент. Глядя на трафарет, видим, что мы получили способ вычисления искомых чисел U_i^n. Для этого решение необходимо начать с первого слоя (нулевой слой по времени задан НУ). Любой из узлов первого слоя определяется



значениями функции в трех узлах нулевого слоя, а они нам известны. Пройдя по всему первому слою от $i = 1$ до $N - 1$, получим значения искомой функции во всех узлах первого слоя. Заметьте, что вычислять значения при $i = 0$ и $i = N$ не нужно, эти значения исвестны, они задаются ГУ. После того, как вы получите первый слой, вы точно также можете рассчитать второй временной слой, затем третий, и т.д. Подобные задачи, когда необходимо последовательно переходить от одного временного слоя к последующему, называют маршевыми задачами. Говорят: маршевый проход, маршевое решение и т.д. В итоге — получено искомое множество чисел U_i^n.

Выражение (*) называют явной разностной схемой (РС) аппроксимации уравнения теплопроводности (или просто — явной схемой аппроксимации). Название "явная" обусловлено тем, что искомая величина — значение функции на последующем временном слое — явным образом выражается только лишь через известные нам величины. На этом простом примере мы показали и основную идею МКР и основные понятия метода, такие как сетка, схема аппроксимации производной, РС аппроксимации уравнения. Кроме этого, увидели, как "участвуют" в решении начальные и граничные условия.

Пойдем немного дальше. На этом же примере покажем, что такое неявная схема аппроксимации ДУ. Левая часть уравнения — аппроксимация производной по времени. Мы использовали аппроксимацию производной в узле (i, n) разностью вперед по времени. Но это же самое выражение можно интерпретировать, как разность назад по времени в узле $(i, n + 1)$. Если принять такую интерпретацию, то правую часть уравнения тоже нужно привязать к $(n + 1)$ слою, а не к n-му. Тогда получим неявную РС.

$$\frac{U_i^{n+1} - U_i^n}{\tau} = \frac{a}{h^2}\left(U_{i+1}^{n+1} - 2U_i^{n+1} + U_{i-1}^{n+1}\right) \ , \qquad i = 1, \ldots, N-1 \ .$$

Перепишем РС в виде, удобном для построения трафарета,

$$0 = -U_i^{n+1} + U_i^n + \lambda\left(U_{i+1}^{n+1} - 2U_i^{n+1} + U_{i-1}^{n+1}\right) \ , \qquad i = 1, \ldots, N-1 \ .$$

Трафарет этой схемы изображен на рисунке. Почему неявная? Потому, что теперь неизвестные величины — значения функции на $(n + 1)$-ом слое — выражаются сами через себя. Уравнение для каждого узла содержит более, чем одно неизвестное. Поэтому поотдельности, как было в случае с явной РС, решить эти уравнения невозможно. Однако полное число уравнений равно числу узлов, для которых мы выписываем эти уравнения,

$N-1$ и это же есть число неизвестных (нам нужно определить значения функции в $N-1$ узлах сетки). Поэтому ситуация не безнадежна, труднее чем в случае явной РС, но не безнадежна. Мы получили не что иное как систему $N-1$ уравнений, линейную алгебраическую систему, с $N-1$ неизвестными. Такие системы мы с вами умеем решать. К вашим услугам весь спектр прямых или итерационных методов, которые мы с вами разбирали. Можете выбирать любой, какой вам по душе. Как получить первый временной слой. Запишем нашу РС для $n=0$ в следующем виде. Перенесем в уравнении для каждого узла неизвестные величины влево.

$$\begin{aligned} i=1\ , \quad &\to \quad (1+2\lambda)U_1^1-\lambda U_2^1=U_1^0+\lambda U_0^1\ , \\ i=2\ , \quad &\to \quad -\lambda U_1^1+(1+2\lambda)U_2^1-\lambda U_3^1=U_2^0\ , \\ i=3\ , \quad &\to \quad -\lambda U_2^1+(1+2\lambda)U_3^1-\lambda U_4^1=U_3^0\ , \end{aligned}$$

и т.д. Первое, что бросается в глаза, мы имеем хорошо обусловленную систему, т.е. матрица нашей системы удовлетворяет условию преобладания диагональных членов. Причем, это достаточно частое явления: когда вы решаете ДУ в ЧП, вы, как правило, будете получать хорошо обусловленные системы линейных уравнений. И это не может не радовать. А если вы имеете хорошо обусловленную линейную систему, то для ее решения сам собой напрашивается метод Гаусса–Зейделя. Как мы уже знаем, этот метод сходится, причем непременно сходится, для подобных систем. Теперь согласно методу Гаусса–Зейделя задаем начальное приближение значениям функции на первом временном слое $U_{i,1}^{(0)}$, и записываем уравнение для первого узла (первый индекс — по x, второй — по t)

$$i=1\ , \quad \to \quad (1+2\lambda)U_{1,1}^{(1)}=\lambda U_{2,1}^{(0)}+U_{1,0}+\lambda U_{0,1}\ .$$

Здесь я индексы и по координате и по времени опустил вниз, а верхний индекс означает номер итерации. Для второго узла используем уже найденное к этому моменту значение $U_{1,1}^{(1)}$

$$i=2\ , \quad \to \quad (1+2\lambda)U_{2,1}^{(1)}=\lambda U_{3,1}^{(0)}+\lambda U_{1,1}^{(1)}+U_{2,0}\ .$$

Для третьего узла мы уже имеем значения U во втором узле:

$$i=3\ , \quad \to \quad (1+2\lambda)U_{3,1}^{(1)}=\lambda U_{4,1}^{(0)}+\lambda U_{2,1}^{(1)}+U_{3,0}\ ,$$

и т.д. Таким образом вы получите некоторое первое приближение на первом слое. Нулевое приближение можно задавать по всякому (в разумных пределах, конечно) — все равно метод сойдется. Разумным нулевым приближением, на мой взгляд, являются значения функции на предыдущем слое. Получив первое приближение, мы аналогично рассчитаем второе приближение, затем третье, и т.д. до тех пор пока не будет достигнута некоторая требуемая точность, например, до выполнения одного из соотношений

$$\max_i\left|U_{i,1}^{(k)}-U_{i,1}^{(k-1)}\right|<\varepsilon\ , \qquad \max_i\left|\frac{U_{i,1}^{(k)}-U_{i,1}^{(k-1)}}{U_{i,1}^{(k)}}\right|<\varepsilon\ .$$

Первое из этих соотношений устанавливает предельное значение абсолютной погрешности расчета, второе — относительной.

В итоге вы получите значения искомой функции на первом временном слое. После этого, совершенно аналогично используя описанный алгоритм, мы можем рассчитать второй слой, отталкиваясь теперь уже от найденных значений на первом слое, затем третий слой и т.д. Как я уже сказал, при решении ДУ в ЧП, как правило, получаются хорошо обусловленные матрицы, поэтому метод Гаусса–Зейделя используется достаточно часто. В

применении к эллиптическим уравнениям этот метод называется также методом Либмана или методом последовательных смещений.

Что же касается нашего конкретного примера, то в этом случае использовать итерационный метод Гаусса–Зейделя нецелесообразно. Есть другой, более эффективный путь, а именно, прямой метод — метод прогонки. Действительно, если повнимательнее посмотреть на нашу систему, то можно заметить, что матрица системы трехдиагональная:

$$\begin{array}{lccccc} i=1 & (1+2\lambda)\,U_1^1 & -\lambda U_2^1 & & & = U_1^0+\lambda U_0^1\ , \\ i=2 & -\lambda U_1^1 & +(1+2\lambda)\,U_2^1 & -\lambda U_3^1 & & = U_2^0\ , \\ \dots & \dots & \dots & \dots & & \dots \\ i=N-1 & & -\lambda U_{N-2}^1 & +(1+2\lambda)\,U_{N-1}^1 & & = U_{N-1}^0+\lambda U_N^1\ . \end{array}$$

Применив метод прогонки, мы получим числа U_i^1 на первом временном слое без всяких итераций. Далее — очевидно: для вычисления U_i^2 на следующем временном слое мы снова должны решать аналогичную по виду систему уравнений. И опять без всяких итераций, используя эффективный метод прогонки.

Метод прогонки настолько эффективен, что, я бы сказал, он является просто спасением для неявных РС. Благодаря этому методу неявные схемы в смысле быстроты расчета уже почти не уступают явным схемам, и поэтому становятся, в этом смысле, вполне конкурентно способными. Вы скажете, что их гораздо труднее запрограммировать — это правда. Но зато они часто имеют ряд значительных преимуществ перед явными схемами, о которых мы будем говорить чуть позже. А сейчас рассмотрим еще один, очень распространенный способ построения РС.

Метод Кранка–Николсона.

Мы продолжаем строить РС для уравнения теплопроводности

$$\frac{\partial U}{\partial t}=a\frac{\partial^2 U}{\partial x^2}\ .$$

Вспомним, при построении явной схемы мы аппроксимировали производную по времени разностью вперед по времени в узле (i,n), и вторую производную по координате привязали к временному слою n:

$$\frac{\partial^2 U}{\partial x^2}=\frac{1}{h^2}\left(U_{i-1}^n-2U_i^n+U_{i+1}^n\right)\ .$$

При построении неявной схемы мы взяли то же самое выражение для аппроксимации производной по времени, но сказали, что это разность назад по времени в узле $(i,n+1)$, и, как следствие, привязали производную по координате к временному слою $(n+1)$:

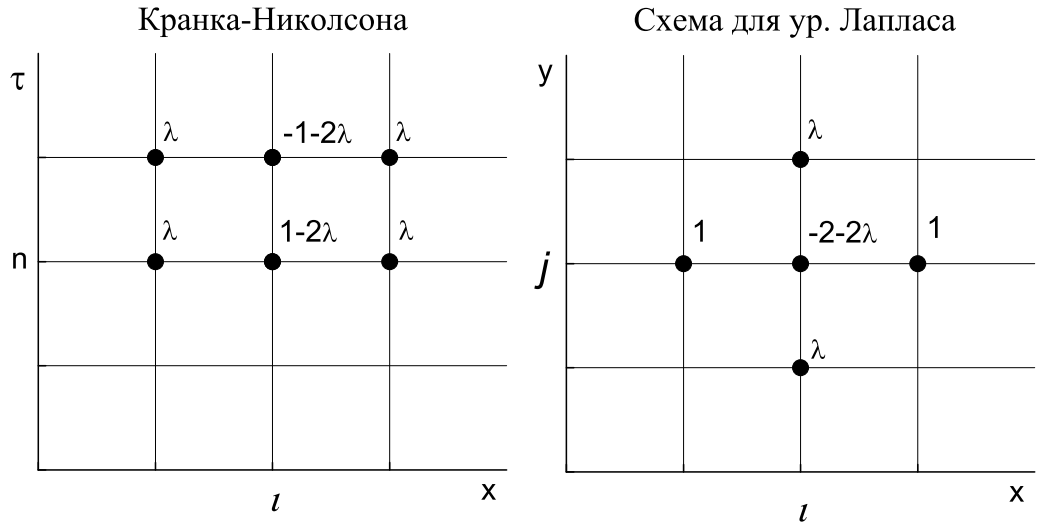
$$\frac{\partial^2 U}{\partial x^2}=\frac{1}{h^2}\left(U_{i-1}^{n+1}-2U_i^{n+1}+U_{i+1}^{n+1}\right)\ .$$

Теперь мы попытаемся соблюсти большую строгость. А именно, мы используем то же самое выражение для производной по времени, но будем теперь считать, что это выражение аппроксимирует производную в точке точно посередине между узлами (i,n) и $(i,n+1)$. Такая интерпретация приводит к тому, что мы должны усреднить записанные выражения для второй производной. В итоге получаем РС Кранка–Николсона.

$$\frac{U_i^{n+1}-U_i^n}{\tau}=\frac{a}{2h^2}\left(U_{i-1}^n-2U_i^n+U_{i+1}^n+U_{i-1}^{n+1}-2U_i^{n+1}+U_{i+1}^{n+1}\right)\ ,\quad i=1\ ,\dots\ ,N-1\ .$$

Опять же обозначая $\lambda=a\tau/2h^2$, и перенося все вправо, можем записать

$$-U_i^{n+1}+U_i^n+\lambda\left(U_{i-1}^n-2U_i^n+U_{i+1}^n+U_{i-1}^{n+1}-2U_i^{n+1}+U_{i-1}^{n+1}\right)=0\ ,\quad i=1\ ,\dots\ ,N-1\ .$$



Зарисуем трафарет этой схемы (см. рис.). Видим, что схема КН относится к неявным схемам, т.к. каждое из уравнений содержит больше чем одно неизвестное. Но при этом получившаяся система хорошо обусловлена, выполняется преобладание диагональных членов, значит можно применять метод Гаусса–Зейделя. Однако опять можно заметить, что матрица системы — трехдиагональная, и, следовательно, самый оптимальный способ для решения такой системы — метод прогонки.

Все перечисленные РС являются двухслойными — при вычислениях используются значения с двух слоев по времени. Довольно часто приходится применять схемы, содержащие большее количество временных слоев. Мы рассмотрим в качестве примера две трехслойные схемы.

РС уравнения Лапласа.

Возьмем уравнение эллиптического типа — уравнение Лапласа

$$\frac{\partial^2 U}{\partial x^2} + a\frac{\partial^2 U}{\partial y^2} = 0 \ .$$

Здесь в отличие от уравнения параболического типа — уравнения теплопроводности — по обоим переменным присутствуют производные второго порядка. Аппроксимируя эти производные центральными разностями в окрестности узла (i, j), получим

$$\frac{U_{i-1}^j - 2U_i^j + U_{i+1}^j}{h_x^2} + a\frac{U_i^{j-1} - 2U_i^j + U_i^{j+1}}{h_y^2} = 0 \ .$$

Т.е. вы волей-неволей приходите к трехслойной РС. Обозначая $\lambda = ah_x^2/h_y^2$, перепишем в виде

$$U_{i-1}^j - 2U_i^j + U_{i+1}^j + \lambda U_i^{j-1} - 2\lambda U_i^j + \lambda U_i^{j+1} = 0 \ , \quad i = 1, \ldots, N_x - 1 \ , \quad j = 1, \ldots, N_y - 1 \ .$$

Зарисуем трафарет этой схемы (см. рис). Теперь, если по переменной y дополнительные условия (а их должно быть два) поставлены при $y = 0$, т.е. мы имеем два НУ (например, $U(y = 0, x) = \phi(x)$ и $u_y(y = 0, x) = \psi(x)$), то РС — явная, и необходим маршевый проход по слоям y. Если же задана конечная область по y и поставлены ГУ (например, $U(y = 0, x) = \phi(x)$ и $u(y = L, x) = \psi(x)$), то РС — неявная. В этом случае имеем систему $(N_x - 1) \cdot (N_y - 1)$ уравнений и, естественно, столько же неизвестных. Значения на краях задаются ГУ. Сразу же можно сказать, что матрица получившейся системы не трехдиагональная. Значит метод прогонки исключен. Но зато система хорошо обусловлена: узел

(i,j) имеет преобладающий коэффициент. Во внутренней области сетки коэффициент при центральном узле по абсолютному значению равен сумме абсолютных значений других коэффициентов, но зато вблизи границ, когда один или два узла заданы ГУ, центральный коэффициент больше (по модулю) суммы других коэффициентов. Значит такую систему нужно решать методом Гаусса–Зейделя. Как я уже упоминал, в случае эллиптических уравнений, метод Гаусса–Зейделя называется методом Либмана или методом последовательных смещений.

Схема Ричардсона.

Многослойные схемы возникают не только благодаря наличию вторых производных. К многослойности может привести и первая производная, если вы аппроксимируете ее с высокой точностью. Попытаемся усовершенствовать РС уравнения теплопроводности. До сих пор мы использовали либо разность вперед по времени, либо разность назад по времени. Как вы помните, при такой аппроксимации производной возникает погрешность первого порядка по величине шага, т.е. $O(\tau)$. Ричардсон (вот ведь умнющий человек) предложил использовать центральную разность и получил следующую РС (которая теперь носит его имя)

$$\frac{U_i^{n+1} - U_i^{n-1}}{2\tau} = \frac{a}{h^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \ , \quad i = 1 \ , \dots \ , N-1 \ .$$

Этот прием весьма распространен и зачастую бывает очень полезен. Однако в случае уравнения теплопроводности эта схема — схема Ричардсона — является историческим примером классической катастрофы. Правдоподобная и на первый взгляд очень точная РС не позволяет получить истинное решение ДУ. Как показывают численные расчеты решение ДУ, полученное по схеме Ричардсона не сходится к истинному решению ДУ. Вы можете спросить, как это показывается. А я вам говорил, что такое ВЭ. Так вот здесь достаточно провести ВЭ на т.н. модельном уравнении. Т.е. взять и запрограммировать задачу, которая имеет аналитическое решение. И сопоставляя то, что вам дает численный расчет, и известное вам аналитическое решение, вы тут же обнаружите, что расчет не дает ничего даже близкого к истинному решению. Говорят: решение не сходится.

Будет или не будет сходится решение — это главный вопрос, который мучает физика, когда он строит РС какого-нибуть ДУ в ЧП. Строить РС вы теперь умеете, а вот оценить, будет ли наблюдаться сходимость, вы пока не можете. Вопрос сходимости решения РС настолько важен, фундаментален, что мы наше дальнейшее внимание посвятим изучению этого вопроса.

IV.2. Сходимость разностных схем.

Математические основы вопросов сходимости РС хорошо развиты только для линейных систем. Результаты линейной теории используются в виде наводящих соображений для нелинейных задач, а их применимость проверяется затем в ходе ВЭ.

Сходящаяся конечно-разностная схема математически определяется как схема, дающее решение, которое стремится к истинному решению ДУ при измельчении сетки. К сожалению, часто бывает не так. Возможны следующие варианты.

Первый вариант. При измельчении сетки значение искомой функции в заданном узле стремится к определенному пределу, но этот предел не совпадает с истинным решением. Говорят: РС не согласована с ДУ (не аппроксимирует ДУ).

Второй вариант. При измельчении сетки значение искомой функции в заданном узле вообще не стремится к определенному значению. Говорят: РС неустойчива. Здесь возможны две ситуации: значение функции в заданном узле постоянно растет (падает) — в этом случае говорят о *статической неустойчивости*; значение функции в заданном узле осциллирует — говорят о *динамической неустойчивости*. Как проявляет себя неустойчивость в численном расчете? Как правило, это быстрое, катастрофическое нарастание чисел $|U_i^n|$.

Для линейных ДУ в ЧП существует т.н. теорема Лакса, которая утверждает, что необходимым и достаточным условием сходимости РС является выполнение согласованности и устойчивости. Для схем аппроксимации линейных ДУ в ЧП:

$$\text{согласованность} + \text{устойчивость} \Longrightarrow \text{сходимость}$$

Необходимо отметить, что во многих работах по вычислительной математике предполагается справедливость этой теоремы и для схем аппроксимации нелинейных ДУ в ЧП. Хотя для них эта теорема строго не доказана. При таком использовании теоремы Лакса предполагается применение принципа "замороженных" коэффициентов. Под методом "замороженных" коэффициентов понимают простой прием, когда коэффициенты в ДУ заменяют некоторыми константами и анализируют полученное ДУ с постоянными коэффициентами. А потом если установлена устойчивость РС для такого линейного ДУ, говорят: ну, возможно, и для исходного нелинейного ДУ тоже будет наблюдаться устойчивость. Просто ничего лучшего пока не придумано. Иными словами, мы не можем доказать устойчивость РС, но можем показать ее неустойчивость для "замороженных" коэффициентов, а это позволяет исключить такую схему из рассмотрения. Если же схема для уравнения с "замороженными" коэффициентами будет согласованной и устойчивой, то есть надежда, что эта схема будет сходящейся и для нелинейного ДУ. Это показывает накопленный опыт численных расчетов.

Мы, за неимением ничего лучшего, тоже будем ориентироваться на теорему Лакса. Стало быть первое, что мы должны сделать, записав какую-либо РС, — это проанализировать ее согласованность и устойчивость. Анализом этих свойств РС мы сейчас и займемся.

IV.2.1. Согласованность разностной схемы.

При рассмотрении основных положений МКР, мы не задумываясь считали, что если заменить все входящие в ДУ производные их аппроксимациями в виде конечных разностей, то полученное выражение будет конечно-разностной аппроксимацией этого ДУ. Однако, это не всегда так. Говорят, что РС аппроксимации ДУ согласована с исходным ДУ (или

аппроксимирует ДУ), если в пределе, когда размеры ячеек сетки стремятся к нулю, РС эквивалентна этому самому ДУ в каждой из узловых точек.

$$\lim_{\tau,h\to 0} (\text{РС})_{x,t} = (\text{ДУ})_{x,t} \ .$$

Количественной характеристикой согласованности является погрешность (ошибка) аппроксимации ДУ данной РС

$$E = (\text{РС} - \text{ДУ})_{i,n} \ .$$

В терминах погрешности аппроксимации: РС согласована со своим ДУ, если погрешность аппроксимации во всех узлах сетки (в произвольном узле сетки) стремится к нулю при любом измельчении шагов сетки ($\tau \to 0$, $h \to 0$). РС и ДУ при этом записывают, перебрасывая все члены в одну сторону (чтобы с другой стороны остался ноль).

Рассмотрим в качестве примера уравнение теплопроводности. ДУ запишем в виде

$$\frac{\partial U}{\partial t} - a\frac{\partial^2 U}{\partial x^2} = 0 \ .$$

Берем РС, которую мы придумали для этого уравнения, например, явную РС ВВЦП

$$\frac{U_i^{n+1} - U_i^n}{\tau} - \frac{a}{h^2}\left(U_{i+1}^n + U_{i-1}^n - 2U_i^n\right) = 0 \ .$$

Подставим левые части этих выражений в определение погрешности аппроксимации

$$E = (\text{РС} - \text{ДУ})_{i,n} = \left[\frac{U_i^{n+1} - U_i^n}{\tau} - \frac{a}{h^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right)\right] - \left(\frac{\partial U}{\partial t} - a\frac{\partial^2 U}{\partial x^2}\right) \ .$$

Устремляем шаги сетки к нулю, и смотрим, к чему стремится E. Все соседние узлы расписываются в виде ряда Тейлора в окрестности заданного узла (i, n).

$$U_i^{n+1} = U_i^n + \left(\frac{\partial U}{\partial t}\right)_{i,n}\tau + \left(\frac{\partial^2 U}{\partial t^2}\right)_{i,n}\frac{\tau^2}{2} + O(\tau^3) \ ,$$

$$U_{i+1}^n = U_i^n + U'_{i,n}h + U''_{i,n}\frac{h^2}{2} + U'''_{i,n}\frac{h^3}{3!} + U^{IV}_{i,n}\frac{h^4}{4!} + U^{V}_{i,n}\frac{h^5}{5!} + U^{VI}_{i,n}\frac{h^6}{6!} + \dots \ ,$$

$$U_{i-1}^n = U_i^n - U'_{i,n}h + U''_{i,n}\frac{h^2}{2} - U'''_{i,n}\frac{h^3}{3!} + U^{IV}_{i,n}\frac{h^4}{4!} - U^{V}_{i,n}\frac{h^5}{5!} + U^{VI}_{i,n}\frac{h^6}{6!} \ .$$

Подставляем эти выражения в определение погрешности E

$$E = \left[\frac{\partial U}{\partial t} - a\frac{\partial^2 U}{\partial x^2} + \left(\frac{\tau}{2}\frac{\partial^2 U}{\partial t^2} - \frac{2ah^2}{4!}\frac{\partial^4 U}{\partial x^4}\right) + O(\tau^2) + O(h^4)\right] - [\text{ДУ}] \ .$$

Величины $O(\tau^2) + O(h^4)$ условно обозначают так $O(\tau^2, h^4)$. Видим, что если τ и $h \to 0$, то погрешность аппроксимации стремится к нулю для любого узла:

$$E = \left(\frac{\tau}{2}\frac{\partial^2 U}{\partial t^2} - \frac{2ah^2}{4!}\frac{\partial^4 U}{\partial x^4}\right) + O(\tau^2, h^4) \to 0 \ .$$

Ведущие члены погрешности, т.е. величины $O(\tau, h^2)$, называются порядком аппроксимации ДУ данной РС. Например, РС ВВЦП аппроксимирует уравнение теплопроводности (или согласована с уравнением теплопроводности) с первым порядком по τ и вторым по h.

Погрешность аппроксимации — необычайно полезный инструмент. Как правило, если есть сходимость, то ошибка численного решения будет уменьшаться подобно погрешности аппроксимации. Так, например, глядя на погрешность аппроксимации, мы можем сказать, что при уменьшении шага по времени в два раза ошика численного решения... неизвестно как изменится. Аналогично, если изменить только h, то это не даст полезной информации. Зато мы видим, что если согласованно уменьшить шаги по времени и по координате, например, в четыре раза уменьшить τ и в два раза уменьшить h, то можно ожидать, что ошибка численного решения уменьшится в четыре раза. А вот это уже очень полезная информация. Имея такую информацию, мы можем, во-первых, сразу оценить размеры нашей ошибки, во-вторых, применить экстраполяционный перход к пределу.

Вы можете сказать, что для получения погрешности аппроксимации РС, не нужно производить никаких дополнительных манипуляций. Действительно, вы ведь уже имеете такой инструмент анализа, как погрешность аппроксимации производной. Когда мы заменяли производные в уравнении теплопроводности их разностными аналогами, мы видели и оценивали какие погрешности при этом вводим. По времени это как раз была погрешность $O(\tau)$, а по координате это как раз была погрешность $O(h^2)$. Все так и есть. Это мы и получили в итоговой погрешности для всей РС. Мы, собственно, для того с вами и изучали такой инструмент, как погрешность аппроксимации производной, чтобы предугадывать в некоторой степени результат, который получится в итоге. Так, например, теперь если вместо разности вперед по времени в РС использовать центральную разность, погрешность аппроксимации которой уже $O(\tau^2)$ (схема Ричардсона), то мы можем ожидать, что погрешность аппроксимации РС будет $O(\tau^2, h^2)$. Но... только ожидать. На самом деле все может статься. Вы всегда должны проверить свои предположения строгим расчетом погрешности итоговой РС по той методике, которую теперь имеете. В результате вы скорее всего получите, что ваши ожидания оправдались, но иногда вы будете приятно или неприятно удивлены неожиданными сюрпризами.

Пример такого сюрприза. Берем схему Ричардсона

$$\frac{U_i^{n+1} - U_i^{n-1}}{2\tau} = \frac{a}{h^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \ ,$$

и в выражении для аппроксимации второй производной заменяем U_i^n на $(U_i^{n-1} + U_i^{n+1})/2$. При этом получим явную схему, называемую схемой Дюфорта–Франкела (1953г.)

$$\frac{U_i^{n+1} - U_i^{n-1}}{2\tau} = \frac{a}{h^2}\left(U_{i+1}^n + U_{i-1}^n - U_i^{n-1} - U_i^{n+1}\right) \ .$$

Эту схему называют трехслойной — при вычислениях используются значения с трех слоев по времени. Если выполнить все вычисления, то получим

$$E = \frac{\tau^2}{6}\frac{\partial^3 U}{\partial t^3} + \left(\frac{\tau}{h}\right)^2 a\frac{\partial^2 U}{\partial t^2} - h^2\frac{a}{12}\frac{\partial^4 U}{\partial x^4} \ .$$

Если $\tau,\ h \to 0$, но так что $\tau/h = \text{const}$, то согласованности нет. Оказывается, что эта схема не согласована с уравнением теплопроводности. Замечу что, последнее не означает полной непригодности подобных схем. Просто в отсутствие согласованности, мы уже не можем положиться на теорему Лакса, и стало быть нужно быть очень осторожным, тщательно протестировать данную схему, провести массу предварительных ВЭ. Вообще, конечно, накопленный опыт позволяет сформулировать рекомендацию — не использовать несогласованные схемы аппроксимации.

И, наконец, даже если ваши ожидания оправдались, т.е. если погрешность аппроксимации всей РС стремится к нулю при измельчении сетки, причем именно так, как это

можно было предполагать, на основе аппроксимации производных, ваш труд все равно не окажется напрасным. Полученное выражение для погрешности аппроксимации РС несет в себе еще массу полезной информации. Так, посмотрев на записанное выражение, иногда (в некоторых РС) вы сможете заметить, что при выполнении какого-то соотношения между шагами сетки и параметрами уравнения может возникнуть более высокий порядок аппроксимации. Например, для уравнения теплопроводности

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial}{\partial t}\frac{\partial U}{\partial t} = \frac{\partial}{\partial t}\left(a\frac{\partial^2 U}{\partial x^2}\right) = a\frac{\partial^2}{\partial x^2}\frac{\partial U}{\partial t} = a^2\frac{\partial^4 U}{\partial x^4} \ .$$

Для погрешности схемы ВВЦП получаем

$$E = \frac{\partial^2 U}{\partial t^2}\left(\frac{\tau}{2} - \frac{h^2}{12a}\right) + O(\tau^2, h^4) \ .$$

При $\tau = h^2/(6a)$ скобка $= 0$.

Напоследок, я приведу уже без детального вывода (выводить вы теперь уже и сами умеете) погрешность аппроксимации неявной РС для уравнения теплопроводности (для нее в качестве исходного узла разумно взять узел $(i, n+1)$)

$$E = \frac{\tau}{2}\frac{\partial^2 U}{\partial t^2} + \frac{2ah^2}{4!}\frac{\partial^4 U}{\partial x^4} \ ,$$

и схемы Кранка–Николсона (узел $(i, n+1)$, хотя для узлов (i, n) и $(i, n+1)$ будет то же самое)

$$E = \frac{\tau^2}{12}\frac{\partial^3 U}{\partial t^3} + \frac{ah^2}{12}\frac{\partial^4 U}{\partial x^4} \ .$$

Можно отметить, что для этих схем уже нет условия перехода погрешности аппроксимации РС в более высокий порядок. Еще можно отметить, что погрешность РС Кранка–Николсона по времени на порядок меньше, чем для предыдущих схем. Получается, что используя схему К-Н не нужно мельчить по времени, чтобы достигнуть достаточной точности решения. Это обстоятельство часто оказывается решающим для выбора схемы, особенно, если вам надо найти решение на достаточно протяженном интервале времени.

IV.3. Устойчивость разностных схем.

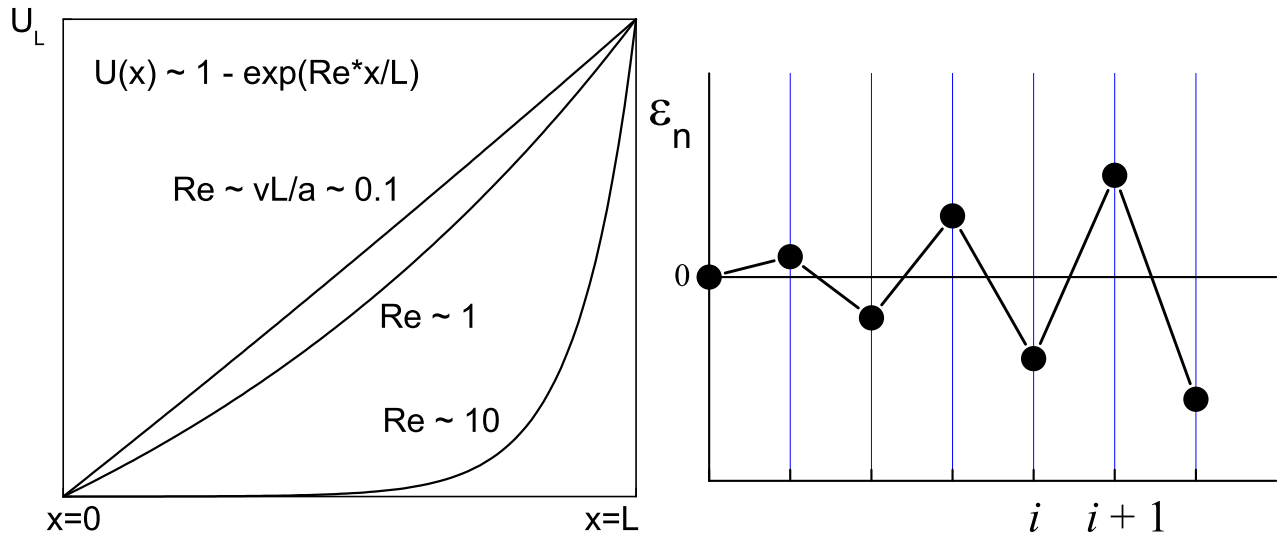
Рассмотрим для начала опять же то, что попроще — параболическое уравнение — уравнение теплопроводности. Но теперь для большей общности введем в него т.н. конвективный член

$$\frac{\partial U}{\partial t} = -v\frac{\partial U}{\partial x} + a\frac{\partial^2 U}{\partial x^2} \ .$$

Это уравнение называют уравнением конвекции и диффузии. Оно описывает перенос некоторой скалярной величины U при заданной скорости конвекции v, a — теперь вы можете интерпретировать либо как коэффициент теплопроводности (U — температура), либо как коэффициент вязкости (U — функция вихря). Если у нас заданы ГУ первого рода, т.е. например,

$$U(x=0) = 0 \ , \quad U(x=L) = U_L \ ,$$

то мы легко можем найти стационарное решение (см. рис.)



$$U(x) = U_L\frac{1-\exp(Re*x/L)}{1-\exp(Re)} \ , \quad Re = \frac{vL}{a} \ .$$

Т.е. для любого НУ после достаточно продолжительного промежутка времени наше решение должно сойтись к известному нам стационарному решению.

Построим для этого уравнения уже известную вам РС ВВЦП

$$\frac{U_i^{n+1} - U_i^n}{\tau} = -\frac{v}{2h}\left(U_{i+1}^n - U_{i-1}^n\right) + \frac{a}{h^2}\left(U_{i+1}^n - 2U_i^n + U_{i-1}^n\right) \ .$$

Пусть мы преодолели достаточно много временных слоев и ожидаем, что решение уже должно выйти на стационарное значение, т.е. U_i^n — очень близко к стационарному значению $\bar{U}_i^n$. Реально в численном расчете мы, конечно, никогда не получим истинного решения $\bar{U}_i^n$ в чистом виде. В каждом узле решение вычисляется с какой-то погрешностью. Это может быть о-очень маленькая погрешность, но она всегда есть. Пусть отклонение расчетных значений от истинного стационарного решения $\varepsilon_n = U_i^n - \bar{U}_i^n$ на (n)-ом временном слое имеет следующий вид (см. рис.). Тогда подставляя в схему ВВЦП возмущенное

решение, можем записать

$$\frac{\bar{U}_i^{n+1}-\bar{U}_i^n}{\tau}+\frac{\varepsilon_i^{n+1}-\varepsilon_i^n}{\tau}= \begin{array}{l} -\dfrac{v}{2h}\left(\bar{U}_{i+1}^n-\bar{U}_{i-1}^n\right)+\dfrac{a}{h^2}\left(\bar{U}_{i+1}^n-2\bar{U}_i^n+\bar{U}_{i-1}^n\right) \\ -\dfrac{v}{2h}\left(\varepsilon_{i+1}^n-\varepsilon_{i-1}^n\right)+\dfrac{a}{h^2}\left(\varepsilon_{i+1}^n-2\varepsilon_i^n+\varepsilon_{i-1}^n\right)\ . \end{array}$$

То что относится к истинному решению сокращается — на то оно и истинное решение. Оставшиеся члены показывают, как поведет себя при переходе к следующему временному слою возмущение ε_i^n

$$\Delta\varepsilon_i=\varepsilon_i^{n+1}-\varepsilon_i^n=-\frac{v\tau}{2h}\left(\varepsilon_{i+1}^n-\varepsilon_{i-1}^n\right)+\frac{a\tau}{h^2}\left(\varepsilon_{i+1}^n-2\varepsilon_i^n+\varepsilon_{i-1}^n\right)\ .$$

Рассмотрим ситуацию, когда у нас имеется только диффузионный член

$$\Delta\varepsilon_i=\frac{a\tau}{h^2}\left(\varepsilon_{i+1}^n-2\varepsilon_i^n+\varepsilon_{i-1}^n\right)\ .$$

Поскольку $\varepsilon_{i+1}^n>0$ и $\varepsilon_{i-1}^n>0$, а в i-ом узле $\varepsilon_i^n<0$, то поправка положительна, т.е. поправка корректирует предыдущее значение. Аналогично можно показать, что для узлов с положительным отклонением от истинного решения поправки будут отрицательны. И все это, конечно, прекрасно, но лишь до тех пор пока шаг по времени τ достаточно мал. Если же окажется, что τ слишком велико, то поправка окажется чрезмерной, и вместо коррекции мы просто получим смену знака у начального возмущения, а при еще несколько большем шаге по времени эти возмущения начнут не просто менять знак, но еще и усиливаться со временем. Вот она какая — в чистом виде динамическая неустойчивость. И мы сразу видим способ борьбы с ней — уменьшение шага по времени. Т.е. существует некоторый критический шаг по времени, начиная с которого РС становится неустойчивой. Такая неустойчивость, которая проявляется лишь при некоторых условиях, называется условная неустойчивость. Говорят: явная схема ВВЦП для уравнения теплопроводности условно устойчива.

Теперь вернемся к нашему уравнению. Рассмотрим ситуацию, когда у нас имеется только конвективный член:

$$\Delta\varepsilon_i=\varepsilon_i^{n+1}-\varepsilon_i^n=-\frac{v\tau}{2h}\left(\varepsilon_{i+1}^n-\varepsilon_{i-1}^n\right)\ .$$

В этом случае мы с вами получили явную схему ВВЦП для т.н. линейного волнового уравнения первого порядка. И что же мы видим. Если $v>0$ (поток идет слева направо) и если возмущения растут в том же направлении: $\varepsilon_{i+1}>\varepsilon_{i-1}$, то поправка отрицательна для всех узлов, т.е. начальное возмущение будет монотонно возрастать. Налицо статическая неустойчивость. Причем эту неустойчивость мы не сможем убрать измельчая нашу сетку по переменным τ и h. Говорят: явная схема ВВЦП для линейного волнового уравнения первого порядка абсолютно неустойчива, а значит совершенно неприменима для численных расчетов.

Итак подведем итоги. По свойствам устойчивости РС аппроксимации ДУ бывают:

- абсолютно (безусловно) неустойчивые,
- условно устойчивые,
- абсолютно (безусловно) устойчивые.

По определению: РС называется устойчивой, если на каждом шаге по маршевой координате любая ошибка (погрешность округления, погрешность аппроксимации, просто

ошибка) не возрастает при переходе от одного шага к другому.
Таким образом понятие неустойчивости, сторого говоря, применимо лишь при решении "маршевых" задач. Т.е. когда решение находится последовательным движением в маршевом направлении от "поверхности" (линии), на которой заданы НУ. Решение при этом ищется в незамкнутой области. Это — уравнения параболического и гиперболического типов. С уравнениями эллиптического типа попроще жить. В них нет понятия неустойчивости и они, как правило, всегда сходятся (лишь бы РС была согласована с исходным ДУ).

Когда же мы имеем маршевые задачи, то тут все сложнее. Обычно для достижения устойчивости РС требуется намного больше усилий, чем для достижения согласованности. Проверить согласованность не очень сложно (кроме этого, согласованность может выполняться автоматически — некоторые методы построения РС дают всегда согласованные РС). Устойчивость — свойство более "тонкое", и обычно доказывается более сложно. Следует сказать, что большинство методов анализа устойчивости применимо лишь к РС, аппроксимирующим линейные ДУ.

Анализ устойчивости РС

Этот обзор методов анализа устойчивости необходим постольку, поскольку исключительно часто в литературе используется терминология, связанная с анализом устойчивости. Разработано несколько методов анализа устойчивости, т.е. методов, позволяющих предсказать, будет ли данная РС аппроксимации ДУ устойчивой или неустойчивой. Матричный метод, метод Неймана, метод дискретных возмущений, метод Хёрта и т.д. Большинство методов разработано только для линейных ДУ.

Мы рассмотрим только два метода: метод дискретных возмущений и метод Неймана. Первый метод применим для схем аппроксимации любых ДУ, а второй — только для линейных. Но это два наиболее часто упоминаемых в литературе метода. Основные идеи этих методов покажем на примере РС аппроксимации одномерного линейного уравнения теплопроводности и одномерного линейного волнового уравнения первого порядка. Начнем с метода дискретных возмущений и применим его для анализа устойчивости явной РС для уравнения теплопроводности.

Метод дискретных возмущений

Суть метода. В исходное уравнение в некоторой точке вводится дискретное возмущение величины ε и прослеживается влияние этого возмущения. РС будет устойчивой, если возмущения затухают (ну или хотя бы не возрастают). В качестве примера рассмотрим уравнение теплопроводности (одномерное) с нулевыми Н и ГУ. Это только для простоты выкладок — все можно сделать и для любых других Н и ГУ.

Что является решением уравнения теплопроводности при указанных НУ и ГУ? Очевидно $\equiv 0$. Но если, например, начальные нулевые условия будут представлены в ЭВМ не чистыми нулями, а некоторыми малыми числами U_i^0? Если РС устойчива, то решение при таких НУ не должно, по крайней мере, возрастать с каждым шагом по времени. Это следует из определения устойчивости.

Конкретизируем задачу:
Пусть $U_j^0 = 0$ для всех j кроме $j = i$. $U_i^0 = \varepsilon$. Если для всех i и n будет выполняться:

$$\left|\frac{U_i^n}{\varepsilon}\right| \le 1 \ ,$$

то РС будет устойчивой (по определению). Вспомним явную РС для одномерного уравнения теплопроводности

$$U_i^{n+1} = \lambda U_{i+1}^n + (1-2\lambda)U_i^n + \lambda U_{i-1}^n \ , \quad \lambda = a\tau/h^2 \ .$$

Теперь вычислим U_i^1 для заданных выше НУ

$$U_i^1 = (1-2\lambda)\varepsilon \ .$$

Потребуем выполнения устойчивости на первом шаге:

$$\left|\frac{U_i^1}{\varepsilon}\right| = \left|\frac{(1-2\lambda)\varepsilon}{\varepsilon}\right| \le 1 \ ,$$

т.е.

$$|(1-2\lambda)| \le 1 \ , \quad \Longrightarrow \quad \begin{cases} 1-2\lambda \le 1 \ , \\ 1-2\lambda \ge -1 \ . \end{cases}$$

Первое неравенство выполняется для всех положительных λ (а λ по своему смыслу — положительное число). Второе неравентство дает

$$\lambda \le 1 \ ,$$

т.е. если потребовать выполнения условия устойчивости только на первом шаге, то из него следует, что $\lambda \le 1$. Но схема будет устойчивой, если условие устойчивости выполнено для любого узла и на любом шаге по времени. Вычислим U_i^2.

$$U_i^2 = \lambda U_{i+1}^1 + (1-2\lambda)U_i^1 + \lambda U_{i-1}^1 \ ,$$

т.е. для вычисления U_i^2 нужны уже U_{i+1}^1 и U_{i-1}^1:

$$U_{i+1}^1 = \lambda U_{i+2}^0 (=0) + (1-2\lambda)U_{i+1}^0 (=0) + \lambda U_i^0 = \lambda\varepsilon \ , \quad U_{i-1}^1 = \ldots = \lambda\varepsilon \ .$$

Поэтому:

$$U_i^2 = \lambda\cdot\lambda\cdot\varepsilon + (1-2\lambda)(1-2\lambda)\varepsilon + \lambda\cdot\lambda\varepsilon = \varepsilon\left(1-4\lambda+6\lambda^2\right) \ .$$

$$\left|\frac{U_i^2}{\varepsilon}\right| = \left|\left(1-4\lambda+6\lambda^2\right)\right| \le 1 \ .$$

Это снова запись системы неравенств

$$\begin{cases} 1-4\lambda+6\lambda^2 \le 1 \ , \\ 1-4\lambda+6\lambda^2 \ge -1 \ . \end{cases} \ .$$

Проанализируем сначала второе неравенство. Легко показать, что это неравенство выполняется для любых λ, т.е. ничего нового мы не получили. Первое же неравенство дает:

$$\lambda \le \frac{2}{3} \ .$$

После этого надо проанализировать, что дают все последующие шаги по времени, т.е. оценить

$$\left|\frac{U_i^3}{\varepsilon}\right| \ , \quad \left|\frac{U_i^4}{\varepsilon}\right| \ , \quad \ldots \ .$$

Показано, что последовательный анализ решений, получающихся неравенств дает в пределе $n \to \infty$ условие: $\lambda \le 1/2$. Т.е. для того, чтобы явная РС для одномерного уравнения теплопроводности была устойчива, требуется выполнение определенного соотношения между параметром a в уравнении и шагами сетки

$$\tau \le \frac{h^2}{2a} \ .$$

Вот он тот самый, критический, размер шага по времени, который превращает РС ВВЦП из устойчивой в неустойчивую. Другими словами, если вы используете РС ВВЦП для уравнения теплопроводности, то вы не должны задавать шаг по времени больше указанного отношения. В этом, как правило, трагедия всех явных РС. Явные схемы чрезвычайно просты, программировать их легко, не надо никаких систем линейных решать, но... они (явные схемы) условно устойчивы. Т.е. вы будете связаны по рукам и ногам ограничениями типа записанного. Вам придется сильно измельчать шаг по времени. Причем это будет связано не с достижением необходимой точности, а, вот ведь незадача, с требованием какой-то там устойчивости. Вам бы может для вашей точности хватило шага в 0.01 по времени и по координате, а вам придется брать τ на два порядка меньше. А это — на два порядка больше время расчета. Чрезвычайно неудобно получается.

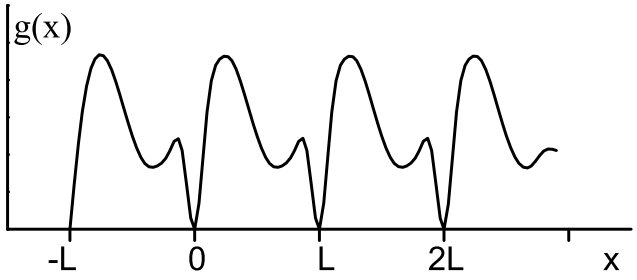
Хотя я может быть слегка перегнул палку. Вообще-то огульно обвинять все явные РС в условной сходимости не стоит. Бывают исключения. Например, трехслойная схема Дюфорта–Франкела обладает несмотря на то, что является явной РС, абсолютной устойчивостью. Но зато, если вы помните, она не согласована с исходным ДУ. За все в нашем мире надо платить.

Итак, мы рассмотрели метод дискретных возмущений. Еще раз напомню, метод применим как для линейных, так и для нелинейных ДУ в ЧП. Весьма универсален. Но уж больно он громоздкий. Чтобы получить окончательный результат (даже в нашем простеньком примере), ручки и бумаги бывает недостаточно. Приходится подключать мощнае современные возможности аналитических вычислений на ЭВМ, т.е. такие пакеты как MAPLE, Methematica, MathCad, MathLab. А ведь их еще надо бы знать для начала. Да и к чему нам такая общность, если теорема Лакса строго выполняется только для линейных ДУ. Давайте тогда уж и устойчивость проверять соответствующим способом, посредством метода, который строго применим для линейных ДУ. А уж нелинейные ДУ, что ж они потерпят. Для них все равно нет строгой теории. В связи с такими соображениями очень широкое распространение получил, гораздо более изящный метод — ....

Метод Неймана

Предложен Дж. фон Нейманом в Лос-Аламосе в 1944 году и получил к сегодняшнему дню наибольшее распространение. Строго говоря, он применим только для анализа устойчивости РС аппроксимации линейных ДУ. Но вместе с методом "замороженных" коэффициентов может быть использован для выявления неустойчивых РС для нелинейных ДУ.

Поясним основные идеи метода. Метод использует представление возмущения в виде ряда Фурье. Как известно, любую (почти любую — удовлетворяющую условиям Дирихле: функция ограничена и имеет конечное число относительных максимумов, минимумов и точек разрыва I-го рода на некотором конечном интервале) периодическую функцию можно представить в виде ряда Фурье.



$$g(x) = \sum_{m=-\infty}^{m=+\infty} C_m \exp\left(j\frac{2\pi}{L}mx\right) \ , \quad j = \sqrt{-1} \ ,$$

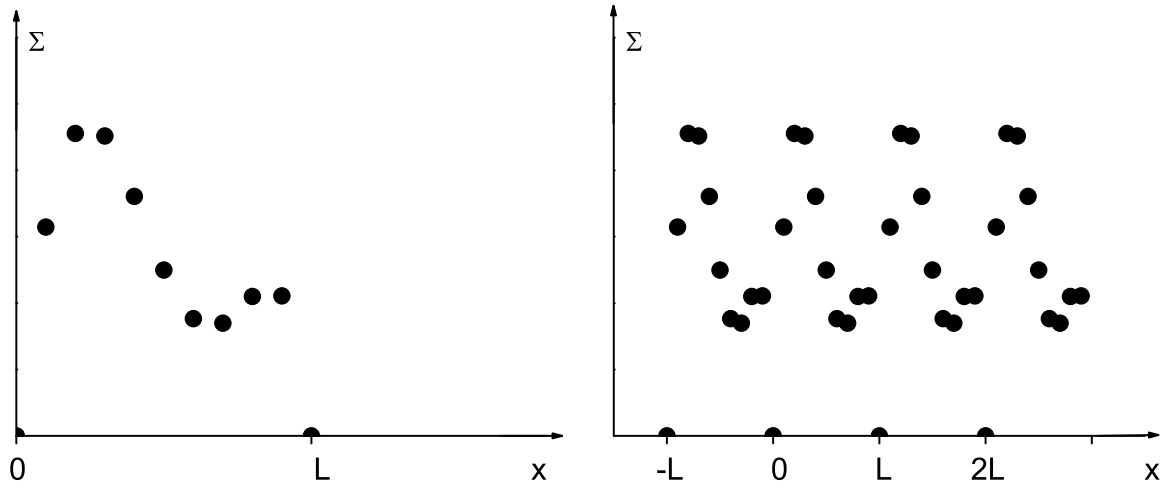
C_m — комплексная амплитуда Фурье-компоненты (или гармоники).

$$C_m = \frac{1}{L}\int_0^L g(x) \exp\left(-j\frac{2\pi}{L}mx\right) \ dx \ .$$

Отдельные слагаемые в сумме называют: фурье-компонента, гармоника, мода.

$$C_m \exp\left(j\frac{2\pi}{L}mx\right) = C_m \exp\left(jk_m x\right) \ , \qquad k_m = \frac{2\pi m}{L} \ \text{ — волновое число .}$$

Вернемся к нашей задаче. В этом методе, как и в методе дискретных возмущений, анализируется реакция (ответ) на внесенное возмущение (ошибку). Пусть вносимуе возмущение, реакция на которое исследуется, выглядит так: Однако ничто не мешает нам считать, что



это возмущение периодично с периодом L. За период можно взять интервал в $2L$, от $-L$ до L. Достроив функцию в интервале $(-L, 0)$ симметрично относительно точки $x = 0$, можно добиться гладкости первой производной. Но в нашем случае это не принципиально. $0 - L$ — расчетная область по оси x. Это никак не отразится на поведении возмущения на интересующем нас расчетном интервале от 0 до L. Но такая функция, которая определена только в конечном числе точек, удовлетворяет условию Дирихле, т.е. ее тоже можно представить в виде ряда Фурье:

$$\varepsilon(x) = \sum_m C_m \exp\left(jk_m x\right) \ .$$

Далее ход рассуждений таков. Уравнение, разностную аппроксимацию которого мы должны проанализировать на устойчивость, линейно. Следовательно, решение, вызванное возмущением в виде суммы, будет суммой решений, вызванных отдельными слагаемыми в возмущении. Поэтому мы можем сделать вывод об устойчивости РС, проанализировав реакцию (отклик), вызванную возмущением в виде одной (но произвольной) моды. Очевидно, что если мы покажем, что решение, вызванное возмущением в виде одной какой-то произвольной моды, не будет возрастать, то не будет возрастать и решение, вызванное возмущением $\varepsilon(x)$.

Следуя этой идее, представим возмущение в виде одной гармоники с произвольным k_m (волновым числом): $\exp(jk_m x)$. Тогда решение на временном слое $n = 1$, соответствующее $t = \tau$, будет иметь вид:

$$G \exp\left(jk_m x\right) \ ,$$

где G — в общем случае комплексное число, называемое множителем перехода (коэффициентом перехода, коэффициентом усиления). Такой вид решение будет иметь вследствие линейности исходного ДУ. Очевидно, что если мы покажем, что

$$\left|\frac{G\exp(jk_mx)}{\exp(jk_mx)}\right| \le 1 \ , \quad \text{или} \quad |G| \le 1 \ ,$$

для любых k_m, то РС будет устойчивой.

Проанализируем устойчивость явной РС ВВЦП для ДУ теплопроводности по методу Неймана.

$$U_i^{n+1} = \lambda U_{i+1}^n + (1-2\lambda)U_i^n + \lambda U_{i-1}^n \ .$$

Берем в качестве возмущения на n-ом временном слое $U(n,x) = \exp(jk_mx)$. Тогда:

$$U_i^n = \exp(jk_mx_i) \ , \quad U_{i+1}^n = \exp(jk_mx_{i+1}) \ , \quad U_{i-1}^n = \exp(jk_mx_{i-1}) \ , \quad U_i^{n+1} = G\exp(jk_mx_i) \ .$$

Подставим в РС

$$G\exp(jk_mx_i) = \lambda\exp(jk_mx_{i+1}) + (1-2\lambda)\exp(jk_mx_i) + \lambda\exp(jk_mx_{i-1}) \ .$$

Вспомним, что $x_{i+1} = x_i + h$, $x_{i-1} = x_i - h$.

$$G\exp(jk_mx_i) - \exp(jk_mx_i) = \lambda\left(\exp\left(jk_m(x_i+h)\right) - 2\exp(jk_mx_i) + \exp\left(jk_m(x_i-h)\right)\right) \ .$$

$$(G-1)\exp(jk_mx_i) = \lambda\left(\exp\left(jk_mx_i\right)\exp\left(jk_mh\right) - 2\exp(jk_mx_i) + \exp\left(jk_mx_i\right)\exp\left(-jk_mh\right)\right) \ .$$

Сокращаем на $\exp\left(jk_mx_i\right)$ и получаем

$$G - 1 = \lambda\left(\exp\left(jk_mh\right) - 2 + \exp\left(-jk_mh\right)\right) \ .$$

Учитывая, что

$$\exp\left(jk_mh\right) + \exp\left(-jk_mh\right) = 2\cos(k_mh) \ ,$$

имеем

$$G - 1 = \lambda\left(2\cos(k_mh) - 2\right) = 2\lambda\left(\cos(k_mh) - 1\right) = -4\lambda\sin^2\frac{k_mh}{2} \ ,$$

$$G = 1 - 4\lambda\sin^2\frac{k_mh}{2} \ .$$

Теперь используем условие устойчивости $|G| \le 1$

$$\left|1 - 4\lambda\sin^2\frac{k_mh}{2}\right| \le 1 \ ,$$

$$\begin{cases} 1 - 4\lambda\sin^2\frac{k_mh}{2} \le 1 \\ 1 - 4\lambda\sin^2\frac{k_mh}{2} \ge -1 \end{cases} \ .$$

Первое неравенство выполняется автоматически, т.к. λ — положительно. Второе неравенство дает

$$2 \ge 4\lambda\sin^2\frac{k_mh}{2} \ , \quad \Longrightarrow \quad \lambda \le 1\Big/2\sin^2\frac{k_mh}{2} \ .$$

Т.к. это неравенство должно выполняться для любых волновых чисел, то получаем

$$\lambda \le \frac{1}{2} \ .$$

Заметьте, что номер рассматриваемого узла i даже не вошел в полученное условие устойчивости, т.е. получено условие для произвольного узла по координате. И второе: мы внесли возмущение в момент времени $t = 0$, т.е. при $n = 0$, и рассмотрели отклик на него при $n = 1$. Однако легко показать, что схема будет устойчивой на любом произвольном шаге по времени. Для этого представим, что возмущение внесено на каком-то n-ом шаге по времени. В этом методе возмущение вносится во все узлы сетки, а не в один, как в методе дискретных возмущений. Тогда в силу линейности исходного уравнения соотношение между возмущением и откликом будет тем же самым

$$\exp(jk_m x_i) \to G\exp(jk_m x_i) \ ,$$

и, если $|G| \le 1$, то возмущение не будет возрастать от шага к шагу.

$$\exp(jk_m x_i) \to \left(G^1 G^2 G^3 \dots G^n\right)\exp(jk_m x_i) \ ,$$

и, если каждое $|G^i| \le 1$, то ....

Аналогично, анализ устойчивости по Нейману можно сделать для неявной РС для уравнения теплопроводности.

$$G\exp(jk_m x_i) - \exp(jk_m x_i) = \lambda\left(G\exp(jk_m(x_i+h)) - 2G\exp(jk_m x_i) + G\exp(jk_m(x_i-h))\right) \ .$$

Если проделать очень простые преобразования, то получим:

$$G = 1\Big/1 + 4\lambda\sin^2\frac{k_m h}{2} \ .$$

Отсюда сразу видно, что $|G| \le 1$ для любых x_i и k_m, т.к. $\lambda > 0$ по своему смыслу. Следовательно, неявная РС для уравнения теплопроводности является абсолютно устойчивой (безусловно устойчивой).

Так же легко показать, что схема Кранка–Николсона для уравнения теплопроводности, как истинный представитель неявных РС, абсолютно устойчива.

Легкость, с которой мы получаем результаты, в рамках метода Неймана, позволяет нам рассмотреть устойчивость РС аппроксимации еще для одного важного ДУ. Я имею в виду

Одномерное волновое уравнение первого порядка.

Линейное одномерное уравнение конвекции, одномерное линейное уранение переноса.

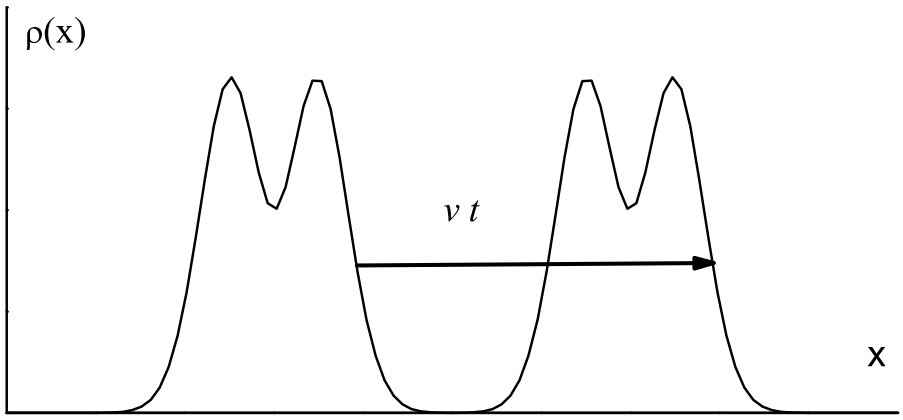
$$\frac{\partial U}{\partial t} + v\cdot\frac{\partial U}{\partial x} = 0 \ ,$$

U — некоторая скалярная величина (плотность, температура), v — некоторая заданная величина (не вычисляемая из этого ур-ия), это скорость в задачах описания конвективного переноса. Это уравнение описывает перенос (распространение в пространстве) со скоростью v некоторой величины, характеризующей сплошную среду.

Понятно как выглядит решение этого уравнения: это просто распространение начального распределения со скоростью v, причем форма начального распределения не меняется. Т.е. если $U(x,0) = F(x)$, то $U(x,t) = F(x-vt)$. Рассмотрим для этого уравнения несколько РС.

РС ВВЦП:

$$\frac{U_i^{n+1} - U_i^n}{\tau} + v\cdot\frac{U_{i+1}^n - U_{i-1}^n}{2h} = 0 \ .$$



Проанализируем устойчивость по Нейману. Как и ранее, возмущение — $\exp(jk_m x_i)$, отклик — $G\exp(jk_m x_i)$.

$$\frac{G\exp(jk_m x_i) - \exp(jk_m x_i)}{\tau} + \frac{v}{h}\frac{\exp\left(jk_m(x_i+h)\right) - \exp\left(jk_m(x_i-h)\right)}{2} = 0 \ ,$$

$$G - 1 + \frac{v\tau}{h}\cdot j\sin(k_m h) = 0 \ .$$

Величина $v\tau/h = C$ называется числом Куранта. Это общепринятое обозначение этого числа.

$$G - 1 + Cj\sin(k_m h) = 0 \ ,$$

$$|G| = \sqrt{1^2 + C^2\sin^2(k_m h)} \ .$$

Сразу же видно, что для всех $C > 0$: $|G| > 1$. Вывод: схема абсолютно неустойчива.

Вторая РС: вперед по времени и с левой разностью по координате.

$$\frac{U_i^{n+1} - U_i^n}{\tau} + v\cdot\frac{U_i^n - U_{i-1}^n}{h} = 0 \ .$$

Проанализируем устойчивость по Нейману. Получим:

$$G - 1 + C\left(1 - \exp(-jk_m x_i)\right) = 0 \ ,$$

$$G = 1 - C + C\cos(k_m x_i) - jC\sin(k_m x_i) = \left(1 - C + C\cos(k_m x_i)\right) + j\cdot\left(-C\sin(k_m x_i)\right) \ ,$$

$$G = \sqrt{\left(1 - C + C\cos(k_m x_i)\right)^2 + C^2\sin^2(k_m x_i)} \ .$$

Итак, нам нужно $|G| \le 1$ — устойчивость схемы. Анализ показывает, что данное неравенство выполняется, если $0 \le C \le 1$, т.е. РС устойчива при выполнении

$$0 \le \frac{v\tau}{h} \le 1 \ .$$

Это условие даже имеет свое название: условие Куранта – Фридрихса – Леви (КФЛ).

Получили, что РС можно применять только в случае, если $v > 0$, т.к. h и τ положительны по определению. А что делать, если $v < 0$? Оказывается, что в этом случае будет условно устойчива другая схема:

$$\frac{U_i^{n+1} - U_i^n}{\tau} + v\cdot\frac{U_{i+1}^n - U_i^n}{h} = 0 \ ,$$

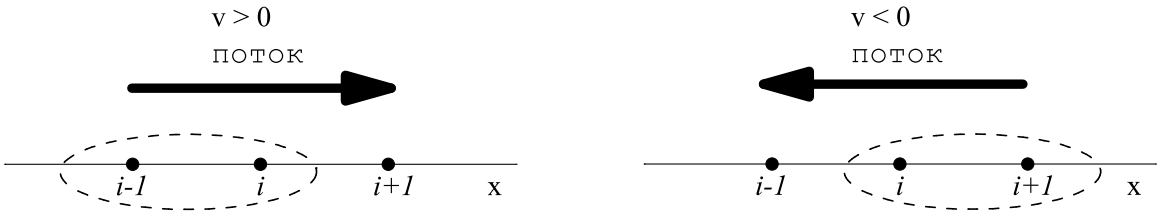
где пространственная производная аппроксимируется правой разностью. Она будет условно устойчивой при $v < 0$ и при выполнении условия:

$$\frac{|v|\tau}{h} \leq 1 \ .$$

Полученным результатам можно дать полезную физическую интерпретацию, которая иногда может помочь оценить устойчивость даже не прибегая к методам анализа устойчивости.

Во-первых:
Как можно интерпретировать величину h/v? $h/v = t_p$, где t_p — время распространения физического свойства на расстояние, равное одному пространственному шагу сетки. А условие КФЛ требует $v\tau/h \leq 1$, т.е. $\tau \leq t_p$. Словами можно так: выбираемый шаг по времени должен быть меньше, чем время распространения физического процесса на один пространственный шаг сетки.

Во-вторых:
Изобразим схематически направление потоков физ. величины U в случае $v > 0$ и в случае $v < 0$. На этой же схеме изобразим узлы пространственной сетки. Если $v > 0$, то устой-



чивой является схема, в которой пространственная производная аппроксимируется левой разностью, т.е. для вычисления используются значения величины U в узлах i и $i-1$. Если $v < 0$, то устойчивой является схема, в которой пространственная производная аппроксимируется правой разностью, т.е. для вычисления используются значения величины U в узлах i и $i+1$. Обратите внимание, в каждом из этих случаев устойчивой оказалась РС, в которой для вычислений используется информация из области "вверх по потоку". Существует целый класс разностных аппроксимаций ДУ называемых "аппроксимациями разностями против потока".

В разных областях течений скорости могут иметь разный знак, поэтому даже сам вид аппроксимации ДУ выбирается в зависимости от того, какой знак имеет скорость.

Итак, если вы построили РС какого-то ДУ в ЧП. Вы должны выполнить следующие обязательные действия (перед тем, как сядете писать программу):
• Рассчитайте с какой погрешностью ваша РС аппроксимирует исходное ДУ. Это позволит вам сделать вывод о том, согласована ваша РС или нет с исходным ДУ, и скоординировать изменение шагов по независимым переменным.
• Оцените по Нейману устойчивость вашей РС. Это позволит сделать вывод, будет ли сходиться численное решение к чему-нибуть полезному.

Консервативность РС

Расчет по любой РС всегда является лишь приближенным. Измельчая сетку вы можете уменьшать погрешность полученного вами решения, т.е. добиваться все большей и большей близости численно полученной функции $U(x,t)$ и истинного решения исходного ДУ. И вот тут появляется один ньюанс, которому физики почему-то любят придавать большое значение. Очень часто ДУ, которое вы решаете численными методами, является следствием какого-нибуть физического закона сохранения (массы, импульса, энергии). Уравнение в ЧП описывает эти законы в точке. Так, например, линейное одномерное уравнение конвекции первого порядка

$$\frac{\partial U}{\partial t} + v \cdot \frac{\partial U}{\partial x} = 0 \ ,$$

если U трактовать как плотность жидкости ρ, — есть не что иное, как закон сохранения массы. Уравнение теплопроводности

$$\frac{\partial U}{\partial t} = a \frac{\partial^2 U}{\partial x^2} \ ,$$

если U — температура T, — есть не что иное, как закон сохранения энергии. Конечно-разностная аппроксимация (схема) аппроксимирует ДУ в некоторой области, содержащей несколько узлов сетки (в области, содержащей узлы сетки, входящие в РС). Поскольку РС аппроксимирует исходное ДУ в окрестности каждого узла лишь приближенно и решение, которое мы находим является лишь приближенным, то естественно ожидать, что законы сохранения, конечно, будут выполняться, но тоже лишь приближенно.

Что это означает. Допустим, вы рассматриваете задачу о естественной конвекции в полностью замкнутом сосуде с непроницаемыми стенками. В начальный момент жидкость, заполняющая сосуд, покоится. Боковые стенки поддерживаются при разной температуре. В сосуде начинается перенос тепла от более горячей стенки к холодной, который сопровождается естественной конвекцией. Вы облекаете эту задачу в форму математической системы ДУ в ЧП, а это такие уравнения, как уравнение неразрывности (закон сохранения массы), уравнения Навье–Стокса (закон сохранения импульса), уравнение теплопроводности (закон сохранения энергии), и начинаете решать вашу систему. Предположим, что в итоге вы получили какое-то весьма правдоподобное решение. И вот тут вы замечаете один прелюбопытный ньюанс. Если вы исходя из полученного вами решения $\rho(\vec{r},t)$, подсчитаете массу жидкости, то обнаружите, что она изменилась. Может быть и не сильно, на какие-то доли процента, но все равно как-то это неприятно: сосуд был закрыт, стенки непроницаемы. Дальше хуже. Если вы, опять же пользуясь вашим решением, подсчитаете полное количество тепла пройденное сквозь вашу жидкость и потраченное на нагрев отдельных частей системы, то вы можете обнаружить, что и с законом сохранения энергии у вас не все в порядке. Конечно, на эти мелочи можно было бы и закрыть глаза, тем более, что при измельчении сетки эти эффекты уменьшаются, но... физики почему-то становятся очень щепетильны, когда речь заходит о законах сохранения. Они говорят: "Вы можете с любой приемлимой для вас погрешностью находить ваше решение, т.е. конкретное распределение плотности в сосуде, или конкретное поле скоростей и температур, но потрудитесь при этом, чтобы законы сохранения выполнялись точно. Законы сохранения — это святое."

Так вот по способности точно удовлетворять интегральным законам сохранения все РС подразделяют на два типа: консервативные и неконсервативные. По определению: консервативной называется РС, которая обеспечивает точное (исключая погрешности округления) выполнение законов сохранения на любой сетке в конечной области, содержащей любое число узлов. Современные тенденции при численном решении системы уравнений

тепломассопереноса (в численном моделировании) — стремление применять консервативные РС.

Как проверить, является ли построенная вами РС консервативной. Покажем все это на примере уравнения неразрывности.

$$\frac{\partial \rho}{\partial t} + v \cdot \frac{\partial \rho}{\partial x} = 0 \ ,$$

Это уравнение является следствием закона сохранения массы (если оно записано для плотности).

$$\frac{\partial}{\partial t} \int_V \rho dV = - \int_S \vec{v} \rho d\vec{s} \ .$$

Нормаль внешняя по отношению к рассматриваемому обьему. Это интегральная формулировка. Чтобы получить ДУ неразрывности, достаточно вспомнить теорему Остроградского–Гаусса. Интеграл от потока некоторой величины на поверхности равен интегралу от дивиргенции этой величины по объему, ограниченному данной поверхностью. В одномерном случае получаем:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (v\rho)}{\partial x} = 0 \ ,$$

ρ — плотность (неизвестная функция), v — скорость (известная функция), т.е. из данного уравнения ищется $\rho(x,t)$. Если v — const, то мы получим, то что писали раньше, но теперь мы будем считать, для большей общности, что v зависит от координаты, и потому оставим ее под знаком производной. Запишем одну из возможных РС для этого ДУ (вперед по времени, вправо по координате).

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} + \frac{\rho_{i+1}^n v_{i+1}^n - \rho_i^n v_i^n}{h} = 0 \ .$$

Теперь чтобы проверить, удовлетворяет ли наша РС интегральному закону сохранения на произвольном множестве узлов сетки, мы должны построить разностный аналог интегрального закона на этом самом произвольном множестве узлов L.

$$\frac{\partial}{\partial t} \int_L \rho dV = - \int_S \vec{v} \rho d\vec{s} \ .$$

В качестве множества L берем узлы с, начиная с номера k по номер m.

$$\frac{\partial}{\partial t} \int_{x_k}^{x_m} \rho dx = -\vec{v} \cdot \vec{n} \rho|_{x=x_k} - \vec{v} \cdot \vec{n} \rho|_{x=x_m} = v_k^n \rho_k^n - v_m^n \rho_m^n \ .$$

Интеграл же заменим на его разностный аналог самым примитивным образом, по методу прямоугольников:

$$\sum_{i=k}^{i=m-1} \frac{\partial \rho}{\partial t} h = v_k^n \rho_k^n - v_m^n \rho_m^n \ .$$

В таком виде закон сохранения записан для произвольной конечной области. Поэтому РС будет консервативной, если она будет удовлетворять этому конечно-разностному аналогу. Проверить это уже понятно как. Подставим в разностный аналог интегрального соотношения производную по времени в том виде, как она вычисляется по РС, и посмотрим получаем ли мы при этом тождество.

$$\sum_{i=k}^{m-1} h \left(- \frac{\rho_{i+1}^n v_{i+1}^n - \rho_i^n v_i^n}{h} \right) = - \sum_{i=k}^{m-1} \left(\rho_{i+1}^n v_{i+1}^n - \rho_i^n v_i^n \right) =$$
$$- \Big(\rho_{k+1}^n v_{k+1}^n - \rho_k^n v_k^n + \rho_{k+2}^n v_{k+2}^n - \rho_{k+1}^n v_{k+1}^n + \dots$$
$$+ \rho_{m-1}^n v_{m-1}^n - \rho_{m-2}^n v_{m-2}^n + \rho_m^n v_m^n - \rho_{m-1}^n v_{m-1}^n \Big) = - \left(- \rho_k^n v_k^n + \rho_m^n v_m^n \right) \ .$$

Структура получившейся суммы такова, что взаимно компенсируются потоки через прилегающие грани ячеек сетки. Т.е. при использовании приведенной выше аппроксимации ДУ выполнился интегральный аналог закона сохранения (для произвольного обьема!, содержащего произвольное число узлов сетки!). Поэтому РС — консервативная.

Может показаться, что это совершенно очевидно и должно выполняться для любой РС аппроксимации ДУ, но это не так. Покажем это на другой РС для этого же уравнения. Запишем это уравнение по-другому, выполнив только формальное преобразование производной от произведения.

$$\frac{\partial \rho}{\partial t} + \rho \frac{\partial v}{\partial x} + v \frac{\partial \rho}{\partial x} = 0 \ .$$

Разностные аппроксимации для производных возьмем те же самые:

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = -\frac{1}{h} \left(\rho_i^n (v_{i+1}^n - v_i^n) + v_i^n (\rho_{i+1}^n - \rho_i^n) \right) \ .$$

Вычислим ту же сумму, подставив в нее разностную аппроксимацию производной по времени уже из этой новой аппроксимации ДУ.

$$- \sum_{i=k}^{m-1} \left[\rho_i^n \left(v_{i+1}^n - v_i^n \right) + v_i^n \left(\rho_{i+1}^n - \rho_i^n \right) \right] = - \sum_{i=k}^{m-1} \left(\rho_i^n v_{i+1}^n - \rho_i^n v_i^n + v_i^n \rho_{i+1}^n - v_i^n \rho_i^n \right) \ .$$

Видно, что получилось что-то другое. Преобразуем, добавляя и вычитая $\rho_{i+1}^n v_{i+1}^n$,

$$= - \sum_{i=k}^{m-1} \left(\rho_{i+1}^n v_{i+1}^n - \rho_i^n v_i^n \right) - \sum_{i=k}^{m-1} \left(-\rho_{i+1}^n v_{i+1}^n + \rho_i^n v_{i+1}^n + v_i^n \rho_{i+1}^n - v_i^n \rho_i^n \right) \ .$$

Т.е. РС удовлетворяет разностному аналогу какого-то другого интегрального соотношения. В такой РС как-бы замаскированы какие-то источники и стоки, что совсем не может быть для закона сохранения массы. Первая сумма в точности совпадает с тем, что уже получалось, т.е. $\rho_k^n v_k^n - \rho_m^n v_m^n$. Но добавилась еще сумма, структура которой явно такая, что в общем случае она не обращается в нуль:

$$- \sum_{i=k}^{m-1} \left(-\rho_{i+1}^n v_{i+1}^n + \rho_i^n v_{i+1}^n + v_i^n \rho_{i+1}^n - v_i^n \rho_i^n \right) = - \sum_{i=k}^{m-1} \left(\rho_{i+1}^n - \rho_i^n \right) \left(v_i^n - v_{i+1}^n \right) \ .$$

Можно проверить, подставляя для $i = 0, 1, \ldots$, что слагаемые в этой сумме не компенсируются. Из второй записи видно, что такая схема будет консервативной только в случае, когда v не зависит от x.

Вывод: вторая схема не является консервативной. Внутри расчетной области могут появиться источники и стоки небольшой (обычно) интенсивности. Еще раз обращаю внимание на то, что понятие консервативности РС никак напрямую не связано с понятием точности. Может оказаться так, что расчет по неконсервативной схеме даст гораздо более высокую точность вычисления неизвестной функции (численное решение будет гораздо точнее). Вопрос о том, нужно ли обеспечивать очень точное выполнение законов сохранения в конечной области (т.е. применять консервативную схему), зависит от поставленной задачи. Некоторые задачи вообще нельзя решить без применения консервативной схемы. Обычно надо применять консервативную схему, если задача связана с описанием процесса в какой-то замкнутой области. Если задача, например, типа пограничного слоя, то часто не надо. Вообще-то, что это давний спор. Он идет всю недолгую пока историю развития численных методов в применении к задачам тепломассопереноса и газовой динамики.

В заключение о способах получения консервативных схем. Обратите внимание на различие в форме записи исходного уравнения. Из одной формы получилась консервативная

схема, а из другой — нет, хотя разностные аппроксимации производных взяты одни и те же. Первая форма записи уравнения называется "дивергентной" или, что то же самое: "консервативной".

Уравнение записано в дивергентной (консервативной) форме, если коэффициенты при производных являются либо константами, либо функциями, производные от которых в уравнение не входят. Если построить РС для уравнения, записанного в дивергентной форме, то схема обязательно будет консервативной.

Последний пример, относящийся к понятию консервативности. Одномерное уравнение теплопроводности в случае, когда плотность, удельная теплоемкость и коэф. теплопроводности зависят от координаты или температуры:

$$\frac{\partial(\rho c T)}{\partial t} = \frac{\partial}{\partial x}\left(\kappa \frac{\partial T}{\partial x}\right) \ .$$

Это дивиргентная (консервативная) форма. Любая РС, придуманная вами для этого ДУ, будет консервативна. В недивергентной форме:

$$\frac{\partial(\rho c T)}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} + \frac{\partial \kappa}{\partial x}\frac{\partial T}{\partial x} \ .$$

Теперь уже нельзя сказать наперед, что у вас получится.

Разностные аппроксимации граничных условий (ГУ)

До этого мы все время рассматривали МКР в применении к ДУ, для которых поставлены ГУ первого рода. Это наиболее простой случай, т.к. реализация ГУ первого рода в МКР наиболее простая — просто значения в граничных узлах принимаются равными значениям при соответствующих значениях координат. ГУ первого рода не требуют аппроксимации.

Напомню, что для получения какого-то конкретного решения уравнения в ЧП в общем случае необходимо задать вспомогательные (дополнительные) условия. Т.е. это какие-то условия, накладываемые на искомую функцию и (или) ее производные на границах расчетной области (области изменения независимых переменных). В случае, если одной из независимых переменных является время, такие дополнительные условия подразделяют на начальные и граничные. Ясно, что с точки зрения математики они представляют собой одни и те же дополнительные условия. Количество и вид вспомогательных условий определяется видом (типом) уравнения и расчетной областью (областью изменения независимых переменных).

Чаще всего выделяют три типа ГУ:

1. ГУ первого рода (условия Дирихле, задача Дирихле).
   $U_G = f$, где f — заданная функция независимых переменных. G — граница расчетной области.

2. ГУ второго рода (условия Неймана, задача Неймана).
   $\left(\dfrac{\partial U}{\partial n}\right)_G = \phi$, где ϕ — функция независимых переменных. n — направление внешней нормали к границе.

3. ГУ третьего рода (смешанная граничная задача, условия Робина).
   $\left(\dfrac{\partial U}{\partial n}\right)_G + kU_G = F$, где F — функция независимых переменных.

В численных методах решения ДУ только ГУ первого рода могут быть реализованы точно.

При постановке ГУ второго и третьего рода всегда вносятся дополнительные ошибки, т.к. реализация этих условий в численных методах всегда связана с аппроксимацией входящих в них производных. Мы рассмотрим самые основные "общие" моменты при реализации ГУ. Можно выделить несколько способов реализации ГУ в численных методах.

Пример возможных реализаций ГУ второго рода.

Возьмем то же нестационарное одномерное уравнение теплопроводности.

$$\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2} \ ,$$

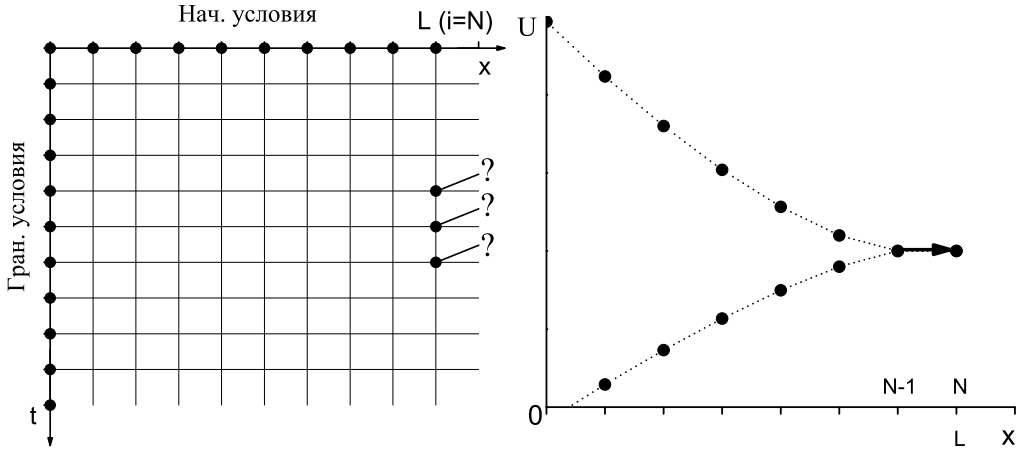
НУ: $U(x,0) = \phi(x)$, а ГУ возьмем второго рода (на одной границе)

$$\begin{cases} U(0,T) = \alpha \ , \\ \left(\frac{\partial U}{\partial x}\right)_{x=L} = 0 \ . \end{cases}$$

Физическая интерпретация такая: правый конец стержня теплоизолирован (тепловой поток равен 0)

$$\vec{q} = -\kappa\vec{\nabla}U \ , \quad \text{в одномерном случае:} \quad q_x = -\kappa\frac{\partial U}{\partial x} \ .$$

Поэтому если $q_x = 0$, то $\left(\frac{\partial U}{\partial x}\right)_G = 0$. Изобразим сетку (см. рис.). Отличие этого случая от



всех предшествующих состоит в том, что значения U_i^n при $i = N$ нам теперь неизвестны и, следовательно, их надо получить.

Предположим, что для аппроксимации мы использовали явную РС.

$$U_i^{n+1} = U_i^n + \frac{a\tau}{h^2}\left(U_{i+1}^n + U_{i-1}^n - 2U_i^n\right) \ , \quad i = 1 \ , \dots \ , N \ .$$

Обратите внимание: теперь надо вычислять и U_N^n, т.е. при $i = N$. Какие возможны способы решения этой задачи? Ясно, что здесь обязательно надо использовать ГУ при $x = L$: $\left(\frac{\partial U}{\partial x}\right)_{x=L} = 0$. Непосредственно в таком виде это условие использовать не удается — надо его аппроксимировать, т.е. заменить производную ее разностной аппроксимацией. Как вам уже известно, сделать это можно многими способами.

Первый способ.

$$\frac{\partial U}{\partial x} = \frac{U_i^n - U_{i-1}^n}{h} \ ,$$

т.е. в точке $x = L$ (при $i = N$) будет

$$\frac{U_i^n - U_{i-1}^n}{h} = 0 \ .$$

Отсюда $U_N^n = U_{N-1}^n$ при всех n, т.е. на каждом шаге по времени! Т.е. в этом случае идет счет по РС аппроксимации ДУ до $i = N - 1$ как и при ГУ первого рода, а значение температуры на границе принимается равным значению, полученному в ближайшем узле, находящемся в расчетной области. Графическая интерпретация: (см. рис.). Пунктир: возможные поведения решения уравнения вблизи границы. Заметьте, что при использованной аппроксимации ГУ мы делаем ошибку $\sim O(h)$.

Второй способ.

$$\left(\frac{\partial U}{\partial x}\right)_{x=L} = \frac{U_{i+1}^n - U_{i-1}^n}{2h} \ .$$

Использована центральноразностная аппроксимация для $i = N$. Значит на каждом шаге по времени

$$\frac{U_{N+1}^n - U_{N-1}^n}{2h} = 0 \ , \quad \text{т.е.} \quad U_{N+1}^n = U_{N-1}^n \ .$$

Сразу нужно отметить, что погрешность такой замены $O(h^2)$. Здесь возможны две интерпретации постановки ГУ в таком виде:

а) Расчет температуры идет по РС ДУ, но теперь не до $i = N - 1$, а до $i = N$, НО при

вычислении U_N^n используется число U_{N+1}^n, которое входит в РС равное (заменяется) U_{N-1}^n. Обратите внимание: для $i = N$ будет несколько другая формула РС.
б) Вторая интерпретация постановки ГУ через центральноразностную аппроксимацию связана с введением линии "фиктивных" узлов для $i = N + 1$, за границей расчетной области. Действительно, если мы введем такую линию и продолжим начальные условия за границу, приняв $U_{N+1}^0 = U_{N-1}^0$, то мы сможем рассчитать U_N^1 по той же РС, что и внутри расчетной области. Рассчитав все U_i^1 вплоть до $i = N$, мы присвоим функции U в фиктивном $(N + 1)$-ом узле значение из узла $(N - 1)$. И так далее ....

Выбор между этими двумя способами интерпретации такой постановки ГУ определяется личным вкусом, опытом работы, наглядностью программирования. Как отмечается в литературе, более нагляден второй способ интерпретации.

Третий способ.

Использование специальных конечно-разностных формул, как правило несимметричных.

$$\left(\frac{\partial U}{\partial x}\right)_{x=L} = \frac{U_{i-2}^n - 4U_{i-1}^n + 3U_i^n}{2h} \ .$$

Погрешность такой замены $\sim O(h^2)$. Используя ГУ, можно получить

$$U_{N-2}^n - 4U_{N-1}^n + 3U_N^n = 0 \ ,$$

и все точно также, как и в первом способе: по разностной аппроксимации считаем при $i = 1, \ldots, N - 1$, и значение: $U_N^n = \frac{1}{3}\left(4U_{N-1}^n - U_{N-2}^n\right)$.

Завершая разговор о ГУ необходимо отметить два момента:

1. Разностная аппроксимация ГУ может испортить точность аппроксимации, понизить порядок аппроксимации всего уравнения. Например. Явная РС аппроксимации уравнения теплопроводности $\sim O(\tau, h^2)$; если же задано ГУ второго рода и оно аппроксимируется первым способом (левой разностью), то в целом точность решения будет уже соответствовать $O(h)$. Считается, хотя это не обязательно надо выполнять, что точность аппроксимации ГУ должна соответствовать порядку аппроксимации ДУ.

2. Разностная аппроксимация ГУ может сделать более жестким условие устойчивости. Рассмотренный ранее анализ устойчивости никак не касался ГУ (для простоты изложения основных идей). С учетом ГУ условия устойчивости в некоторых случаях могут ужесточаться.