В.И.РАЩИКОВ А.С.РОШАЛЬ

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ФИЗИЧЕСКИХ ЗАДАЧ

УДК 519.6

Р а щ и к о в В. И., Рошаль А. С.. Численные методы решения физических задач. Учебное пособие. СПб: Лань, 2004.-124 с.

В пособии изложены общий подход к решению физических задач численными методами и элементы теории погрешностей; методы интерполирования и аппроксимирования, численного решения нелинейных уравнений, дифференцирования и интегрирования; численные методы линейной алгебры, методы численного решения обыкновенных дифференциальных уравнений и простейших уравнений в частных производных. Особое внимание уделяется применению численных методов к решению инженерно-физических задач. Введение, главы 1—4 написаны В. И. Ращиковым, главы 5—10 и заключение— А.С. Рошалем.

Предназначено для физических и инженерно-технических специальностей.

ВВЕДЕНИЕ

Трудно представить область научных исследований, где бы в той или иной мере не использовался компьютер. В процессе решения физических задач компьютер является мощным инструментом исследования, и как всякий сложный инструмент требует специальной подготовки для квалифицированного использования. Успех решения задачи связан с правильным и корректным построением всего вычислительного процесса получения численного результата и его интерпретации. Целью расчетов в подавляющем большинстве физических задач является понимание изучаемого явления, а не конкретные числа, поэтому прежде чем приступить к решению задачи, необходимо ясно представлять себе, как может быть использовано полученное решение.

При изучении интересующих нас физических объектов и процессов с помощью компьютера будем проводить не физический, а математический или численный эксперимент, суть которого заключается в изучении процессов и систем с помощью математического моделирования на компьютере на основе уравнений и алгоритмов, описывающих эти процессы или системы.

Попробуем определить место, занимаемое численным экспериментом в ряду других способов изучения реального мира. Процесс любого исследования можно представить себе как цепь последовательных взаимодействий теории и эксперимента. Теория, опираясь на ряд фундаментальных законов, стремится при помощи математического аппарата извлечь из них информацию для удовлетворительного описания эксперимента и составления прогнозов. В этой связи численное моделирование рассматривается как некий инструмент, позволяющий упростить используемый математический аппарат или полностью заменить аналитическое решение численным, если аналитическое решение становится очень сложным. Таким образом, численное моделирование оказывается как бы частью теории, однако это в большей степени относится к фундаментальной физике и в меньшей к прикладной, где работа на компьютере гораздо ближе к экспериментальной, чем к теоретической.

Физический эксперимент представляет собой некую модель действительности, которую мы и изучаем. Если же эта действительность слишком сложна, то мы упрощаем экспериментальную модель, отбросив несущественные на наш взгляд эффекты. Другой причиной упрощения исследуемой экспериментальной модели может служить ее чрезмерная дороговизна и, как следствие, невозможность проведения эксперимента необходимое число раз и т.д. Численное моделирование напоминает эксперимент такого рода, и поэтому часто говорят о численном эксперименте.

Как в физическом, так и в численном эксперименте, когда изучаемое явление достаточно сложно, получаемые результаты могут сильно отличаться от ожидаемых. Поэтому вычислитель, как и экспериментатор, должен детально проанализировать весь ход процесса, чтобы понять появление данного результата. В ходе такого анализа экспериментатор опирается на показания приборов, вычислитель же — на промежуточные результаты и вспомогательные величины. Как физик-экспериментатор, так и физик-вычислитель должны заранее спланировать эксперимент, т. е. определить, какие основные характеристики необходимы для понимания данного явления, а, следовательно, поставить необходимые измерительные приборы в физическом эксперименте и предусмотреть вывод промежуточных данных в численном, так как программа без выводов так же бесполезна, как и эксперимент без измерений.

Таким образом, математическое моделирование является то инструментом теории, то неким типом эксперимента. Оно может подтверждать или опровергать те или иные упрощенные гипотезы, либо использоваться как средство прогнозирования в эксперименте. Проверка численной модели, как и любой теории, осуществляется экспериментально.

К недостаткам численного эксперимента следует отнести возможность появления нефизических результатов, вызванных ограничениями используемых численных методов и машинными ошибками. Поэтому физик, занимающийся моделированием, должен знать, как происходят вычисления и представлять себе пределы применимости данной конкретной модели. В противном случае он видит лишь "голые" цифры, истинный смысл которых может быть скрыт в вычислениях. У Эдингтона есть история о человеке, который ловил в море рыбу сетью с ячейками определенного размера. Проанализировав улов, он сделал вывод о том, что самые маленькие рыбы в сети и есть самые маленькие рыбы в море. Таким образом, он допустил ошибку, поскольку не учел, как проходила ловля рыбы. Так же и результат вычислений может зависеть от того, каким методом он получен.

К преимуществам численного эксперимента следует отнести возможность выделения основных факторов, определяющих поведение изучаемого объекта и тем самым оправдать те или иные упрощающие предположения, положенные в основу описывающих его теоретических моделей. Кроме того, результаты численного эксперимента воспроизводимы, что не всегда верно в отношении физического эксперимента. Чтобы более детально исследовать какую-нибудь часть процесса в численном эксперименте, достаточно

вывести лишь дополнительную информацию, в то время как в физическом эксперименте необходима, как правило, установка дополнительной аппаратуры, которая в свою очередь может повлиять на изучаемое явление.

Являясь мощным инструментом исследования, численное моделирование требует от физика знания в определенном объеме численных методов, умения ориентироваться в них и использовать наиболее подходящие для решения поставленной задачи.

А. Основные этапы решения физических задач на компьютере

Процесс решения физических задач на компьютере можно разбить на ряд последовательных этапов, необходимых для получения конечного результата. Проанализируем эти этапы на примере простой физической задачи падения тел у поверхности Земли.

1. Изучение и постановка задачи.

Итак, мы рассматриваем падение предметов у земной поверхности, причем нас будут интересовать лишь их скорость и положение в заданные моменты времени, если они известны в начале. Таким образом, кроме описательной части задачи первый этап включает в себя и определение конечных целей решения.

2. Выбор физической модели

В качестве физической модели будем исследовать движение материальной точки по вертикали под действием сил тяготения и сопротивления воздуха, которое описывается вторым законом Ньютона:

$$m\,\vec{a}(t) = \vec{F}(y, v, t),$$

где a, v и t — соответственно ускорение, скорость и время, m — масса тела, $\vec{F}$ —равнодействующая всех сил, приложенных к телу

В зависимости от наших допущений мы можем прийти к нескольким физическим моделям. Так, например, пренебрегая сопротивлением воздуха и рассматривая лишь движение вблизи поверхности Земли, приходим к модели свободно падающего тела:

$$v(t) = v_0 + gt$$
$$y(t) = y_0 + v_0 t + \frac{1}{2} g t^2 \qquad \text{(B1)}$$

где g — ускорение свободного падения вблизи поверхности Земли, а y_0 и v_0—соответственно начальные координата и скорость тела.

Если нас будет интересовать в том числе и падение тел на достаточном удалении от поверхности Земли, физическая модель должна быть уточнена введением зависимости силы тяжести от расстояния до центра Земли:

$$F_T = GMm/(R+y)^2 = mg/(1+y/R)^2,$$

где R— радиус Земли, G — гравитационная постоянная, M — масса Земли, $g = GM / R^2$.

В случае учета сопротивления воздуха, физическая модель еще более усложняется. Теперь помимо силы тяготения $\vec{F}_T$ будем учитывать и силу сопротивления воздуха $\vec{F}_c$,т.е. равнодействующая равна сумме этих сил:

$$\vec{F} = \vec{F}_T + \vec{F}_c \qquad \text{(B2)}$$

В общем случае зависимость сопротивления среды от скорости может быть достаточно сложной и задаваться в виде некой экспериментально снятой характеристики. Однако наиболее удобно задать ее, например, в виде:

$$F_c(v) = k \cdot v,$$

а параметр k будет определяться в зависимости от свойств среды и геометрии тела. Понятно, что такой вид зависимости может быть использован лишь в определенном диапазоне скоростей и возможно потребуется его замена, например, на $F_c(v) = kv^2$ в другом диапазоне.

Физическую модель можно усложнять и дальше, вводя всевозможные вращения, и т.д. Таким образом убедились, что физическая модель будет меняться в зависимости от поставленной задачи и конечных целей решения, а ее применимость строго ограничена.

3. Разработка математической модели

Математическая модель представляет собой математическую формулировку задачи, описывающей поведение исследуемого физического явления или объекта. Она должна быть по возможности максимально простой и доступной для исследования.

В нашем случае для описания одномерного движения

материальной точки необходимо решить систему двух дифференциальных уравнений:

$$\frac{dy}{dt} = v(t);$$
$$\frac{dv(t)}{dt} = \frac{F}{m}(y, v, t), \qquad (B3)$$

где $F(y,v,t)$— равнодействующая сил, приложенных к телу. Начальные условия: $y(0)=y_0$, $v(0)=v_0$.

В случае свободного падения тел у поверхности Земли в качестве математической модели может быть использована система уравнений (В1). Как мы видим, для решения поставленной задачи важно правильно определить область применения той или иной модели.

4. Численное решение задачи

Численное решение задачи можно разбить на ряд подэтапов.

Выбор численного метода. Этот подэтап состоит в переводе математической модели на язык, понятный машине. Иными словами необходимо записать систему уравнений (В1) или (В3) в виде последовательности элементарных операций (алгоритм решения), приводящих к конечному результату. Если в случае системы (В1) сделать это достаточно просто, то для решения системы (В3) нужно воспользоваться одним из численных методов решения дифференциальных уравнений, например, методом Эйлера. Обозначим через Δt шаг по времени, тогда момент времени t_m соответствующий n-му шагу, будет определяться как

$$t_n = t_0 + n\Delta t,$$

а необходимые нам значения координаты y и скорости v из системы:

$$v_{n+1} = v_n + \frac{F_n}{m}\Delta t; \quad y_{n+1} = y_n + v_n \Delta t. \qquad (B 4)$$

Здесь скорость v_{n+1} и координата y_{n+1} в конечной точке интервала вычисляются через значения этих величин и силы, действующей на тело в начальной точке интервала. Использованный метод решения дифференциальных уравнений не является единственным. Существует множество методов, которые будут рассмотрены ниже, позволяющих наиболее рационально и с необходимой точностью решать соответствующие уравнения движения. Здесь мы лишь отметили необходимость знания численных методов и корректного их использования.

Разработка программы. На этом этапе алгоритм решения задачи записывается на понятном машине языке в виде строго определенной последовательности операций, что и называется программой.

Написание программы непосредственно на машинном языке требует от программиста знания большого количества машинных команд и весьма трудоемко. Поэтому значительно чаще для составления программ используется некоторый промежуточный язык, более близкий к математическому, после чего специальной программой (транслятором) она переводится на язык машины.

Анализ и обработка входной информации. Для решения конкретной задачи необходимы исходные данные, **с** помощью которых по программе будет вычислен результат. **В** нашем случае это начальные значения координаты и скорости тела, а также действующей на него силы (В 2). Поскольку способ задания силы, а также значения координаты и скорости могут меняться, то это необходимо заранее предусмотреть в программе.

Проведение математического эксперимента на компьютере. Составленная программа, как правило, содержит разного рода описки и ошибки. Целью данного этапа является исправление функциональных ошибок, препятствующих нормальному завершению программы. Это могут быть как синтаксические ошибки (ошибки в языке программирования), так и арифметические, логические и т.д. Исправление (отладка) осуществляется запуском программы на компьютере.

Обработка и предварительный анализ информации. После того, как **в** результате работы программы получены результаты, необходимо оценить, достаточна ли выведенная из машины информация, т.е. хватает ли ее, чтобы сделать необходимые выводы; удовлетворяет ли степень точности выведенного, удобна ли она для дальнейшей обработки.

5. Анализ результатов эксперимента

Итак, в результате работы программы получены необходимые численные значения неизвестных величин либо соответствующие графики зависимостей. В нашем случае это могут быть графики зависимостей $v(t)$ и $y(t)$ либо $v(y)$. Однако вероятность того, что эти числа или графики окажутся неверными, весьма велика. Поэтому необходимо подвергнуть их критическому анализу. Для проверки наиболее часто используются следующие способы:

проверка порядка величин, который позволяет устранить лишь только самые грубые ошибки;

моделирование простых задач, имеющих аналитические решения (в нашем случае модель, составленную по системе уравнений (В3) можно проверить по формулам (В1), положив *F/m=g);*

сравнение с программой, пригодность которой доказана независимо;

проверка на выполнение законов сохранения;

сравнение с экспериментальными результатами и т.д.

Приведенные выше способы, конечно же, не исчерпывают всего многообразия тестов, являющихся важнейшей частью численного эксперимента, поскольку они гарантируют правильность модели и позволяют достаточно просто интерпретировать ее результаты с точки зрения заложенных в нее предположений

6. Усовершенствование численной модели

Если в результате проведенного на этапе 5 анализа результатов численного эксперимента оказалось, что настоящая модель нас не устраивает, необходимо её усовершенствование. В этом случае мы должны вернуться ко второму этапу нашего решения, провести необходимую корректировку физической модели, которая в свою очередь приведет к изменениям в математической модели (3-й этап) и численном решении (4-й этап). Дальнейший анализ результатов (5-й этап) покажет, следует ли дополнительно модернизировать модель, либо она нас удовлетворяет, и тогда можно перейти к последнему этапу.

7. Реализация эксперимента

Этот этап является конечной целью численного моделирования и состоит в изучении на численной модели интересующего нас явления.

Б. Погрешности вычислений

Решая задачу на компьютере, мы получаем не точное решение, а лишь некоторое приближенное, так как в решении присутствуют, как правило, три основные типа погрешностей.

1. Исходные или неустранимые погрешности

Исходные или неустранимые погрешности появляются в результате несоответствия математической модели изучаемому физическому явлению. Так, например, могут быть не учтены какие-либо важные черты исследуемого процесса, либо нарушены границы применимости данной модели. Возвращаясь к задаче о падении тел у поверхности Земли, попытка использовать систему уравнений (В1) для исследования динамики пенопластового шарика неизбежно приведет к такой ошибке.

К тому же типу погрешностей можно отнести и погрешности исходных данных, которые в большинстве случаев являясь экспериментальными, уже содержат в себе исходную погрешность, Это могут быть начальные и граничные условия задачи, коэффициенты и правые части (силы) моделируемых уравнений и т.д.

Если обозначить через $y(t)$ координату падающего тела в физической модели, то переход к математической модели, описываемой системой уравнений (В3), может привести к неустранимым погрешностям вследствие неточного задания начальных значений координаты и скорости тела, а также действующей на него силы. В силу этого значения координаты в математической и физической моделях будут отличаться. Обозначая через $y(t)$ координату в математической модели, придем к следующему определению неустранимой погрешности:

$$\varepsilon_1 = \tilde{y} - y.$$

Во многих случаях под погрешностью понимают не рассмотренную выше разность между приближениями, а некоторые меры близости между ними. Например, в скалярном случае полагают

$$\rho_1 = |\tilde{y} - y|.$$

2. Погрешность численного метода или остаточная погрешность

Погрешность численного метода или остаточная погрешность является следствием того, что численным методом решается уже другая, упрощенная задача, которая является приближением исходной. Так, в нашем примере при решении системы дифференциальных уравнений движения тела (В3) методом Эйлера (В 4), скорость тела и действующая на него сила на шаге интегрирования постоянны, что вносит дополнительную погрешность в вычисляемое значение координаты тела y_n. Кроме того, погрешность численного метода может быть связана с заменой интеграла суммой с конечным числом членов, усечением рядов при вычислении функций, интерполированием по таблицам и т.д. Погрешность численного метода

$$\varepsilon_2 = \tilde{y}_n - \tilde{y}$$

или

$$\rho_2 = |\tilde{y}_n - \tilde{y}|.$$

как правило регулируема, так как можно изменить шаг задачи, число итераций, число точек интерполяции, число учитываемых

членов ряда и т.д. При моделировании стараются уменьшить величину этой погрешности так, чтобы она оказалась в несколько раз меньше исходной погрешности, так как дальнейшее уменьшение приводит лишь к удорожанию расчета за счет роста затрачиваемого машинного времени, не давая выигрыша в точности результата

3. Погрешности округления или вычислительные погрешности

Погрешности округления или вычислительные погрешности связаны с ограниченностью разрядной сетки компьютер. В силу этого при вводе данных в машину, при выполнении арифметических операций и выводе данных производится округление, которое и приводит к дополнительной погрешности. В нашем примере в машине имеется значение координаты тела y_n, в которое входит и ошибка округления

$$\varepsilon_3=\tilde{y}_n^*-\tilde{y}_n,\quad \rho_3=\left|\tilde{y}_n^*-\tilde{y}_n\right|.$$

Полная погрешность, **т.** е. разность между реально получаемым и точным решением задачи ε_0 удовлетворяет следующему равенству:

$$\varepsilon_0=\varepsilon_1+\varepsilon_2+\varepsilon_3=\tilde{y}_n^*-y;$$

$$\rho_0=\left|\tilde{y}_n^*-y\right|;$$

$$\rho_0\leqslant\rho_1+\rho_2+\rho_3.$$

Абсолютной погрешностью приближения $\tilde{u}$ к точному значению величины u называют величину

$$\Delta_u=|\tilde{u}-u|.$$

Относительную погрешность определяют как

$$\delta_u=\frac{\Delta_u}{|u|}\simeq\frac{|\tilde{u}-u|}{|\tilde{u}|}.$$

При сложении или вычитании чисел их абсолютные погрешности складываются, а относительная погрешность суммы заключена между наибольшим и наименьшим значениями относительных погрешностей слагаемых. При умножении или делении чисел друг на друга их относительные погрешности складываются. При возведении в степень приближенного числа его относительная погрешность умножается на показатель степени.

В случае функции многих независимых переменных

$$\Delta_u=\sum_{i=1}^{n}\frac{\partial u}{\partial x_i}\Delta x_i;$$

$$\delta_u=\sum_{i=1}^{n}\left|\frac{1}{u}\frac{\partial u}{\partial x_i}\right|\Delta x_i=\sum_{i=1}^{n}\left|\frac{\partial}{\partial x_i}\ln u\right|\Delta x_i.$$

абсолютная и относительная погрешности определяются выражениями:

В. Корректность и устойчивость

Задача $y=Ax$ называется корректно поставленной, если для любых входных данных $x\in X$ (из некоторого класса X) решение $y\in Y$ существует, единственно и устойчиво по входным данным. Первые два требования естественны, так как приступая к численному решению задачи нужно быть уверенным, что решение существует, а однозначная последовательность действий в численном алгоритме диктует требование единственности. Кроме того, если исходные данные имеют небольшую погрешность, то и погрешность решения должна быть невелика. Иными словами, задача называется устойчивой по входным данным x, если решение у непрерывно зависит от входных данных.

Отсутствие устойчивости означает, что даже незначительные погрешности в исходных данных могут привести к большим погрешностям решения или вовсе неверному результату. О таких задачах говорят, что они чувствительны к погрешностям исходных данных.

1. ИНТЕРПОЛИРОВАНИЕ В ФИЗИЧЕСКИХ ЗАДАЧАХ

Численное моделирование большинства физических задач сопряжено, как правило, с необходимостью учета факторов, которые не могут быть описаны аналитически. Имеется лишь ряд экспериментальных зависимостей, полученных в фиксированном числе точек интересующего нас диапазона переменных. Так, например, при решении широко распространенной задачи о динамике макро- и микрообъектов во, внешних гравитационных или электромагнитных полях информацию о поле часто бывает невозможно получить в виде аналитических функций без ввода дополнительных упрощающих предположений, которые могут существенно повлиять на результат. В этом случае необходимо прибегнуть к экспериментальным характеристикам, причем эксперимент может быть проведен лишь конечное число раз Таким образом, мы приходим к физической задаче, в которой ряд функций задан на конечном числе точек x_i фиксированной области изменения аргумента $x_i\in[a, b]$. Численный метод, однако, может потребовать знания этих функций для всех значений аргумента из этой области. В этом случае возникает задача восстановления функции $y(x)$ для всех значений $x\in[a, b]$, если известны ее

значения в некотором фиксированном числе точек x_i этого отрезка.

Наиболее простым и распространенным способом решения этой задачи является интерполяция между соседними значениями, которая сводится к построению функции $\varphi(x)$ совпадающей с функцией $y(x)$ в точках x_i, то есть

$$\varphi(x_i)=y(x_i)=y_i, \quad i=0,1,2,\ldots,n, \qquad (1.1)$$

где $n+1$ — число заданных на отрезке $[a, b]$ точек, а x_i—узлы интерполяции.

Таким образом, будем говорить об интерполяции, когда значение x заключено между крайними узлами интерполяции $x_0=a$ и $x_n=b$. Если же x выходит за пределы отрезка $[a, b]$, то говорят об экстраполяции. При экстраполяции далеко за пределы отрезка $[a, b]$, ошибка может оказаться очень большой.

При выборе интерполирующей функции $\varphi(x)$ необходимо ограничить поиск функциями, легко и быстро вычисляющимися на компьютере, так как их, как правило, приходится вычислять многократно.

Будем искать интерполирующую функцию $\varphi(x)$ в виде линейной комбинации некоторых элементарных функций:

$$\varphi(x)=\sum_{k=0}^{n} c_k\varphi_k(x),$$

где $\{\varphi_k(x)\}$ — фиксированные линейно независимые функции, а $\{c_k\}$ неизвестные пока коэффициенты.

Чтобы найти неизвестные коэффициенты $\{c_k\}$, воспользуемся условием (1.1), которое приводит к системе $(n+1)$ уравнений

$$\sum_{k=0}^{n} c_k\varphi_k(x_i)=y_i, \quad i=0,1,\ldots,n.$$

Для однозначного вычисления коэффициентов $\{c_k\}$ необходимо, чтобы при любом выборе узлов интерполяции x_i внутри заданного отрезка был отличен от нуля определитель

$$|\Phi|=\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0.$$

Система функций $\{\varphi_k\}$ при этом условии называется чебышевской. Зная значения функции в узлах интерполяции y_i можно определить значения коэффициентов $\{c_k\}$. Наиболее часто в качестве системы линейно независимых функций $\{\varphi_k\}$ используются степенные полиномы x^k либо тригонометрические полиномы $\sin(\omega_k x), \cos(\omega_k x), (k=0,1,\ldots,n)$.

1.1. Полиномиальная интерполяция

Полином степени n

$$P_n(x)=c_0+c_1x+c_2x^2+\ldots+c_nx^n=\sum_{k=0}^{n} c_kx^k$$

имеет $(n+1)$ коэффициент. Естественно полагать, что $(n+1)$ условие, наложенное на многочлен в общем виде, позволит однозначно вычислить коэффициенты. Потребуем, чтобы полином проходил через $(n+1)$ точку (x_i, y_i), $(i=0,1\ldots,n)$, где $y_i=\varphi(x_i)$ и $x_i\neq x_j$. Тогда приходим к системе линейных уравнений

$$\sum_{k=0}^{n} c_kx_i^k=y_i, \quad i=0,1,\ldots,n,$$

или, раскрывая сумму,

$$\begin{aligned} &c_0+c_1x_0+c_2x_0^2+\ldots+c_nx_0^n=y_0;\\ &c_0+c_1x_1+c_2x_1^2+\ldots+c_nx_1^n=y_1;\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ &c_0+c_1x_n+c_2x_n^2+\ldots+c_nx_n^n=y_n. \end{aligned}$$

Определителем этой системы линейных уравнений относительно неизвестных c_k является определитель Вандермонда:

$$W=\begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}=f(x_0,x_1,\ldots,x_n).$$

Ясно, что определитель есть функция $x_0,x_1\ldots,x_n$. Если считать его функцией от x_n, то он — многочлен степени n и обращается в нуль при

$$x_n=x_j, \quad j=0,1,\ldots,n-1, \quad \text{т. е.} \quad f(x_0,x_1,\ldots,x_n)$$

содержит множители

$$\prod_{i=0}^{n-1}(x_n - x_i) = (x_n - x_1)(x_n - x_2)\dots(x_n - x_{n-1}).$$

Рассматривая определитель как функцию от x_{n-1} мы точно так же приходим к существованию множителей

$$\prod_{i=0}^{n-2}(x_{n-1} - x_i)$$

или окончательно для всех остальных степеней

$$\prod_{j>i=0}^{n}(x_j - x_i).$$

Суммируя изложенное выше, определитель Вандермонда

$$W = \prod_{j>i=0}^{n}(x_j - x_i),$$

то есть при $x_i \neq x_j$ для $i \neq j$ интерполяционный полином существует и единственен.

Таким образом, если функция $y(x)$ задана в $(n + 1)$ узлах, то можно построить полином степени n, который совпадает (пренебрегая ошибками округления) с функцией в узловых точках. Предполагая, что полином и функция близки между узловыми точками, мы можем использовать полином вместо функции в дальнейших вычислениях или аналитических оценках.

1.2. Полином Лагранжа

В описанном выше способе построения интерполяционного полинома в качестве базисных функций $\{\varphi_k\}$ были взяты одночлены вида: $1, x, x^2, \dots, x^n$. Другим возможным подходом является использование базиса интерполяционных полиномов Лагранжа. Идея метода состоит в нахождении многочлена $L_k(x)$, который принимает значение единицы в одной узловой точке и нуль во всех остальных:

$$\delta_{ki} = \begin{cases} 1, & \text{если } k = i, \\ 0, & \text{если } k \neq i, \end{cases}$$

где δ_{ki}, —символ Кронекера.

Легко убедиться в том, что указанным свойством обладает полином степени n

$$L_k(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}.$$

Многочлен $L_k(x)y_k$ принимает значение y_k в k-й узловой точке и равен нулю во всех других узлах. Из этого следует, что интерполяционный полином Лагранжа

(1.3)

$$P_n(x) = \sum_{k=0}^{n} L_k(x)\, y_k = \sum_{k=0}^{n} y_k \prod_{i \neq k} \frac{x - x_i}{x_k - x_i}$$

имеет степень не выше n и $P_n(x_i) = y_i$.

Если в каждом узле интерполяции известно не только значение функции y_i, но и ее производной y'_i, то можно построить многочлен (2n+1) степени, так как должны быть удовлетворены $(2n + 2)$ условия:

$$P(x_i) = y_i;$$
$$P'(x_i) = y'_i,$$

где $i = 0, 1,2, \dots, n$. Если узлы x_i различны, то существует единственное решение, называемое многочленом Эрмита, и находится оно способом, аналогичным методу Лагранжа. Так, например, многочлен Эрмита для случая двух узловых точек x_0 и x_1, в которых заданы значения функции y_0, y_1 и ее производной y_0', y_1', имеет вид:

$$H(x)=y_0+(x-x_0)\left\{y_0'+\frac{x-x_0}{x_0-x_1}\left[\left(y_0-\frac{y_0-y_1}{x_0-x_1}\times\right.\right.\right.$$
$$\left.\left.\left.\times\left(y_0'-\frac{2(y_0-y_1)}{x_0-x_1}+y_1'\right)\right]\right\}.$$

1.3. Интерполяционная формула ньютона

При построении интерполяционных многочленов мы подразумевали, что множество используемых узлов известно. Однако часто бывает известна лишь требуемая точность, а число узлов не фиксировано. Построим другой интерполяционный полином, число используемых узлов которого можно было бы легко увеличить или уменьшить без повторения всего цикла вычислений, изменяя тем самым точность интерполяции.

Введем понятие разделенных разностей:

$$y(x_i, x_j)=[y(x_i)-y(x_j)]/(x_i-x_j)$$

—разделенная разность первого порядка;

$$y(x_i, x_j, x_k)=[y(x_i, x_j)-y(x_j, x_k)]/(x-x_k)$$

--разделенная разность второго порядка.

Если $y(x)=P_n(x)$ — полином степени n, то для него первая разделенная разность

$$P(x, x_0)=[P(x)-P(x_0)]/(x-x_0) \qquad (1.4)$$

есть, как видно, полином $(n-1)$ степени, а вторая разделенная разность

$$P(x, x_0\ x_1)=[P(x, x_0)-P(x_0, x_1)]/(x-x_1) \qquad (1.5)$$

— полином $(n-2)$ степени и т.д. Таким образом, мы приходим к выводу, что $(n+1)$ разделенная разность равна нулю. По определению разделенных разностей из (1.4) и (1.5) следует:

$$P(x)=P(x_0)+(x-x_0)P(x, x_0); \qquad (1.6)$$
$$P(x, x_0)=P(x_0, x_1)+(x-x_1)P(x, x_0, x_1); \qquad (1.7)$$
$$P(x, x_0, x_1)=P(x_0, x_1, x_2)+(x-x_2)P(x, x_0, x_1, x_2); \qquad (1.8)$$

и так далее до n-й разделенной разности, в которую будет входить $(n+1)$ разделенная разность, равная нулю.

Заменяя в правой части уравнения (1.6) первую разделенную разность уравнением (1.7), в котором вторая разделенная разность взята из (1.8) и т.д., приходим к интерполяционному полиному вида

$$P(x)=P(x_0)+(x-x_0)P(x_0, x_1)+(x-x_0)(x-x_1)P(x_0, x_1, x_2)+$$
$$+\ldots+(x-x_0)(x-x_1)\ldots(x-x_{n-1})P(x_0, x_1, \ldots, x_n). \qquad (1.9)$$

Поскольку $P(x)$— интерполяционный полином для функции $y(x)$, то его значения в узлах совпадают со значениями функции $y(x_i)$, а значит совпадают и разделенные разности. Следовательно, можно записать полином

$$y(x)=y(x_0)+\sum_{k=1}^{n}(x-x_0)(x-x_1)\ldots(x-x_{k-1})y(x_0, x_1, \ldots, x_k), \qquad (1.10)$$

называемый полиномом Ньютона. Когда разделенные разности вычислены, полином Ньютона удобно вычислять, используя схему Горнера, сократив число операций умножения _____

$$y(x)=y(x_0)+(x-x_0)\left[y(x_0,x_1)+(x-x_1)\left[y(x_0,x_1,x_2)+\ldots\right]\right]. \quad (1.11)$$

Вычисление полинома по такой схеме требует n операций умножения и $2n$ сложений, в то время как для построения полинома Лагранжа требуется $\sim n^2$ операций, т.е. при большом числе узлов последний метод оказывается существенно более быстрым.

1.4. Точность интерполяции

Рассмотрим вопрос о том, как сильно могут отличаться функция и многочлен в точках, отличных от узловых. Для оценки близости полинома $P_n(x)$ к функции $f(x)$ предполагают, что существует $(n + 1)$ непрерывная производная $f^{n+1}(x)$, тогда погрешность определяется следующим выражением:

$$f(x)-P_n(x)=\frac{f^{n+1}(\xi)}{(n+1)!}\prod^{n}(x-x_i), \quad \xi\in[a,b] \quad (1.12)$$

Верхняя граница погрешности определяется выражением

$$\varepsilon_{max}=h^{n+1}\frac{M}{(n+1)!},$$

где $h=|x_n-x_0|$, а $M=\max|f^{n+1}(\xi)|$.

Эта ошибка определяется тем, что для приближения функции нельзя использовать бесконечное число операций: разложение в ряд должно быть усечено до некоторого порядка, откуда и происходит название этой ошибки — ошибка усечения. Заметим, что если $(n + 1)$ производная функции не ограничена, ошибка не уменьшается при уменьшении шага сетки.

К ошибке усечения добавляются ошибки округления. Повышая порядок разложения, мы уменьшаем ошибку усечения, но увеличиваем ошибку округления, так как вырастает число операций. Если число операций невелико, то ошибками округления можно пренебречь.

1.5. Интерполяция сплайнами

Сплайном называется функция, которая вместе с несколькими производными непрерывна на всем отрезке $[a, b]$, а на частичном отрезке $[x_i,x_{i+1}]$ является алгебраическим многочленом. Максимальная по всем частичным отрезкам степень многочлена называется степенью сплайна.

Рассмотрим случай, когда между любыми соседними узлами функция $y(x)$ интерполируется кубическим полиномом — кубическим сплайном. Его коэффициенты на каждом интервале определяются из условия сопряжения в узлах

$$\varphi(x_i)=y(x_i); \quad \varphi'(x_i-0)=\varphi'(x_i+0);$$
$$\varphi''(x_i-0)=\varphi''(x_i+0). \quad (1.13)$$

Кроме этого, на границе при $x = x_0$ и $x = x_n$ ставятся условия

$$\varphi''(x_0)=0, \quad \varphi''(x_n)=0.$$

Приведенные выше уравнения представляют собой математическую модель гибкого тонкого стержня из упругого материала, закрепленного в двух соседних узлах интерполяции с заданными углами наклона. Стержень принимает форму, минимизирующую его потенциальную энергию. Пусть функция $\varphi(x)$ описывает форму стержня. Уравнение свободного равновесия имеет вид $\varphi^{IV} = 0$. т.е. для каждой пары узлов функция $\varphi(x)$ является многочленом третьей степени.

Будем искать кубический сплайн в виде

$$\varphi(x)=a_i+b_i(x-x_{i-1})+c_i(x-x_{i-1})^2+d_i(x-x_{i-1})^3,$$
$$x_{i-1}\leqslant x\leqslant x_i.$$

Воспользовавшись первым условием в (1.13), получим

$$\varphi(x_{i-1})=a_i=y_{i-1};$$
$$\varphi(x_i)=a_i+b_ih_i+c_ih_i^2+d_ih_i^3=y_i, \quad (1.14)$$

где $h_i=x_i-x_{i-1}$, $i= 1,2, \ldots,n$.

Вычисляя затем производные

$$\varphi'(x)=b_i+2c_i(x-x_{i-1})+3d_i(x-x_{i-1})^2;$$
$$\varphi''(x)=2c_i+6d_i(x-x_{i-1}),\quad x_{i-1}\leqslant x\leqslant x_i$$

и требуя их непрерывности при $x = x_i$, получим

$$b_{i+1}=b_i+2c_ih_i+3d_ih_i^2;$$
$$c_{i+1}=c_i+3d_ih_i,\quad (i=1,2,\ldots,n-1). \qquad (1.15)$$

Общее число неизвестных коэффициентов равно 4 n, а число уравнений для их определения (1.14) и (1.15) равно ($4n$ —2). Недостающие два уравнения получаем из условия равенства нулю вторых производных на концах при $x = x_0$ и $x=x_n$ (условия не закрепленности концов):

$$c_1=0,\quad c_n+3d_nh_n=0.$$

Кубический сплайн, полученный при условии незакрепленных концов (условие нулевой кривизны на концах), является самым гладким среди всех интерполяционных функций данного класса.

Выражая из (1.15) $d_i=(c_{i+1}-c_i)/3h_i$ и подставляя в (1.14), получаем, исключив $a_i=y_{i-1}$,

$$b_i=[(y_i-y_{i-1})/h_i]-\frac{1}{3}h_i(c_{i+1}+2c_i),\quad (i=1,2,\ldots,n-1);$$
$$b_n=[(y_n-y_{n-1})/h_n]-\frac{2}{3}h_nc_n.$$

Подставив теперь выражения для b_i, b_{i+1} и d_i, в первое уравнение (1.15), после несложных преобразований получаем для определения c_i уравнение второго порядка:

$$h_ic_i+2(h_i+h_{i+1})c_{i+1}+h_{i+1}c_{i+2}=3\left(\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i}\right)$$

с условиями на концах $c_1 = 0$ и $c_{n+1}=0$.

Условие $c_{n+1} = 0$ эквивалентно условию $c_n+3d_nh_n=0$ и уравнению $c_{i+1} = c_i+3d_ih_i$ при $i = n$.

Матрица этой системы трехдиагональная, т.е. отличны от нуля лишь элементы, находящиеся на главной и двух соседних с ней диагоналях сверху и снизу. Для ее решения целесообразно использовать метод прогонки, который будет описан ниже.

Можно показать, что для построенного таким образом сплайна в случае равноотстоящих узлов максимальные по модулю отклонения $\varphi(x)$от $y(x)$, $\varphi'(x)$от $y'(x)$ и $\varphi''(x)$от $y''(x)$ на отрезке $[a,b]$ равны соответственно $0(h^4)$, $0(h^3)$, $0(h^2)$. Таким образом, точность аппроксимации функции сплайном можно подобрать, изменяя величину шага интерполяции h.

Можно ввести понятие сплайна любого порядка m как функцию, которая является на каждом отрезке полиномом степени m, и во всех внутренних узлах удовлетворяет условиям непрерывности функции и производных до (m —1) порядка включительно. При m = 3 приходим к рассмотренному выше кубическому сплайну, а m = 1 соответствует замене функции ломаной, проходящей через точки (x_i ,y_i). Сплайны более удобны, чем интерполяционные многочлены для аппроксимации функций на больших отрезках с большим числом узлов, так как в этом случае для достижения заданной точности может потребоваться интерполяционный многочлен очень высокой степени, что на практике неприемлемо.

В физике достаточно часто приходится аппроксимировать многомерные функции, в силу чего широко распространена задача интерполяции по многомерным (как правило двух- и трехмерным) таблицам. Примером могут служить экспериментально полученные распределения полей физических величин в двух- и трехмерных областях, которые затем используются в численных моделях. В этом случае в качестве независимых переменных, выступают пространственные координаты. Для сокращения объема таблиц приходится использовать большие шаги по аргументам, что приводит к жестким требованиям к способу интерполяции. Достаточно часто применяют метод выравнивания, изложенный в главе 2, что позволяет использовать интерполяционные многочлены невысоких степеней. Кроме того, в многомерных случаях налагаются дополнительные условия на число узлов интерполяции и их расположение Многомерная интерполяция настолько громоздка, что используется, как правило, многочлен первой или второй степени. По этой же причине интерполяция эрмитова типа практически не употребляется, а сплайновая используется в основном при разностном решении уравнений в частных производных

1.6. Интерполирование тригонометрическими полиномами

Для интерполирования периодических функций естественно

воспользоваться тригонометрическими полиномами. Рассмотрим для общности комплексную функцию $u(x)$ с периодом l. Коэффициенты $\bar{u}(n)$ тригонометрического полинома

$$T(x)=\sum_{n=-N/2}^{N/2-1}\bar{u}(n)\,e^{i\omega_n x},\quad \omega_n=\omega_1 n,\quad \omega_1=\frac{2\pi}{l},\quad i\equiv\sqrt{-1}, \tag{1.16}$$

интерполирующего функцию $u(x)$, можно найти из системы комплексных линейных алгебраических уравнений

$$T(x_j)=u(j),\ u(j)\equiv u(x_j),\ 0\le x_j<l\ \ (j=0,1,..,N-1) \tag{1.17}$$

где x_j узлы интерполирования на полуоткрытом интервале $(0,l)$. Если все узлы x_j различны, система (1.17) имеет единственное решение, которое приводит к полиному

$$T(x)=\sum_{j=0}^{N-1}u(j)\,T_j(x), \tag{1.18}$$

где

$$T_i(x)=\frac{\pi}{l}\cdot\frac{A(x)}{A'(x_j)\sin\frac{\pi}{l}(x-x_j)},\ A(x)=\prod_{i=0}^{N-1}\sin\frac{\pi}{l}(x-x_i) \tag{1.19}$$

аналогичному интерполяционному полиному Лагранжа (1.3). Здесь штрих означает производную по x, а фундаментальные тригонометрические интерполяционные полиномы $T_j(x)$ обладают свойством

$$T_j(x_i)=\delta_{ji}\quad (j,i=0,1,...,N-1), \tag{1.20}$$

где δ_{ji} —символ Кронекера.

В дальнейшем ограничимся наиболее распространенным в физических приложениях случаем равноотстоящих узлов

$$x_j=jh,\quad h=l/N\quad (j=0,1,...,N-1). \tag{1.21}$$

При этом формулы (1.19) упрощаются и полином (1.18) приобретает вид

$$T(x)=\frac{1}{N}\sum_{j=0}^{N-1}u_j\frac{\sin[\pi N(x-x_j)/l]}{\sin[\pi(x-x_j)/l]}. \tag{1.22}$$

Для функций, заданных с шагом h на бесконечной прямой, тригонометрический интерполяционный полином

(1.23)

$$T(x)=\sum_{j=-\infty}^{\infty}u(j)\,s(x-x_j),\quad s(x)=\frac{\sin(\pi x/h)}{(\pi x/h)}\quad x_j=jh.$$

Если спектр периодической функции $u(x)$ не содержит частот выше максимальной частоты

$$f_{max}=\frac{\omega_{max}}{2\pi}=\frac{\omega N/2}{2\pi}=\frac{N}{2l}=\frac{1}{2h} \tag{1.24}$$

(такие функции называются функциями с ограниченным спектром), то в силу единственности тригонометрического интерполяционного полинома, формулы (1.22), (1.23) являются точными, т. е. точное равенство $u(x)=T(x)$ выполняется всюду, а не только в узлах x_j. Следовательно, функция $u(x)$, спектр которой ограничен частотой f_{max}, может быть точно синтезирована по своим отсчетам, выбранным с шагом $h=1/(2f_{max})$. Это утверждение составляет содержание теоремы Котельникова (или Котельникова — Найквиста) Представление

$$u(x)=\sum_{n=-N/2}^{N/2-1}\bar{u}(n)e^{i2\pi nx/t}\ (i=\sqrt{-1}) \tag{1.25}$$

для комплексной функции $u(x)$ с периодом l, заданной в равно-

отстоящих узлах (1.21), называют конечным или дискретным рядом Фурье с коэффициентами $u(n)$. При $x = x_j = jh$ (I.25) приобретает вид

$$u(j) = \sum_{n=-N/2}^{N/2-1} \bar{u}(n)\, e^{i2\pi nj/N}. \quad (1.26)$$

Домножая (1.26) на $\exp(-i2\pi nj/N)$, суммируя по всем j и учитывая дискретное соотношение ортогональности

$$\frac{1}{N}\sum_{}^{N-1} e^{i2\pi n(j-i)/N} = \delta_{ij}, \quad (1.27)$$

получаем формулу для коэффициентов Фурье

$$\bar{u}(n) = \frac{1}{N}\sum_{j=0}^{N-1} u(j)\, e^{-i2\pi nj/N} \quad \left(-\frac{N}{2} \leqslant n < \frac{N}{2} - 1\right). \quad (1.28)$$

Вычисление всех N коэффициентов $\bar{u}(n)$ по формуле (1.28) требует около N^2 комплексных умножений и сложений, не считая операций для вычисления комплексных экспонент. В физических приложениях число коэффициентов N может составлять 10^3—10^5, и объем вычислений оказывается слишком большим. Предположим теперь, что N—составное (не простое) число, т.е. $N = N_1 N_2$, где N_1, N_2 — множители N. Подставим в (1.28) j и n в виде

$$j = j_1 + j_2 N_1, \quad n = n_1 N_2 + n_2$$

$$(j_1, n_1 = 0, 1, \ldots, N_1 - 1; \quad j_2, n_2 = 0, 1, \ldots, N_2 - 1). \quad (1\text{-}29)$$

Тогда, подставляя (1.29) в (1.28), можно представить (1.28) в виде разложения

$$\bar{u}(n) = \frac{1}{N}\sum_{j_1=0}^{N_1-1} e^{-i2\pi nj_1/N} \sum_{j_2=0}^{N_2-1} u(j)\, e^{-i2\pi nj_2/N_2}. \quad (1.30)$$

В (1.30) требуется $N N_2$ операций для вычисления внутренних сумм и NN_1— для внешних; всего требуется

$N(N_1 + N_2) < N^2$ операций.

Далее, если число N_2 также составное, редукцию типа (1.30) можно аналогично применить к вычислению каждой из N_1 независимых внутренних сумм в (1.30) и т.д. В общем случае, когда N представляется в виде произведения $N = N_1 N_2 \ldots N_m$ число арифметических операций можно сократить до величины

$$N(N_1 + N_2 + \ldots + N_m) \ll N^2. \quad (1.31)$$

В частности, когда все N_k одинаковые и $N = q^m$, где q, m — целые числа, выигрыш в числе операций составляет

$$p = N^2/(m\, q\, N) = N/(q \log_q N) \quad (1.32)$$

и достигает максимума при $q = 2$ или $q = 3$. Например, при $q = 2$, $N = 2^{12} = 4096$, m=12 выигрыш оказывается равным 170. В действительности выигрыш оказывается еще большим, так как при этом многие экспоненты в формулах типа (1.30) обращаются в ±1, ± i и соответствующие умножения пропадают.

Описанный метод, основанный на сведении преобразования Фурье длинной выборки к суперпозиции независимых преобразований Фурье многих коротких выборок, называется быстрым преобразованием Фурье (БПФ) и широко используется во многих физических задачах. Найденные коэффициенты $\bar{u}(n)$ можно затем использовать в интерполяционном многочлене (1.16) или (1.25). Заметим, что учитывая вытекающее из (1.28) свойство периодичности коэффициентов Фурье $\bar{u}(n+N) = \bar{u}(n)$ можно преобразовать (1.26) к виду

$$u(j) = \sum_{n=0}^{N-1} \bar{u}(N-n)\, e^{-i2\pi nj/N}. \quad (1.33$$

Поскольку формула (1.33) аналогична (1.28), значения $u(j)$ тоже можно вычислить с помощью быстрого преобразования Фурье по коэффициентам $\bar{u}(n)$, взятым в обратном порядке

1.7. Применение быстрого преобразования Фурье (БПФ) в физике

Наиболее часто быстрое преобразование Фурье используется в двух областях:

при обработке результатов физического эксперимента;

при численном решении дифференциальных уравнений,

описывающих различные физические процессы

В качестве примера первого приложения рассмотрим корреляционный анализ. Пусть $u_1(x)$, $u_2(x)$—комплексные функции с периодом l. Их взаимная корреляционная функция $K(x)$, являющаяся важной характеристикой процессов $u_1(x)$, $u_2(x)$ определяется формулой типа свертки (иногда называемой круговой сверткой)

$$K(x)=u_1(x)\circledast \mathring{u}_2(x)=\frac{1}{l}\int_0^l \mathring{u}_1(x')\,\mathring{u}_2^*(x'-x)\,dx'; \qquad (1.34)$$

$$\mathring{u}_i(x)=u_i(x)-\langle u_i(x)\rangle, \quad \langle u_i(x)\rangle=\frac{1}{l}\int_0^l u_i(x)\,dx\ (i=1,2). \qquad (1.35)$$

Коэффициенты Фурье $\bar{K}(n)$ представляют собой взаимную спектральную плотность мощности взаимодействия (кросс-спектр) процессов $u_1(x)$ и $u_2(x)$. Если $u_1(x) = u_2(x) = u(x)$, получаем автокорреляционную функцию, коэффициенты которой $\bar{K}(n)$представляют собой спектральную плотность мощности процесса $u(x)$.

Имеется теорема о свертке, согласно которой коэффициенты Фурье $\bar{K}(n)$, $\bar{\mathring{u}}_1(n)$, $\bar{\mathring{u}}_2(n)$ связаны равенством

$$\bar{K}(n)=\bar{\mathring{u}}_1(n)\,\bar{\mathring{u}}_2^*(-n). \qquad (1.36)$$

Для функций, заданных в узлах (1.21), аналогично:

$$K(j)=u_1(j)\circledast u_2^*(j)=\frac{1}{N}\sum_{m=0}^{N-1} u_1(m)\,\mathring{u}_2^*(m-j); \qquad (1.37)$$

$$u_i(j)=u_i(j)-\langle u_i(j)\rangle, \quad \langle u_i(j)\rangle=\frac{1}{N}\sum_{j=0}^{N-1} u_i(j) \quad (i=1,2), \qquad (1.38)$$

причем, согласно теореме о свертке, по-прежнему

$$\bar{K}(n)=\bar{\mathring{u}}_1(n)\,\bar{\mathring{u}}_2^*(-n). \qquad (1.39)$$

Таким образом, для быстрого вычисления $K(j)$ сначала вычисляются коэффициенты Фурье $\bar{u}_1(n)$, $\bar{u}_2(n)$, затем $\bar{K}(n)$ (1.39) и, наконец, $K(j)$ по формуле типа (1.33). Использование при этом алгоритмов быстрого преобразования Фурье дает экономию операций более, чем на два порядка, по сравнению с вычислением непосредственно по формуле (1.37).

Примеры приложений быстрого преобразования Фурье к решению дифференциальных уравнений будут даны в следующих главах.

2. АППРОКСИМИРОВАНИЕ В ФИЗИЧЕСКИХ ЗАДАЧАХ

При замене функции интерполяционным многочленом необходимым условием является прохождение интерполяционного многочлена через значения функции в узлах интерполяции. В случае использования экспериментальных зависимостей, значения функции в узлах получены с определенной погрешностью (часто достаточно большой), поэтому нецелесообразно прибегать к интерполяции, заставляя интерполяционный полином повторять эти ошибки. В этом случае лучше воспользоваться аппроксимацией, т. е. подбором функции близко проходящей от заданных точек, заранее определив критерии «близости». В зависимости от выбранного способа приближения можно получить сильно отличающиеся друг от друга результаты: кривая может точно проходить через все заданные точки и в то же время сильно отличаться от сглаженной аппроксимирующей функции (рис. 1).

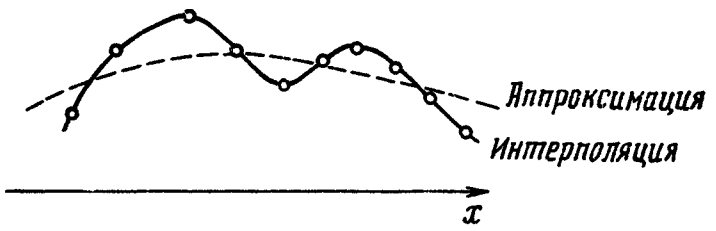


Рис. 1. Интерполяция и аппроксимация

2.1. Среднеквадратичное и равномерное приближения

Будем аппроксимировать функции многочленом степени m:

$$\varphi(x)=c_0+c_1x+c_2x^2+\ldots+c_mx^m, \qquad (2.1)$$

коэффициенты которого c_i подберем так, чтобы минимизировать отклонение многочлена от данной функции.

Воспользуемся среднеквадратичным приближением функции $y(x)$ многочленом $\varphi(x)$ на множестве (x_i, y_i), $(i = 0, 1, 2 \ldots\ldots n)$, при котором мерой отклонения является величина S, равная сумме квадратов разностей между значениями многочлена и функции в данных точках

(2.2)

$$S = \sum_{i=0}^{n} [\varphi(x_i) - y_i]^2.$$

Для построения аппроксимирующего многочлена нужно подобрать коэффициенты $c_0, c_1 \ldots, c_m$ так, чтобы величина S была наименьшей. В этом и состоит метод наименьших квадратов. Как уже упоминалось выше, среднеквадратичное приближение сглаживает неточности функции, давая правильное представление о ней.

Иногда требуются более жесткие условия на аппроксимирующий многочлен: во всех точках отрезка $[a, b]$ отклонение многочлена $\varphi(x)$ от функции $y(x)$ должно по абсолютной величине быть меньше $\varepsilon > 0$:

$$|y(x) - \varphi(x)| < \varepsilon, \quad a \leqslant x \leqslant b.$$

Такую аппроксимацию будем называть равномерной с погрешностью г на отрезке $[a, b]$.

Введем понятие абсолютного отклонения Δ многочлена $\varphi(x)$ от функции $y(x)$ на отрезке $[a, b]$:

$$\Delta = \max_{a \leqslant x \leqslant b} |y(x) - \varphi(x)|, \quad (2.3)$$

и среднеквадратичного отклонения при среднеквадратичном приближении функции

$$\Delta = \sqrt{\frac{S}{n}}. \quad (2.4)$$

Рис.2 иллюстрирует принципиальные различия среднеквадратичного *(а)* и равномерного (б) приближений.

Существует понятие наилучшего приближения функции $y(x)$ многочленом $\varphi(x)$ (2.1), когда его коэффициенты c_i выбирают так, чтобы абсолютное отклонение (2.3) было минимальным.

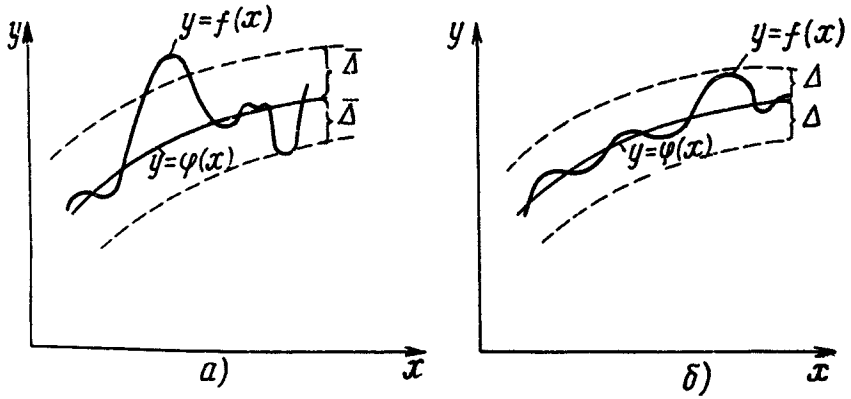


Рис. 2. Приближения: *а* — среднеквадратичное; *б* — равномерное.

В этом случае многочлен $\varphi(x)$ называют многочленом наилучшего равномерного приближения. Доказаны существование и единственность такого многочлена для функции, непрерывной на замкнутом, ограниченном множестве.

2.2. Разложение в степенные ряды

Достаточно часто при вычислении значений функций на компьютере используется разложение их в степенные ряды. Так, например, хорошо известна аппроксимация sin x степенным рядом:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

Погрешность такой аппроксимации определяется упоминавшимися в предыдущей главе погрешностями усечения и округления. Погрешность усечения сильно зависит от значения аргумента и растет с его ростом, т.е. неравномерно распределена по исследуемому интервалу. Одним из способов, позволяющим равномерно распределить ошибку по всему интервалу,

является использование многочленов Чебышева

$$T_n(x)=\frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^n+\left(x-\sqrt{x^2-1}\right)^n\right],$$

ортогональных на отрезке [-1, 1], n = 0, 1... На практике часто используют многочлены Чебышева для повышения точности аппроксимации функций с помощью ряда Тейлора. Вычислим для примера функцию sin x на стандартном отрезке [-1, 1] В этом случае ряд для $sin\ x$ приводится к виду:

$$\sin\frac{\pi x}{2}=\frac{\pi x}{2}-\frac{1}{3!}\left(\frac{\pi x}{2}\right)^3+\frac{1}{5!}\left(\frac{\pi x}{2}\right)^5-\dots.$$

При таком вычислении погрешность сильно возрастает к концам отрезка $x=\pm 1$. Если же использовать ряд

$$\sin\left(\frac{\pi x}{2}\right)=c_0+c_1T_1(x)+c_2T_2(x)+\dots,$$

членами которого являются многочлены Чебышева, то погрешность будет равномерно распределена по всему отрезку

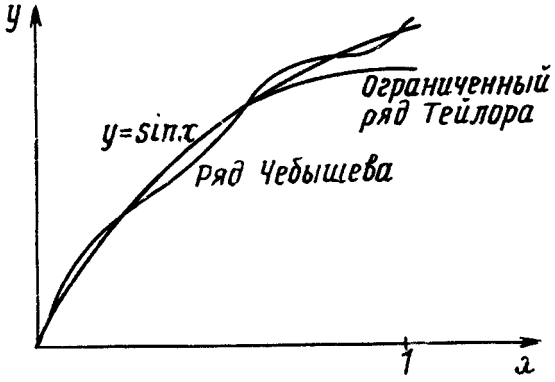


Рис 3 Аппроксимация рядами Тейлора и Чебышева

(рис.3). При построении равномерных приближений аналитических функций хорошие результаты дают так называемые аппроксимации Паде, основанные на использовании дробно-рациональных функций. При построении аппроксимации Паде для функции $u(x)$ сначала строится ее ряд Тейлора:

$$u(x)\simeq\sum_{j=0}^{J}c_jx^j,\quad c_j=\frac{1}{j!}\cdot\frac{d^ju(0)}{dx^j},\qquad(2.5)$$

который затем аппроксимируется дробно-рациональной функцией вида

$$F_{nm}(x)=\frac{P_n(x)}{Q_m(x)},\ \text{где}\ P_N(x)=\sum_{n=0}^{N}a_nx^n,\ Q_m(x)=\sum_{m=0}^{m}b_mx_m$$

Один из коэффициентов $F_{NM}(x)$, например b_0, можно положить равным 1, а остальные $N + M + 1$ коэффициентов выберем так, чтобы сохранить $N + M + 1$ коэффициентов ряда Тейлора (2.5). Из этого условия, полагая $J = N + M$, находим

$$\sum_{j=0}^{N+M}c_jx^j=P_N(x)/Q_M(x),$$

откуда

$$\left(\sum_{j=0}^{N+M}c_jx^j\right)\left(\sum_{m=0}^{M}b_mx^m\right)=\sum_{n=0}^{N}a_nx^n.\qquad(2.6)$$

Приравнивая в (2.6) коэффициенты при одинаковых степенях x получаем систему $N + M + 1$ линейных алгебраических уравнении для нахождения

$$a_n(n=0,1,\dots,N),\ b_m(m=1,2,\dots,M).$$

Достоинство аппроксимации Паде—высокая точность при относительно невысоких степенях N, M, а следовательно, и высокая

скорость вычислений. Поэтому аппроксимации Паде часто используются для вычисления элементарных функций в системном программном обеспечении компьютеров.

2.3. Регрессионный анализ — метод выравнивания

Регрессионным анализом называется изучение связи между зависимой переменной y и одной или несколькими независимыми переменными. Ограничимся случаем одной независимой переменной x. Будем для определенности полагать, что имеется некоторое множество значений независимой переменной $\{x_j, \ j=1,2, \ \ldots,n\}$ известных (измеренных) точно, и соответствующее множество значений $\{y_j, \ j—1,2,.. \ ,n\}$, которые представляют собой искаженные случайными измерительными погрешностями значения функции этих x_j. Зависимость

$$y=\varphi(x,\vec{c}),$$

где $\vec{c}=(c_0,c_1, \ \ldots,c_m)$— вектор параметров функции φ, называют регрессией y на x, компоненты $c_1,c_2,\ldots,c_m$ вектора $\vec{c}$ — параметрами регрессии, а кривую $\varphi(x, \ \vec{c})$ — линией (кривой) регрессии. Основная задача регрессионного анализа — отыскание функции, которая в некотором смысле наилучшим образом описывает (аппроксимирует) зависимость y от x. Обычно заранее задаются видом функции $\varphi(x,\vec{c})$ и отыскивают компоненты вектора $\vec{c}$ - параметры регрессии. Рассмотрим простейший случай двухпараметрических функций $\varphi(x,a,b)$, т. е. предполагается, что в отсутствие погрешностей

$$y=\varphi(x,a,b).$$

Для многих часто встречающихся в физических задачах зависимостей можно подобрать такую замену переменных и параметров

$$X=X(x,y), \quad Y=Y(x,y),$$
$$A=A(a,b), \quad B=B(a,b),$$

что по новым переменным зависимость становится линейной вида

$$Y=AX+B. \qquad (2.7)$$

Например, зависимость

$$y=\frac{x}{ax^k+b}$$

преобразуется к линейной (2.7) заменой переменных

$$X=x^k, \quad Y=\frac{x}{y}, \quad A=a, \quad B=b,$$

а зависимость —

$$y=e^{ax}$$

заменой

$$\lambda=x, \quad Y=\ln y, \ A=a, \quad B=\ln b.$$

Такая замена переменных в регрессионном анализе называется выравниванием.

Уравнение (2.7) в j-м узле дает (2.8)

$$Y_j=AX_j+B,$$

где $X_j=X(x_j,y_j), \quad Y_j=Y(x_j, \ y_j).$

Введем обозначения

$$\bar{X}=\frac{1}{n}\sum_{j=1}^{n}X_j, \quad \bar{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_j, \quad \sigma_x^2=\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\bar{X})^2,$$

$$K_{xy}=\frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\bar{Y})(X_j-\bar{X}). \qquad (2.9)$$

Здесь $\bar{X}$ — среднее (по всем j) значение X, $\bar{Y}$ — среднее значение Y, σ_x^2 представляет собой дисперсию X, K_{xy} называют ковариацией X и Y. (Конечно, из-за ограниченности n и существования экспериментальных погрешностей, формулы (2.9) дают лишь приближённые значения, т.е. оценки средних, дисперсии и

ковариации).

Суммируя (2.8) по всем j, находим

$$\bar{Y} = A\,\bar{X} + B. \qquad (2.10)$$

Вычитая (2.10) из уравнения (2.8), получаем

$$Y_j - \bar{Y} = A\left(X_j - \bar{X}\right).$$

Домножая последнее уравнение на (X,— *X)* и суммируя по j, находим

$$K_{xy} = A\,\sigma_x^2,$$

откуда получаем

$$A = \frac{K_{xy}}{\sigma_x^2},$$

а затем из уравнения (2.10)

$$B = \bar{Y} - A\,\bar{X}.$$

Найдя *A и B* можно затем вычислить искомые параметры *a* и *b* функции φ.

2.4. Метод наименьших квадратов

Как уже упоминалось ранее, метод наименьших квадратов состоит в минимизации квадратов отклонений значений многочлена и функции в данных точках (2.2):

$$S = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} \left|\varphi\left(x_i, c_0, c_1, \ldots, c_m\right) - y_i\right|^2. \qquad (2.11)$$

Параметры $c_1, c_2, \ldots, c_m$, найдем из условия минимума функции $S(c_0, c_1, \ldots, c_m)$. Если отклонение ε_i, подчиняется нормальному закону распределения, то полученные таким образом значения параметров наиболее вероятны.

Поскольку c_i, выступают в роли независимых переменных функции *S*, то минимум найдем, приравнивая нулю частные производные по этим переменным:

$$\frac{\partial S}{\partial c_0} = 0, \qquad \frac{\partial S}{\partial c_1} = 0, \quad \frac{\partial S}{\partial c_2} = 0, \ \ldots, \ \frac{\partial S}{\partial c_n} = 0, \qquad (2.12)$$

т.е. приходим к системе уравнений для определения c_i. Если в качестве аппроксимирующей функции взять многочлен (2.1), то выражение для квадратов отклонений (2.11) примет вид:

$$S = \sum_{i=0}^{n} \left(c_0 + c_1 x_i + c_2 x_i + \ldots + c_m x_i - y_i\right)^2.$$

Приравнивая нулю частные производные (2.12), приходим к системе:

$$\frac{\partial S}{\partial c_0} = 2\sum_{i=0}^{n} \left(c_0 + c_1 x_i + \ldots + c_m x_i^m - y_i\right) = 0;$$

$$\frac{\partial S}{\partial c_1} = 2\sum_{i=0}^{n} \left(c_0 + c_1 x_i + \ldots + c_m x_i^m - y_i\right) x_i = 0;$$

. . . . . .

$$\frac{\partial S}{\partial c_m} = 2\sum_{i=0}^{n} \left(c_0 + c_1 x_i + \ldots + c_m x_i^m - y_i\right) x_i^m = 0.$$

Собирая коэффициенты при неизвестных, $c_0, c_1, \ldots, c_m$

Получаем систему уравнений:

$$(n+1)c_0 + c_1\sum_{i=0}^{n} x_i + c_2\sum_{i=0}^{n} x_i^2 + \ldots + c_m\sum_{i=0}^{n} x_i^m = \sum_{i=0}^{n} y_i;$$

$$c_0\sum_{i=0}^{n} x_i + c_1\sum_{i=0}^{n} x_i^2 + c_2\sum_{i=0}^{n} x_i^3 + \ldots + c_m\sum_{i=0}^{n} x_i^{m+1} = \sum_{i=0}^{n} x_i y_i;$$

. . . . . .

$$c_0\sum_{i=0}^{n} x_i^m + c_1\sum_{i=0}^{n} x_i^{m+1} + c_2\sum_{i=0}^{n} x_i^{m+2} + \ldots + c_m\sum_{i=0}^{n} x_i^{2m} = \sum_{i=0}^{n} x_i^m y_i.$$

Решая систему, находим неизвестные параметры $c_0, c_1, ..., c_m$.

В более компактном виде можно записать:

$$\begin{aligned} b_{00}c_0 + b_{01}c_1 + \ldots + b_{0m}c_m &= a_0; \\ b_{10}c_0 + b_{11}c_1 + \ldots + b_{1m}c_m &= a_1; \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ b_{m0}c_0 + b_{m1}c_1 + \ldots + b_{mm}c_m &= a_m, \end{aligned}$$

если ввести обозначения $b_{kl} = \sum_{i=0}^{n} x_i^{k+l}$, $a_k = \sum_{i=0}^{n} x_i^k y_i$; $k, l = 0, 1, ..., m$.

2.5. Цифровая фильтрация экспериментальных результатов

При исследовании физических процессов $u(t)$ наблюдаемые величины $u_i = u(t_i)$ обычно являются суммой неизвестных точных значений $\bar{u}_i$ (полезного сигнала) и случайных погрешностей (помехи) ε_j:

$$u_j = \bar{u}_j + \varepsilon_j. \quad (2.13)$$

В таких случаях возникает задача фильтрации результатов измерений, под которой понимают восстановление значений $\bar{u}$ на фоне помех ε. При обработке экспериментальных результатов на компьютере для этого применяются цифровые фильтры, представляющие собой специальные программы, выполняющие фильтрацию результатов эксперимента. Цифровые фильтры вычисляют отфильтрованные значения по формуле

$$\hat{u}_j = F(u_{j+m}; m = 0, \pm 1, \pm 2..) \quad (2.14)$$

описывающей некоторое преобразование результатов измерений Если функция F (2.14) линейная, то фильтрация называется линейной; если же F — нелинейная функция, то фильтрация называется нелинейной. Полученное значение u, (2.14) принимается затем в качестве оценки неизвестной величины $\bar{u}_j$.

При выборе функции F (2.14) обычно руководствуются некоторыми априорными соображениями о характере функций $\bar{u}(t)$ и ε(t), например, полагают, что сигнал $\bar{u}(t)$ — более гладкая функция, чем помеха $\varepsilon(t)$. В дальнейшем ограничимся более простой линейной фильтрацией результатов измерений, полученных с постоянным шагом τ. В этом случае обычно используется метод наименьших квадратов (§2.4). Возьмем, например, пять значений (узлов) с номерами j, $j \pm 1$, $j \pm 2$ и аппроксимируем этот отрезок кривой $u(t)$ квадратичной кривой (параболой) $\hat{u}(t)$ так, чтобы отклонение

$$I = \sum_{m=-2}^{2} \left(\hat{u}_{j+m} - u_{j+m}\right)^2 = \min. \quad (2.15)$$

Из условия (2.15) по методу наименьших квадратов находим три коэффициента параболы, а затем

$$\hat{u}_j = \hat{u}(t_j) = u_j - \frac{3}{35}\delta^4 u_j, \quad (2.16)$$

где $\delta^4 u_j$ — четвертая центральная разность u(t) в j-м узле:

$$\delta^4 u_j = u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j-2}. \quad (2.17)$$

Из (2.16), (2.17) следует, что

$$\hat{u}_j = \frac{1}{35}\left(-3u_{j-2} + 12u_{j-1} + 17u_j + 12u_{j+1} - 3u_{j+2}\right) \quad (2.18)$$

Такую фильтрацию называют фильтрацией при помощи четвертых разностей.

Для оценки эффекта фильтрации предположим, что все измерения u_j независимы и имеют одинаковую среднеквадратичную погрешность $\overline{\varepsilon_j^2} = \sigma^2$.

Тогда среднеквадратичная погрешность $\hat{\sigma}^2$ величины $\hat{u}_j$

равна сумме среднеквадратичных погрешностей величин в правой части (2.18), т.е.

$$\overline{\sigma^2}=[(3^2+12^2+17^2+12^2+3^2)/35^2]\sigma^2\simeq 0{,}5\,\sigma^2. \quad (2.19)$$

Следовательно, благодаря привлечению при фильтрации дополнительной информации (пять узлов вместо одного) среднеквадратичная погрешность снижается вдвое, экспериментальные результаты сглаживаются.

Описанное сглаживание является локальным, так как в нем используется лишь несколько узлов, соседних с *j*. Более сложным является глобальное сглаживание или сглаживание в целом, при котором используется вся информация о процессе. В качестве относительно простого примера глобального сглаживания рассмотрим усечение спектра. В этом методе с помощью быстрого преобразования Фурье вычислим, как описано в главе 1, достаточно большое число *N* коэффициентов Фурье $\bar{u}(n)$ функции *u(t)*, где $\bar{u}(n)$ задаются формулой (1.28). Полагая, что высшие гармоники Фурье создаются помехой, оценим по некоторому числу $2m \gg 1$ высших гармоник средний квадрат их амплитуды:

$$s_0=\frac{1}{2m}\sum_{i=0}^{m-1}\left[\left|\bar{u}\left(-\frac{N}{2}+i\right)\right|^2+\left|\bar{u}\left(\frac{N}{2}-1-i\right)\right|^2\right] \quad (2.20)$$

и отбросим (занулим) все гармоники $\bar{u}(n)$ с амплитудами $|\bar{u}(n)|^2 \leqslant a\,s_0$, где $a>0$ — некоторая величина, близкая к 1. После этого с помощью быстрого преобразования Фурье синтезируем сглаженную функцию *u(t)* по оставшимся коэффициентам $\bar{u}(n)$.

Как показывает опыт, значительно лучшие результаты могут быть получены с помощью нелинейной фильтрации экспериментальных данных, которую называют также адаптивной, поскольку в ней функциональная зависимость *F* (2.14) подбирается в соответствии с характером измеряемой функции *u(t)*. Ввиду сложности эти проблемы здесь не рассматриваются.

3. ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

В данной главе будут рассмотрены наиболее часто используемые в физических задачах методы решения нелинейных алгебраических уравнений и систем. Их можно разделить на прямые и итерационные. Прямые методы, позволяющие записать выражение для корней в виде формулы, в большинстве практически важных физических задач использовать не удается. Поэтому решения ищутся методами последовательных приближений или итерационными методами.

Итерационный метод можно разбить на два этапа:

отыскание приближенного значения корня или содержащего его отрезка;

уточнение приближенного значения до заданной степени точности.

Начальное приближение может находиться из физических соображений, из опыта решения аналогичных задач, с помощью графических методов и т.д.

3.1. Уравнения с одним неизвестным

Одной из наиболее частых задач, с которыми сталкивается физик — это решение уравнений вида

$$f(x)=0. \quad (3.1)$$

Отыскание корней такого уравнения, как уже упоминалось выше, должно состоять из двух этапов:

1. Общий анализ уравнения, позволяющий узнать, существуют ли решения, каково их количество и, возможно, каковы их приближенные значения. Результатом такого рассмотрения является выделение интервалов, на которых функция *f(x)* меняет знак лишь один раз. Эти интервалы могут быть определены математическими или, в ряде сложных случаев, численными методами. Анализ физической основы данной задачи часто позволяет сделать важные выводы, которые в дальнейшем могут быть использованы для упрощения решения. Так, например, при решении уравнения дисперсии

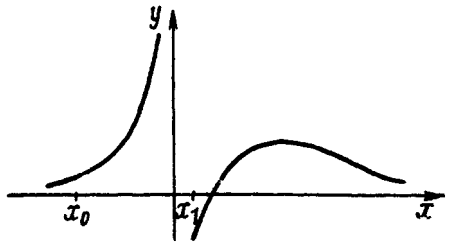
$$D(\omega, k)=0,$$

можно рассматривать лишь решения вида $\omega=\omega(k)$, которые, как известно, существуют для некоторых предельных случаев, легко решаемых аналитически ($k=0$ или $k\to\infty$). Не следует начинать

решение какого-либо уравнения, предварительно не определив число и порядок его решений. Более того, должны быть известны и свойства функции *f(x),* чтобы избежать случая, когда изменение знака функции не соответствует корню (отрезок $[x_0,x_1]$, рис. 4).

2. На втором этапе на каждом интервале организуется поиск корня одним из методов локального поиска корней. Необходимо помнить о своеобразии представления корня уравнения (3.1) на компьютере. При вычислении функции в окрестности ее математического нуля практически в любом алгоритме вычитаются друг из друга близкие величины, что ведет к росту ошибок округления и возможной потери точности. Таким образом, вместо четко выраженного пересечения оси в математических функциях (рис. 5, а) мы приходим к более или менее ограниченной области значений функции, которая определяется ошибками округления (рис. 5, б).

Рис. 4. Изменение знака функции, не соответствующее полюсу



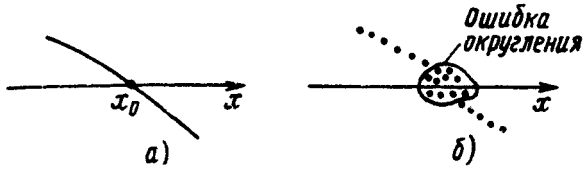


Рис. 5. Представление корня уравнения: *а* — математическое; *б* — машинное

Таким образом, при машинном решении речь идет о том, чтобы найти одну из точек, в которых модуль функции имеет наименьшее из возможных значение.

Поиск корня уравнения математически осуществляется при помощи построения последовательности Коши $\{x_i\}$, когда при заданном ε существует такое *N,* что для всех *n и p* превышающие *N,* выполняется $|x_n - x_p| < \varepsilon$.

3.1.1. Метод половинного деления (дихотомия)

Пусть мы знаем, что на отрезке [*a, b]* находится искомое значение корня уравнения (3.1), т.е. $x \in [a, b]$. В качестве начального приближения возьмем середину отрезка.

$$x_0 = \frac{a+b}{2}.$$

Теперь исследуем значения функции $f(x)$ на концах образовавшихся отрезков $[a,x_0]$ и $[x_0, b]$ (рис.6). Выберем из них тот, на концах которого функция принимает значения разного знака, так как он и содержит искомый корень. Вторую половину отрезка можно не рассматривать. Затем делим новый отрезок пополам и приходим вновь к двум отрезкам, на концах одного из которых функция меняет знак, т.е. содержит корень. Таким образом, после каждой итерации исходный отрезок сокращается вдвое, т. е. после *n* итераций он сократится в 2^n раз. Процесс итераций будет продолжаться до тех пор, пока значение модуля функции не окажется меньше заданной точности ε .

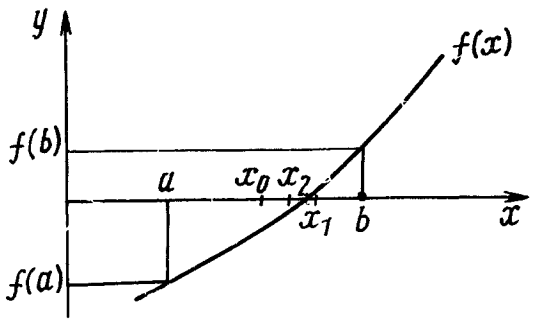


Рис. 6 Геометрическое представление метода деления отрезка пополам

$$|f(x_n)| < \varepsilon,$$

либо длина исследуемого отрезка станет меньше допустимой:

$$|x_{n+1} - x_n| < \varepsilon.$$

Как видно из вышесказанного, метод довольно медленный, однако он всегда сходится к корню.

3.1.2. Удаление корней

Если функция *f(x)* имеет на отрезке [*a, b*] несколько корней и непрерывна на [*a, b*], то вспомогательная функция $g(x) = f(x)/(x - x_1)$ непрерывна, причем все корни функции *g(x)* совпадают с корнями функции *f(x)*, за исключением x_j*. Другими словами, если на отрезке [*a, b*] функция имеет несколько корней, то можно найти любой из них, исключить его, перейдя к новой функции *g(x)* и повторить процесс нахождения корней этой функции. Таким образом можно найти все корни функции *f(x)*.

3.1.3. Метод простой итерации

Приведем исходное уравнение (3.1) к виду

$$x = \varphi(x), \qquad (3.2)$$

где $\varphi(x)$ определяется, например, следующим способом:

$$\varphi(x) = x + \rho(x) f(x),$$

где $\rho(x)$—произвольная функция, не имеющая на [a,*b*] корней или константа. Метод простой итерации задается следующим алгоритмом:

$$x_{n+1} = \varphi(x_n), \qquad (3.3)$$

где n=0, 1,2…—номер итерации.

Нам необходимо найти приближенное значение корня x^* уравнения (3.2) с относительной погрешностью $\varepsilon > 0$ так, чтобы при $n > n_0(\varepsilon)$ выполнялось следующее условие:

$$|x_n - x^*| \leqslant \varepsilon |x_0 - x^*|, \text{ при } n \geqslant n_0(\varepsilon),$$

или

$$|x_{n+1} - x_n| < \varepsilon, \qquad (3.4)$$

то есть результаты последовательных итераций близки.

Достаточным условием сходимости итераций является следующее:

$$|\varphi'(x_n)| < 1. \qquad (3.5)$$

Это условие является гарантией сходимости последовательности $\{x_i\}$ к решению уравнения. Оно не является необходимым, т. е. существуют функции, для которых это условие не выполняется, и тем не менее этот метод приводит к решению. Рассмотрим геометрическую интерпретацию метода, т.е. найдем точку пересечения двух кривых $u = \varphi(x)$ и $y = x$ (рис.7).

Производная изображенной функции удовлетворяет условию (3.5). Задавая начальное приближение x_0, найдем $x_1 = \varphi(x_0)$. Геометрически это означает, что необходимо провести горизонтальную прямую через точку $\varphi(x_0), x_0)$ до пересечения с прямой $y = x$ и вертикальную через эту точку. Эта вертикальная прямая пересекая ось x дает нам значение x_1. Далее процесс повторяется, т.е. находится следующее приближение $x_2 = \varphi(x_1)$ и т.д. Из рисунка видно, что последовательность $\{x_i\}$ сходится к корню x*. Изображенный на рис. 7 случай соответствует $\varphi'(x) > 0$, когда все приближения $\{x_i\}$ находятся с одной стороны от корня. Если же $\varphi'(x) < 0$ (продолжая удовлетворять условию (3.5)), то $\{x_i\}$ будут последовательно располагаться с разных сторон от корня x* (рис.8). Если же условие (3.5) не выполнено $|\varphi'(x)| > 1$, то процесс расходится (рис.9). Получим условие (3.5) из условия сходимости последовательности итераций (3.4).

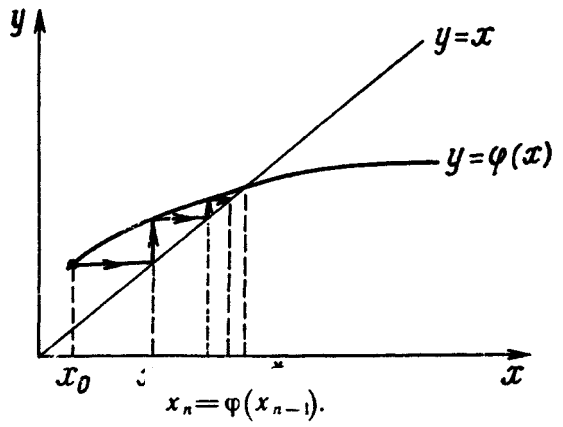
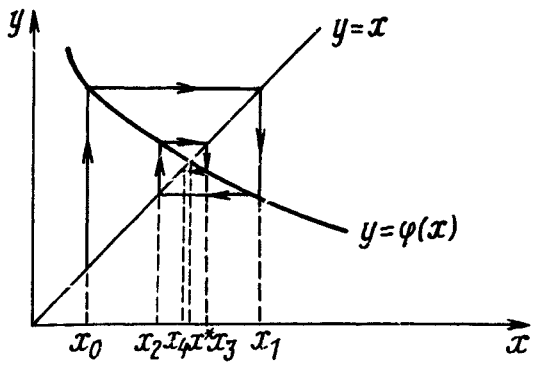


Рис. 7. Геометрическое представление метода последовательных приближений $(0<\varphi'(x)<1)$

Рис. 8.



Геометрическое представление метода последовательных прибли жений $(0>\varphi'(x)>-1)$

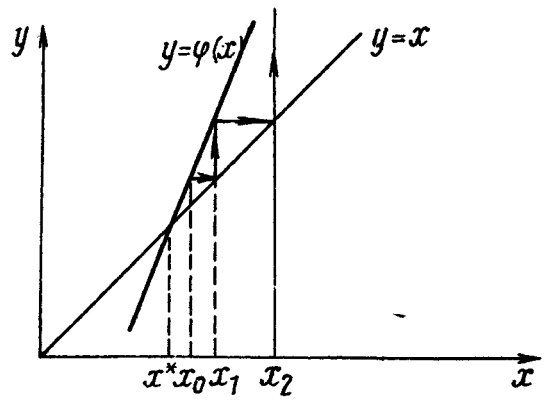


Рис. 9. Геометрическое представление метода последовательных приближений $(\varphi'(x)>1)$

Пусть x^* — искомое значение корня, тогда из (3.2)

$$x^* = \varphi(x^*),$$

а согласно алгоритму простой итерации (3.3)

$$x_n = \varphi(x_{n-1})$$

Вычитая два последних уравнения друг из друга, имеем

$$x_n - x^* = \varphi(x_{n-1}) - \varphi(x^*).$$

Умножим правую часть полученного уравнения на $\frac{x_{n-1}-x^*}{x_{n-1}-x^*}=1$,

тогда

$$x_n - x^* = \frac{\varphi(x_{n-1})-\varphi(x^*)}{x_{n-1}-x^*}\cdot(x_{n-1}-x^*),$$

или согласно теореме о среднем

$$x_n - x^* = \varphi'(\xi)(x_{n-1}-x^*), \quad \text{где } \xi \in [x_{n-1}, x^*].$$

Если ввести некое число $m = \max\limits_{x\in[a,b]} \varphi'(x)$, то

$$|x_n - x^*| \leqslant m\,|x_{n-1}-x^*|.$$

Аналогичным образом можно получить:

$$|x_{n-1}-x^*| \leqslant m\,|x_{n-2}-x^*|,$$
$$|x_n-x^*| \leqslant m^2|x_{n-2}-x^*|, \quad \text{или}$$
. . . . . .
$$|x_n-x^*| \leqslant m^n|x_0-x^*|.$$

Если величина максимальной на [*a, b*] производной $m<1$, то независимо от выбора начального приближения x_0 с ростом n правая часть становится малой и $\{x_n\}$ сходится к x^*. Если же $m>1$, то $|x_n-x^*|$ неограниченно возрастает с ростом n.

Методы итераций, как и большинство других итерационных методов, имеют существенное достоинство — они не накапливают ошибок вычисления, так как в этом случае ошибки вычислений приводят лишь к ошибке на данной итерации и, как следствие, влияют на число итераций, но не на точность конечного результата.

3.1.4. Метод касательных

Метод касательных (метод Ньютона) можно отнести к аналитическим методам, в основе которых лежит замена исследуемой функции $f(x)$ более простой $A(x)$ выбранной так, чтобы легко решалось уравнение

$$A(x)=0. \quad (3.6)$$

Таким образом, уравнение (3.1) заменяется уравнением (3.6), причем на практике в качестве функции $A(x)$ берут линейную

$$A(x)=ax+b.$$

Метод состоит в построении последовательности координат точек (x_n, y_n), через которые проводятся прямые, все более приближающиеся к исходной кривой $f(x)$.

Если обозначить через K_n наклон n-й прямой, то переход от точки $(x_n, y_n$ к точке (x_{n+1}, y_{n+1}) осуществляется следующим преобразованием:

$$x_{n+1}=x_n-\frac{f(x_n)}{k_n};$$
$$y_{n+1}=f(x_{n+1}). \quad (3.7)$$

Геометрически этот переход можно изобразить, как показано на рис. 10.

Заменяя в (3.7) K_n на производную в точке x_n, т.е. заменяя участок исследуемой кривой касательной, приходим к формуле метода касательных:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}, \quad n=0, 1, 2,... \quad (3.8)$$

Геометрически (рис.11) в точке x_0 проводится касательная к кривой $f(x)$ и находится точка ее пересечения с осью абсцисс.

Уравнение касательной к кривой $f(x)$ в точке $M_0(x_0, f(x_0))$ имеет вид

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

а следующее приближение x_1 являющееся точкой пересечения касательной с осью абсцисс

$$x_1=x_0-\frac{f(x_0)}{f'(x_0)}.$$

Аналогичным образом можно найти и приближения

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

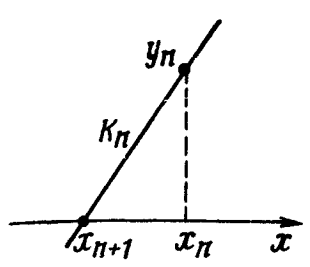


Рис. 10. Геометрическая интерпретация шага решения

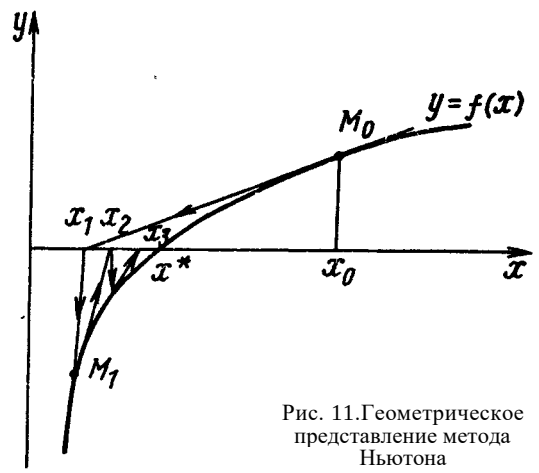


Рис. 11.Геометрическое представление метода Ньютона

не забывая, что $f'(x_n)=0$.

Метод Ньютона можно рассматривать как частный случай метода простой итерации. Аналогично (3.2) запишем выражение для $\varphi(x)$ в виде

$$\varphi(x)=x+\rho(x)f(x).$$

Будем искать функцию $\varphi(x)$ такую, чтобы $|\varphi'|$ был бы минимален

$$\varphi'(x)=1+\rho'(x)f(x)+\rho(x)f'(x).$$

Так как вблизи корня $f(x)\approx 0$ и поскольку мы ищем минимальное значение модуля производной, то

$$\varphi'(x)=1+\rho(x)f'(x)=0.$$

Последнее уравнение дает

$$\rho(x)=-\frac{1}{f'(x)}.$$

Таким образом, мы приходим к окончательной формуле

$$\varphi(x)=x-\frac{f(x)}{f'(x)}.$$

В этом случае условие сходимости (3.5) принимает вид

$$\left|1-\frac{f'(x)}{f'(x)}+\frac{f(x)f''(x)}{(f'(x))^2}\right|<1,$$

или преобразуя

$$|f(x)f''(x)|<(f'(x))^2.$$

При произвольном нулевом приближении итерации будут сходиться, если всюду будет выполнено полученное выше условие. В противном случае сходимость будет лишь в некоторой окрестности корня.

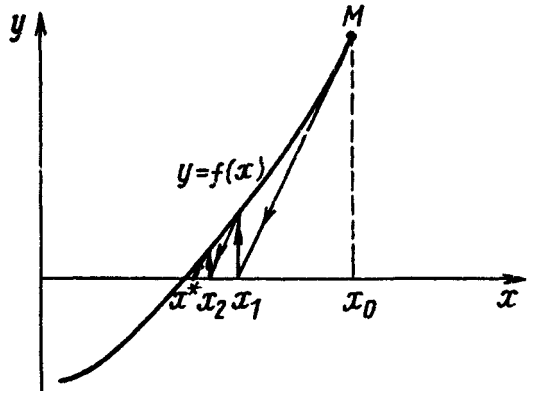


Рис.12. Геометрическое представление метода фиксированной касательной.

Разновидностью метода Ньютона является метод фиксированной касательной, когда коэффициент наклона K_n не меняется в ходе итераций: $K=K_n=K_0=f'(x_0)$, т.е. аппроксимирующая прямая остается параллельной касательной к кривой в начальной точке (рис. 12).

Для окончания итерационного процесса могут быть использованы следующие критерии.

1. Максимальное число итераций. Этот критерий необходим в случае, если методы не сходятся. Тем не менее трудно заранее определить, сколько итераций будет необходимо для получения удовлетворительной точности.

2. Слабая вариация приближения к корню:

$$|x_{n+1}-x_n|<\varepsilon \text{ или } |x_{n+1}-x_n|<\varepsilon|x_n|.$$

3. Достаточно малое значение функции

$$|f(x_n)|<\varepsilon.$$

Использование метода Ньютона требует тщательного анализа

поведения функции в окрестности корня, поскольку данный метод может дать ошибочные результаты (рис. 13).

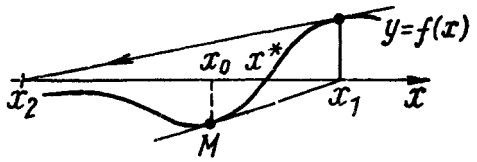


РИС. 13. Иллюстрация случая, когда метод Ньютона не приводит к правильному результату

Как видно из рис.13 в случае такого задания начального приближения x_0 метод не приводит к правильному результату. Таким образом, метод Ньютона очень чувствителен **к** выбору начального приближения, которое должно находиться вблизи корня. Поэтому в ряде случаев целесообразно использовать комбинированный метод, заключающийся в использовании сначала всегда сходящегося метода (например, дихотомии), а после некоторого числа итераций - быстросходящегося метода Ньютона

3.1.5. Метод секущих

Вычисление производной функции $f'(x)$, необходимое в методе Ньютона, не всегда удобно или возможно. Замена производной первой разделенной разностью, которую находят по двум последним итерациям (т. е. замена производной на секущую) приводит нас к методу секущих. **С** точки зрения аналитических методов (3.6) в качестве аппроксимирующей прямой взята прямая, проходящая через две последние точки x_n и x_{n-1}, т.е. вместо K_n в (3.7) необходимо подставить

$$K_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

тогда придем к формуле метода секущих:

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

$$y_{n+1} = f(x_{n+1}). \qquad (3.9)$$

Метод секущих является двухшаговым методом, т .е. требует двух начальных (разгонных) точек x_0 и x_1. Графически метод иллюстрируется рис. 14.

Сначала через выбранные точки $(x_0, f(x_0)), (x_1, f(x_1))$ проводим прямую до пересечения с осью абсцисс и определяем x_2, а вертикальная прямая в точке x_2 дает f(x_2). Далее прямая проводится через точки $(x_1, f(x_1))$ и $(x_2, f(x_2))$ и т.д., пока не будет выполнено одно из трех условий окончания итерационного процесса, изложенных при описании метода Ньютона.

В знаменателе формулы (3.9) при приближении *к* корню стоит разность двух малых чисел, вследствие чего возможно накопление ошибки из-за потери значащих цифр. В этом случае необходимо использовать прием Гарвика, заключающийся в контроле величины $|x_{n+1} - x_n|$. Если эта величина начинает расти, то нужно закончить итерационный процесс, если же убывает, то можно продолжать итерации.

Описанный выше метод также называют методом подвижной секущей, в отличие от метода фиксированной секущей, когда аппроксимирующая прямая остается параллельной постоянному направлению *AB* (рис. 15).

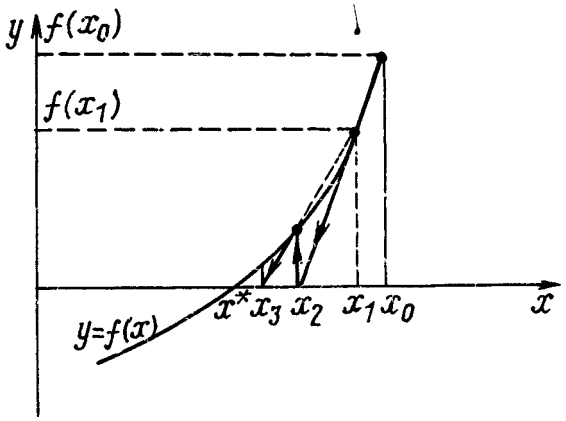


Рис. 14. Геометрическое представление метода секущих

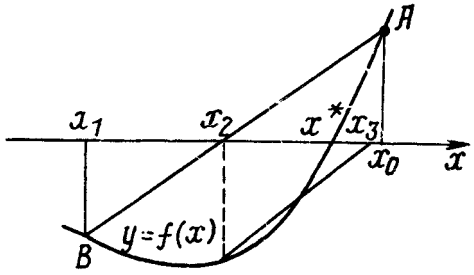


Рис. 15. Геометрическое представление метода фиксированной секущей

В этом случае выражение для коэффициента наклона K_n в (3.7) имеет следующий вид:

$$K_n = K_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Этот метод является одним из самых надежных среди аналитических методов в смысле сходимости при условии выбора начальных точек по обе стороны от корня.

3.1.6. Метод парабол

Выберем в качестве аппроксимирующей функции (3.6) итерационный многочлен Ньютона второй степени, построенный по трем последним итерациям:

$$A(x) = f(x_n) + (x - x_n) f(x_n, x_{n-1}) + (x - x_n)(x - x_{n-1}) \times \times f(x_n, x_{n-1}, x_{n-2}),$$

где $f(x_n,x_{n-1}\_)$ и $/(x_n,x_{n-1}, x_{n-2}\_)$— соответственно первая и вторая разделенные разности. В этом случае исходное уравнение (3.1) заменяется следующим:

$$a z^2 + b z + c = 0, \qquad (3.10)$$

где

$$z = x - x_n, \quad a = f(x_n, x_{n-1}, x_{n-2}), \quad b = a(x_n - x_{n-1}) + f(x_n, x_{n-1}),$$
$$c = f(x_n).$$

Таким образом, в отличие от рассмотренных ранее других аналитических методов, аппроксимирующая функция является не линейной, а параболической.

Решение уравнения (3.10) дает два корня, меньший из которых по абсолютной величине определяет новое приближение

$$x_{n+1} = x_n + z.$$

Метод парабол является трехшаговым методом, так как для начала счета нужно знать три начальных приближения x_0 ,x_1 и x_2. Достоинством метода является его способность сходиться к комплексному корню, даже если все предыдущие приближения были действительными.

По скорости сходимости среди аналитических методов самым быстрым является метод Ньютона, чуть уступают ему методы парабол и подвижной секущей, и вдвое более медленной сходимостью обладают методы фиксированных касательной и секущей, а также метод простой итерации. Однако по числу операций, которые необходимо выполнить для отыскания корня с заданной точностью в зависимости от вида функции *f(x)* и ее производной *f'(x)* метод подвижной секущей может оказаться самым быстрым.

Таким образом, выбор между эффективными методами и методами надежными является постоянной проблемой при численном решении задач на компьютере.

3.2. Системы нелинейных уравнений

Для систем нелинейных уравнений вида

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0; \\ f_2(x_1, x_2, \ldots, x_n) &= 0; \\ &\ldots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned} \qquad (3.11)$$

практически не существует прямых методов решения, поэтому необходимо использовать итерационные.

3.2.1. Метод простой итерации

Приведем систему (3.11) к виду:

$$\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n); \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n); \\ &\ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}$$

воспользовавшись одним из описанных ранее способов (3.2). Тогда процесс перехода от начального приближения $x_1^0, x_2^0, \ldots x_n^0$ к последующим будет осуществляться следующим образом:

$$\begin{aligned} x_1 &= \varphi_1(x_1^0, x_2^0, \ldots, x_n^0); \\ x_2 &= \varphi_2(x_1, x_2^0, \ldots, x_n^0); \\ &\ldots \\ x_i &= \varphi_i(x_1, x_2, \ldots, x_{i-1}, x_i^0, \ldots, x_n^0); \\ &\ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_{n-1}, x_n^0). \end{aligned}$$

Итерационный процесс продолжается до тех пор, пока изменения всех неизвестных на двух последовательных итерациях не станут малыми:

$$\left|x_i^{n+1} - x_i^n\right| < \varepsilon, \quad i = 1, 2, 3, \ldots, n.$$

Сходимость метода сильно зависит от выбора начального приближения, которое должно быть достаточно близким к решению.

3.2.2. Метод Ньютона

Метод Ньютона, как и для случая одного уравнения, обладает гораздо более быстрой сходимостью по сравнению с методом простой итерации. Формулу для метода Ньютона в случае системы уравнений можно получить, разложив в ряд Тейлора функцию $f(x_1, x_2, \ldots, x_n)$ и отбросив все члены с производными выше первой. Переход от начального приближения к последующему осуществляется по формулам:

$$x_1 = x_1^0 + \Delta x_1, x_2 = x_2^0 + \Delta x_2, \ldots, x_n = x_n^0 + \Delta x_n. \qquad (3.12)$$

Разложив правые части в (3.11) в ряд Тейлора, имеем:

$$f_1(x_1, x_2, \ldots, x_n) \approx f_1(x_1^0, x_2^0, \ldots, x_n^0) + \frac{\partial f_1}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial f_1}{\partial x_n}\Delta x_n = 0;$$

$$f_2(x_1, x_2, \ldots, x_n) \approx f_2(x_1^0, x_2^0, \ldots, x_n^0) + \frac{\partial f_2}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial f_2}{\partial x_n}\Delta x_n = 0;$$

. . . . . .

$$f_n(x_1, x_2, \ldots, x_n) \approx f_n(x_1^0, x_2^0, \ldots, x_n^0) + \frac{\partial f_n}{\partial x_1}\Delta x_1 + \ldots + \frac{\partial f_n}{\partial x_n}\Delta x_n = 0.$$

Или, преобразуя к более удобному виду:

$$\frac{\partial f_1}{\partial x_1}\Delta x_1 + \frac{\partial f_1}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial f_1}{\partial x_n}\Delta x_n = -f_1(x_1^0, x_2^0, \ldots, x_n^0);$$

$$\frac{\partial f_2}{\partial x_1}\Delta x_1 + \frac{\partial f_2}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial f_2}{\partial x_n}\Delta x_n = -f_2(x_1^0, x_2^0, \ldots, x_n^0);$$

. . . . . .

$$\frac{\partial f_n}{\partial x_1}\Delta x_1 + \frac{\partial f_n}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial f_n}{\partial x_n}\Delta x_n = -f_n(x_1^0, x_2^0, \ldots, x_n^0), \qquad (3.13)$$

причем значения функции **и ее** производной вычисляются в точках $x_1^0, x_2^0, \ldots, x_n^0$.

Определителем системы (3.13) является якобиан

$$J = \begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \ldots & \frac{\partial f_2}{\partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \ldots & \frac{\partial f_n}{\partial x_n} \end{vmatrix}.$$

Условием существования единственного решения на каждой итерации является отличие определителя от нуля

$$J \neq 0.$$

Таким образом, итерационный процесс в методе Ньютона заключается в нахождении приращений $\Delta x_1, \ldots, \Delta x_n$,к значениям неизвестных $x_1^0, \ldots, x_n^0$. Условием завершения процесса является следующее:

$$\max_i |\Delta x_i| < \varepsilon,$$

т.е. приращения достаточно малы. Как и в случае одного уравнения, сходимость сильно зависит от выбора начального приближения, причем с ростом числа уравнений в системе она ухудшается.

4. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

4.1. Численное дифференцирование

По определению производной функции $f(x)$ в точке x называется предел следующего отношения

$$y' = f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}. \qquad (4.1)$$

Если по каким-либо причинам производную от функции в данной точке аналитически найти не удается, либо функция $f(x)$ задана на конечном множестве точек $\{x_i\}$, $i = 0, 1, \ldots, n$, то необходимо перейти к численному дифференцированию. Естественно было бы заменить Δy и Δx разностями значений функции и аргумента в соседних узлах. Тогда вместо (4.1) имеем

$$y' \approx \frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Если таблица функции задана с некоторым постоянным шагом аргумента h, то вводя обозначения $\Delta y = y_1 - y_0$, $\Delta x = h$, $y_i = f(x_i)$ получим

$$y_1' \simeq \frac{(y_1 - y_0)}{h}. \qquad (4.2)$$

Таким образом, производную в точке x_1 выразили через разность значений функции в этой точке y_1\ и слева от нее y_0, т.е. через левые разности. Аналогично можно поступить, выразив производную с помощью правых разностей

$$y_1' = \frac{y_2 - y_1}{h}, \qquad (4.3)$$

или с помощью центральных разностей

$$y_1' = \frac{y_2 - y_0}{2h}. \qquad (4.4)$$

Аналогичным образом можно прийти к формулам для

производных более высокого порядка

$$y_1''=(y_1')'\simeq\frac{y_2'-y_1}{h}\simeq\frac{y_2-2y_1+y_0}{h^2}. \qquad (4.5)$$

Численное дифференцирование может быть представлено заменой функции $f(x)$ некой аппроксимирующей функцией $\varphi(\div)$ с последующим ее дифференцированием $\varphi'(x)$.

Заменив, например, функцию интерполяционным полиномом Ньютона с равноотстоящими узлами и ограничившись первым членом придем к полученным ранее формулам (4.2) — (4.4). Погрешность аппроксимирования может быть представлена в виде

$$R(x)=f(x)-\varphi(x). \qquad (4.6)$$

Продифференцировав ее нужное число раз, находим

$$R^{(k)}(x)=f^{(k)}(x)-\varphi^{(k)}(x). \qquad (4.7)$$

Величину $R^{(k)}x)$ будем называть погрешностью аппроксимации производной, так как она показывает ее отклонение от истинного значения.

В случае таблично заданной функции **с** равномерным шагом h погрешность аппроксимации зависит от шага h и записывается в виде $0(h^k)$, где k — порядок аппроксимации.

Оценим погрешность формул (4.2) — (4.4) с помощью ряда Тейлора. Разложим функцию $y_i=f(x_i)$, $i=0, 1, \ldots, n$ **в** ряд **в** точке x_i с шагом $\Delta x=-h$. Отбросив члены **со** второй производной, имеем

$$y(x_1-h)=y(x_1)-y'(x_1)h+0(h^2) \qquad (4.8)$$

или

$$y_0=y_1-y_1'h+0(h^2).$$

Преобразуем полученное **выше** уравнение **к** виду

$$y_1'=\frac{y_1-y_0}{h}+0(h). \qquad (4.9)$$

Сравнивая теперь это уравнение с (4.2) приходим к выводу, что аппроксимация производной с помощью левых разностей имеет первый порядок точности относительно шага h.

Заменив в разложении (4.8) шаг $\Delta x=h$, приходим к аппроксимации производной правыми разностями с тем же порядком аппроксимации

$$y_1'=\frac{y_2-y_1}{h}+0(h).$$

Если же оставить в разложении (4.8) и в соответствующем ему разложении с положительным шагом $\Delta x=h$ члены до $O(h^4)$, то получим:

$$\begin{aligned} y_0&=y_1-y_1'h+\frac{y_1''}{2!}h^2-\frac{y_1''}{3!}h^3+0(h^4),\\ y_2&=y_1+y_1'h+\frac{y_1''}{2!}h^2+\frac{y_1''}{3!}h^3+0(h^4). \end{aligned} \qquad (4.10)$$

Вычитая друг из друга уравнения (4 10), приходим к выражению производной через центральные разности (4.4)

$$y_1'=\frac{y_2-y_0}{2h}+0(h^2),$$

имеющему второй порядок точности.

Складывая уравнения (4.10) получаем формулу для вычисления второй производной, аналогичной (4.5), с аппроксимацией второго порядка

$$y_1''=\frac{y_0-2y_1+y_2}{h^2}+0(h^2).$$

Рассмотренная выше погрешность является упоминавшейся ранее погрешностью усечения ряда Тейлора, которая уменьшается с уменьшением шага, и определяется величиной остаточного члена. Не надо, однако, забывать о присутствии в наших вычислениях погрешности округления, которая с уменьшением шага увеличивается. Поэтому для получения оптимальной точности необходимо выбрать такой шаг h, чтобы исключить вычитание близких по величине чисел v_i задав в алгоритме следующую процедуру проверки $/y_i-y_{i-1}/>\varepsilon$, где ε — некоторое малое число.

4.1.1. Улучшение аппроксимации

Порядок точности производных в формулах аппроксимации прямо пропорционален числу узлов, используемых в аппроксимации. Однако с ростом числа узлов соотношения становятся громоздкими, и растет объем вычислений. Усложняется оценка точности. Есть простой способ уточнить решение при фиксированном числе узлов. Это метод Рунге—Ромберга.

Пусть $y'(x)$ — производная, которую надо аппроксимировать, $\varphi'(x, h)$ — конечно-разностная аппроксимация этой производной

на равномерной сетке с шагом h, R — погрешность (остаточный член) аппроксимации с главным членом $h^p r(x)$:

$$R = h^p r(x) + 0(h^{p+1}).$$

Тогда в общем случае имеем:

$$y'(x) = \varphi'(x, h) + h^p r(x) + 0(h^{p+1}). \quad (4.11)$$

Записав то же выражение в той же точке x при другом шаге $h_1 = kh$, получим:

$$y'(x) = \varphi'(x, kh) + (kh)^p r(x) + 0((kh)^{p+1}). \quad (4.12)$$

Приравнивая (4.11) и (4.12), находим выражение для главного члена погрешности:

$$h^p r(x) = \frac{\varphi'(x, h) - \varphi'(x, kh)}{k^p - 1} + 0(h^{p+1}).$$

Подставляя теперь найденное выражение в (4.11), получаем формулу Рунге:

$$y'(x) = \varphi'(x, h) + \frac{\varphi'(x, h) - \varphi'(x, kh)}{k^p - 1} + 0(h^{p+1}).$$

Эта формула позволяет по результатам двух расчетов значений производной $\varphi'(x, h)$ и $\varphi'(x, kh)$ с шагами h и kh с порядком точности p найти ее уточненное значение с порядком точности $p+1$.

Если можно провести расчеты с шагами $h_1, h_2, \ldots, h_q$, то можно получить уточненное решение для производной $y'(x)$ по формуле Ромберга:

$$y'(x) = \begin{vmatrix} \varphi'(x, h_1) & h_1^p & h_1^{p+1} \ldots h_1^{p+q-2} \\ \varphi'(x, h_2) & h_2^p & h_2^{p+1} \ldots h_2^{p+q-2} \\ \cdots & \cdots & \cdots \\ \varphi'(x, h_q) & h_q^p & h_q^{p+1} \ldots h_q^{p+q-2} \end{vmatrix} \times$$

$$\times \begin{vmatrix} 1 & h_1^p & h_1^{p+1} \ldots h_1^{p+q-2} \\ 1 & h_2^p & h_2^{p+1} \ldots h_2^{p+q-2} \\ \cdots & \cdots & \cdots \\ 1 & h_q^p & h_q^{p+1} \ldots h_q^{p+q-2} \end{vmatrix} + 0(h^{p+q-1}). \quad (4.13)$$

Таким образом, порядок точности возрастает на q—1, но функция должна иметь непрерывные производные достаточно высокого порядка.

4.1.2. Дифференцирование со сглаживанием

Довольно часто в физических задачах при дифференцировании исходную функцию f(x) заменяют некой сглаженной $\varphi(x)$ Сглаживание, как правило, выполняется методом наименьших квадратов. Рассмотрим простейший случай такого сглаживания. Пусть линейная аппроксимация $\varphi(x) = c_0 + c_1 x$ дает удовлетворительную точность. Тогда система уравнений (2.12) для определения неизвестных коэффициентов c_0 и c_1 принимает вид:

$$\sum_{i=0}^{n} (c_0 + c_1 x_i - y_i) = 0;$$

$$\sum_{i=0}^{n} (c_0 + c_1 x_i - y_i) x_i = 0, \quad (4.14)$$

где сумма берется по всем узлам сетки x_i лежащим в исследуемом интервале. Определив среднее значение как

$$\bar{x}=\frac{\sum_{i=0}^{n} x_i}{n+1}, \quad \bar{y}=\frac{\sum_{i=0}^{n} y_i}{n+1}, \qquad (4.15)$$

можно показать, что $c_0+c_1\bar{x}=\bar{y}$, или домножив на x

$$(c_0+c_1\bar{x}-\bar{y})\,\bar{x}=0.$$

Вычитая последнее уравнение из второго уравнения системы (4.14), имеем

$$c_1\sum_{0}^{n}\left(x_i^2-\bar{x}^2\right)-\sum_{0}^{n}(y_i x_i-\bar{y}\,\bar{x})=0. \qquad (4.16)$$

Поскольку $y'(x)\approx\varphi'(x)=c_j$, то в нашем случае

$$y'(x)\simeq\frac{\sum_{i=0}^{n}(y_i x_i-\bar{y}\,\bar{x})}{\sum_{i=0}^{n}\left(x_i^2-\bar{x}^2\right)}. \qquad (4.17)$$

Используя вводимое выше определение среднего (4.15), преобразуем знаменатель и числитель (4.17) к виду

$$\sum_{i=0}^{n}\left(x_i^2-\bar{x}^2\right)=\sum_{i=0}^{n}\left(x_i^2+\bar{x}^2\right)-\sum_{i=0}^{n}2\bar{x}^2=\sum_{i=0}^{n}\left(x_i^2+\bar{x}^2\right)-$$
$$-2\bar{x}\sum_{i=0}^{n}x_i=\sum_{i=0}^{n}(x_i-\bar{x})^2,$$

и аналогично

$$\sum_{i=0}^{n}(y_i x_i-\bar{y}\,\bar{x})=\sum_{i=0}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

Тогда, окончательно для (4.14) имеем:

$$y'(x)\simeq\frac{\sum_{i=0}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=0}^{n}(x_i-\bar{x})^2}.$$

Существуют критерии, позволяющие построить сглаживающие аппроксимации с необходимым количеством членов. Использование такого сглаживания позволяет существенно ослабить влияние ошибок, заложенных в экспериментально заданные табличные значения функции $y(x)$, а также вести расчет с более крупным шагом или с малым числом свободных параметров.

4.1.3. Частные производные

В случае функции ДВУХ переменных, заданной таблицей

$u_{ij}=f(x_i,y_j)$, где $x_i=x_0+ih_1$ и $y_j=y_0+jh_2$, $i=0,1,\ldots,I$; $j=0,1,\ldots,J$, для малых шагов h_1 и h_2 выражения для частных производных можно записать в виде:

$$\frac{\partial u}{\partial x}\simeq\frac{f(x+h,y)-f(x,y)}{h_1}, \qquad \frac{\partial u}{\partial y}\simeq\frac{f(x,y+h_2)-f(x,y)}{h_2};$$

или

$$\left(\frac{\partial u}{\partial x}\right)_{ij}\approx\frac{u_{i+1,j}-u_{ij}}{h_1}, \quad \left(\frac{\partial u}{\partial y}\right)_{ij}\simeq\frac{u_{i,j+1}-u_{ij}}{h_2}.$$

Для численного дифференцирования таких функций также можно построить аппроксимирующие многочлены. Как и в случае одной переменной, построим ряд Тейлора, и таким образом найдем выражения для соответствующих производных и порядка аппроксимации. Так, например, имеем для центральной производной по x

$$\left(\frac{\partial u}{\partial x}\right)_{ij} = \frac{u_{i+1,j} - u_{i-1,j}}{2h_1} + O(h_1^2),$$

и для второй производной

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{ij} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_1^2} + O(h_1^2).$$

Аналогично аппроксимируются производные по другой переменной.

4.2 Численное интегрирование

Пусть на отрезке [a, b] в точках $x_0 = a, x_1, \ldots, x_n = b$ задана функция $y_i = f(x_i)$. Нам необходимо вычислить определенный интеграл вида

$$\int_a^b f(x)\,dx. \qquad (4.18)$$

Используя определение интеграла как предел интегральной суммы, имеем:

$$\int_a^b f(x)\,dx = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(\xi_i)\,\Delta x_i,$$

где $x_i < \xi_i, < x_{i+1}$ — некая средняя точка интервала x_i, x_{i+1}. Задача интегрирования графически сводится к нахождению площади под графиком функции $f(x)$ на заданном отрезке (рис. 16). Ось x делится на n отрезков длиной Δx_i и на каждом отрезке по определенному критерию выбирается точка ξ_i и вычисляется в этой точке значение функции $f(\xi_i)$. Площадь определяется суммой площадей полученных прямоугольников. Когда длины отрезков $\Delta x_i \to 0$, площадь прямоугольников стремится к значению интеграла.

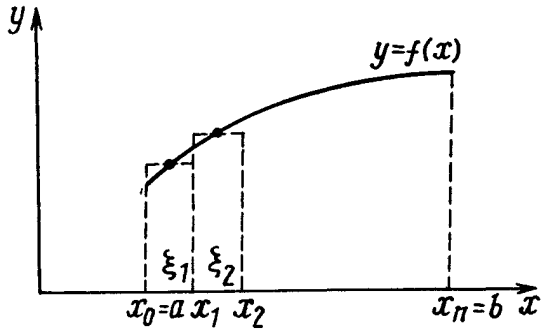


Рис. 16. Геометрическое представление задачи о численном интегрировании

Помимо табличного задания подынтегральной функции задача численного интегрирования возникает также и в том случае, если нельзя проинтегрировать функцию $f(x)$:

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a), \qquad \textbf{(4.19)}$$

т е. если первообразная не выражена в элементарных функциях. Для численного интегрирования функцию $f(x)$ заменяют такой аппроксимирующей функцией $\varphi(x)$ интеграл от которой легко бы вычислялся. Наиболее часто в качестве аппроксимирующих выступают обобщенные интерполяционные многочлены. Поскольку такая аппроксимация линейна относительно параметров, то функцию при этом заменяют неким линейным выражением, коэффициентами которого служат значения функции в узлах

$$f(x) = \sum_{i=0}^{n} f(x_i)\,\Phi_i(x) + r(x), \qquad (4.20)$$

где $r(x)$— остаточный член аппроксимации.
Подставляя (4.20) в (4.18), получим

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} q_i f(x_i) + R, \qquad (4.21)$$

где $q_i = \int_a^b \Phi_i(x)\,dx, \quad R = \int_a^b r(x)\,dx.$

Формула (4.21) называется квадратурной формулой с весами q_i и узлами x_i. Как видно из формулы, веса q_i зависят лишь от расположения узлов, но не от вида функции $f(x)$. Говорят, что квадратурная формула точна для многочленов степени m, если при замене функции $f(x)$ произвольным алгебраическим

многочленом степени m остаточный член становится равным нулю.

4.2.1. Интерполяционные квадратуры

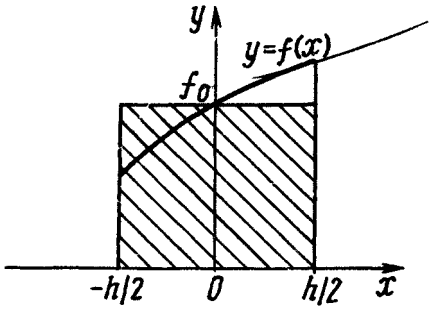
Аппроксимируем функцию $f(x)$ полиномом кулевой степени

$$f(x) \simeq f(x_0) = f_0. \qquad (4.22)$$

Для вычисления интеграла на отрезке $[a, b]$ разобьем его на маленькие отрезки длиной h, а интеграл — на сумму интегралов на отдельных участках.

Тогда для одного участка

$$\int_{-h/2}^{h/2} f(x)\,dx \simeq h f_0, \qquad (4.23)$$



где f_0 — значение функции в середине отрезка. Таким образом, площадь криволинейной трапеции (рис. 17) аппроксимируется прямоугольником, причем функция вычислена в средней точке отрезка.

ля i — го отрезка

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \simeq h f_{i+1/2},$$

Рис. 17 Геометрическое представление метода прямоугольников

где $f_{i+1/2} = f(a + (i + 1/2)h)$.

Тогда, окончательно, значение интеграла на $[a, b]$

$$\int_a^b f(x)\,dx = h(f_{1/2} + f_{3/2} + \ldots + f_{n-1/2}) + r(x). \qquad (4.24)$$

Определим остаточный член в формуле прямоугольников. Пусть $F(x) = \int_0^x f(t)\,dt$, а $F_{\pm 1/2} = F(\pm h/2)$. Так как $F(0)=0$, $F'(0) = f_0$, $F''(0) = f_0' = f'(0)$ и $F'''(x) = f''(x)$, то согласно формуле Тейлора

$$F_{\pm 1/2} = F(0 \pm h/2) = F(0) \pm \frac{h}{2} f_0 + \frac{h^2}{8} f_0' \pm \frac{h^3}{48} f''(\xi_\pm),$$

где $\xi_\pm$ — некоторая точка отрезка $[-h/2, h/2]$.

Однако по формуле Ньютона—Лейбница

$$\int_{-h/2}^{h/2} f(x)\,dx = F_{1/2} - F_{-1/2} = h f_0 + \frac{h^3}{24} \frac{f''(\xi_-) + f''(\xi_+)}{2} =$$

$$= h f_0 + \frac{h^3}{24} f''(\xi), \quad |\xi| \leqslant \frac{h}{2}. \qquad (4.25)$$

Сравнивая (4.23) и (4.25), находим остаточный член формулы прямоугольников для одного отрезка:

$$\frac{h^3}{24} f''(\xi).$$

На всем отрезке $[a, b]$, с учетом

$$\sum_{i=0}^{n} f''(\xi_i) = n f''(\xi) = \frac{b-a}{h} f''(\xi), \quad \xi \in [a, b],$$

имеем:

$$r(x) = h^2 \frac{b-a}{24} f''(\xi). \qquad (4.26)$$

Заменим теперь функцию $f(x)$ в (4.18) интерполяционным многочленом первой степени вида

$$f(x) = f_0 + \frac{f_1 - f_0}{h} x, \qquad (4.27)$$

где $f_0 = f(0)$, $f_1 = f(h)$. Т.е. в отличие от (4.22), где криволинейная трапеция заменялась прямоугольником (см. рис. 17), заменим теперь ее трапецией (рис. 18). Подставив (4.27) в (4.18), получим

$$\int_0^h \left(f_0 + \frac{f_1 - f_0}{h} x \right) dx = h \frac{f_0 + f_1}{2},$$

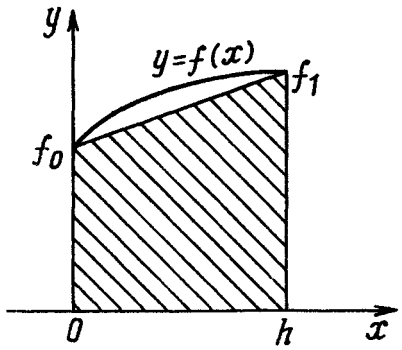


Рис 18. Геометрическое представление метода трапеций

и окончательно формула трапеций для всего отрезка [*a, b]*

$$\xi \in [a, b]. \quad (4.28)$$

$$\int_a^b f(x)\,dx = h\left(f_{1/2}+f_1+f_2+\ldots+f_{N-1}+f_{N/2}\right)-h^2\frac{b-a}{12}f''(\xi),$$

Замена f(*x*) в (4.18) интерполяционным многочленом второй степени вида

$$f(x)=f_0+\frac{f_1+f_{-1}}{2h}+\frac{f_{+1}-2f_0+f_{-1}}{2h^2}x^2, \quad (4.29)$$

где

$f_{-1}=f(-h),\ f_0=f(0),\ f_1=f(h),$

приводит нас к формуле Симпсона (рис. 19).

Действительно, подставив (4.30) в (4.18). получим:

$$\int_{-h}^{h}\left(f_0+\frac{f_1x+f_{-1}}{2h}x+\frac{f_{+1}-2f_0+f_{-1}}{2h^2}x^2\right)dx=\frac{h}{3}\times$$

$$\times(f_{-1}+4f_0+f_1). \quad (4.30)$$

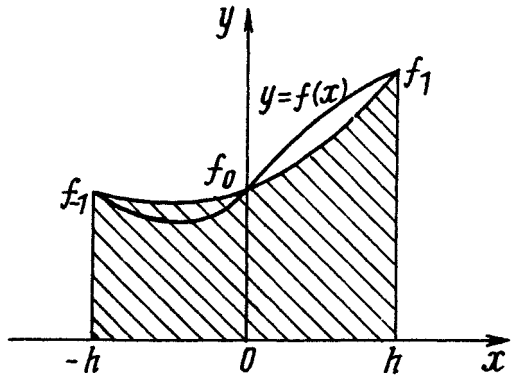


Рис. 19. Геометрическое представление метода Симпсона

Окончательное выражение для всего отрезка

$$\int_a^b f(x)\,dx=\frac{h}{3}(f_0+4f_1+2f_2+4f_3+\ldots+4f_{2n-1}+f_{2n})+$$

$$+\frac{h^4(b-a)}{180}f^{IV}(\xi). \quad (4.31)$$

Анализ остаточных членов формул (4.26), (4.28) и (4.31) показывает, что формулы прямоугольников и трапеций точны для многочленов первой степени, т. е. для линейных функций, в то время как формула Симпсона точна для многочленов третьей степени.

4.2.2. Квадратурная формула Гаусса

В рассмотренных ранее формулах узлы x_i были фиксированы (расположены равномерно на [*a,b,*]) и в квадратурной формуле (4.21) q_i— фиксированы. Тогда для построения интерполяционного полинома, аппроксимирующего функцию f (x) на [a, b] остается лишь (n + 1) независимое условие, т.е. известные значения функции в узлах интерполяции $f(x_i)$. Таким образом, используя эти условия, можно построить многочлен не выше n-й степени Если же не фиксировать положение узлов, а следовательно и q_i то в нашем распоряжении оказываются ($2n$+2) условия, с помощью которых можно построить многочлен (2 n + 1) степени.

Так возникла задача нахождения среди всех квадратурных формул с $(n+1)$ узлами формулы с таким распоряжением узлов x_i на $[a,b]$ и с такими весами q_i при которых она точна для многочленов максимальной степени. Интуитивно ясно, что погрешность метода тем меньше, чем выше порядок многочлена, при численном интегрировании которого получается точный результат.

Приведем интеграл (4.18) к виду

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{+1} f(\bar{x})\,d\bar{x}, \qquad (4.32)$$

где $x = \frac{1}{2}(b-a)\bar{x} + 1/2\,(b+a)$.

Далее будем рассматривать стандартный отрезок [-1,+1], опуская штрих над x. Пусть $x_i \in [-1,1]$, $i=0,1,...,n$ попарно не пересекающиеся произвольные узлы. Тогда, если выбирать **веса в** виде

$$q_i = \int_{-1}^{1} L_{n,i}(x)\,dx, \quad i=0,1,2,..,n, \qquad (4.33)$$

где $L_{n,i}(x)$ — коэффициенты Лагранжа для интерполяционного многочлена n-й степени, а в качестве узлов взять корни многочлена Лежандра

$$X_n(x) = \frac{1}{n!\,2^n}\cdot\frac{d^n}{dx^n}(x^2-1)^n,$$

то полученная квадратурная формул;а $\sum_i q_i f(x_i)$ будет точна для многочленов (2n -1) степени. Это и будет квадратурной формулой Гаусса.

Построим такую квадратурную формулу для случая двух узлов на отрезке [-1,1] (n = 2):

$$X_2(x) = \frac{1}{8}\cdot\frac{d^2}{dx^2}(x^2-1)^2 = 0,$$

или

$$12x^2 - 4 = 0,\ x_{1,2} = \pm\frac{1}{\sqrt{3}}$$

Таким образом, мы определили положение узлов x_1 и x_2, которые расположены симметрично на отрезке [-1,1]. Вычислим теперь веса q_i:

$$q_0 = \int_{-1}^{+1}\frac{x-x_1}{x_0-x_1}dx = 1; \qquad q_1 = \int_{-1}^{+1}\frac{x-x_0}{x_1-x_0}dx = 1.$$

Значение интеграла будет вычисляться в нашем случае по формуле

$$\sum_{i=0}^{1} q_i f(x_i) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Правильный выбор узлов на заданном отрезке позволяет провести аппроксимирующую прямую таким образом, чтобы площади вертикально и горизонтально заштрихованных областей равнялись бы друг другу (рис.20).

В нашем случае площади под кривой $f(x)$ и аппроксимирующей прямой на отрезке [-1,1] будут максимально близки.

Можно показать, что узлы x_i всегда расположены симметрично относительно точки x=0, а веса положительны и в

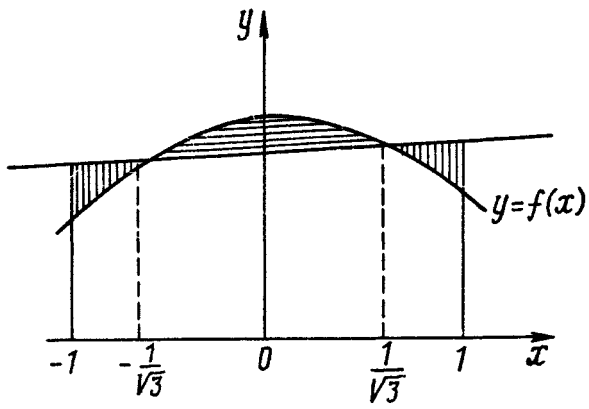


Рис 20 Геометрическое представление метода Гаусса с двумя ординатами

симметричных точках совпадают при любом n. Квадратурную формулу Гаусса наиболее целесообразно использовать для вычисления интегралов от функций с высокой гладкостью и при небольшом числе узлов.

5. ЧИСЛЕННЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

Численные методы линейной алгебры можно условно разделить на четыре группы: решение систем линейных алгебраических уравнений, обращение матриц, вычисление детерминантов, решение проблемы собственных значений. Поскольку обращение матриц и вычисление детерминантов практически сводятся к решению систем линейных алгебраических уравнений, основными численными методами линейной алгебры можно считать рассматриваемые в этой главе методы решения систем линейных уравнений и решения проблемы собственных значений.

Задачи линейной алгебры составляют обширный, хорошо разработанный раздел вычислительной математики, широко используемый в физических расчетах. Размерность возникающих при этом систем уравнений n составляет от нескольких единиц до нескольких миллионов и, конечно, влияет на выбор метода решения. Теоретической основой численных методов линейной алгебры служат теория линейных векторных пространств и теория матриц.

5.1.Прямые методы решения систем линейных алгебраических уравнений

Прямыми в вычислительной математике называют численные методы, в которых:

требующееся для решения задачи число арифметических операций можно оценить по расчетным формулам заранее, до начала решения;

решение является точным.

При этом надо иметь в виду, что в вычислительной математике «точное решение» понимается как «решение с точностью до погрешностей округления», т. е. как решение, которое можно было бы получить по точным формулам на идеализированном компьютере с бесконечной разрядностью машинного слова. В действительности, погрешности округления, конечно, имеются, так что прямые методы также дают приближенное решение.

Рассмотрим задачу численного решения системы n линейных алгебраических уравнений:

$$\begin{cases} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1; \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2; \\ \cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=b_n, \end{cases} \qquad (5.1)$$

которую можно короче записать в матричном виде

$$Ax=b, \qquad (5.2)$$

где A — квадратная матрица размера $n\times n$, x и b — n-мерные векторы:

$$A=(a_{ij};\ i,j=1,2,\ldots,n),\quad x=(x_i;\ i=1,2,\ldots,n),$$
$$b=(b_i;\ i=1,2,\ldots,n). \qquad (5.3)$$

Будем полагать, что матрица A невырожденная, т.е. детерминант $|A|\neq 0$ **и,** следовательно, решение системы (5.1) существует **и** единственно. Наиболее распространенными прямыми **методами** решения систем линейных алгебраических уравнений **являются методы** исключения или, как их часто называют, методы исключения Гаусса. Преобразуем первое уравнение системы **(5.1)** к виду

$$x_1+a_{12}^{(1)}x_2+\ldots+a_{1n}^{(1)}x_n=b_1^{(1)}; \qquad (5.4)$$

$$a_{1j}^{(1)}=a_{1j}/a_{11}\ (j=2,3,\ldots,n),\ b_1^{(1)}=b_1/a_{11}. \qquad (5.5)$$

Домножая **затем** (5.4) на $-a_{il}$ **и** складывая с i-м уравнением (i=2,3, ...,n), исключим x_i из всех уравнений, получая систему (n -1) уравнений. Далее аналогично исключим x_2, x_3 и т.д. Этот процесс исключения называется прямым ходом решения. Наконец, остается одно уравнение, из которого находим x_n, а затем последовательно остальные неизвестные $x_{n-1}, x_{n-2}, \ldots, x_1$; этот этап решения называют обратным ходом. Общие формулы прямого хода имеют вид:

$$a_{kj}^{(k)}=a_{kj}^{(k-1)}/a_{kk}^{(k-1)}\ (j\geqslant k+1),\ b_k^{(k)}=b_k^{(k-1)}/a_{kk}^{(k-1)}; \qquad 5.6)$$

$$\begin{cases} a_{ij}^{(k)}=a_{ij}^{(k-1)}-a_{ik}^{(k-1)}a_{kj}^{(k)} & (i,j\geqslant k+1); \\ b_i^{k}=b_i^{(k-1)}-a_{ik}^{(k-1)}b_k^{(k)} & (i\geqslant k+1); \\ (k=1,2,\ldots,n),\ a_{ij}^{(0)}=a_{ij}, & b_i^{(0)}=b_i\ (i,j=1,2,\ldots,n). \end{cases} \qquad 5.7)$$

Здесь формулы (5.6)—коэффициенты в самом верхнем уравнении на k-ом шаге исключения, формулы (5.7) —остальные

коэффициенты. В результате прямого хода матрица A приводится к верхней треугольной матрице, и система (5.1) приобретает вид:

$$\begin{bmatrix} 1 & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & 1 & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & 1 & \dots & a_{3n}^{(3)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \cdot \\ b_n^{(n)} \end{bmatrix}. \quad (5.8)$$

В результате обратного хода

$$x_i = b_i^{(i)} - \sum_{k=i+1}^{n} a_{ik}^{(i)} x_k \quad (i = n, n-1, \dots 1) \quad (5.9)$$

находятся все неизвестные.

Описанный алгоритм иногда называют схемой «единственного деления», а элемент a_{kk}^{k-1} на который производится деление в формулах (5.6) — ведущим элементом. Если очередной ведущий элемент окажется нулевым, решение прерывается; если же ведущий элемент будет близок к 0, решение будет продолжено, но погрешность может сильно возрасти. Назовем главным максимальный по модулю элемент матрицы. Для устранения указанных выше трудностей рекомендуется на каждом шаге прямого хода исключать неизвестное при главном элементе. Такой метод, называемый методом исключения с выбором главных элементов, имеет меньшую погрешность, но более сложный алгоритм и большую длительность счета из-за поиска главных элементов.

Представим теперь матрицу A *в* факторизованном виде, т. е.

$$A = L\,U, \quad (5.10)$$

где L — нижняя треугольная матрица, U — верхняя треугольная матрица **с** единицами вдоль главной диагонали:

$$L = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}, \quad U = \begin{bmatrix} 1 & u_{12} & \dots & u_{1n} \\ 0 & 1 & \dots & u_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix}. \quad (5.11)$$

Элементы матриц L и U находятся из матричного уравнения (5.10). После этого исходная система (5.2) сводится к последовательному решению двух систем

$$L\,y = b; \quad (5.12)$$

$$U\,x = y \quad (5.13)$$

с треугольными матрицами**.**

Введем расширенные матрицы, дополняя A и U еще одним столбцом:

$$A + b,\ a_{i,n+1} = b_i; \quad U + y,\ u_{i,n+1} = y_i. \quad (5.14)$$

Решение выполняется по формулам**:**

$$l_{mj} = a_{mj} - \sum_{k=1}^{j-1} l_{mk} u_{kj} \quad (j = 1, 2, \dots, n;\ m = j, j+1, \dots, n); \quad (5.15)$$

$$u_{im} = \left(a_{im} - \sum_{k=1}^{i-1} l_{ik} u_{km} \right) / l_{ii} \quad (i = 1, 2, \dots, n;$$
$$m = i+1, i+2, \dots, n+1); \quad (5.16)$$

$$x_i = u_{i,n+1} - \sum_{k=i+1}^{n} u_{ik} x_k \quad (i = n, n-1, \dots, 1). \quad (5.17)$$

Формулы (5.15), (5.16) описывают факторизацию матрицы A и решение уравнения (5.12) (прямой ход), формула (5.17) — решение уравнения (5.13) (обратный ход). Алгоритм решения показан на рис.21. Описанный метод называют методом факторизации, схемой Халецкого (Холесского) или компактной схемой Гаусса. Решение этим методом также встречается с трудностями, если в формуле (5.16) $l_{ii} \simeq 0$.

В физических задачах очень часто встречаются системы линейных алгебраических уравнений с симметричной матрицей A, т. е.

$$A=A^{\mathrm{T}}, \quad a_{ij}=a_{ji} \quad (5.18)$$

$$(i, j=1, 2, \ldots, n)$$

(значок т означает транспонирование). Для таких систем наиболее эффективным прямым методом является метод квадратных корней, представляющий собой разновидность метода факторизации. В этом методе матрица A разлагается в произведение

$$A=A^{T}S \quad (5.19)$$

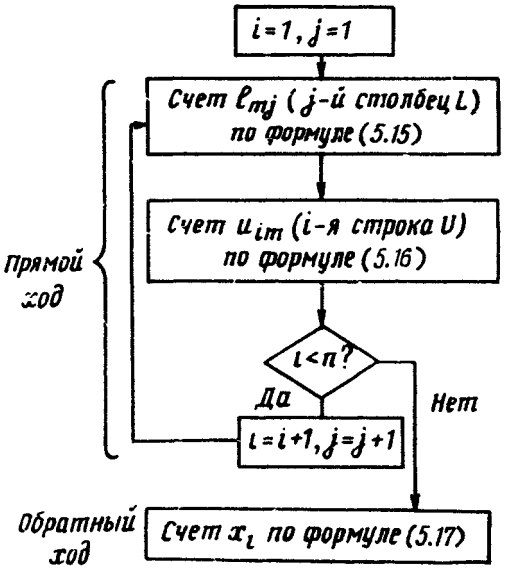


Рис.21. Схема алгоритма метода факторизации

где S — верхняя треугольная матрица с нулями ниже главной диагонали. Из матричного уравнения (5.19) находим:

$$s_{ii}=\left(a_{ii}-\sum_{k=1}^{i-1} s_{ki}^2\right)^{1/2} \quad (i=1, 2, \ldots, n), \quad - \quad (5.20)$$

$$s_{ij}=\left(a_{ij}-\sum_{k=1}^{i-1} s_{ki}s_{kj}\right)/s_{ii} (i<j), \quad s_{ij}=0 \ (i>j), \quad (5.21)$$

а затем из уравнений

$$S^{\mathrm{T}}y=b, \quad Sx=y \quad (5.22)$$

получаем

$$y_i=\left(b_i-\sum_{k=1}^{i-1} s_{ki}y_k\right)/s_{ii} \quad (i=1, 2, \ldots, n), \quad (5.23)$$

$$x_i=\left(y_i-\sum_{k=i+1}^{n} s_{ik}x_k\right)/s_{ii} \quad (i=n, n-1, \ldots, 1). \quad (5.24)$$

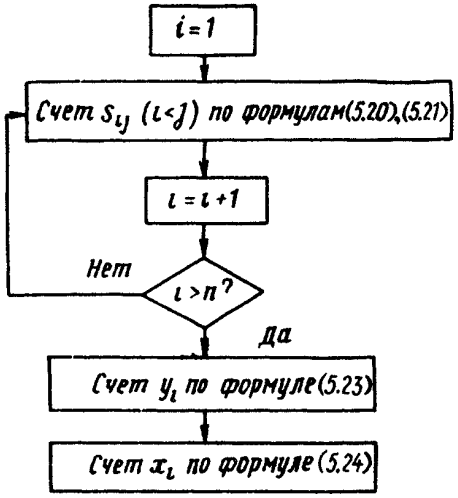


Рис 22 Схема алгоритма метода квадратных корней

Алгоритм решения показан на рис. 22.

Решение методом квадратных корней может быть выполнено в пространстве действительных чисел, если все подкоренные выражения в формуле (5.20) положительны, что в свою очередь, обеспечивается положительной определенностью матрицы A.

В физических приложениях иногда встречаются переопределенные системы линейных алгебраических уравнений:

$$\begin{cases} Ax=b, \quad A=(a_{ij};\ i=1, 2, \ldots, m;\ j=1, 2, \ldots, n); \\ x=(x_1, x_2, \ldots, x_n), \quad b=(b_1, b_2, \ldots, b_m), \end{cases} \quad (5.25)$$

когда число уравнений m больше числа неизвестных n, A —

прямоугольная матрица с m строками и n столбцами ($n<m$). Решение системы (5.25) возможно лишь в среднеквадратичном, т.е. из условия минимума среднеквадратичной погрешности

$$I=||Ax-b||^2=(Ax-b, Ax-b)=\min. \quad (5.26)$$

Составленная в соответствии с методом наименьших квадратов (§ 2.4) система уравнений

$$\frac{\partial I}{\partial x_i}=0 \quad (i=1,2,\ldots,n) \quad (5.27)$$

приводится к системе n линейных алгебраических уравнений

$$A^TAx=A^Tb \quad (5.28)$$

с квадратной, симметричной и положительно определенной матрицей A^TA размера $n\times n$. Система (5.28) затем решается одним из рассмотренных выше методов, например, методом квадратных корней.

Все прямые методы решения систем n линейных алгебраических уравнений требуют $O(n^3)$ арифметических операций и $O(n^2)$ ячеек памяти для хранения промежуточных результатов. Эти свойства, в особенности последнее. обычно ограничивают применимость прямых методов значениями $n\lesssim 10^3$.

Кроме того, большое число $O(n^3)$ арифметических операций может приводить к значительному накоплению вычислительных погрешностей, которые в прямых методах никак не корректируются. Поэтому погрешность прямых методов, формально являющихся точными, часто оказывается выше, чем итерационных, рассматриваемых в следующем параграфе

5.2. Итерационные методы решения систем линейных алгебраических уравнений

Итерационными называются численные методы, обладающие следующими особенностями:

решение находится с помощью последовательных приближений— итераций, начиная с некоторого начального приближения («0-й итерации»), которое должно быть задано (выбрано) заранее;

решение является приближенным, но с любой заданной погрешностью;

требуемое для достижения заданной погрешности число итераций, а следовательно, и число арифметических операций определяется в ходе счета и заранее неизвестно.

Итерационные методы обычно устойчивее прямых к погрешностям округлений, так как эти погрешности постоянно корректируются в ходе итераций наряду с другими типами погрешностей, различают одношаговые и многошаговые итерационные методы. В одношаговых методах для вычисления очередной итерации достаточно знания одной, предыдущей итерации, тогда как в многошаговых методах используется несколько предыдущих итераций.

Одношаговые итерационные методы решения системы уравнений (5.2) .можно записать в общей, так называемой канонической форме:

$$B^{(k)}\frac{x^{(k)}-x^{(k-1)}}{\tau_k}+Ax^{(k-1)}=b \quad (k=1,2,\ldots). \quad (5.29)$$

Здесь k — номер итерации, $x^{(0)}$— начальное приближение, а матрицы $B^{(k)}$ и числа τ_k, называемые итерационным параметром, задают тот или иной итерационный процесс. Очередное приближение $x^{(k)}$ находится из уравнения

$$B^{(k)}x^{(k)}=g^{(k)}, \quad g^{(k)}\equiv(B^{(k)}-\tau_kA)x^{(k-1)}+\tau_kb \quad (5.30)$$

или

$$x^{(k)}=(B^{(k)})^{-1}g^{(k)}=C^{(k)}x^{(k-1)}+\tau_k(B^{(k)})^{-1}b, \quad (5.31)$$

$$C^{(k)}\equiv E-\tau_k(B^{(k)})^{-1}A, \quad (5.32)$$

где E — единичная матрица.

Обозначим $\bar{x}$ - точное решение,

$$\varepsilon^{(k)}=x^{(k)}-\bar{x} \quad (5.33)$$

— погрешность на k-й итерации. Итерационный процесс (5.31) сходится, если

$$\lim_{k\to\infty}||\varepsilon^{(k)}||=0; \quad (5.34)$$

под нормой $||\cdot||$ здесь будем понимать среднеквадратичную норму

$$||u||=\left(\sum_{i=1}^{n}u_i^2\right)^{1/2}=(u,u)^{1/2}. \quad (5.35)$$

Так как точное решение $\bar{x}$ неизвестно, о величине погрешности приходится судить по невязке

$$\xi^{(k)} = Ax^{(k)} - b, \tag{5.36}$$

связанной с погрешностью $\varepsilon^{(k)}$ равенством

$$\xi^{(k)} = A\varepsilon^{(k)}. \tag{5.37}$$

Подставляя в (5.31) $x = \bar{x} + \varepsilon$, находим, что погрешность и невязка на соседних итерациях в одношаговых методах связаны равенствами

$$\varepsilon^{(k)} = C^{(k)}\varepsilon^{(k-1)}, \quad \xi^{(k)} = AC^{(k)}A^{-1}\xi^{(k-1)}. \tag{5.38}$$

Учитывая, что $\|C\varepsilon\| \leqslant \|C\|\,\|\varepsilon\|$, находим из (5.38) достаточное условие сходимости итерационного процесса (5.31):

$$\|C^{(k)}\| \leqslant \rho < 1 \quad (k = 1, 2, \ldots). \tag{5.39}$$

При этом норма погрешности $\|\varepsilon^{(k)}\|$ убывает не медленнее, чем со скоростью геометрической прогрессии со знаменателем ρ. Такая сходимость называется линейной.

Если все матрицы $B^{(k)} = E (k = 1, 2, ..)$, итерационный метод называется явным, в противном случае метод называется неявным. В неявных методах матрицы $B^{(k)}$ должны, конечно, обращаться легче, чем А, иначе использование итерационного метода теряет смысл. Если $B^{(k)} \equiv B$, $\tau_k = \tau$, т.е. не зависят от k, метод называется стационарным, иначе —нестационарным.

. Нестационарный метод называется циклическим, если $B^{(k)}$ и τ_k повторяются с некоторым периодом K.

В дальнейшем ограничимся наиболее распространенным в физических задачах случаем, когда A — симметричная, положительно определенная матрица. Для таких матриц стационарный одношаговый итерационный процесс

$$B\frac{x^{(k)} - x^{(k-1)}}{\tau} + Ax^{(k-1)} = b \quad (k = 1, 2, \ldots) \tag{5.40}$$

сходится, если

$$B - 0{,}5\tau A > 0, \tag{5.41}$$

где неравенство вида $C>0$ означает, что матрица С положительно определенная, т.е. квадратичная форма $(Cu,u)>0$ для всех $u \neq 0$.

Представим матрицу A в виде

$$A = L + D + U, \tag{5.42}$$

где L — нижняя, а U — верхняя треугольные матрицы с нулями вдоль главных диагоналей, D — диагональная матрица:

$$L = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & 0 \end{bmatrix}, \quad D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix},$$

$$U = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{bmatrix}. \tag{5.43}$$

Полагая в канонической форме (5.40) $B = D$, $\tau = 1$, получаем итерационный метод Якоби, имеющий в компонентах вид

$$x_i^{(k)} = -\sum_{j \neq i} a_{ij}x_j^{(k-1)}/a_{ii} + b_i/a_{ii} \quad (i = 1, 2, \ldots, n). \tag{5.44}$$

Из условия (5.41), принимающего в данном случае вид $L + U<D$, следует, что метод Якоби сходится, если матрица A обладает диагональным преобладанием, т. е.

$$a_{ii} > \sum_{j \neq i} |a_{ij}| \quad (i = 1, 2, \ldots, n) \tag{5.45}$$

(все диагональные элементы $a_{ii} > 0$, так как матрица A положительно определенная).

Полагая в (5.40) $B = D + L$, $\tau = 1$, получаем итерационный метод Зейделя

$$x^{(k)} = -D^{-1}(Lx^{(k)} + Ux^{(k-1)} - b), \tag{5.46}$$

который в компонентах имеет вид

$$x_i^{(k)} = -a_{ii}^{-1}\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - b_i\right). \tag{5.47}$$

Обобщением метода Зейделя является метод последовательной верхней релаксации

$$x^{(k)}=(1-\omega)\,x^{(k-1)}-\omega D^{-1}(L\,x^{(k)}+U\,x^{(k-1)}-b), \quad (5.48)$$

которой получается из (5.40) при $B=D+\omega L$, $\tau=\omega$, где $\omega>0$ — релаксационный параметр, и в компонентах записывается в виде

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

$$(i=1,2,\ldots,n). \quad (5.49)$$

Релаксацию по формуле (5.49) называют поточечной, так как она последовательно выполняется по точкам *(i,j)*. Этот метод сходится при $0<\omega<2$; отсюда, в частности, следует, что сходится и метод Зейделя (5.46), который получается из (5.48) при $\omega=1$.

В явных методах $B=E$,

$$\frac{x^{(k)}-x^{(k-1)}}{\tau_k}+A\,x^{(k-1)}=b \quad (k=1,2,\ldots) \quad (5.50)$$

и равенства (5.38) несколько упрощаются:

$$\xi^{(k)}=C^{(k)}\xi^{(k-1)}, \quad \varepsilon^{(k)}=C^{(k)}\varepsilon^{(k-1)}, \quad C^{(k)}=E-\tau_k A\ (k=1,2,\ldots). \quad (5.51)$$

Стационарный явный метод

$$\frac{x^{(k)}-x^{(k-1)}}{\tau}+A\,x^{(k-1)}=b, \quad x^{(k)}=C\,x^{(k-1)}+\tau\,b \quad (552)$$

$B=E$, τ - согласно (5.53)

$$(C=E-\tau A;\ k=1,2,\ldots)$$

называется простой итерацией. Необходимое и достаточное условие сходимости простых итераций:

$$0<\tau<2/\lambda_{\max}(A), \quad (5.53)$$

где $\lambda_{\max}(A)$ — максимальное собственное значение положительно определенной матрицы *Ф*. Практически удобнее пользоваться достаточным условием сходимости (5.39), которое для простых итераций принимает вид

$$||C||\leqslant\rho<1, \quad (5.54)$$

где $||C||$ —любая норма матрицы *C*, например

$$\|C\|=\max_{1\le i\le n}\sum_{j=1}^{n}|c_{ij}| \quad (5.55)$$

Более общим является нестационарный явный итерационный метод Ричардсона (5.50) с переменным итерационным параметром $\tau_k(k=1,2,..)$.

Начальное приближение $x^{(0)}$ в итерационных методах может быть выбрано произвольно, например, можно принять $x^{(0)}=0$ или $x^{(0)}=b$. Итерации проводятся до выполнения одного из условий:

$$1)\ \max|x_i^{(k)}-x_i^{(k-1)}|<\delta; \quad 2)\ ||\xi^{(k)}||<\delta; \quad 3)\ k\geqslant k_{\max}, \quad (5.56)$$

где δ — заданная малая величина, $k_{\max}$ — предельное число итераций. Число арифметических операций в итерационных методах имеет порядок $O(n^2k)$ где k — число выполненных итераций. Скорость сходимости итераций S оценивают по формуле

$$S=-\ln(||\xi^{(k)}||/||\xi^{(k-1)}||). \quad (5.57)$$

Оценка (5.57) является асимптотической, т.е. ею можно пользоваться при $k\gg1$, когда уже выполнено много итераций; при этом величина $1/S$ представляет собой число итераций, требующееся для уменьшения нормы невязки в e раз.

Асимптотическая скорость сходимости S в решающей степени определяется обусловленностью матрицы A, которую можно характеризовать числом обусловленности

$$\rho=||A||\cdot||A^{-1}||=\lambda_{\max}(A)/\lambda_{\min}(A), \quad (5.58)$$

где $\lambda_{\max}(A),\lambda_{\min}(A)$ — границы спектра собственных значений матрицы A. Предположим сначала, что известна лишь верхняя граница спектра $\lambda_{max}(A)$. Полагая в методе простых итераций (5.52) $\tau=1/\lambda_{\max}(A)$, можно найти, что асимптотическая скорость сходимости S = l/p. Для плохо обусловленных, т е. близких к вырожденным матриц $p>>1$ и сходимость простых итераций будет очень медленной.

Если известна и нижняя граница спектра $\lambda_{min}(A)$, то оптимальную скорость сходимости $S = 2/p$ метода простых итераций (5.52) можно получить при

$$\tau=\tau_0=2/[\lambda_{\max}(A)+\lambda_{\min}(A)]. \quad (5.59)$$

Метод простых итераций (5.52) с итерационным параметром τ_0 (5.59) иногда называют методом смещений. Такой же скоростью сходимости обладает и метод Зейделя (5.46),

причем для его использования не требуется знание границ спектра Для циклического варианта метода Ричардсона (5.50) можно подобрать оптимальный набор итерационных параметров:

$$\tau_k = 2/\left[\lambda_{\max}(A)+\lambda_{\min}(A)+(\lambda_{\max}(A)-\lambda_{\min}(A))\cos\frac{\pi(2k-1)}{2K}\right]$$

$$(k=1,2,\ldots,K), \quad (5.60)$$

который при $K \gg 1$ обеспечивает высокую асимптотическую скорость сходимости $S=2/\sqrt{p}$. Поскольку при нахождении набора (5.60) используется теория полиномов Чебышева, его называют чебышевским. При этом параметры τ_k из набора (5.60) следует перебирать не по возрастанию номера k, а в некотором строго определенном порядке для обеспечения вычислительной устойчивости.

В методе последовательной верхней релаксации (5.48) также можно получить быструю асимптотическую сходимость со скоростью $S=4/\sqrt{p}$ при оптимальном значении релаксационного параметра

$$\omega=\omega_0=\frac{2}{1+\sqrt{1-\lambda^2_{\max}(D^{-1}(L+U))}}. \quad (5.61)$$

Входящее в формулу (5.61) максимальное собственное значение матрицы $D^{-1}(L+U)$ можно вычислить заранее, пользуясь методами, изложенными в следующем параграфе, или получить его асимптотическую (при $K \gg 1$) оценку непосредственно в ходе итераций по формуле

$$\lambda^{(k)}_{\max} \simeq \|x^{(k)}-x^{(k-1)}\| / \|x^{(k-1)}-x^{(k-2)}\| =$$

$$=\left[\sum_{i=1}^{n}\left(x_i^{(k)}-x_i^{(k-1)}\right)^2 / \sum_{i=1}^{n}\left(x_i^{(k-1)}-x_i^{(k-2)}\right)^2\right]^{1/2}, \quad (5.62)$$

используя затем это значение в (5.61).

Отдельную группу составляют вариационные нестационарные методы решения систем линейных алгебраических уравнений. В этих методах решение системы (5.2) заменяется эквивалентной задачей поиска нулевого минимума некоторого квадратичного функционала типа (5.26). Так, в методе минимальных невязок

$$x^{(k)}=x^{(k-1)}-\tau_k\xi^{(k-1)} \quad (k=1,2,\ldots) \quad (5.63)$$

итерационный параметр τ_k выбирается из условия минимизации $\|\xi^{(k)}\|$ равным

$$\tau_k=\|\xi^{(k-1)}\|^2_A/\|A\xi^{(k-1)}\|^2 \quad (k=1,2,\ldots), \quad (5.64)$$

где обозначение вида $\|u\|_A$ —так называемая A — норма:

$$\|u\|_A=(Au,u)^{1/2}. \quad (5.65)$$

Хотя асимптотическая скорость сходимости этого метода $S = 2/p$ такая же, как метода Зейделя или метода смещений, он обеспечивает быструю сходимость на начальных итерациях и рекомендуется к применению на начальном этапе итерационного процесса

В явном вариационном методе скорейшего спуска вида (5.50) выбираемый из условия минимизации $\|\varepsilon^{(k)}\|_A$ итерационный параметр

$$\tau_k=\|\xi^{(k-1)}\|^2/\|\xi^{(k-1)}\|^2_A \quad (5.66)$$

обеспечивает уменьшение $\|\varepsilon^{(k)}\|_A$ со скоростью геометрической прогрессии (т е линейное):

$$\|\varepsilon^{(k)}\|_A \leqslant \rho_0^k\|\varepsilon^{(0)}\|_A, \quad \rho_0=\frac{1-\eta}{1+\eta}, \quad \eta=\frac{1}{p}=\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}. \quad (5.67)$$

При $K \gg 1$ можно на каждой итерации уточнять полученное методом простых итераций приближение $x^{(k)}$, заменяя его уточненной величиной

$$\hat{x}^{(k)}=\left[x^{(k)}-\lambda_{\max}(C)\,x^{(k-1)}\right]/\left[1-\lambda_{\max}(C)\right], \quad (5.68)$$

где $\lambda_{\max}(C)$ — максимальное собственное значение матрицы $C=E-\tau A$. Такое уточнение называют ускорением сходимости итераций «по Люстернику». Если значение $\lambda_{\max}(C)$ неизвестно, оно может оцениваться непосредственно в ходе итераций Это приводит к уточненной оценке компонент k-й итерации:

$$\hat{x}^{(k)}=\frac{x_i^{(k)}x_i^{(k-2)}-\left(x_i^{(k-1)}\right)^2}{x_i^{(k)}-2x_i^{(k-1)}+x_i^{(k-2)}} \quad (i=1,2\ldots,n;\ k \gg 1). \quad (5.69)$$

Такое уточнение называют «δ^2—процессом Эйткена». Уточнения

(5.68), (5.69) могут заметно ускорить итерационный процесс. однако при $\lambda_{max}(C)\sim 1$ эти формулы близки к машинной неопределенности вида 0/0 и ими следует пользоваться с осторожностью.

В формуле (5.69) используется уже не одна, а две предыдущие итерации, так что она относится к двухшаговым методам. Рассмотрим неявные двухшаговые итерационные методы вида

$$B\frac{x^{(k)}-x^{(k-1)}+(1-\alpha_k)(x^{(k-1)}-x^{(k-2)})}{\tau_k\alpha_k}+Ax^{(k-1)}=b$$

$$(k=2,3....), \quad (5.70)$$

где B — симметричная, положительно определенная матрица;

τ_k, a_k —итерационные параметры. Очередное приближение находится по формуле

$$\begin{cases} x^{(k)}=\alpha_k x^{(k-1)}+(1-\alpha_k)x^{(k-2)}-\tau_k\alpha_k w^{(k-1)}; \\ w^{(k-1)}=B^{-1}\xi^{(k-1)}, \xi^{(k-1)}=Ax^{(k-1)}-b \quad (k=2,3,...). \end{cases} \quad (5.71)$$

Для начала итерационного процесса (5.71) необходимы $x^{(0)}$, $x^{(1)}$. $x^{(0)}$ по-прежнему выбирается произвольно, а $x^{(1)}$ находится по одношаговой формуле, которая получается из (5.71) при k = 1, $\alpha_1 = 1$. Двухшаговые методы требуют дополнительной памяти компьютера для хранения $x^{(k-2)}$, но зато при правильном выборе итерационных параметров могут обеспечить более быструю сходимость, чем одношаговые.

Наиболее распространенным двухшаговым методом является вариационный метод сопряженных градиентов. В этом методе из условия минимизации $\|\varepsilon^{(k)}\|_A$ итерационные параметры на каждой итерации определяются по формулам

$$\tau_k=(w^{(k-1)},\xi^{(k-1)})/(Aw^{(k-1)},w^{(k-1)}) \quad (k=1,2,...); \quad (5.72)$$

$$w^{(k)}=B^{-1}\xi^{(k)} \quad (k=0,1,2,...), \quad \alpha_1=1; \quad (5.73)$$

$$\alpha_k=\left[1-\frac{\tau_k}{\tau_{k-1}\alpha_{k-1}}\cdot\frac{(w^{(k-1)},\xi^{(k-1)})}{(Aw^{(k-2)},w^{(k-2)})}\right]^{-1} \quad (k=2,3,...). \quad (5.74)$$

Показано, что метод сопряженных градиентов дает точное решение системы (5.2) за n итераций, т е. он может при этом считаться прямым методом. Если же фактически выполненное число итераций $l<<k<n$, то асимптотическая скорость сходимости метода не хуже, чем у одношагового метода с чебышевским набором параметров, но в отличие от последнего, в нем не требуется знание границ спектра собственных значений матрицы А. В целом, итерационные методы позволяют достаточно эффективно решать большие системы линейных алгебраических уравнений и поэтому применимы для $n< 10^6$.

5.3. Численные методы решения проблемы собственных значений

Проблемой собственных значений называют задачу нахождения собственных значений λ и собственных векторов x квадратной матрицы A из однородного уравнения

$$Ax=\lambda x. \quad (5.75)$$

Различают частичную и полную проблемы собственных значений. Частичной называют проблему нахождения некоторой части собственных значений матрицы A, например, максимального по модулю, двух максимальных по модулю, максимального и минимального собственных значений и т.п., а также, возможно, соответствующих собственных векторов. Полной называют проблему нахождения всех собственных значений, а иногда еще и всех собственных векторов матрицы A.

Рассмотрим сначала более простую частичную проблему, полагая, что матрица A имеет полную систему (базис) нормированных собственных векторов e_i (i= 1,2, ...,n), являющихся решением задачи

$$Ae_i=\lambda_i e_i, \quad (e_i,e_i)^{1/2}=\|e_i\|=1 \quad (i=1,2,...,n). \quad (5.76)$$

Такой базис существует, например, если A — симметричная матрица. (Заметим, что действительная симметричная матрица имеет действительные собственные значения). Предположим, что максимальное по модулю собственное значение λ_1 — простое, т. е. является простым корнем характеристического уравнения

$$|A-\lambda E|=0. \quad (5.77)$$

Это, в частности, справедливо для действительной матрицы с положительными элементами. Тогда λ_1 можно найти с помощью итерационного процесса, называемого степенным методом:

$$\begin{cases} \|x^{(k-1)}\|=(x^{(k-1)},x^{(k-1)})^{1/2}, \ e_1^{(k-1)}=x^{(k-1)}/\|x^{(k-1)}\|; \\ x^{(k)}=Ae_1^{(k-1)}, \quad \lambda_1^{(k)}=(x^{(k)},e_1^{(k-1)}) \quad (k=1,2,...), \end{cases} \quad (5.78)$$

где k — номер итерации; в качестве $x^{(0)}$ принимается произвольный, например случайный, вектор. При $k\to\infty$ итерационный процесс (5.78) сходится линейно, т.е. $\lambda_1^{(k)}\to\lambda_1$ с погрешностью,

убывающей со скоростью геометрической прогрессии со знаменателем $|\lambda_2/\lambda_1|$, где λ_2 — следующее по модулю собственное значение матрицы A.

Если λ_2 также простое, его можно затем найти, организуя аналогичный (5.78) итерационный процесс для вектора $x^{(k)} - \lambda_1 x^{(k-1)}$, однако погрешность при этом значительно возрастает. Далее можно аналогично вычислить λ_3 погрешность станет еще больше и т.д. Пусть теперь

$$|\lambda_1(A)| > |\lambda_2(A)| > \ldots > |\lambda_n(A)|. \quad (5.79)$$

Для нахождения максимального и минимального (в алгебраическом смысле) собственных значений матрицы A

$$\lambda_{max}(A) = \max_i \lambda_i(A), \quad \lambda_{min}(A) = \min_i \lambda_i(A) \quad (5.80)$$

найдем описанным выше методом максимальные по модулю собственные значения $\lambda_1(A)$ матрицы A и $\lambda_1(B)$ матрицы

$$B = A - \lambda_1(A)\,E. \quad (5.81)$$

Можно показать, что

$$\begin{cases} \lambda_{max}(A) = \lambda_1(A), & \lambda_{min}(A) = \lambda_1(A) + \lambda_1(B) \quad (\lambda_1(A) > 0); \\ \lambda_{min}(A) = \lambda_1(A), & \lambda_{max}(A) = \lambda_1(A) + \lambda_1(B) \quad (\lambda_1(A) < 0). \end{cases} \quad (5.82)$$

Перейдем теперь к существенно более сложной полной проблеме собственных значений. При относительно небольших значениях $n \le 20$ можно непосредственно находить корни характеристического полинома

$$P_n(\lambda) = |A - \lambda E|$$

степени n. В методе интерполяции, например, для этого выбираются $n+1$ произвольных различных значений Λ_i вычисляются $P_n(\Lambda_i)$ и по значениям

$(\Lambda_i, P_n(\Lambda_i); \quad i = 0, 1, \ldots, n)$ строится интерполяционный многочлен n-й степени, который в силу единственности будет всюду совпадать с $P_n(\lambda)$. Собственные значения $\lambda_i(A)$ находятся затем как корни этого интерполяционного многочлена; наиболее удобным при этом является метод парабол (§ 3.6), позволяющий отыскивать и комплексные корни.

В методе вращений, пригодном только для симметричных матриц, матрица A последовательно преобразуется к диагональному виду с помощью преобразований подобия

$$A^{(k)} = (T^{(i,j)})^{-1} A^{(k-1)} T^{(i,j)} \quad (k = 1, 2, \ldots), \quad A^{(0)} = A, \quad (5.83)$$

где $T^{(i,j)}$ — ортогональные матрицы, имеющие структуру

$$\begin{cases} t_{ii}^{(i,j)} = t_{jj}^{(i,j)} = \cos\varphi, \quad t_{kk}^{(i,j)} = 1 \quad (k \ne i, j); \\ t_{ij}^{(i,j)} = -t_{ji}^{(i,j)} = \sin\varphi \end{cases} \quad (5.84)$$

(остальные элементы нулевые), обладающие свойством

$$(T^{(i,j)})^{-1} = (T^{(i,j)})^{T} \quad (5.85)$$

и называемые матрицами вращения. Поскольку матрицы $A^{(k)}$ (k=0,1,2,...) связаны преобразованием подобия (5.83), они имеют одинаковые собственные значения. Значение φ в (5.84) можно подобрать так, что внедиагональные элементы $a_{ij}=a_{ji}$ матрицы A в результате преобразования (5.83) обращаются в нуль. После преобразования A к диагональному виду, вдоль главной диагонали будут располагаться собственные значения $\lambda_i(A)$ (i=1,2,...,n).

Большей общностью обладают LR — и QR — алгоритмы, применимые для действительных и комплексных систем с несамосопряженными матрицами A, имеющими комплексные собственные значения. В LR — алгоритме, предложенном Рутисхаузером, строится итерационный процесс

$$A^{(k-1)} = L^{(k-1)} R^{(k-1)}, \quad A^{(0)} = A; \quad (5.86)$$

$$A^{(k)} = R^{(k-1)} L^{(k-1)} = (L^{(k-1)})^{-1} A^{(k-1)} L^{(k-1)} \quad (k = 1, 2, \ldots), \quad (5.87)$$

в котором в качестве первого этапа (5.86) матрица $A^{(k-1)}$ факторизуется в произведение левой треугольной матрицы $L^{(k-1)}$ ("left") с 1 на главной диагонали и правой треугольной матрицы $R^{(k-1)}$ («right») с отличными от 1 элементами на главной диагонали. Разложение (5.86) выполняется аналогично разложению в методе факторизации (§ 5.1). Матрицы $A^{(k)}$ (k = 0, 1, 2,...) связаны преобразованием подобия (5.87). Доказано, что

$$\lim_{k\to\infty} L^{(k)} = E, \quad \lim_{k\to\infty} A^{(k)} = R^{(k)}. \quad (5.88)$$

так что в пределе $k \to \infty$ матрица $A^{(k)}$ приводится к правой треугольной матрице с собственными значениями $\lambda_i(A)$ вдоль главной диагонали.

В *QR*-алгоритме Френсиса и Кублановской строится процесс

$$A^{(k-1)}=Q^{(k-1)}R^{(k-1)},\quad A^{(0)}=A; \qquad (5.89)$$

$$A^{(k)}=R^{(k-1)}Q^{(k-1)}=(Q^{(k-1)})^{+}A^{(k-1)}Q^{(k-1)} \quad (k=1,2,\ldots), \qquad (5.90)$$

где $R^{(k-1)}$ —правая треугольная матрица, $Q^{(k-1)}$—унитарная матрица т. е. матрица, обладающая свойством

$$Q^{-1}=Q^{+},\quad q_{ij}^{(-1)}=q_{ji}^{*} \quad (i,j=1,2,\ldots,n), \qquad (5.91)$$

значок « + » означает сопряженную матрицу, (- 1) —элементы обратной матрицы, * — комплексно-сопряженную величину. Здесь, в силу унитарности (5.91), матрицы $A^{(k)}$ (k = 0, 1,..) также связаны преобразованием подобия (5.90), так что набор их собственных значений сохраняется. Доказано, что при $k\to\infty$ матрица $A^{(k)}$ стремится к диагональной, так что в пределе

$k\to\infty$ собственные значения $\lambda_i(A)$ легко находятся из (5.89). *LR* и *QR* — алгоритмы наиболее эффективно решают полную проблему собственных значений. Их применимость ограничивается, в основном, емкостью оперативной памяти компьютера; с ее увеличением эти алгоритмы могут стать основными.

После нахождения собственных значений $\lambda_i(A)$, соответствующие собственные векторы матрицы A могут быть формально вычислены из однородного уравнения

$$(A-\lambda_i(A)E)\,x=0 \quad (i=1,2,\ldots,n). \qquad (5.92)$$

Однако из-за того, что собственные значения $\lambda_i(A)$ найдены с погрешностью, детерминант $|A-\lambda_i(A)E|$ системы (5.92) оказывается отличным от 0, так что (5.92) будет иметь лишь тривиальное нулевое решение. Поэтому вместо (5.92) решается уравнение

$$(A-\lambda_i(A)E)\,x=b, \qquad (5.93)$$

где b — произвольный, например случайный вектор. Доказано, что получаемое таким методом, называемым методом обратных итераций, решение x приблизительно равно i-му собственному вектору матрицы A

5.4. Применение в физических задачах

Наиболее часто задачи линейной алгебры возникают при аппроксимации линейных физических процессов, определенных в некоторой пространственно-временной области, дискретными (сеточными) процессами, определенными в узлах конечной пространственно-временной сетки. Таковы, например, задачи решения обыкновенных дифференциальных уравнений, уравнений в частных производных, интегральных уравнений. Пусть, например,

$$L\,u(\vec{r})=-f(\vec{r}) \quad (\vec{r}\in V) \qquad (5.94)$$

— не зависящее от времени дифференциальное уравнение с линейным дифференциальным оператором L, определенное в области V ($\vec{r}$ —радиус-вектор). Вводя в V некоторую сетку и заменяя производные в операторе L некоторыми конечно-разностными выражениями, можно привести уравнение (5.94) к системе линейных алгебраических уравнений относительно значений функции u в узлах сетки. Размерность этой системы n равна числу сеточных узлов, в которых отыскивается решение, и быстро увеличивается с ростом размерности области V.

Вместо таких конечно-разностных методов можно также использовать метод, называемый методом Галеркина. В этом методе решение уравнения (5.94) отыскивается в виде обобщенного полинома

$$u(\vec{r})=\sum_{j=1}^{n}c_j\varphi_j(\vec{r}), \qquad (5.95)$$

где $\varphi_i(\vec{r})$ — заданные функции, а c_j (j = 1,2, ...,n)— коэффициенты, которые находятся из системы линейных алгебраических уравнений

$$(L\,u+f,\varphi_i)=0 \quad (i=1,2,\ldots,n), \qquad (5.96)$$

имеющей вид

$$\begin{cases} Ac=b,\ A=(a_{ij}),\ a_{ij}=(\varphi_i,L\,\varphi_j)\ (i,j=1,2,\ldots,n);\\ c=(c_1,\ldots,c_n),\ b=(b_1,\ldots,b_n),\ b_i=-(f,\varphi_i)\ (i=1,2,\ldots,n). \end{cases} \qquad (5.97)$$

Здесь круглые скобки означают скалярное произведение вида

$$(v,w)=\int_{V} v\,w\,d\vec{r}. \qquad (5.98)$$

Достаточно полно эти методы рассматриваются в следующих главах. Системы линейных алгебраических уравнений возникали также в уже рассмотренных в главах 2 и 4 задачах интерполирования, аппроксимирования, численного интегрирования

и дифференцирования.

При исследовании стационарных колебательных и волновых процессов в механических, электродинамических и других системах часто возникает задача решения проблемы собственных значений вида

$$L\,u=\lambda\,u, \quad (5.99)$$

где L— линейный дифференциальный оператор, λ —его собственные значения. Отыскание решения задачи (5.99) с помощью конечно-разностных сеточных методов или метода Галеркина (5.95) позволяет свести ее к рассмотренной в §5.3 проблеме собственных значений некоторой матрицы.

Аналогичные задачи возникают и при решении интегральных уравнений. Поскольку вычислительные методы линейной алгебры находят с голь широкое применение, они продолжают интенсивно развиваться и совершенствоваться.

6. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

При исследовании многих физических процессов возникают задачи решения систем обыкновенных дифференциальных уравнений. Будем предполагать, что решение рассматриваемых далее задач существует и единственно, так что необходимо лишь получить это решение численно. Различают задачи с начальными условиями (задача Коши) и краевые задачи. Численные методы решения задачи Коши, в свою очередь, принято делить на одношаговые и многошаговые.

6.1. Одношаговые численные методы решения задачи Коши

Рассмотрим нормальную систему n обыкновенных дифференциальных уравнений, т. е. систему первого порядка, представленную в виде

$$\frac{d\,y}{d\,x}=f\,(x,y), \quad y=(y_1,y_2,\ldots,y_n), \quad f=(f_1,f_2,\ldots,f_n). \quad (6.1)$$

Задача Коши для нормальной системы (6.1) состоит в отыскании решения на интервале $x_0\le x\le x_n$ при заданном начальном условии

$$y(x_0)=y_0. \quad (6.2)$$

Решение будем отыскивать в дискретных точках (узлах) $x=x_j$ (j=1,2, ..,N) с шагом

$$h_j=x_{j+1}-x_j \quad (j=0,1,2,\ldots), \quad (6.3)$$

обозначая

$$y(x_j)=y_j. \quad (6.4)$$

Одношаговыми называют численные методы, в которых для нахождения решения y_{j+1} на следующем шаге достаточно иметь решение лишь на одном предыдущем шаге, т. е. y_j Простейшим одношаговым методом является метод степенных рядов, в котором решение на j-м шаге отыскивается в виде ряда Тейлора:

$$y_{j+1}=\sum_{p=0}^{m} c_p h_j^p, \quad c_p=y^{(p)}(x_j)/p! \quad (p=0,1,\ldots,m), \quad (6.5)$$

где m — порядок метода. Входящие в (6.5) производные $y^{(p)}(x)$ любого порядка p можно найти, последовательно дифференцируя уравнение (6.1). Например, для одного уравнения (n=l)

$$y'=f(x,y),\ y''=f_x+f_y y',\ y'''=f_{x^2}+2f_{xy}y'+f_{y^2}(y')^2+f_y y'' \quad .(6.6)$$

и т.д. (штрихи означают дифференцирование по x). Формула (6.5) имеет погрешность $O(h^{m+1})$ на шаге h_j; в предположении, что погрешности на отдельных шагах суммируются, полная погрешность на всем интервале интегрирования $[x_0,x_n]$ будет выше на порядок, т. е. $O(h^m)$, где

$$h=\max(h_j). \quad (67)$$

Следовательно, (6.5) — это метод m-го порядка. При m = 1 получаем метод первого порядка

$$y_{j+1}=y_j+h_j f_j, \quad f_j=f(x_j,y_j), \quad (6.8)$$

называемый методом Эйлера и имеющий простую геометрическую интерпретацию: на каждом шаге отрезок интегральной кривой $y(x)$ заменяется отрезком касательной к ней. Поскольку в результате интегральная кривая заменяется ломаной линией, метод Эйлера называют также методом ломаных. При $m>$\ недостатком метода степенных рядов является необходимость получения аналитических выражений для многих производных и их вычисления на каждом шаге. Поэтому этот метод обычно используется при $m<3$ и в сочетании с другими методами. Рассмотрим теперь метод последовательных приближений, называемый также методом Пикара. В этом методе решение на j-м шаге отыскивается с помощью итераций вида

$$y^{(i)}(x)=y(x_j)+\int_{x_j}^{x} f(\xi, y^{(i-1)}(\xi))\,d\xi \quad (i=1,2,\ldots), \quad (6.9)$$

где i — номер приближения, ξ — переменная интегрирования. В качестве начального приближения можно принять

$$y^{(0)}(x)=y(x_j)=\text{const}.$$

Получив после нескольких приближений функциональную зависимость $y(x)$ и подставляя в нее $x=x_{j+1}$, находим значение в следующем узле $y(x_{j+1})$ и т.д.

Практически, методом можно пользоваться лишь **в том** случае, если все интегралы в (6.9) берутся аналитически. Это имеет место, например, когда правая часть уравнения (6.1) —линейная функция y с постоянными коэффициентами, т. е.

$$y'=A\,y+b, \quad (6.10)$$

где A —матрица размера $n\times n$, b — вектор размера n, не зависящие от x, y. Обозначая

$$u=A\,y+b, \quad (6.11)$$

приведем уравнение (6.10) к виду

$$u'=A\,u. \quad (6.12)$$

Получаемое методом последовательных приближений (6.8) решение уравнения (6.12) на шаге h_j можно записать в виде

$$u_{j+1}=e^{A h_j}u_j, \quad e^{A h_j}=\sum_{s=0}^{\infty}\frac{(A h_j)^s}{s!}, \quad (6.13)$$

где матричная экспонента называется «экспоненциал». Учитывая затем (6.11), окончательно получаем

$$y_{j+1}=y_j+h_j\sum_{s=1}^{\infty}\frac{(A h_j)^{s-1}}{s!}u_j. \quad (6\text{-}14)$$

Слагаемые суммы в (6.14), начиная с некоторого номера s, быстро убывают до машинного 0 из-за s в знаменателе, после чего суммирование можно оборвать. Достоинство этого метода — возможность вычислений с высокой точностью при большом шаге h.

Основными одношаговыми методами являются методы "предиктор-корректор", наиболее известны из которых методы Рунге-Кутга. Для построения этих методов будем отыскивать решение (6.1) (полагая пока n = 1) в виде

$$y_{j+1}=y_j+h_j[\beta f(x_j,y_j)+\alpha f(x_j+\gamma h_j, y_j+\delta h_j)]. \quad (6.15)$$

Коэффициенты $\alpha,\beta,\gamma,\delta$ определим так, чтобы как можно больше членов в (6.15) совпали с членами ряда Тейлора

$$y_{j+1}=y_j+h_j y_j+0{,}5\,h_j^2 y_j''+\ldots=y_j+h_j f_j+$$
$$+0{,}5\,h_j^2(f_x+f_y f)_j+\ldots \quad (6.16)$$

Используя приближенные равенства

(6.17)

$$f(x_j+\gamma h_j,\ y_j+\delta h_j)\simeq f_j+h_j(f_x\gamma+f_y\delta)_j;$$
$$y_{j+1}\simeq y_j+h_j(\alpha+\beta)f_j+\alpha h_j^2(f_x\gamma+f_y\delta)_j; \quad (6.18)$$

и приравнивая почленно (6.18) и ряд Тейлора (6.16), находим

$$\alpha+\beta=1, \quad \alpha\gamma=0{,}5, \quad \alpha\delta=0{,}5 f_j. \quad (6.19)$$

Поскольку для четырех неизвестных $\alpha,\beta,\gamma,\delta$ имеется лишь три уравнения (6.19), одно из неизвестных, например α, может быть выбрано произвольным. Так, полагая α=1, получаем метод, в котором каждый шаг решения складывается из двух этапов. Первый этап, называемый «предиктор», состоит в вычислении предварительного («предсказанного») значения на полушаге $h_j/2$:

$$\tilde{y}_{j+1/2}=y_j+0{,}5\,h_j f(x_j,y_j)+O(h_j^2). \quad (6.20)$$

На втором этапе, называемом «корректор>, вычисляется окончательное, «исправленное» значение на целом шаге h, с использованием предсказанного значения $\tilde{y}_{j+1/2}$

$$y_{i+1}=y_i+h_i f(x_i+h_i/2,\ \tilde{y}_{i+1/2})+O(h_i^3). \quad (6.21)$$

Таким образом строится решение на всем интервале $[x_0, x_n]$ Полная погрешность метода имеет порядок $O(h^2)$, т.е. это метод второго порядка.

Описанный метод имеет простую интерпретацию. Формально инт ым

$$y_{i+1}=y_i+\int_{x_i}^{x_{i+1}} f(x,y)\,dx. \quad (6.22)$$

Однако интеграл в (6.22) аналитически не берется, так как не известна входящая в f функция $y(x)$. Вычисляя интеграл приближенно по формуле прямоугольников, получаем:

$$y_{i+1}=y_i+h_i f(x_i+h_i/2,\ y_{i+1/2})+O(h_i^3),$$
$$y_{i+1/2}=y(x_i+h_i/2). \quad (6.23)$$

В (6.23) также входит неизвестное значение $y_{j+1/2}$, но, разлагая $f(x,y)$ в ряд Тейлора, нетрудно показать, что $y_{j+i/2}$ в (6.23) можно заменить на $\tilde{y}_{j+1/2}$ (6.20) с сохранением порядка погрешности $O(h_i^3)$. В результате получаем формулу (6.21).

Выбирая теперь a = 0,5, получаем другой метод «предиктор-корректор» второго порядка, в котором каждый шаг решения также складывается из двух этапов:

$$\tilde{y}_{i+1}=y_i+h_i f_i+O(h_i^2),\quad f_i=f(x_i,y_i); \quad (6.24)$$
$$y_{i+1}=y_i+0{,}5\,h_i(f_i+\tilde{f}_{i+1})+O(h_i^3),$$
$$\tilde{f}_{j+1}=f(x_{j+1},\tilde{y}_{j+1}). \quad (6.25)$$

Этот метод можно также получить, вычисляя интеграл в (6.22) по формуле трапеций

$$y_{i+1}=y_i+0{,}5\,h_i(f_i+f_{i+1})+O(h_i^3) \quad (6.26)$$

и заменяя в (6.26) неизвестное значение f_{j+1} на $\tilde{f}_{i+1}$ с сохранением порядка погрешности.

Описанные методы входят в группу методов Рунге-Кугга второго порядка. Название «предиктор-корректор» этих методов связано с тем, что первоначальное, предсказанное значение в них имеет невысокую точность, но последующая коррекция повышает порядок погрешности на 1. При α=0 из уравнения (6.18) можно получить рассмотренный ранее метод Эйлера, в котором вообще отсутствует коррекция, и который имеет поэтому первый порядок погрешности. Наоборот, увеличивая число коррекций можно строить методы Рунге-Кутта более высокого порядка. Наиболее часто из этих методов используется метод Рунге-Кутта четвертого порядка, который описывается формулами:

$$y_{i+1}=y_i+\frac{1}{6}(g_1+2g_2+2g_3+g_4)+O(h_i^5), \quad (6.27)$$

где $g_1=h_i f_i$,

$$g_2=h_i f(x_i+0{,}5h_i,\ y_i+0{,}5g_1),$$
$$g_3=h_i f(x_i+0{,}5h_i,\ y_i+0{,}5g_2),$$
$$g_4=h_i f(x_i+h_i,\ y_i+g_3).$$

Хотя формулы (6.20), (6.21), (6.24), (6.25), (6.27) получены для одного уравнения, их можно применять и к системам уравнений (6.1), полагая, что в этих формулах y и f — n-мерные векторы. Достоинствами методов Рунге-Кутта являются высокая точность, возможность вычислений с переменным шагом h_i, хорошая устойчивость к погрешностям округлений благодаря коррекциям, экономное использование памяти компьютере Недостаток этих методов — замедление счета из-за неоднократного вычисления правой части f.

6.2. Многошаговые численные методы решения задачи Коши

Многошаговыми называют численные методы, в которых для нахождения решения y_{j+1} на следующем шаге необходимо знать решение на нескольких предыдущих шагах y_{i+p}(p = 0,1,..., m; $m \geq 1$). В отличие от одношаговых методов, многошаговые методы обычно используются с постоянным шагом h, так как иначе расчетные формулы слишком усложняются.

Простейшим и часто применяемым является многошаговый метод «с перешагиванием» второго порядка. В этом методе производная y' в уравнении (6.1) заменяется центрально-разностной производной с погрешностью второго порядка:

$$y_i'=0{,}5(y_{i+1}-y_{i-1})/h+O(h^2). \quad (6.28)$$ Подставляя

(6.28) в (6.1), находим

$$0{,}5(y_{i+1}-y_{i-1})/h+O(h^2)=f_i,\quad f_i=f(x_i,y_i), \quad (6.29)$$

откуда получаем расчетную формулу

$$y_{i+1}=y_{i-1}+2h f_i+O(h^3). \quad (6.30)$$

Наиболее распространенной является группа многошаговых методов, называемых методами Адамса. Заметим, что поскольку известны значения

$$y_{i-p}=y(x_{i-p}),\quad x_{i-p}=x_i-ph\quad (p=0,1,...,m), \quad (6.31)$$

то, следовательно, известны и значения

$$f_{j-p}=f(x_{j-p}, y_{j-p}). \quad (6.32)$$

Построим интерполяционный полином m-й степени $P_m(x)$ проходящий через узлы (x_{j-p}, f_{j-p}, $p = 0,1 \ldots\ldots\ldots m$) и имеющий, как известно, погрешность $O(h^{m+1})$ (§ 1.4). Тогда

$$y' = P_m(x) + O(h^{m+1}). \quad (6,33)$$

Интегрируя теперь уравнение (6.33) на шаге h, находим

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} P_m(x)\,dx + O(h^{m+2}), \quad (6.34)$$

где $x_{j+1}=x_j+h$, $y_{j+1}=y(x_{j+1})$. Вычисляя аналитически интеграл в (6.34), получаем y_{j+1}, после чего можно сделать следующий шаг и т.д.

Получаемая таким образом формула называется экстраполяционной формулой Адамса, так как в ней используется экстраполирование многочлена $P_m(x)$ на один шаг h за пределы области расположения интерполяционных узлов $[x_{j-m}, x_j]$. Полная погрешность такого метода $O(h^{m+1})$, т.е. порядок метода $m + 1$ равен числу узлов интерполяционного полинома $P_m(x)$ и числу правых частей f_{j-p} участвующих в вычислениях.

Формулы Адамса принято записывать в виде

$$y_{j+1} = y_j + h \sum_{p=0}^{m} \alpha_p f_{j-p} + O(h^{m+2}); \quad (6.35)$$

коэффициенты a_p даются во многих книгах. Например:

$m=1$: $\alpha_0=1{,}5$, $\alpha_1=-0{,}5$;

$m=2$: $\alpha_0=23/12$, $\alpha_1=-16/12$, $\alpha_2=5/12$;

$m=3$: $\alpha_0=55/24$, $\alpha_1=-59/24$, $\alpha_2=37/24$, $\alpha_3=-9/24$;

$m=4$: $\alpha_0=1901/720$, $\alpha_1=-2774/720$, $\alpha_2=2616/720$, $\alpha_3=-1274/720$, $\alpha_4=251/720$.

Достоинствами многошаговых методов являются хорошая точность и высокая скорость счета вследствие того, что на каждом шаге правая часть f вычисляется однократно. К недостаткам относятся невозможность счета с переменным шагом и необходимость хранения в памяти компьютера нескольких предыдущих значений y_{j-p} или правой части f_{j-p}(p= 1,2, ...,m). Для одного уравнения (n=l) это, конечно, несущественно. Но в физических задачах часто приходится решать системы из сотен тысяч и даже миллионов уравнений, например, при анализе динамики ансамбля молекул, электронов, протонов; тогда дополнительные затраты памяти будут очень велики. Кроме того, для начала счета в многошаговых методах требуются начальные, так называемые разгонные значения y_{-p}(p=1.2,..., m). Эти значения должны быть предварительно получены каким-либо одношаговым методом.

Из-за отсутствия коррекций многошаговые методы менее устойчивы к вычислительным погрешностям, чем методы «предиктор-корректор», причем устойчивость снижается с ростом m. Меньшая устойчивость вынуждает проводить вычисления с достаточно малым шагом h, что существенно замедляет решение и может лишить многошаговые методы их основного преимущества — высокой скорости. С целью улучшения устойчивости можно провести одну коррекцию. Для этого, считая предсказанным решение $\tilde{y}_{i+1}$, вычисленное по экстраполяционной формуле (6.35), найдем новое значение правой части $f_{j+1} = f(x_{j+1}, y_{j+1})$, построим новый интерполяционный многочлен m-й степени $Q_m(x)$, проходящий через узлы

(x_{j+1-p}, f_{j+1-p}; p = 0,1, ...,m), и с его помощью вычислим окончательное, скорректированное значение

$$y_{j+1} = y_j + h \sum_{p=0}^{m} \beta_p f_{j+1-p} + O(h^{m+2}), \quad (f_{j+1}=\tilde{f}_{j+1}). \quad (6.36)$$

Формула (6.36) называется интерполяционной формулой Адамса, значения β_p для m = 1—4 приводятся ниже:

$m=1$: $\beta_0=0{,}5$, $\beta_1=0{,}5$;

$m=2$: $\beta_0=5/12$, $\beta_1=8/12$, $\beta_2=-1/12$;

$m=3$: $\beta_0=9/24$, $\beta_1=19/24$, $\beta_2=-5/24$, $\beta_3=1/24$;

$m=4$: $\beta_0=251/720$, $\beta_1=646/720$, $\beta_2=-264/720$, $\beta_3=106/720$, $\beta_4=-19/720$.

Заметим, что коррекция не увеличивает порядок погрешности $O(h^{m+1})$, а лишь улучшает устойчивость метода. Более полно проблема устойчивости обсуждается в следующем параграфе.

6.3. Устойчивость численных методов решения задачи Коши. Неявные методы

Проблему устойчивости численных методов решения задачи Коши исследуем на примерах. Рассмотрим сначала линейное уравнение

$$y' = -\lambda y + \varphi(x) \quad (\lambda > 0), \quad y' + \lambda y = \varphi(x). \quad (6.37)$$

Из-за вычислительных погрешностей вместо y будет вычисляться $y + \varepsilon$, где ε — погрешность. Подставляя $y + \varepsilon$ в (6.37) и учитывая, что для точного решения y выполняется уравнение (6.37), получаем уравнение для погрешности

$$\varepsilon' + \lambda \varepsilon = 0, \quad (6.38)$$

которое отличается от исходного уравнения лишь отсутствием правой части $\varphi(x)$. (Это естественно, так как мы считаем правую часть известной точно.) Уравнение (6.38) имеет аналитическое решение

$$\varepsilon = e^{-\lambda x} \varepsilon_0 \quad (6.39)$$

(ε_0—начальная погрешность), из которого видно, что погрешность не нарастает.

Предположим теперь, что уравнение (6.38) решается численно методом Эйлера (6.8), т.е.

$$\varepsilon_{j+1} = \varepsilon_j - \lambda h_j \varepsilon_j = (1 - \lambda h_j) \varepsilon_j \quad (j = 0, 1, 2, \ldots). \quad (6.40)$$

Чтобы погрешность не нарастала от шага к шагу, должно выполняться условие

$$|1 - \lambda h_j| < 1 \quad (j = 0, 1, 2, \ldots), \quad (6.41)$$

откуда следует ограничение на шаг

$$0 < h < 2/\lambda, \quad (6.42)$$

где h дается формулой (6.7). Методы, устойчивые при некоторых ограничениях на шаг, называются условно устойчивыми. Следовательно, метод Эйлера (6.8) —условно устойчивый.

Все рассмотренные до сих пор методы относятся к так называемым явным методам, так как в них неизвестное y_{j+1} явно выражается через уже известные значения.

Предположим теперь, что в методе Эйлера производная y' берется не в начале, а в конце шага:

$$y_{j+1} = y_j + h_j f_{j+1}, \quad f_{j+1} = f(x_{j+1}, \quad y_{j+1}). \quad (6.43)$$

Этот метод является уже неявным, так как в нем y_{j+1} выражается через известные значения неявно, как решение некоторого, в общем случае нелинейного уравнения. Решение этого уравнения можно получить методами, описанными в главе 3.

Применяя метод (6.43) к уравнению (6.37), находим для погрешности

$$\varepsilon_{j+1} = \varepsilon_j - h_j \lambda \varepsilon_{j+1}, \quad \varepsilon_{j+1} = (1 + h_j \lambda)^{-1} \varepsilon_j, \quad (6.44)$$

откуда видно, что $|\varepsilon_{j+1}| < |\varepsilon_j|$ при всех $h_j > 0$. Методы, устойчивые при любых положительных значениях шага, называются безусловно (или абсолютно) устойчивыми. Общий анализ численных методов решения дифференциальных уравнений показывает, что все явные методы условно устойчивы, тогда как среди неявных методов существуют безусловно устойчивые.

В качестве следующего примера рассмотрим линейное уравнение вынужденных колебаний:

$$\ddot{x} + \omega^2 x = \varphi(t). \quad (6.45)$$

Здесь x — смещение, ω— частота колебаний, $\varphi(t)$ — вынуждающая сила, точка означает дифференцирование по времени t. Уравнение для погрешности

$$\ddot{\varepsilon} + \omega^2 \varepsilon = 0 \quad (6.46)$$

можно свести к системе двух уравнений первого порядка

$$\dot{\varepsilon} = \varepsilon_1, \quad \dot{\varepsilon}_1 = -\omega^2 \varepsilon. \quad (6.47)$$

Домножая первое из этих уравнений на $i\omega \left(i = \sqrt{-1}\right)$ и складывая со вторым, получаем одно комплексное уравнение

$$\dot{z} = i\omega z \quad (z = \varepsilon_1 + i\omega \varepsilon), \quad z(0) = z_0. \quad (6.48)$$

Заметим, что (6.48) имеет аналитическое решение

$$z = e^{i\omega t} z_0, \quad (6.49)$$

из которого видно, что амплитуда погрешности постоянна: $|z| = |z_0| = \text{const}$.

Решение (6.48) явным методом Эйлера (6.8) с шагом $\tau = \text{const}$ дает:

$$z_{j+1} = z_j + i\omega\tau z_j = (1 + i\omega\tau) z_j \quad (j = 0, 1, 2, \ldots, N), \quad (6.50)$$

(6.51)

$$|z_{j+1}|^2 = (1 + \omega^2\tau^2) |z_j|^2, \quad |z_{j+1}|^2 > |z_j|^2.$$

Из (6.51) следует, что погрешность нарастает при любых значениях

τ. Такие численные методы называют безусловно неустойчивыми. Видно, что для уравнения колебаний явный метод Эйлера безусловно неустойчив. Нарастание погрешности можно оценить по формулам:

$$\begin{cases} |z_{j+1}|^2 \simeq e^{\omega^2\tau^2}|z_j|^2 = e^{(j+1)\omega^2\tau^2}|z_0|^2, \\ |z_{j+1}| \simeq e^{0{,}5(j+1)\omega^2\tau^2}|z_0|. \end{cases} \quad (6.52)$$

Так, например, при $\omega\tau = 0{,}1\pi$ (20 шагов на период колебаний) погрешность $|z|$ увеличивается за $N = 200$ шагов (10 периодов) в отношении e^{10}, т.е. неустойчивость является очень сильной.

Происхождение неустойчивости в данном примере имеет простую геометрическую интерпретацию. Аналитическое решение (6.49) описывает движение по окружности радиусом z_0 в комплексной плоскости ($\varepsilon_i, i\omega\varepsilon$). В явном методе Эйлера движение на каждом шаге осуществляется по отрезку касательной к окружности и превращается в движение по ломаной спирали нарастающим радиусом.

Неявный метод Эйлера (6.43) дает решение уравнения (6.48) с убывающей погрешностью

$$|z_{j+1}| = z_j + i\omega\tau z_{j+1} = (1 - i\omega\tau)^{-1} z_j,$$
$$|z_{j+1}|^2 = |z_j|^2/(1 + \omega^2\tau^2) \quad (6.53)$$

и, следовательно, безусловно устойчив. Геометрически это объясняется тем, что движение осуществляется из точки z_j по отрезку прямой, параллельной касательной $\dot{z}_{j+1}$ к окружности в точке z_{j+1} и направленной внутрь круга, т.е. по ломаной спирали убывающим радиусом.

Рассмотрим, наконец, решение уравнения (6.45) методом предиктор-корректор (6.20), (6.21). В этом методе для погрешностиполучаем:

$$\begin{cases} z_{j+1} = [1 - 0{,}5(\omega\tau)^2 + i\omega\tau] z_j, \quad |z_{j+1}|^2 = [1 + 0{,}25(\omega\tau)^4]|z_j|^2, \\ |z_{j+1}| \simeq e^{(j+1)\omega^4\tau^4/8}|z_0|. \end{cases} \quad (6.54)$$

Из (6.54) видно, что для уравнения колебаний рассматриваемый явный метод, как и явный метод Эйлера, безусловно неустойчив, но благодаря коррекции его неустойчивость значительно слабее. Так, при тех же значениях $\omega\tau = 0{,}1\pi$, $N = 200$, относительный рост погрешности составляет сего $e^{0{,}24} = 1{,}28$, т.е. погрешность практически не увеличивается.

В заключение этого параграфа заметим, что в численном решении важен не абсолютный, а относительный рост погрешности по сравнению с точным решением. Так, даже экспоненциально нарастающая погрешность может оставаться малой, если точное решение дифференциальной задачи также экспоненциально нарастает. Поэтому для корректного вывода об устойчивости или неустойчивости численного метода требуется тщательный анализ поведения как погрешности, так и точного решения для каждой дифференциальной задачи.

6.4. Численное решение краевых задач для обыкновенных дифференциальных уравнений

Краевой задачей называется задача нахождения решения системы (6.1) ($n \geq 2$), для которой дополнительные условия задаются более чем в одной точке. Здесь ограничимся линейной краевой задачей (линейными называются краевые задачи для линейных уравнений с линейными краевыми условиями) для одного уравнения второго порядка

$$y'' + p(x)y' + q(x)y = f(x) \quad (x_0 \leqslant x \leqslant x_N), \quad (6.55)$$

$$y(x_0) = y_0, \quad y(x_N) = y_N. \quad (6.56)$$

(Возможны и краевые условия общего вида, например, $\alpha y_0' + \beta y_0 = \gamma$). Решение будем отыскивать на сетке с постоянным шагом h (такие сетки называются равномерными) и узлами

$$x_j = x_0 + jh\ (j = 0, 1, \ldots, N), \quad h = (x_N - x_0)/N, \quad (6.57)$$

где N— число шагов.

Аппроксимируя производные в (6.55) конечно-разностными формулами

$$y_j'' = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + O(h^2), \quad y_j' = \frac{y_{j+1} - y_{j-1}}{2h} + O(h^2), \quad (6.58)$$

сведем задачу (6.55), (6.56) к системе линейных алгебраических уравнений:

$$a_j y_{j-1} - b_j y_j + c_j y_{j+1} = d_j \quad (j = 1, 2, \ldots, N-1); \quad (6.59)$$

$$a_j = 1 - 0{,}5p_j h, \quad b_j = 2 - q_j h^2, \quad c_j = 1 + 0{,}5p_j h^2, \quad d_j = f_j h^2, \quad (6.60)$$

$$p_j = p(x_j), \quad q_j = q(x_j), \quad f_j = f(x_j), \quad (6.61)$$

где y_0, y_n— заданные краевые условия.

Система уравнений (6.59) имеет трехдиагональную матрицу, так как каждое уравнение содержит лишь три соседних неизвестных y_j, $y_{j\pm1}$. Экономным методом решения таких систем является метод прогонки, представляющий собой разновидность метода исключения Гаусса. В этом методе решение отыскивается в

виде

$$y_j=\xi_{j+1}y_{j+1}+\eta_{j+1} \quad (j=1,2,...,N-1), \quad (6.62)$$

где ξ_{j+1}, η_{j+1}—пока неизвестные коэффициенты, называемые прогоночными. Заменим в (6.62) j на j—1 и, подставляя

$$y_{j-1}=\xi_j y_j+\eta_j \quad (6.63)$$

в уравнение (6.59), преобразуем его к виду

$$y_j=\frac{c_j}{b_j-a_j\xi_j}y_{j+1}+\frac{a_j\eta_j-d_j}{b_j-a_j\xi_j}. \quad (6.64)$$

Сравнивая теперь (6.62) и (6,64), получаем рекуррентные формулы для прогоночных коэффициентов

$$\xi_{j+1}=\frac{c_j}{b_j-a_j\xi_j}, \quad \eta_{j+1}=\frac{a_j\eta_j-d_j}{b_j-a_j\xi_j} \quad (j=1,2,...,N-1). \quad (6.65)$$

Применяя (6.62) к j = 0, находим, что

$$\xi_1=0, \quad \eta_1=y_0. \quad (6.66)$$

Таким образом, метод решения складывается из двух этапов. На первом этапе (прямой ход прогонки) вычисляются прогоночные коэффициенты по рекуррентным формулам (6.65) с начальными значениями (6.66). На втором этапе (обратный ход прогонки) вычисляются y_j(j = N-1, N — 2,..., 1) по формуле (6.62) с учетом заданного краевого условия y_N. Погрешность решения имеет тот же порядок, с которым аппроксимируются производные, т. е. $O(h^2)$.

Согласно формулам (6.65), (6.62), прямой ход требует 6 арифметических операций, а обратный — 2 операции на каждый узел; общее число операций $8(N—1)\sim 8N$ линейно зависит от числа узлов. Анализ показывает, что прогонка устойчива при выполнении условий

$$a_j, b_j, c_j>0, \quad b_j\geqslant a_j+c_j \quad (6.67)$$

для всех j, причем хотя бы в одном узле должно выполняться строгое неравенство $b_j>a_j+c_j$. В рассматриваемой задаче условия (6.67) приводят к требованиям

$$h<2/\max|p_j|, \quad q_j\leqslant 0 \quad (j=1,2,..,N-1). \quad (6.68)$$

Помимо метода прогонки для решения краевых задач (6.55), (6.56), могут также использоваться более точные, но алгоритмически более сложные метод Галеркина и метод сплайновых аппроксимаций, а для уравнений с постоянными коэффициентами p = const, q = const, еще и метод преобразования Фурье. Эти методы здесь не рассматриваются.

6.5. Применение численных методов решения обыкновенных дифференциальных уравнений в физических задачах

Численное решение больших систем обыкновенных дифференциальных уравнений часто встречается при исследовании динамики различных физических объектов. Например, моделирование динамики ансамблей $M\gg 1$ заряженных частиц (электронов, ионов) в электрическом поле $\vec{E}(\vec{r},t)$ сводится к решению задачи Коши для системы $6M$ уравнений:

$$\begin{cases} \vec{r}_i=\vec{v}_i, \quad \vec{p}=\gamma m_0\vec{V}, \quad \gamma=\sqrt{1+\left(\frac{p}{m_0c}\right)^2}. \\ \vec{p}_i=q\vec{E}_i, \end{cases} \quad (6.69)$$

Здесь $\vec{r}$,—радиус-вектор, $\vec{p}_i$— импульс i-й частицы (i= 1,2,...,M); q, m_0—заряд и масса покоя частиц, γ — релятивистский множитель, c — скорость света.

Экономным методом решения системы (6.69) является разновидность метода " с перешагиванием", в которой импульсы вычисляются в целые моменты времени $t = k\tau$ (k = 0,1,2,...), а координаты — в полуцелые $t + \tau/2$ (τ — временной шаг решения). Пусть в текущий момент t в памяти компьютера хранятся значения $\vec{p}_i(t)$, $\vec{r}_i(t+\tau/2)$ для всех M частиц. Тогда шаг решения выполняется по формулам:

$$\begin{cases} \vec{p}_i(t+\tau)=\vec{p}_i(t)+\tau q\vec{E}(\vec{r}_i(t+\tau/2),t+\tau/2); \\ \vec{r}_i\left(t+\frac{3}{2}\tau\right)=\vec{r}_i\left(t+\frac{\tau}{2}\right)+\tau\frac{\vec{p}_i(t+\tau)}{\gamma_i(t+\tau)m_0} \quad (i=1,2,...,M). \end{cases} \quad (6.70)$$

В качестве следующего характерного примера рассмотрим одномерное уравнение теплопроводности:

$$\frac{\partial u}{\partial t}=a^2\frac{\partial^2 u}{\partial x^2}+f(x) \quad (0<x<l, 0\leqslant t\leqslant T); \quad (6.71)$$

$$u(x,0)=u^0(x), \quad u(0,t)=\mu_1, \quad u(l,t)=\mu_2, \quad (6.72)$$

описывающее распределение температуры $u(x,$ t$)$ в области [0,l]. Здесь a^2—-коэффициент теплопроводности, $f(x)$ описывает стационарные источники тепла, $u°(x)$—начальное распределение

температуры. Вводя на отрезке [0,l] сетку с постоянным шагом $h = l/N$ и узлами x_j ($j = 0,1, ,N$) аппроксимируем пространственную производную в (671) конечно-разностной формулой

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x=x_j} = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + O(h^2), \quad u_j = u(x_j, t). \qquad (6.73)$$

Будем полагать, что u_i(j= 1, 2, .., N- 1) — новые неизвестные, зависящие только от t. Тогда задача (6.71), (6.72) сводится к задаче Коши для системы N— 1 линейных уравнений с постоянными коэффициентами:

$$\begin{cases} \dot{u}_j = (a/h)^2 (u_{j+1} - 2u_j + u_{j-1}) + f_j \quad (j = 1, 2, \ldots, N-1), \\ f_j = f(x_j), \quad u_0 = \mu_1, \quad u_N = \mu_2, \quad u_j(0) = u_j^0 = u^0(x_j). \end{cases} \qquad (6.74)$$

Такой способ решения уравнений в частных производных называется методом прямых, так как решение $u(x,t)$ отыскиваегся в области (x,t), вдоль прямых $x = x_j$, параллельных оси t. Для решения задачи Коши (6.74) можно, например, воспользоваться методом Пикара (6.14). Поскольку в данном случае входящая в (6 14) матрица A редкая, т.е. имеет мало ненулевых элементов, решение будет достаточно быстрым.

Наконец, если нас интересует лишь установившееся распределение температуры при $t \to \infty$, то полагая в (6.71) $\partial u/\partial t = 0$, получаем линейную краевую задачу

$$a^2 \frac{d^2 u}{dx^2} = -f(x), \quad u(0) = \mu_1, \quad u(l) = \mu_2. \qquad (6.75)$$

(В (6.75) вместо частной производной $\partial^2 u/\partial x^2$, записана полная производная d^2u/dx^2 , так как установившаяся температура уже не зависит от t.) Краевая задача (6.75) может быть решена методом прогонки (§ 6.4).

6.6 Численные методы решения жёстких систем обыкновенных дифференциальных уравнений

Рассмотрим задачу Коши для однородной системы n линейных обыкновенных дифференциальных уравнений

$$\frac{du}{dx} = Au \quad (0 \le x \le l) \quad (6.76)$$

$$u(0) = u_0 \ , \quad (6.77)$$

где u – n-мерный вектор, $A = (a_{ij})$ - матрица размера $n \times n$.

Решение задачи имеет вид:

$$u = \sum_{k=1}^{n} c_k e^{\lambda_k x} u^{(k)}, \qquad (6.78)$$

где λ_k, $u^{(k)}$ $(k = 1, 2, \ldots, n)$ -собственные значения и собственные векторы матрицы A (в общем случае комплексные), а коэффициенты c_k находятся из начальных условий (6.77).

Будем полагать, что все действительные части собственных значений $\mathrm{Re}\,\lambda_k < 0$, т.е. решение устойчиво и все компоненты суммы (6.78) убывают с ростом x.

В физических задачах (например, в физической кинетике) встречаются случаи, когда $\mathrm{Re}\,\lambda_k$ сильно отличаются (иногда в отношении до 10^{10}), что сильно затрудняет численное решение системы дифференциальных уравнений. Систему обыкновенных дифференциальных уравнений вида (6.76) с постоянной матрицей A называют жесткой, если:

1) $\mathrm{Re}\,\lambda_k < 0 \qquad (k = 1, 2, \ldots, n),$ (6.79)

2) $S = \left| \max_k (-\mathrm{Re}\,\lambda_k) / \min_k (-\mathrm{Re}\,\lambda_k) \right| >> 1$ (6.80)

Число S-жесткость (stiffness) системы (6.76). Если сама матрица A зависит от x, то ее собственные значения и жесткость также являются функциями x, т.е.

$\lambda_k = \lambda_k(x) \qquad (k = 1, 2, \ldots, n),$ (6.81)

$S = S(x).$ (6.82)

Тогда система уравнений считается жесткой, если хотя бы в одной точке $x \in [0, l]$ выполняются условия (6.79), (6.80).

Численно решать жесткую систему с помощью явных методов практически невозможно, потому что шаг решения h приходится выбирать из условия устойчивости, учитывая самую быстро меняющуюся компоненту решения:

$$h < 2 / \max_k \mathrm{Re}(-\lambda_k), \qquad (6.83)$$

а интервал интегрирования

$$l = O\left(1/\min_k \mathrm{Re}(-\lambda_k)\right). \qquad (6.84)$$

И хотя компоненты решения с большими значениями $\mathrm{Re}(-\lambda_k)$ быстро затухают почти до 0, условие устойчивости (6.83) не позволяет увеличить шаг. Поэтому численные методы решения жестких систем обыкновенных дифференциальных уравнений должны обязательно обладать следующими свойствами:

1) метод (схема) является неявным;
2) метод должен обеспечивать автоматический выбор шага *h* в зависимости от скорости изменения решения *u(x)*

Свойство 1 необходимо для сохранения устойчивости решения, а свойство 2 требуется потому, что в процессе интегрирования системы шаг *h* может возрастать в $10^6 - 10^7$ раз и более.

Выше рассматривались линейные системы, но значительно чаще встречаются жесткие нелинейные системы обыкновенных дифференциальных уравнений, когда матрица *A* зависит не только от *x*, но и от решения *u(x)*. В этих случаях рекомендуется приближенно линеаризовать систему, рассматривая малые отклонения *z=u-v* от некоторого вспомогательного решения *v(x)*, которое можно выразить через элементарные функции. Построение *v(x)* может быть сопряжено с необходимостью многократного решения полной проблемы собственных значений для матрицы *A*. Определение жесткости (6.79), (6.80) сохраняется. Целесообразно также ввести термин “жесткая устойчивость”, понимая под этим, что для “паразитных” компонент решения с большими $-\mathrm{Re}\,\lambda_k$ требуется устойчивость, а для главных компонент (с малыми $-\mathrm{Re}\,\lambda_k$) гарантируется точность.

В целом теория и численные методы решения жестких систем обыкновенных дифференциальных уравнений являются довольно сложными, но в мощных современных пакетах прикладных программ обычно имеется несколько хороших программ их численного решения.

6.7. Численное решение обыкновенных дифференциальных уравнений, не приведенных к нормальному виду

Пусть имеется задача Коши для обыкновенного дифференциального уравнения в общем виде

$$F(x, y, y') = 0, \quad y(x_0) = y_0, \qquad (6.85)$$

не разрешенном относительно первой производной y'. Для решения задачи (6.85) можно формально применять одну из одношаговых схем, рассмотренных в 6.1.

Пусть, например, на *j*-ом шаге известны x_j *и* y_j (в начале решения это x_0, y_0), и будем полагать, что

$$y' = f(x, y), \qquad (6.86)$$

где *f* – некоторая неявно определенная функция. Численно решая уравнение

$$F(x_j, y_j, f_j) = 0, \qquad (6.87)$$

где $f_j = f(x_j, y_j)$, находим f_j, а затем по формуле (6.20) – предсказанное значение $\tilde{y}_{j+1/2}$. Затем из уравнения

$$F\left(x_j + h_j/2,\ \tilde{y}_{j+1/2},\ f_{j+1/2}\right) = 0 \qquad (6.88)$$

вычисляем $f_{j+1/2}$ и, наконец, по формуле (6.21) получаем исправленное значение y_{j+1} на следующем шаге. При численном решении уравнений вида (6.87), (6.88) хорошее начальное приближение для f_j, $f_{j+1/2}$ можно получить, экстраполируя ранее найденные значения f_{j-1}, f_{j-2}, $f_{j-1/2}$,…, так что число итераций будет небольшим. Аналогично можно пользоваться и другими формулами одношаговых методов.

Другим подходом может служить метод степенных рядов. В этом методе на *j*-ом шаге сначала из уравнения (6.87) вычисляется f_j. Дифференцируя затем (6.85) получаем

$$\frac{dF}{dx} = F_{x,j} + F_{y,j} f_j + F_{f,j} y''_j = 0, \qquad (6.89)$$

откуда явно определяется y''_j. Если взять затем вторую производную функции F (6.85), то из уравнения

$$\frac{d^2F}{dx^2} = 0 \qquad (6.90)$$

можно получить y'''_j и т.д., а затем по формуле метода степенных рядов вида (6.5) найти значение y_{j+1} на следующем шаге.

6.8. Численное решение обыкновенных дифференциальных уравнений второго порядка

В задачах механики часто встречается решение

обыкновенных дифференциальных уравнений вида

$$y'' = f(x, y). \qquad (6.91)$$

Хотя это уравнение всегда можно свести к системе двух уравнений первого порядка, иногда предпочтительнее воспользоваться многошаговым методом Штёрмера. Явный метод Штёрмера представляется в виде

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = \sum_{i=0}^{k} b_i f(x_{j-i}, y_{j-i}), \qquad (6.92)$$

где h – шаг решения, а коэффициенты b_i подбираются так, чтобы разность

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - \sum_{i=0}^{k} b_i f(x_{j-i}, y_{j-i}) \qquad (6.93)$$

как можно меньше отличалась от 0. Если суммирование в формуле (6.92) начать с $i = -1$, получится неявный метод Штёрмера:

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} = \sum_{i=-1}^{k} b_i f(x_{j-i}, y_{y-i}) \qquad (6.94)$$

Коэффициенты b_i даются в некоторых справочниках и учебных пособиях.

Недостаток методов Штёрмера, как и других многошаговых методов – необходимость в разгонных значениях $y_1, y_2, \ldots, y_k$ для проведения расчетов по формуле (6.93) начиная со значения $j=k$. Разгонные значения должны быть предварительно найдены каким-либо одношаговым методом.

7. ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ

Многие сложные физические процессы гидродинамики, теории поля, электродинамики, физической кинетики, статистической физики, квантовой механики и других разделов физики описываются уравнениями в частных производных. Методы численного решения этих уравнений составляют наиболее сложную и, пожалуй, наиболее важную область вычислительной математики, опирающуюся на методы из всех остальных областей. Вследствие большого разнообразия задач и численных методов их решения, существующее сейчас в этой области стандартное программное обеспечение недостаточно, и требующиеся программы обычно приходится разрабатывать самостоятельно применительно к конкретным задачам.

В этой главе рассматривается общий подход к численному решению уравнений в частных производных, и на его основе даются методы решения некоторых простейших физических задач.

7.1. Разностные схемы

При решении многих задач встречаются два типа функций. Функции первого типа определяются на континууме точек **и** считаются точно известными (аналитическими), тогда как функции второго типа определены на дискретном множестве точек, называются сеточными и представляют собой аппроксимации функций первого типа. Так, например, элементарные функции *lgx*, e^x, *sinx* и другие являются точными, а хорошо известные таблицы этих функций — их сеточными аппроксимациями. Чтобы подчеркнуть это различие, в данном параграфе точные функции будем обозначать прописными буквами, а приближенные сеточные — соответствующими строчными.

Будем рассматривать дифференциальные задачи вида

$$L\,U = F \qquad (7.1)$$

в некоторой области пространства G с границей Г. В уравнении (7.1) L — дифференциальный оператор, U — подлежащая вычислению функция, F — правая часть (иногда F включают в оператор L). В общем случае U зависит от координат $\vec{r}$ и времени t, т.е. $U = U(\vec{r}, t)$.

Уравнение (7.1) может иметь множество решений. Для выделения из этого множества единственного решения, уравнение (7.1) дополняется краевым и начальным условиями:

$$B\,U\big|_{\Gamma} = \Psi(\vec{r}, t); \qquad (7.2)$$

$$U(\vec{r}, 0) = \Phi(\vec{r}). \qquad (7.3)$$

Оператор B в краевом условии (7.2) в общем случае также может быть дифференциальным, например, **в** него может входить производная $\partial U/\partial n$ по нормали к границе Г**.** Вместо краевого условия (7.2) на некоторых участках границы Г может быть задано условие периодичности решения U по одной или нескольким координатам. Если в оператор L входит вторая производная по времени, то, помимо начального условия (7.3), должно быть задано начальное условие для первой производной по времени

$$U_t(\vec{r}, 0) = \Theta(\vec{r}). \qquad (7.4)$$

Уравнения (7.1)—(7.4) составляют так называемую смешанную краевую задачу.

При исследовании статических (стационарных, установившихся) явлений функция U зависит лишь от координат, и дополнительно для уравнения (7.1) задается лишь краевое условие

$$B\,U\big|_{\Gamma} = \Psi(\vec{r}). \qquad (7.5)$$

Задача (7.1), (7.5) называется краевой. В дальнейшем при решении задач ограничимся краевыми условиями первого рода, когда B — единичный оператор, т.е. функция U принимает на границе Г заданные значения Ψ.

Будем полагать, что решение всех рассматриваемых задач существует и единственно, так что нашей целью является лишь нахождение этого решения численными методами. В области G введем сетку Ω; множество внутренних узлов сетки обозначим ω, а множество граничных узлов — γ Для простоты будем полагать, что граничные узлы сетки располагаются точно на границе (рис. 23) Это снимает трудности переноса краевых значений с Г на сеточную границу γ. Будем использовать только равномерные сетки, т. е. сетки с постоянными шагами по каждой координате

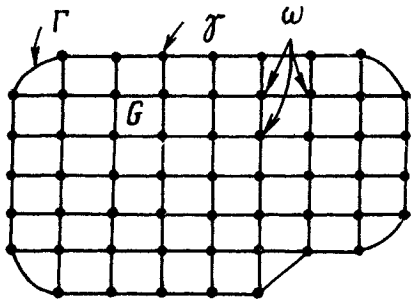


Рис 23 Сетка Ω в области G с границей Г

(сетки с переменными шагами называют неравномерными). Пространственный шаг для краткости обозначим h (можно представлять h как вектор, компоненты которого—шаги по от дельным координатам). По времени t также введем дискретизацию с шагом τ. Все встречающиеся в задаче функции, как заданные, так и искомые, будем приближенно определять в узлах введенной пространственно-временной сетки, называть сеточными и, как уже говорилось, обозначать строчными буквами в отличие от аналитических функций и точного решения, обозначаемых соответствующими прописными буквами.

Дифференциальную задачу аппроксимируем системой конечно-разностных уравнений. Получаемые в результате конечно-разностные уравнения называют разностной схемой. Имеется несколько методов составления разностных схем. Самым простым и распространенным является метод разностной аппроксимации, основанный на замене дифференциальных выражений в операторе L (а если требуется, то и в B) конечно-разностными отношениями. В дальнейшем будем пользоваться этим методом. Полученную разностную схему в общем случае запишем в виде:

$$L_h u = f; \quad (7.6)$$

$$B_h u|_\gamma = \psi; \quad (7.7)$$

$$u|_{t=0} = \varphi, \quad u_t|_{t=0} = \theta. \quad (7.8)$$

Здесь L_h,,B_h — разностные операторы, u—решение разностных уравнений, а функции f, ψ,...—приближенные значения функций F, Ψ, .. на сетке, иногда называемые проекциями F, Ψ, …на сетку. Предположим, что известно точное решение U дифференциальной задачи (7.1) — (7.4). Если подставить это решение в разностную схему (7.6), получится некоторая, отличная от нуля величина ξ, которая называется невязкой и используется для оценки качества разностных схем:

$$\xi = f - L_h U = L U - F - (L_h U - f) = (L - L_h) U - (F - f). \quad (7.9)$$

Аналогично для краевых условий

$$\eta = \psi - B_h U = B U - \Psi - (B_h U - \psi) = (B - B_h) U - (\Psi - \psi). \quad (7.10)$$

Из (7.9), (7.10) видно, что если $F = f$, $\Psi = \psi$, т. е. F и Ψ проецируются на сетку точно, то невязка равна результату действия разности дифференциального и конечно-разностного операторов на точное решение U.

Основными свойствами разностных схем являются аппроксимация, погрешность, устойчивость, сходимость и эффективность. Под свойством аппроксимации понимается, что невязка стремится к нулю с уменьшением шагов сетки h, τ, т.е. согласно (7.9), (7 10) смысл аппроксимации состоит в том, что с уменьшением шагов сетки конечно-разностные уравнения приближаются к дифференциальным:

$$||\xi|| \to 0, \quad L_h \to L, \quad f \to F, \quad ||\eta|| \to 0, \quad B_h \to B, \quad \psi \to \Psi \quad \text{при} \quad h, \tau \to 0. \quad (7.11)$$

Количественной характеристикой аппроксимации служит погрешность аппроксимации, которая представляет собой меру отклонения L_h, B_h, f, ψ соответственно от L, B, F, Ψ и оценивается по величине нормы $||\xi|| + ||\eta||$ в зависимости от шагов h, τ. Обычно погрешность аппроксимации указывается в виде $O(h^k, \tau^s)$ или $O(h^k + \tau^s)$, где k, s — порядок аппроксимации. Можно показать, что аппроксимация имеется, если аппроксимированы производные в дифференциальных операторах, причем порядок аппроксимации разностной схемы обычно совпадает с порядком аппроксимации производных, а в некоторых случаях может быть и выше, как будет показано далее.

Пусть p — некоторые исходные данные разностной схемы, например, значения функций f, ψ,... или коэффициентов оператора

L_k, причем значению p, отвечает решение u_i. Говорят, что разностная схема устойчива, если малые изменения исходных данных приводят к малым изменениям решения, т.е., если $\|p_1 - p_2\| < \varepsilon$, то и $\|u_1 - u_2\| < \delta(\varepsilon)$, где ε, δ — некоторые малые положительные величины. Анализ устойчивости разностных схем основан на использовании специальных критериев; некоторые критерии устойчивости разностных схем будут рассмотрены в дальнейшем.

Пусть (7.1) — (7.4) — корректно поставленная дифференциальная задача. Аппроксимирующая ее разностная схема (7.6) — (7.8) называется корректной, если ее решение существует, единственно и устойчиво. Говорят, что разностная схема сходится (к точному решению), если $u \to U$ при $h, \tau \to 0$. Если погрешность решения

$$||u - U|| \leqslant M_1 h^k + M_2 \tau^s = O(h^k, \tau^s), \quad (7.12)$$

где M_1, M_2, k, s — постоянные, не зависящие от h, τ, то говорят, что разностная схема сходится с k-м порядком по h и s-м порядком по τ. Доказано, что если разностная схема аппроксимирует дифференциальную задачу с порядком $O(h^k, \tau^s)$ и устойчива, то она сходится к точному решению с тем же порядком. Следовательно, свойство сходимости является следствием аппроксимации и устойчивости и не требует специальных доказательств.

В данных выше формулировках аппроксимации, устойчивости, сходимости используются нормы сеточных функций. В вычислительной математике пользуются теми нормами, которые позволяют легче доказать требуемые свойства. Некоторые результаты удается получить с использованием сеточного аналога нормы C (чебышевской нормы), но в большинстве случаев приходится пользоваться сеточным аналогом более слабой среднеквадратичной нормы L_2 (гильбертовой). При этом сеточные нормы вводятся так, чтобы они были согласованы с обычными нормами, т. е. переходили в них при бесконечном уменьшении шагов сетки. Например, для одномерной краевой задачи

$$||u||_C = \max_j |u_j|, \quad ||u||_{L_2} = \left(h \sum_j u_j^2 \right)^{1/2}, \quad (7.13)$$

максимум и сумма находятся по всем узлам сетки с шагом h и номерами j.

Наконец, эффективность разностных схем характеризуется запросами на машинные ресурсы. Важной характеристикой, в частности, является зависимость времени счета от числа узлов пространственной сетки.

7.2. Численное решение одномерного уравнения переноса

Далее в этой главе, если это не вызовет недоразумений, все встречающиеся в задачах функции будем обозначать строчными буквами, но по-прежнему имея в виду, что в дифференциальные уравнения и их решения входят точные, аналитические функции, а в разностные схемы и их решения — приближенные сеточные.

В физических приложениях часто встречаются уравнения эволюционного типа. Уравнениями эволюционного типа называются уравнения вида

$$\frac{\partial u}{\partial t} + L_0 u = f, \quad (7.14)$$

где дифференциальный оператор L_0 уже не содержит производных по времени t; такие уравнения как бы описывают эволюцию начальных условий. Если в уравнении (7.14) L_0 — оператор первого порядка, то оно называется уравнением переноса и типично для задач механики сплошной среды. Примером уравнения переноса может служить известное уравнение непрерывности (закон сохранения заряда)

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \vec{j} = 0, \quad \vec{j} = \rho \vec{v}, \quad (7.15)$$

где ρ —плотность объемного заряда, $\vec{j}$ —плотность тока, $\vec{v}$ — скорость движения заряженной среды.

Простейшая линейная одномерная задача для уравнения переноса:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = f(x, t), \quad c = \text{const} \quad (0 < x < l);$$

$$u(0, t) = u_0 = \text{const}, \quad u(x, 0) = \varphi(x) \quad (7.16)$$

описывает движение сплошной среды с плотностью u вдоль оси x со скоростью c; правая часть f описывает приток (отток) вещества вследствие каких-либо сторонних процессов. Будем также для простоты полагать, что $\varphi(0) = u_0$.

Задача (7.16) имеет аналитическое решение

$$u(x, t) = \varphi(x - c t) + \int_0^t f(x - c(t - t'), t')\, dt', \quad (7.17)$$

где t' — переменная интегрирования. В частности, если $f = 0$, то

$$u(x, t) = \varphi(x - c t). \quad (7.18)$$

Видно, что в этом случае решение уравнения, зависящего от двух переменных, представляется функцией всего одного переменного $z = x - ct$. Такие решения уравнений, зависящие от меньшего числа переменных, называются автомодельными. В рассматриваемой задаче автомодельное решение (7.18) описывает перенос начального распределения $\varphi(x)$ вдоль оси x со скоростью c без изменения его формы. (Этим и объясняется его название — решение как бы воспроизводит, моделирует само себя.)

Для получения численного решения введем равномерную сетку (рис.24) с шагами h, τ и узлами (x_j, t^m), где

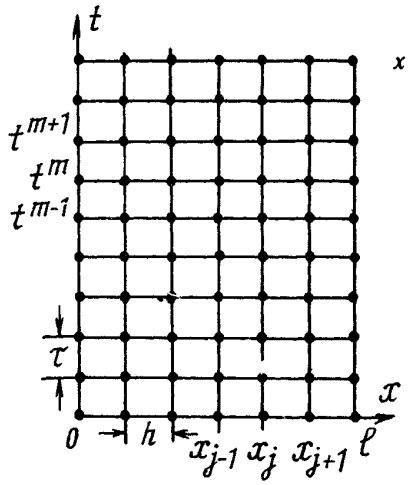


Рис.24 Равномерная пространственно-временная сетка

$$x_j = jh \ (j=0,1,...,N), \quad h=l/N, \qquad (7.19)$$

$$t^m = m\tau \ (m=0,1,2,...).$$

Значения сеточных функций принято обозначать как

$$u_j^m = u(x_j, t^m), \qquad (7.20)$$

т. е. нижние индексы относятся к пространственным переменным, а верхний указывает номер момента времени. Для краткости принято в разностных схемах верхний индекс, по возможности, опускать, используя Обозначения:

$$\begin{cases} u_j = u(x_j, t), \quad \hat{u}_j = u(x_j, t+\tau), \quad \check{u}_j = u(x_j, t-\tau); \\ u_j^{\pm} = u(x_j, t \pm \tau/2). \end{cases} \qquad (7.21)$$

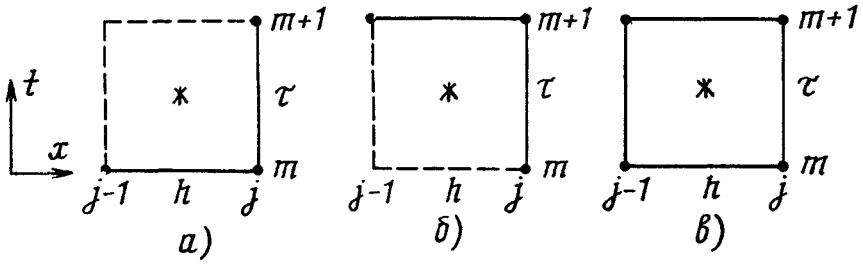


Рис 25 Расположение сеточных узлов разностных схем для одномерного уравнения переноса: а — явная схема; б — неявная схема, в — схема второго порядка

Расположение узлов, на котором строится разностная схема, называют шаблоном. Рассмотрим три шаблона, показанные на рис.25; кружками отмечены узлы шаблона, а звездочкой — центр соответствующей сеточной ячейки, ограниченной штриховыми линиями. Соответствующие трем шаблонам разностные схемы имеют вид:

$$\frac{\hat{u}_j - u_j}{\tau} + c\frac{u_j - u_{j-1}}{h} = f_j \quad (j=1,2,...,N); \qquad (7.22)$$

$$\frac{\hat{u}_j - u_j}{\tau} + c\frac{\hat{u}_j - \hat{u}_{j-1}}{h} = f_j \quad (j=1,2,...,N); \qquad (7.23)$$

$$\frac{1}{2}\left(\frac{\hat{u}_j - u_j}{\tau} + \frac{\hat{u}_{j-1} - u_{j-1}}{\tau}\right) + \frac{c}{2}\left(\frac{\hat{u}_j - \hat{u}_{j-1}}{h} + \frac{u_j - u_{j-1}}{h}\right) = f_j \qquad (7.24)$$

$$(j=1,2,...,N).$$

Рассмотрим теперь, что принять в (7.22) — (7.24) в качестве сеточной проекции f_j, правой части f. Поскольку f зависит от x, t, наименьшая погрешность аппроксимации получается при замене $f(x,t)$ на среднее значение $\bar{f}_j$, по площади сеточной ячейки, т.е. на значение

$$\bar{f}_j = \frac{1}{h\tau}\int_{x_{j-1}}^{x_j} dx \int_{t}^{t+\tau} f(x,t)\,dt = f(x_j - h/2, t+\tau/2) + O(h^2, \tau^2). \qquad (7.25)$$

Следовательно, в уравнениях (7.22) — (7.24) в качестве f_j будем принимать $f(x_j - h/2, t+\tau/2) = f_{j-1/2}^{+}$, т.е. правая часть в разностных схемах вычисляется в отмеченном звездочкой центре сеточной ячейки на рис. 25. Заметим, что если в качестве f_j просто принять значение в j-м узле, т.е. $f(x_j,t)$, то погрешность аппроксимации $\bar{f}_j$ сразу возрастает на порядок, поскольку $\bar{f}_j = f(x_j,t) + O(h,t)$.

Множество сеточных узлов, отвечающих одному моменту времени, составляют так называемый временной слой. Схемы (7.22) — (7.24) связывают два соседних временных слоя и называются поэтому двухслойными. Используя разложения сеточных функций в ряды Тейлора нетрудно показать, что трехузловые схемы (7.22), (7.23) имеют аппроксимацию $O(h,\tau)$, а более сложная чегырехузловая схема (7.24) — аппроксимацию более высокого порядка $O(h^2, \tau^2)$.

Разностная схема (7.22) явная, так как из нее можно явно выразить неизвестное $\hat{u}_j$, через уже известные значения u_j, u_{j-1} и правую часть f_j. Схема (7.23) формально неявная, так как включает два неизвестных значения $\hat{u}_j$, $\hat{u}_{j-1}$, на $(m+1)$ временном слое. Однако преобразовав ее к виду

$$\hat{u}_j = \frac{1}{1+r}(u_j + r\hat{u}_{j-1} + \tau f_j) \quad (j=1,2,...,N), \quad r = c\tau/h, \qquad (7.26)$$

можно явно вычислять $\hat{u}_j$ последовательно, начиная с левой границы, на которой $\hat{u}_0$ известно (это заданное краевое условие). Такие схемы называют схемами «бегущего счета». Неявная схема (7.24) также легко преобразуется в схему бегущего счета. Для анализа устойчивости разностных схем воспользуемся критерием устойчивости, который называется «принцип максимума» и применим к любым двухслойным разностным схемам. Запишем двухслойную разностную схему в общем виде:

$$\sum_i \alpha_i \hat{u}_{j+i} = \sum_k \beta_k u_{j+k} + f_j, \quad (7.27)$$

где j — центральный узел шаблона, причем предварительно узлы перенумеруем так, что

$$|\alpha_0| = \max_i |\alpha_i|. \quad (7.28)$$

Принцип максимума гласит, что для устойчивости разностной схемы достаточно выполнения условия

$$|\alpha_0| \geqslant \sum_{i \neq 0} |\alpha_i| + \sum_k |\beta_k|. \quad (7.29)$$

Для схемы (7.22)

$$\alpha_0 = 1/\tau, \quad \beta_0 = (1-r)/\tau, \quad \beta_{-1} = r/\tau \quad (r = c\tau/h), \quad (7.30)$$

и согласно принципу максимума (7.29) схема устойчива при выполнении условия

$$0 \leqslant r \leqslant 1 \quad (r = c\tau/h), \quad (7.31)$$

т.е. это условно устойчивая схема. Условие устойчивости (7.31) имеет простой физический смысл: вещество среды распространяется в положительном направлении оси абсцисс, причем расстояние $c\tau$, на которое смещается вещество за один временной шаг τ, не должно превышать шаг сетки h.

Сформулированный таким образом критерий (7.31) иногда называют критерием Куранта, а число $r = c\tau/h$ — числом Куранта.

Для неявной разностной схемы (7.23)

$$\alpha_0 = (1+r)/\tau, \quad \alpha_{-1} = -r/\tau, \quad \beta_0 = 1/\tau, \quad (7.32)$$

так что, согласно принципу максимума (7.29), схема устойчива для всех $c > 0$. Следовательно, схема (7.23) безусловно устойчивая. Безусловно устойчивой является и неявная разностная схема второго порядка (7.24).

Анализ разностных схем в целом показывает, что все явные разностные схемы условно устойчивы, тогда как среди неявных схем существуют безусловно устойчивые. Безусловная устойчивость— очень полезное качество разностной схемы, так как позволяет независимо подбирать временной и пространственный шаги сетки, руководствуясь лишь требуемой погрешностью решения. Благодаря этому, можно, например, выбрать мелкий шаг h и крупный шаг τ при исследовании мелкомасштабных (быстро меняющихся по x), но медленно меняющихся со временем процессов переноса; тем самым достигается высокая скорость вычислений.

Из проведенного анализа следует, что разностная схема (7.22) сходится при $0 < r < 1$ с порядком $O(h, \tau)$, схема (7.23) сходится при любых $c > 0$ с порядком $O(h, \tau)$, а схема (7.24) также сходится при любых $c > 0$ с порядком $O(h^2, \tau^2)$ Явная схема (7.22) требует для решения

$$O(MN) = O(1/(h\tau)) = O(h^{-2})$$

арифметических операций, где M — число временных шагов τ, а неявные схемы (7.23), (7.24) — $O(MN) = O(1/(h\tau))$ операций.

7.3. Численное решение одномерного уравнения теплопроводности

Распространение тепла в одномерной области (например, стержне) описывается уравнением эволюционного типа

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x,t) \quad (0 < x < l), \quad a^2 = \text{const}. \quad (7.33)$$

Здесь u — температура, a^2— коэффициент теплопроводности, функция $f(x,t)$ описывает внутренние источники тепла. Уравнением (7.33) описывается также процесс диффузии газа; при этом u — плотность газа, $a^2 = D$—коэффициент диффузии, $f(x,t)$ внутренние источники газа (выделение и поглощение). Уравнение (7.33) дополним начальным и краевыми условиями

$$u(x,0) = \varphi(x), \quad u(0,t) = \psi_1(t), \quad u(l,t) = \psi_2(t). \quad (7.34)$$

Заметим, что для существования и единственности решения уравнения (7.33) не требуется согласования начального и краевых условий, т.е. не требуется выполнения равенств

$\varphi(0) = \psi_1(0)$, $\varphi(l) = \psi_2(0)$, иначе говоря, допускаются разрывы

$u(x, 0)$ на границах в начальный момент $t=0$.

Выбрав на сетке (7.19) чегырехузловой шаблон (рис. 26, *а*), построим двухслойную явную разностную схему:

$$\frac{u_j^{m+1}-u_j^m}{\tau}=a^2\frac{u_{j+1}^m-2u_j^m+u_{j-1}^m}{h^2}+f_j^{m+1/2} \quad (j=1,2,\ldots,N-1),$$

$$f_j^{m+1/2}=f\left(x_j, t^m+\frac{\tau}{2}\right); \quad (7.35)$$

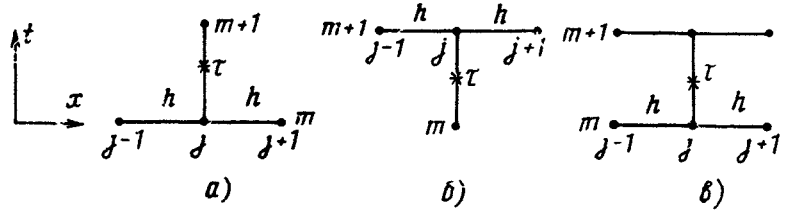


Рис. 26. Расположение узлов разностных схем для одномерного уравнения теплопроводности: *а* — явная схема; *б* — неявная схема; *в* — схема «с полусуммой>

$$u_j^0=\varphi_j \ (j=1,2,\ldots,N-1); \ u_0^m=\psi_1^m, \ u_N^m=\psi_2^m \ (m=0,1,\ldots). \quad (7.36)$$

Здесь, как и в предыдущем параграфе, правая часть f для уменьшения погрешности берется на полуцелом временном шаге в момент $t^m+\tau/2$, отмеченный звездочкой на рис.26, а. Схему (7.35), (7.36) можно преобразовать к удобному для расчетов виду:

$$\hat{u}_j=(1-2\sigma)u_j+\sigma(u_{j+1}+u_{j-1})+\tau f_j^+ \quad (j=1,2,\ldots,N-1), \quad (7.37)$$

$$u_{0=}\psi_1, u_N=\psi_2, \sigma=\tau a^2/h^2 \quad (7.38)$$

Схема (7.35) имеет в общем случае погрешность аппроксимации $O(\tau, h^2)$. Однако, если при $f=0$ выбрать такое отношение шагов τ/h^2, что $\sigma=1/6$, то погрешности конечно-разностных производных по x и t взаимно компенсируются и порядок аппроксимации возрастает до $O(h^4)$. Согласно принципу максимума (7.29), явная схема (7.37) устойчива при ограничении на шаг

$$0\leqslant\sigma\leqslant 0{,}5, \quad 0\leqslant\tau\leqslant 0{,}5h^2/a^2. \quad (7.39)$$

Следовательно, если для увеличения пространственного разрешения потребуется уменьшить шаг h на порядок, то шаг τ для сохранения устойчивости придется уменьшить на два порядка, а число арифметических операций и длительность счета, составляющие $O(1/(\tau h)) = O(h^{-3}) = O(N^3)$, возрастут на три порядка. Поэтому, несмотря на простоту вычислений, явными схемами стараются не пользоваться в задачах, в которых требуется высокое пространственное разрешение, т.е. малый шаг h.

Рассмотрим теперь двухслойную неявную разностную схему, построенную на шаблоне (рис. 26, б)

$$\frac{\hat{u}_j-u_j}{\tau}=a^2\frac{\hat{u}_{j+1}-2\hat{u}_j+\hat{u}_{j-1}}{h^2}+f_j^+ \quad (j=1,2,\ldots,N-1), \quad (7.40)$$

которая отличается от схемы (7.35) лишь тем, что вторая производная по x взята не в момент t, а в момент $t+\tau$. Погрешности аппроксимации схем (7.35) и (7.40) одинаковы. Уравнение (740) преобразуется к виду:

$$\sigma\hat{u}_{j+1}-(1+2\sigma)\hat{u}_j+\sigma\hat{u}_{j-1}=-u_j-\tau f_j^+ \quad (7.41)$$

$$(j=1,2,\ldots,N-1), \quad \sigma=\tau a^2/h^2, \quad \hat{u}_0=\hat{\psi}_1, \quad \hat{u}_N=\hat{\psi}_2, \quad (7.42)$$

и затем решается методом прогонки. В результате получаем решение $\hat{u}_j$ (j= 1,2,..., N—1) на следующем временном слое в момент $t+\tau$ и т.д. Пользуясь принципом максимума (7.29), нетрудно установить, что неявная разностная схема (7.41) безусловно устойчива; условие устойчивости выполняется, конечно, и для прогонки (§ 6.4).

Недостаток рассмотренных разностных схем (7.35) и (7.40) — низкий, т.е. первый порядок погрешности по τ. Рассмотрим теперь полусумму уравнений (7.35) и (7.40), которую называют разностной схемой «с полусуммой» или схемой Кранка — Николсона:

$$\frac{\hat{u}_j-u_j}{\tau}=\frac{a^2}{2}\left(\frac{\hat{u}_{j+1}-2\hat{u}_j+\hat{u}_{j-1}}{h^2}+\frac{u_{j+1}-2u_j+u_{j-1}}{h^2}\right)+f_j^+. \quad (7.43)$$

Можно показать, что неявная разностная схема (7.43), построенная на шестиузловом шаблоне (рис. 26, *в*), имеет бол ее высокий порядок погрешности аппроксимации $O(\tau^2, h^2)$. Уравнение (7.43) преобразуется к виду:

$$\hat{u}_{j+1}-2(1+\sigma^{-1})\hat{u}_j+\hat{u}_{j-1}=g_j \quad (j=1,2,\ldots,N-1); \quad (7.44)$$

$$g_j=-(2\tau/\sigma)f_j^+-[u_{j+1}+2(\sigma^{-1}-1)u_j+u_{j-1}], \quad (7.45)$$

и затем также решается методом прогонки на каждом временном шаге.

Из принципа максимума можно заключить, что разностная схема (7.43) устойчива при условии $\sigma \le 1$, т.е. $\tau < h^2/a^2$. Однако принцип максимума является лишь достаточным критерием, но не является необходимым. Поэтому воспользуемся для анализа устойчивости схемы более сильным, необходимым и достаточным критерием, который называется спектральным или критерием Неймана. Для этого запишем уравнение для погрешности вычислений ε:

$$\frac{\hat{\varepsilon}_j-\varepsilon_j}{\tau}=\frac{a^2}{2}\left(\frac{\hat{\varepsilon}_{j+1}-2\hat{\varepsilon}_j+\hat{\varepsilon}_{j-1}}{h^2}+\frac{\varepsilon_{j+1}-2\varepsilon_j+\varepsilon_{j-1}}{h^2}\right). \quad (7.46)$$

В (7.46), в отличие от (7.43), отсутствует правая часть f, поскольку мы считаем ее известной точно.

Решение уравнения (7.46) будем отыскивать как сумму гармоник вида

$$\varepsilon^m \sim \varepsilon^0 \rho^m e^{-ikx} \quad (i=\sqrt{-1}), \quad (7.47)$$

где m — номер временного шага, ρ — коэффициент роста гармоники на временном шаге, k — волновое число, ε° — начальная амплитуда гармоники в момент $t = 0$. Согласно спектральному критерию, для устойчивости разностной схемы необходимо и достаточно, чтобы выполнялось условие

$$|\rho| \le 1 \quad (7.48)$$

для любых волновых чисел k. Подставляя (7.47) в (7.46) и учитывая, что $\hat{\varepsilon}=\rho\varepsilon$, $x_j=jh$, можно вывести формулу для ρ:

$$\rho=\left(1-2\sigma\sin^2\frac{kh}{2}\right)\left(1+2\sigma\sin^2\frac{kh}{2}\right)^{-1}, \quad \sigma=\frac{\tau a^2}{h^2}, \quad (7.49)$$

из которой видно, что $|\rho| \le 1$ для всех k, т.е. схема "с полусуммой» безусловно устойчива. Поскольку к тому же, эффективность схемы «с полусуммой» не уступает эффективности неявной схемы (7.40), а погрешность ее ниже, она наиболее предпочтительна для решения одномерного уравнения теплопроводности.

7.4. Численное решение двумерного уравнения теплопроводности

Распространение тепла или диффузия газа в двумерной области G описываются уравнением

$$\frac{\partial u}{\partial t}=a^2\Lambda u+f. \quad (7.50)$$

Здесь u, a^2, f обозначают те же величины, что и в предыдущем параграфе. Уравнение (7.50) дополняется краевым и начальным условиями вида (7.2), (7.3). В дальнейшем для простоты будем полагать, что G — прямоугольная область в декартовых координатах:

$$G=\{0 \le x \le l_x, 0 \le y \le l_y\}. \quad (7.51)$$

Соответственно,

$$\Delta u=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}. \quad (7.52)$$

Введем равномерную сетку (рис. 27) с узлами

$$x_j=jh_x, \quad y_i=ih_y, \quad t^m=m\tau, \quad h_x=l_x/N, \quad h_y=l_y/M$$
$$(j=0,1,\ldots,N; \quad i=0,1,\ldots,M; \quad m=0,1,2,\ldots) \quad (7.53)$$

и, аппроксимируя производные в уравнении (7.50) конечно-разностными формулами на шестиузловом шаблоне (рис.28,а), построим явную двухслойную разностную схему

$$\frac{\hat{u}_{ji}-u_{ji}}{\tau}=a^2\left(\frac{u_{j+1,i}-2u_{ji}+u_{j-1,i}}{h_x^2}+\frac{u_{j,i+1}-2u_{ji}+u_{j,i-1}}{h_y^2}\right)+f_{ji}^+ \quad (7.54)$$
$$(j=1,2,\ldots,N-1; \; i=1,2,\ldots,M-1).$$

Значения функции u на границах j=0, N и i=0,M задаются краевыми условиями. Погрешность аппроксимации производных в схеме (7.54) имеет порядок $O(h_x^2, h_y^2, \tau)$. Правая часть

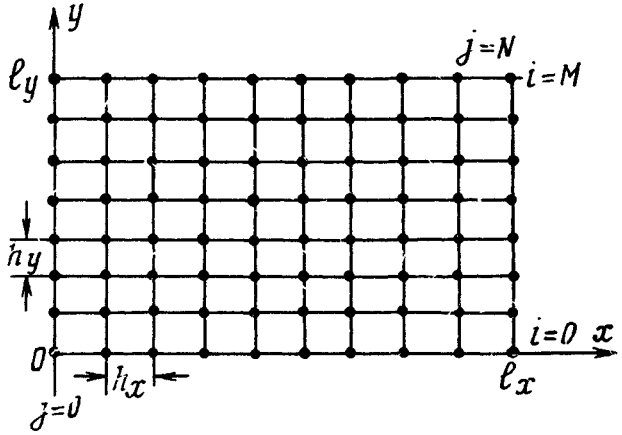


Рис 27 Равномерная пространственная сетка в прямоугольнике

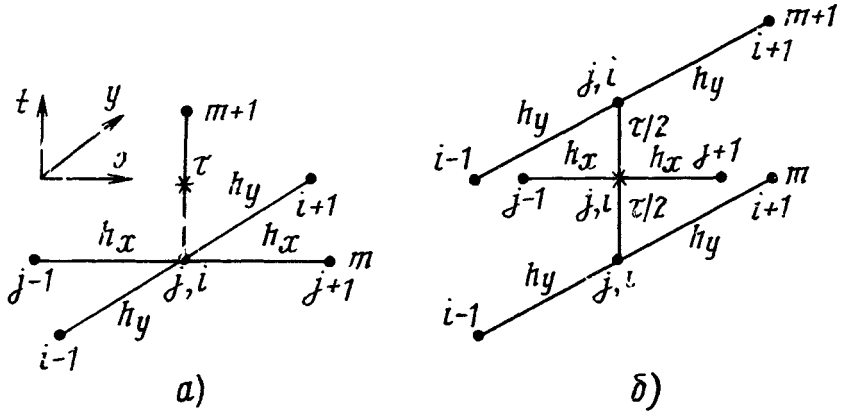


Рис 28 Расположение узлов разностных схем для двумерного уравнения теплопроводности: *а* — явная схема, б — экономичная схема

f_{ii}^{+}, как и в § 7 2, 7.3, вычисляется на полуцелом временном шаге $t+\tau/2$ для уменьшения невязки. Из уравнений (7.54) можно явно вычислять значения $\hat{u}_{ji}$ на последовательности временных слоев в моменты времени $t^m=m\tau$ (m—1,2,...), используя заданные начальные и краевые условия.

Для анализа устойчивости явной разностной схемы (7.54) воспользуемся спектральным критерием. Решение однородного уравнения для вычислительной погрешности

$$\frac{\hat{\varepsilon}_{ji}-\varepsilon_{ji}}{\tau}=a^2\left(\frac{\varepsilon_{j+1,i}-2\varepsilon_{ji}+\varepsilon_{j-1,i}}{h_x^2}+\frac{\varepsilon_{j,i+1}-2\varepsilon_{ji}+\varepsilon_{j,i-1}}{h_y^2}\right) \quad (7.55)$$

будем отыскивать, как и в § 7.3, в виде суммы гармоник

$$\varepsilon\sim\varepsilon^0\rho^m e^{-i(k_x x+k_y y)}, \quad (7.56)$$

аналогичных (7.47); здесь k_x,k_y—компоненты волнового вектора k. Для устойчивости разностной схемы при выбранном шаге τ необходимо и достаточно, чтобы условие (7.48) выполнялось при любых k_x,k_y. Подставляя (7.56) в (7.55) и учитывая (7.52), находим

$$\rho=1-4a^2\tau\left[h_x^{-2}\sin^2(0.5k_x h_x)+h_y^{-2}\sin^2(0.5k_y h_y)\right] \quad (7.57)$$

Когда синусы в *(7.57)* принимают единичные значения, модуль множителя роста достигает максимума

$$\max_{k_x,k_y}|\rho|=4a^2\tau\left(h_x^{-2}+h_y^{-2}\right)-1, \quad (7.58)$$

и, следовательно, согласно спектральному критерию (7.48), разностная схема (7.54) устойчива при условии

$$\tau\leqslant 0{,}5\,a^{-2}\left(h_x^{-2}+h_y^{-2}\right)^{-1}. \quad (7.59)$$

В частном случае, когда область G — квадрат, $N=M$, $h_x=h_y=h$, условие устойчивости (7.59) приобретает вид

$$\tau\leqslant 0{,}25\,(h/a)^2. \quad (7.60)$$

Требующееся для решения число арифметических операций имеет порядок $O(\tau^{-1},h^{-2})$, а с учетом условия устойчивости (7.60) — порядок $O(h^{-4})=O(N^4)$, т.е. очень быстро увеличивается с ростом числа шагов N пространственной сетки. Поэтому простейшие явные схемы редко используются для решения уравнения теплопроводности.

Предположим теперь, что в разностной схеме (7.54) производная по времени в левой части сохраняется в дифференциальной форме, т. е. задача сводится к задаче Коши для системы линейных обыкновенных дифференциальных уравнений

$$\dot{u}_{ji} = a^2\left(\frac{u_{j+1,i}-2u_{ji}+u_{j-1,i}}{h_x^2}+\frac{u_{j,i+1}-2u_{ji}+u_{j,i-1}}{h_y^2}\right)+f_{ji}^{+},$$

$$u_{ji}(0)=u_{ji}^{0} \quad (j=1,2,\ldots,N-1;\ i=1,2,\ldots,M-1) \qquad (7.61)$$

(точка означает дифференцирование по времени t). Задача (7.61) может быть записана в матричном виде

$$\dot{u}=Au+f, \quad u(0)=u^{0}, \quad u=\{u_{ji}\}, \quad f^{+}=\{f_{ji}^{+}\}, \qquad (7.62)$$

где и, f^{+}— векторы значений во внутренних узлах пространственной сетки, A — редкая матрица, связывающая неизвестные u_{ji} в пяти соседних узлах. Далее задача Коши (7.62) решается одним из методов, описанных в предыдущем разделе, например, методом последовательных приближений Пикара.

Такой метод решения называют методом прямых, так как зависимости $u_{ji}(t)$ отыскиваются вдоль прямых, исходящих из внутренних узлов пространственной сетки параллельно оси t.

Метод безусловно устойчив и имеет погрешность $O(h_x^2 h_y^2 \tau^k)$, где порядок временного шага $k = 2$, если правая часть f зависит от t; если же f не зависит от t, то k определяется выбранным численным методом решения системы (7.62). В частности, при использовании метода Пикара решение по t может быть точным при $k \to \infty$.

Явную разностную схему (7.54) можно также преобразовать в неявную с той же погрешностью аппроксимации $O(\tau, h_x^2, h_y^2)$,

если отнести все узловые значения функции u в правой части к моменту $t+\tau$, т.е. снабдить их значком ^. Однако решение полученной таким образом системы линейных алгебраических уравнений относительно неизвестных $\hat{u}_{ji}$ потребует $O((NM)^3)$ арифметических операций при использовании прямых методов (§ 5.1) или $O(IMN)$ операций при использовании итерационных методов (§ 5.2) с $I \gg 1$ итерациями. В результате решение сильно замедляется и разностная схема становится неэффективной.

Важный класс разностных схем для многомерных уравнений в частных производных составляют так называемые экономичные разностные схемы. Экономичными называются разностные схемы, которые сочетают два лучших свойства неявных и явных схем: во-первых, они обладают безусловной устойчивостью, как некоторые неявные схемы, а во-вторых, требующееся для их решения число арифметических операций пропорционально числу узлов сетки, как у явных схем.

Одной из наиболее известных экономичных разностных схем для двумерного уравнения теплопроводности является продольно-поперечная схема, которая строится на шаблоне рис. 28.6 и складывается из двух временных полушагов:

$$\frac{u_{ji}^{+}-u_{ji}}{\tau/2}=a^2\left(\frac{u_{j+1,i}^{+}-2u_{ji}^{+}+u_{j-1,i}^{+}}{h_x^2}+\frac{u_{j,i+1}-2u_{ji}+u_{j,i-1}}{h_y^2}\right)+f_{ji}^{+}; \qquad (7.63)$$

$$\frac{\hat{u}_{ji}-u_{ji}^{+}}{\tau/2}=a^2\left(\frac{u_{j+1,i}^{+}-2u_{ji}^{+}+u_{j-1,i}^{+}}{h_x^2}+\frac{\hat{u}_{j,i+1}-2\hat{u}_{ji}+\hat{u}_{j,i-1}}{h_y^2}\right)+f_{ji}^{+}; \qquad (7.64)$$

$$(j=1,2,\ldots,N-1;\ i=1,2,\ldots,M-1).$$

Как видно, уравнение (7.63) — неявное по x и явное по y, а уравнение (7.64), наоборот — явное по x и неявное по y. Каждое из уравнений (7.63), (7.64) в отдельности имеет аппроксимацию $O(h_x^2, h_y^2, \tau)$, однако можно показать, что на шаге τ их погрешности частично компенсируются и суммарная погрешность продольно-поперечной схемы составляет $0(h_x^2, h_y^2, \tau^2)$, т.е. это схема второго порядка.

Решение схемы на каждом временном шаге складывается из двух этапов, каждый из которых представляет собой цикл прогонок. На первом этапе из уравнения (7.63) с помощью независимых прогонок по всем продольным линиям сетки с номерами i=1,2, …,M — 1 вычисляются значения u_{ji}^{+} в момент $t + \tau/2$, а на втором этапе по этим значениям из уравнения (7.64) с помощью независимых прогонок по всем поперечным линиям сетки с номерами j= 1,2, …,N—1 вычисляются окончательные значения u_{ji}^{+} *в* момент $t + \tau$. На этом временной шаг завершается.

Анализ устойчивости продольно-поперечной схемы с помощью спектрального критерия показывает, что схема безусловно устойчива и, следовательно, сходится к точному решению со вторым порядком. Название схемы указывает на чередование при решении продольных и поперечных прогонок; иногда рассмотренную схему называют также по фамилиям авторов схемой Писмена — Рекфорда. Отметим еще, что использованное в схеме дробление временного шага (на два полушага) применяется и при решении многих других уравнений в частных производных; все такие методы называют методами дробных шагов.

7.5. Численные решения уравнений эллиптического типа

Типичным уравнением в частных производных эллиптического типа является уравнение вида

$$-\Delta u + \mu u = f. \quad (7.65)$$

При $\mu \neq 0$, $f \neq 0$ уравнение (7.65) представляет собой неоднородное уравнение Гельмгольца, при $\mu \neq 0$, $f=0$ — однородное уравнение Гельмгольца, при $\mu=0$ — уравнение Пуассона

$$\Delta u = -f \quad (7.66)$$

и, наконец, при $\mu=0$ и $f=0$ —уравнение Лапласа. Так, например, однородным уравнением Гельмгольца

$$\Delta \vec{E} + k^2 \vec{E} = 0 \quad (7.67)$$

описывается стационарное распределение в резонаторе электрического поля $\vec{E}$ с волновым числом k; уравнение Пуассона

$$\Delta \varphi = -\rho/\varepsilon \quad (7.68)$$

описывает распределение электрического потенциала φ в среде с плотностью заряда ρ и электрической проницаемостью ε и т. п. Установившееся распределение температуры (или плотности газа) также описывается уравнением Пуассона, которое получается из уравнения теплопроводности (диффузии) вида (7.50) при $\partial u / \partial t = 0$.

Поскольку уравнения эллиптического типа часто встречаются в важных физических приложениях, для их решения разработано много численных методов, которые можно разделить на конечно-разностные и проекционные. (В проекционных методах также часто используется некоторая сетка; такие методы называют проекционно-сеточными.) В дальнейшем в этом параграфе ограничимся так называемой задачей Дирихле, или первой краевой задачей для уравнения Пуассона

$$\Delta u(\vec{r}) = -f(\vec{r}) \ (\vec{r} \in G), \quad u(\vec{r}) = \psi(\vec{r}) \ (\vec{r} \in \Gamma) \quad (7.69)$$

в прямоугольнике G (7.51) с границей Г, используя равномерную сетку (7.53).

Рассмотрим сначала конечно-разностные методы. Аппроксимируем вторые производные в операторе $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$ конечно-разностными формулами на пятиточечном шаблоне «крест» (рис. 29) и построим разностную схему

$$\frac{u_{j+1,i} - 2u_{ji} + u_{j-1,i}}{h^2} + \frac{u_{j,i+1} - 2u_{ji} + u_{j,i-1}}{h^2} = -f_{ji} \quad (7.70)$$

$(j = 1, 2, \ldots, N-1;$

$i = 1, 2, \ldots, M-1).$

Схема (7.70) имеет погрешность аппроксимации $O(h_x^2, h_y^2)$,т. е. это схема второго порядка; значения u_{ji}, при $j = 0, N$ и $i=0, M$ задаются краевыми условиями. Уравнения (7.70). представляют собой систему*(N-1)(M-1)* линейных алгебраических уравнений, которую можно записать в виде

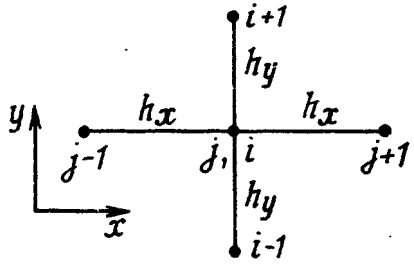


Рис.29. Расположение узлов пространственной сетки для двумерного уравнения эллиптического типа

$$Au = b \quad (7.71)$$

Здесь u — вектор значений u_{ji} во всех внутренних узлах сетки, имеющий $(N-1)(M-1)$ элементов; b — вектор правой части, включающий как значения f_{ji} во внутренних узлах сетки, так и краевые значения; A — симметричная редкая матрица размерности $(N-1)(M-1) \times (N-1)(M-1)$.

Для решения полученной системы (7.71) можно использовать методы, описанные в главе 5, например, метод Зейделя, метод последовательной верхней релаксации, ускорение сходимости с использованием чебышевского набора итерационных параметров и др. Поскольку в некоторых методах используются собственные значения матрицы А, заметим, что простота структуры матрицы A в рассматриваемой задаче (7.70) позволяет найти ее собственные значения аналитически. Отыскивая для этого решение уравнения собственных значений матрицы A

$$Au = \lambda(A) u, \quad (7.72)$$

где $(Au)_{ji}$ представляет собой левую часть уравнения (7.70), в виде

$$u_{ji} \sim \sin \frac{\pi n j}{N} \sin \frac{\pi m i}{M}, \quad (7.73)$$

можно после элементарных преобразований найти

$$\lambda(A) = -4(h_x^{-2}\sin^2\frac{\pi n}{2N} + h_y^{-2}\sin^2\frac{\pi m}{2M})$$

$$(n = 1, 2, ..., N-1;\ m = 1, 2, ..., M-1). \quad ((7.74)$$

(Решение (7.73) обращается в нуль на границе Г прямоугольника G (7.51), как и должно быть, поскольку мы перенесли краевые значения в правую часть b уравнения (7.71), после чего значения на границе стали нулевыми. В частности, из (7.74) находим

$$\lambda(A))_{\min} = [(\frac{\pi}{l_x})^2 + (\frac{\pi}{l_y})^2], (\lambda(A))_{\max} \simeq 4(h_x^{-2} + h_y^{-2}) \qquad (7.75)$$

В квадрате при $l_x = l_y = l, \quad N = M, \quad h_x = h_y = h$

$$\lambda(A))_{\max} = 8/h^{2,}, \lambda(A))_{\min} = 2(\frac{\pi}{l})^2 \qquad (7.76)$$

а число обусловленности

$$p = \lambda(A))_{\max} / \lambda(A))_{\min} = \frac{4}{\pi^2} N^2 \qquad (7.77)$$

Из (7.77) видно, что число обусловленности p пропорционально числу узлов сетки N^2, и при $N \sim 10^2$ составляет несколько тысяч, т. е. матрица A плохо обусловленная. Поэтому для решения системы (7.71) следует использовать методы с повышенной скоростью сходимости.

Будем теперь рассматривать уравнение (7.69) как установившийся предел при $t \to \infty$ уравнения теплопроводности вида (7.50), когда *du/dt~0.* Соответственно, решение задачи (7.69) будем рассматривать как предел при $t \to \infty$ решения уравнения теплопроводности. Конечно-разностные методы, основанные на таком подходе, называют методами установления. Достоинство методов установления в том, что они открывают возможности использовать для решения уравнений эллиптического типа эффективные методы решения уравнения теплопроводности, например, описанную в предыдущем параграфе экономичную продольно-поперечную разностную схему. Решение при этом выполняется шагами по времени τ до тех пор, пока значения u полученные на соседних шагах, не станут достаточно близкими.

Число итераций в конечно-разностных методах (или, соответственно, число временных шагов в методах установления) можно уменьшить, если удастся выбрать начальное приближение (начальное условие), достаточно близкое к ожидаемому решению. Так, в задачах физики часто приходится моделировать изменение потенциала φ (7.68) при малых изменениях плотности заряда ρ, Тогда в качестве начального приближения можно принять распределение потенциала, полученное в предыдущем варианте; в результате для сходимости потребуется немного итераций, решение будет быстрым.

Перейдем теперь к проекционным методам. Для простоты при рассмотрении проекционных методов краевые условия будем полагать нулевыми, хотя это ограничение можно снять. Таким образом, будем рассматривать задачу

$$\Delta u(\vec{r}) + f(\vec{r}) = 0\ (\vec{r} \in G), \quad u(\vec{r}) = 0\ (\vec{r} \in \Gamma). \qquad (7.78)$$

В проекционных методах решение этой задачи отыскивается в виде

$$u(\vec{r}) = \sum_k c_k v_k(\vec{r}), \qquad (7.79)$$

где $\{v_k(\vec{r})\}$ — некоторая выбранная система функций, обращающихся в 0 на границе Г, а $\{c_k\}$ — неизвестные пока коэффициенты.

Наиболее распространенными из проекционных методов являются методы Галеркина (или Бубнова — Галеркина). В этих методах коэффициенты $\{c_к\}$ находятся из системы уравнений

$$(\Delta u + f, v_m) = 0 \quad (m = 1, 2, 3, ...), \qquad (7.80)$$

где круглые скобки, как обычно, обозначают скалярное

произведение, т.е. интеграл по области G. (По аналогии с векторным анализом можно представлять скалярное произведение (7.80) как «проекцию» $A\,u + f$ на координатную функцию v_m.) Подставляя разложение (7.79) в (7.80) и меняя порядок суммирования и интегрирования, получаем систему линейных алгебраических уравнений

$$\sum_k c_k(\Delta v_k, v_n) = -(f, v_m) \quad (m = 1, 2, 3, \ldots) \tag{7.81}$$

для нахождения коэффициентов $\{c_k\}$. Решив систему (7.81), получаем и решение (7.79).

Основное достоинство проекционных методов в том, что в них отсутствует разностная схема, и поэтому отсутствуют погрешности аппроксимации дифференциальных выражений конечно-разностными. Вместо этого имеются погрешности усечения рядов, возникающие из-за того, что в ряде (7.79) используется не бесконечное, а конечное число членов; но обычно погрешность усечения существенно меньше погрешности конечно-разностной аппроксимации. Кроме того, решением вида (7.79) можно пользоваться как обычной аналитической функцией, т.е. вычислять $u(\vec{r})$ в произвольных точках, а не только в узлах сетки, дифференцировать, интегрировать и т. д.

В качестве примера методов Галеркина рассмотрим часто применяемый метод двукратного преобразования Фурье для задачи (7.78). В этом методе решение $u(x,y)$ и правая часть $f(x,y)$ представляются двухкратными конечными рядами Фурье:

$$u(x, y) = \sum_{n=1}^{N-1} \sum_{m=1}^{M-1} \bar{u}_{nm} \sin \frac{\pi n x}{l_x} \sin \frac{\pi m y}{l_y}; \tag{7.82}$$

$$f(x, y) = \sum_{n=1}^{N-1} \sum_{m=1}^{M-1} \bar{f}_{nm} \sin \frac{\pi n x}{l_x} \sin \frac{\pi m y}{l_y} \tag{7.83}$$

(черточкой отмечаются коэффициенты рядов). В силу ортогональности тригонометрических функций в прямоугольнике G (7.51), система (7.81) имеет диагональную матрицу; благодаря этому легко находим связь коэффициентов $\bar{u}_{nm}$ с $\bar{f}_{nm}$:

$$\bar{u}_{nm} = \left[(\pi n/l_x)^2 + (\pi m/l_y)^2\right]^{-1} \bar{f}_{nm}. \tag{7.84}$$

Коэффициенты Фурье $\bar{f}_{nm}$, входящие в равенство (7.84), можно получить, домножая ряд (7.83) на $\sin(\pi n' x/l_x) \sin \times (\pi m' y/l_y)$ и суммируя затем по всем узлам сетки с номерами j=1,2,…,N-I; i=l,2,…,M-1:

$$\bar{f}_{nm} = \frac{4}{NM} \sum_{j=1}^{N-1} \sum_{i=1}^{M-1} f_{ji} \sin \frac{\pi n j}{N} \sin \frac{\pi m i}{M} \tag{7.85}$$

$$(n = 1, 2, \ldots, N-1; \quad m = 1, 2, \ldots, M-1).$$

Решение, таким образом, складывается из трех этапов: 1)вычисление $\bar{f}_{nm}$ (7.85) с помощью алгоритмов быстрого преобразования Фурье (глава 1), 2) вычисление $\bar{u}_{nm}$ (7.84), 3) вычисление u_{ji} по формуле (7.82) во всех узлах сетки:

$$u_{ji} = \sum_{n=1}^{N-1} \sum_{m=1}^{M-1} \bar{u}_{nm} \sin \frac{\pi n j}{N} \sin \frac{\pi m i}{M}$$

$$(j = 1, 2, \ldots, N-1; \; i = 1, 2, \ldots, M-1) \tag{7.86}$$

с помощью быстрого преобразования Фурье. Требуемое для решения число арифметических операций имеет порядок

$O(NM \times \log_2(NM))$. Поскольку логарифм — медленно меняющаяся функция, число операций почти линейно зависит от числа узлов сетки $N\,M$. Если $u(x, y)$ — функция с ограниченным спектром, имеющая не более N гармоник по x и M гармоник по y, то полученное решение будет точным. При этом функция $u(x,y)$ (7.82) и ее производные будут точными во всех точках области G. Если же спектр $u(x,y)$ не ограничен, т.е. имеет «высокочастотный хвост», то решение обладает погрешностью

вследствие усечения ряда Фурье, которая тем больше, чем менее гладкой является функция *u(x,y)* (чем длиннее «хвост» спектра).

Особенно широкое применение методы Галеркина получили в последнее время в форме так называемых методов конечных элементов. В этих методах функции $v_k(\vec{r})$ в (7.79) выбираются так, что они имеют конечный носитель, т. е. отличны от нуля в некоторой малой части $\triangle G_k$ области *G*.

Такие функции $v_k(\vec{r})$,а иногда и сами носители $\triangle G_k$ называют конечными элементами. Умелый выбор конечных элементов очень облегчает и ускоряет численное решение многих задач теории потенциала, теории упругости, теории пластичности и других прикладных задач физики. Более детально, однако, методы конечных элементов здесь рассматривать не будем, так как они выходят за рамки нашего небольшого учебного пособия и требуют дополнительного изучения.

7.6. Дисперсия, диссипация и монотонность разностных схем.

При решении дифференциальных уравнений с помощью разностных схем возможны некоторые нефизические явления, т.е. явления, обусловленные не физикой процесса, а методом, схемой решения. Поэтому такие явления иногда называют схемными. Одно из таких явлений – схемная дисперсия. Как известно, распространение волновых процессов в какой-либо среде характеризуется фазовой v_Φ и групповой $v_{\Gamma Р}$ скоростями:

$$v_\Phi = \frac{\omega}{k},\ v_{\Gamma Р} = \frac{d\omega}{dk},\ k = \frac{2\pi}{\lambda}. \qquad (7.87)$$

Здесь ω, k, λ - частота, волновое число и длина волны соответственно. Если фазовая скорость v_Φ зависит от частоты ω, то говорят, что среда обладает дисперсией, которая называется нормальной, если

$$\frac{dv_\Phi}{d\omega} < 0,\ v_\Phi > v_{\Gamma Р}, \qquad (7.88)$$

и аномальной, если

$$\frac{dv_\Phi}{d\omega} > 0,\ v_\Phi < v_{\Gamma Р}. \qquad (7.89)$$

Даже если в физической среде (например, в вакууме) дисперсия отсутствует, введение расчетной пространственной сетки почти всегда приводит к появлению схемной дисперсии, обычно нормальной. Рассмотрим, например, распространение прямоугольного импульса в среде. Крутые фронты импульса создаются высокочастотными компонентами. Если дисперсия отсутствует, импульс будет перемещаться с фазовой скоростью $v_\Phi = \omega / k$, неограниченно долго не меняя формы (рис. 30, а). В среде с нормальной дисперсией крутые фронты по мере движения сглаживаются, а позади импульса возникает колебательный “хвост”, создаваемый отставшими высокочастотными компонентами (рис. 30, б). В среде с аномальной дисперсией перед сглаженным импульсом постепенно появляется колебательный фронт, созданный быстрыми высокочастотными компонентами (рис. 30, в). Но даже если импульс первоначально имел достаточно гладкую форму, колебательные задний или передний фронты неизбежно возникают при его распространении вследствие вычислительных погрешностей, которые примешиваются к решению и обычно имеют сплошной спектр (“белый шум”).

Схемная дисперсия особенно сказывается на высокочастотных (коротковолновых) компонентах решения, тогда как низкочастотные (длинноволновые) компоненты практически не испытывают схемной дисперсии (как бы “не замечают” введенную сетку) и распространяются с той же скоростью, что и в физической системе. Для уменьшения влияния схемной дисперсии желательно, по возможности, использовать малый пространственный h и временной τ шаги, что, конечно, соответственно увеличивает затраты машинных ресурсов (длительности счета и памяти). Полностью устранить нефизическую дисперсию можно, устранив конечно-разностные производные по времени и пространству, например, отыскивая решение в виде ряда Фурье и вычисляя производные членов ряда аналитически.

Другим нефизическим эффектом является схемная диссипация – постепенное затухание (уменьшение амплитуды) волны по мере распространения. В зарубежной литературе эффекты схемной дисперсии и диссипации часто называют схемной диффузией. Как и схемная дисперсия, схемная диссипация сильнее сказывается на высокочастотных гармониках решения, которые довольно быстро затухают, тогда как низкочастотные гармоники практически не испытываю диссипации, как и в без диссипативной физической среде. Иногда диссипация отсутствует лишь на частоте $\omega = 0$ (на постоянной составляющей решения). В таких случаях говорят, что разностная схема обладает “аппроксимационной вязкостью”. В результате дисперсии и диссипации первоначально прямоугольный импульс по мере распространения приобретает форму колокола (рис. 31). Существуют, однако, разностные схемы, не обладающие диссипацией. В этом случае, колебательные задний или передний фронты будут неограниченно долго сопровождать импульс,

распространяющийся в среде с дисперсией.

Для анализа нефизических явлений решение одномерной разностной схемы отыскивается в виде

$$u_j^m \sim e^{\mathrm{i}(\omega m\tau - kjh)}, \qquad (7.90)$$

причем $\omega = \omega' + \mathrm{i}\omega'',\ \omega' = \mathrm{Re}\,\omega,\ \omega'' = \mathrm{Im}\,\omega.$ (7.91)

Здесь m=0,1,2… - номер временного шага, j=0,1,2,… - номер узла пространственной сетки, ω' - действительная часть частоты ω, ω'' - мнимая часть частоты.

Подставляя вид решения (7.90) в разностную схему и разделяя в ней действительную и мнимую части, можно получить два уравнения, из которых найти $\omega'(k)$ и $\omega''(k)$, а следовательно, фазовую скорость

$$v_{\phi} = \frac{\omega'(k)}{k}, \qquad (7.92)$$

и так называемый “множитель роста”

$$\rho = e^{-\omega''(k)\tau} \approx 1 - \omega''(k)\tau, \qquad (7.93)$$

характеризующий изменение амплитуды волны на шаге τ. В зависимости от значения ω'' и ρ возможны следующие случаи:

1) $\omega''(k) > 0,\ \rho < 1$ для любых k – устойчивая разностная схема с диссипацией.
2) $\omega''(k) > 0,\ \rho < 1$ для любых $\omega' > 0$ и $\omega'' = 0,\ \rho = 1$ на частоте $\omega = 0$ - устойчивая разностная схема с аппроксимационной вязкостью.
3) $\omega''(k) = 0,\ \rho = 1$ для любых k – устойчивая разностная схема без диссипации.
4) $\omega''(k) < 0,\ \rho > 1$ - неустойчивая разностная схема.

Более строгий анализ показывает, что устойчивость возможна и при значениях

$$\rho = 1 + C\tau \qquad (C=\mathrm{const}) \qquad (7.94)$$

но если хотя бы для одной гармоники C>0, то это так называемая “слабая устойчивость”.

Отсутствие диссипации в разностной схеме, обладающей дисперсией, надо понимать в следующем смысле: амплитуды пространственных гармоник (с разными k) не меняются со временем, но их относительные фазы меняются из-за дисперсии; поэтому и форма волны, определяемая суммой гармоник ряда Фурье, может меняться со временем по сложному закону.

Решения некоторых уравнений в частных производных, например, уравнения переноса или теплопроводности могут обладать свойствами монотонности, которое означает: если профиль $u(x,0)$ в начальный момент времени t=0 является монотонным, то эта монотонность сохраняется в ходе эволюции для $u(x,t)$ при любых t>0. Если начальная монотонность сохраняется и для решения разностной схемы, то такая схема называется монотонной, если же монотонность не сохраняется, схему называют немонотонной. На рис.32 показано, что первоначально прямоугольный (ступенчатый) профиль при счете по монотонной разностной схеме сглаживается, но остается монотонным, тогда как при немонотонной схеме приобретает колебательную форму. Монотонность – полезное свойство разностной схемы. Она предохраняет, например, от возможности возникновения в ходе решения немонотонных схем таких нежелательных, нефизических явлений, как отрицательные значения концентрации среды или абсолютной температуры, которые сложно устранить другими способами.

Имеются признаки монотонности разностных схем, хотя для неявных схем ими воспользоваться довольно сложно. Недостаток монотонных разностных схем – более высокая погрешность. Так, например, доказано, что монотонная двухслойная разностная схема для уравнения переноса (7.16) обязательно должна иметь порядок погрешности ниже второго. Поэтому при использовании монотонных разностных схем для сохранения приемлемой погрешности требуются достаточно малые шаги h, τ и повышенная загрузка компьютера.

Анализ рассмотренных ранее разностных схем для уравнения переноса показывает, например, что явная трехузловая, условно устойчивая разностная схема (7.22) – монотонная с нормальной дисперсией и аппроксимационной вязкостью; неявная трехузловая, безусловно устойчивая схема бегущего счета (7.23) – также монотонная с нормальной дисперсией, обладающая аппроксимационной вязкостью при $r = c\tau/h > 1$ и диссипативная при $n \le 1$. Неявная четырехузловая, безусловно устойчивая схема бегущего счета (7.24) – немонотонная (поскольку имеет второй порядок погрешности), бездиссипативная (т.е. $\rho = 1$ для любых k), с нормальной дисперсией при $r = c\tau/h > 1$, аномальной дисперсией при r<1 и без дисперсии при r =1.

Особую проблему составляет необходимое иногда сохранение при счете крутых (например, ступенчатых) фронтов при распространении сплошной среды. Монотонность разностной схемы не обеспечивает сохранение крутого фронта, как показано на рис.32, и для поддержания крутизны необходима специальная, достаточно сложная коррекция решения на каждом временном шаге. Из-за схемных эффектов часто приходится ограничивать длительность счета разностных схем такими временами, на которых нефизические искажения распространяющихся импульсов или волн сохраняют приемлемую величину.

7.7. Численные решения уравнений гиперболического типа

Распространение колебаний и волн (электромагнитных, упругих, акустических и др.) в направлении оси x в неограниченной среде описывается одномерным уравнением гиперболического типа

$$\frac{\partial^2 u}{\partial x^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x,t) \quad (0 \le t \le T) \qquad (7.95)$$

с начальными условиями

$$u(x,0) = \mu_1(x), \quad \frac{\partial u(x,0)}{\partial t} = \mu_2(x). \qquad (7.96)$$

Здесь a^2 - квадрат фазовой скорости волны, а функция $f(x,t)$ - описывает влияние среды, т.е. возбуждение или поглощение волн средой. В ограниченной области $(0 \le x \le l)$ следует добавить еще краевые условия

$$u(0,t)=\mu_3(t),\ u(l,t)=\mu_4(t). \qquad (7.97)$$

Уравнения (7.95) описывает как волны, распространяющиеся направо с фазовой скоростью $a>0$, так и волны, распространяющиеся налево с фазовой скоростью $-a<0$.

Для численного решения уравнений гиперболического типа обычно используются явные разностные схемы. Простейшая явная, трехслойная разностная схема для уравнения (7.95), построенная на 5-узловом шаблоне «крест» (рис.33), имеет вид:

$$\frac{\hat{u}_j - 2u_j + \check{u}_{j-1}}{\tau^2} = a^2 \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + f_j \quad (j = 1,2,...,N-1), \qquad (7.98)$$

$$u_0=\mu_3,\ u_N=\mu_4. \qquad (7.99)$$

Схема имеет погрешность $O(\tau^2, h^2)$; согласно спектральному критерию устойчивости является устойчивой и бездиссипативной при условии

$$r = \frac{a\tau}{h} < 1, \qquad (7.100)$$

где r - число Куранта. Схема обладает нормальной дисперсией, и низкочастотные (длинноволновые) компоненты решения распространяются по сетке с фазовой скоростью $v_{\hat{o}} \approx a$, тогда как высокочастотные (коротковолновые) компоненты испытывают нефизическое замедление до значения

$$v_{\hat{o}} \approx \frac{2a}{\pi} \approx 0.637a. \qquad (7.101)$$

При выборе шагов h, τ следует учитывать, что минимальная длина волны, различимая на сетке с шагом h , составляет $2h$, а максимальная частота равна $0.5/\tau$.

Чтобы начать вычисления по формуле (7.98), необходимо располагать значениями u^0 и u^1:

$$u^0=u(x,0)=\mu_1(x),\ u^1=u(x,\tau), \qquad (7.102)$$

причем для сохранения второго порядка погрешности надо иметь u^1 с погрешностью третьего порядка. Составляя для этого ряд Тейлора

$$u^1 = u^0 + \tau u_t^0 + 0{,}5\tau^2 u_{tt}^0 + O(\tau^3) \qquad (7.103)$$

и подставляя в него u_{tt}^0, найденное из уравнения (7.95):

$$u_{tt}^0 = a^2 u_{xx}^0 + f(x,0) = a^2 \frac{d^2\mu_1}{dx^2} + f, \qquad (7.104)$$

получаем:

$$u^1 = \mu_1 + \tau\mu_2 + \frac{\tau^2}{2}\left(a^2 \frac{d^2\mu_1}{dx^2} + f\right). \qquad (7.105)$$

Более сложная неявная, трехслойная разностная схема второго порядка, построенная на 9-узловом шаблоне (рис.34), имеет вид

$$\frac{\hat{u} - 2u + \check{u}}{\tau^2} = \Lambda\left[\sigma\hat{u} + (1-2\sigma)u + \sigma\check{u}\right] + f, \qquad (7.106)$$

где обозначено:

$$\Lambda u_j = \frac{a^2}{h^2}\left(u_{j+1} - 2u_j + u_{j-1}\right), \qquad (7{,}107)$$

а σ - некоторая константа. Можно доказать, что схема безусловно устойчива при значениях

$$0{,}25 \le \sigma \le 0{,}5. \qquad (7.108)$$

В частности, при σ=0,25 получаем схему

$$\frac{\hat{u} - 2u + \check{u}}{\tau^2} = \frac{1}{2}\Lambda\left[\frac{\hat{u} + \check{u}}{2} + u\right], \qquad (7.109)$$

а при σ=0,5 - схему

$$\frac{\hat{u} - 2u + \check{u}}{\tau^2} = \Lambda\left(\frac{\hat{u} + \check{u}}{2} + f\right). \qquad (7.110)$$

Значение u^1 здесь вычисляется по формуле (7.105), как и для явной схемы, а решения u^2, u^3,… на следующих временных шагах находятся с помощью

прогонки по верхнему временному слою схемы рис.34.

Как и для обыкновенных дифференциальных уравнений, уравнение второго порядка (7.95) можно свести к системе уравнений первого порядка:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial v}{\partial x}, \\ \dfrac{\partial v}{\partial t} = a^2 \dfrac{\partial u}{\partial x} + F(x,t), \end{cases} \qquad (7.111)$$

где $v(x,t) = \int_0^x u_t(s,t)ds,$ (7.112)

$$F(x,t) = \int_0^x f(s,t)ds. \quad (7.113)$$

При этом краевые условия (7.97) сохраняются, а начальные условия приобретают вид:

$$u(x,0) = \mu_1(x), \quad v(x,0) = \int_0^x \mu_2(s)ds. \quad (7.114)$$

Систему уравнений (7.111) часто называют уравнениями акустики, в которой она широко используется.

Вводя вектор неизвестных значений $u_j(t)$, $v_j(t)$ в узлах x_j пространственной сетки, можно привести (7.111) к системе обыкновенных дифференциальных уравнений первого порядка для u_j, v_j и воспользоваться для решения полученной системы безусловно устойчивым методом прямых в модификации Пикара, аналогичным методу (7.62).

Для численного решения уравнений гиперболического типа часто используются явные разностные схемы «на сдвинутых сетках». В качестве примера рассмотрим распространение вдоль оси x в однородной среде плоской электромагнитной волны с компонентами E_y и H_z. Обозначая

$$u = ZE_y, \quad v = H_z, \quad Z = \sqrt{\varepsilon/\mu}, \quad a = 1/\sqrt{\varepsilon\mu}, \quad (7.115)$$

где ε,μ - диэлектрическая и магнитная проницаемости, Z - волновое сопротивление среды, a - фазовая скорость волны, можно привести однородные вихревые уравнения Максвелла
к виду:

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial v}{\partial x} = 0, \\ \dfrac{\partial v}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 \end{cases} \qquad (7.116)$$

с начальными условиями (7.96) для u, u_t. Краевые условия для u в ограниченной среде (в зазоре между двумя пластинами) могут задаваться в виде (7.97).

На рис.35 показан участок сдвинутых сеток для системы (7.116); на рисунке точки – узлы сетки u, кружочки – узлы сетки v. Два узла, обведенные штриховой линией, будем считать единым сдвоенным узлом (j,m), где m - номер временного шага. Тогда разностную схему для системы (7.116) можно записать в виде:

$$\frac{\hat{u}_j - u_j}{\tau} + a\frac{v_j - v_{j-1}}{h} = 0, \qquad (7.117)$$

$$\frac{\hat{v}_j - v_j}{\tau} + a\frac{\hat{u}_{j+1} - \hat{u}_j}{h} = 0. \qquad (7.118)$$

Поскольку систему (7.117), (7.118) можно легко свести к разностной схеме вида (7.98), она также имеет погрешность $O(h^2,\tau^2)$ и нормальную дисперсию с замедлением высших гармоник до скорости (7.101), бездиссипативна и устойчива при условии Куранта $r \le 1$.

При численном решении на каждом временном шаге сначала находится $\hat{u}$ из уравнения (7.117), а затем $\hat{v}$ из уравнения (7.118). Для начала счета следует найти $v(x,\tau/2)$ по формулам:

$$\begin{cases} v(x,\tau/2) = v(x,0) + 0{,}5\tau v_t(x,0) + 0{,}125\tau^2 v_{tt}(x,0) \\ v_t(x,0) = -au_x^0 = -a\dfrac{d\mu_1}{dx} \\ v_{tt}(x,0) = -au_{xt}(x,0) = -a\dfrac{\partial}{\partial x}u_t^0 = -a\dfrac{d\mu_2}{dx} \end{cases} . \qquad (7.119)$$

Явные разностные схемы, в том числе и на сдвинутых сетках, можно аналогично записать для двумерных и трехмерных задач теории поля. При этом число сдвинутых сеток равняется числу учитываемых в задаче компонент электромагнитного поля и составляет от 2 до 6. При построении сдвинутых сеток удобно придерживаться следующего правила: участки границы области, на которых для некоторой компоненты поля задано краевое условие первого рода, должны располагаться на линиях её сетки, а участки сетки, на которых задано

краевое условие второго рода, должны быть смещены на полшага от границы. Такое расположение сеточных линий облегчает задание краевых условий в разностных уравнениях. Пример такой сетки показан на рис.36. Условие устойчивости Куранта для трехмерной сетки имеет вид:

$$v_\phi \tau\left(h_x^{-2}+h_y^{-2}+h_z^{-2}\right)\le 1, \qquad (7.120)$$

где h_x, h_y, h_z - шаги сетки *по x, y, z* соответственно.

При численном решении уравнений гиперболического типа для электромагнитного поля большие вычислительные трудности создают открытые участки границы, т.е. участки, на которых отсутствует отражение и имеется свободное прохождение волны через границу. На этих участках границы надо моделировать известный в физике «принцип излучения», т.е. отсутствие волн, поступающих извне в систему. Для этого надо ввести некоторые предположения о характере поля вне расчётной сетки и аккуратно «сшить» поля на открытой границе. Иногда при этом на «сшивание» расходуется основное время счёта.

Наконец, отметим ещё один нефизический эффект, который может возникать при моделировании взаимодействия электромагнитных волн с релятивистскими электронными пучками, распространяющимися со скоростью, близкой к скорости света *c*. В разностных схемах с нормальной дисперсией (а таких большинство) пучок будет распространяться со скоростью, заметно превышающей фазовую скорость паразитных высших гармоник поля, которые обязательно возникают вследствие вычислительных погрешностей, обычно имеющих сплошной спектр. Вследствие этого становятся возможными «схемное черенковское излучение» и вызванные им неустойчивости пучка. В принципе, паразитные высшие гармоники поля можно подавить путём фильтрации решения на каждом временном шаге, но это требует осторожности, поскольку может иметь следствием нефизическое нарушение закона сохранения энергии (консервативности) и нефизическое изменение фазовой и групповой скорости электромагнитной волны.

7.8.Особенности численного решения квазилинейных уравнений в частных производных.

Квазилинейными называют дифференциальные уравнения, в которых производные входят линейно, но коэффициенты при них являются функциями искомого решения. Простейшим примером может служить одномерное квазилинейное уравнение переноса

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=f(x,t),\quad (0\le x\le l) \qquad (7.121)$$

с начальным и краевым условиями

$$u(x,0)=\varphi(x),\quad u(0,t)=\psi(t). \qquad (7.122)$$

Правая часть уравнения *f(x,t)* описывает внутренние источники сплошной среды, влияющие на её концентрацию *u(x,t)*, например, натекание извне, абсорбцию на стенках канала и т.п. Уравнение (7.121) можно записать и в так называемой дивергентной форме: $\frac{\partial u}{\partial t}+\frac{\partial F}{\partial x}=f(x,t),$ (7.123)

$$F(x,u,t)=u^2/2 \qquad (7.124)$$

где *F* – «поток» (flux).

Однородное квазилинейное уравнение переноса имеет вид

$$\frac{\partial u}{\partial t}+\frac{\partial F}{\partial x}=0. \qquad (7.125)$$

Примерами такого уравнения являются уравнение непрерывности в механике сплошной среды или закон сохранения заряда в теории поля

$$\frac{\partial \rho}{\partial t}+\frac{\partial j}{\partial x}=0 \quad , \quad j=\rho u, \qquad (7.126)$$

где $\rho(x,t)$ - плотность объёмного заряда, *j(x,t)* - плотность тока, *u(x,t)* - средняя скорость заряженного потока в точке *x* в момент времени *t*. Уравнение (7.126) – квазилинейное, так как скорость *u* зависит от электрического поля, которое, в свою очередь, зависит от плотности объёмного заряда ρ.

Уравнение (7.125) с потоком (7.124) имеет аналитическое решение

$$u(x,t)=\varphi(x-ut), \qquad (7.127)$$

аналогичное (7.18), но с переменной скоростью *u(x,t)*, что сразу приводит к ряду физических явлений, сильно осложняющих численное решение. Из этих явлений наибольшие трудности создают сильные и слабые разрывы, которые могут возникать в решении со временем *t*. Сильными разрывами, или ударными волнами называют разрывы функции *u(x,t)*, а слабые разрывы – это разрывы производной u_x. Физически разрывы возникают из-за того, что частицы среды, движущиеся с разными скоростями, могут отставать от других частиц, догонять и обгонять их, так что в некоторых точках или вдоль некоторых линий концентрация среды может меняться скачком. При возникновении ударных волн непрерывное решение построить не удаётся, решение становится неоднозначным, и можно лишь отыскивать некоторое обобщённое решение, используя какой-либо физический закон сохранения. Для сглаживания разрывов и получения однозначного решения в дифференциальные уравнения иногда вводят дополнительные члены – «искусственную вязкость» («псевдовязкость»), которые малы вдали от разрывов, но велики вблизи разрывов и препятствуют их развитию. Этой же цели может служить «схемная вязкость», являющаяся свойством используемой разностной схемы или

численного метода.

Прежде чем перейти к рассмотрению методов численного решения квазилинейных уравнений, отметим одно из важных свойств некоторых дифференциальных уравнений физики – свойство консервативности, под которым понимается выполнение физических законов сохранения. Для однородного уравнения переноса (7.125) свойство консервативности представляет собой закон сохранения полного заряда (или массы) и имеет вид:

$$\int_0^l [u(x,T)-u(x,0)]dx+\int_0^T [F(l,t)-F(0,t)]dt=0\,. \quad (7.128)$$

Здесь первый интеграл представляет собой изменение количества вещества (или заряда) в области [0,*l*] за время *T*, а второй – суммарный приток вещества через границы *x=0* и *x=l* за то же время. В разностной схеме интегрирование надо заменить суммированием по всем ячейкам разностной сетки. Если для разностной схемы удовлетворяется равенство

$$h\sum_j \left(u_j^M-u_j^0\right)+\tau\sum_m \left(F_N^m-F_0^m\right)=0\,, \quad (7.129)$$

где j - номера пространственных узлов x_j *(j=0,1,2,…,N),* m - номера временных слоёв t^m *(m=0,1,2,…,M),* то говорят, что разностная схема обладает свойством консервативности. (Заметим, что в суммах исчезли все внутренние линии пространственно-временной сетки). Если же равенство (7.129) не выполняется, то схема считается неконсервативной.

Если разностная схема удовлетворяет нескольким физическим законам сохранения, то она называется полностью консервативной. Консервативность – полезное свойство разностной схемы, так как оно правильно отображает физические свойства исследуемой системы. Однако обычно трудно добиться, чтобы разностная схема имела одновременно хорошую аппроксимацию, монотонность и консервативность, и из этих свойств приходится выбирать наиболее важное в рассматриваемой задаче.

Для уравнения

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0 \quad (7.130)$$

наиболее простая явная, неконсервативная разностная схема имеет вид

$$\frac{\hat{u}_j-u_j}{\tau}+u_j\frac{u_j-u_{j-1}}{h}=0\,. \quad (7.131)$$

Из дивергентной формы (7.125) уравнения можно получить явную, консервативную разностную схему

$$\frac{\hat{u}_j-u_j}{\tau}+\frac{u_j^2-u_{j-1}^2}{2h}=0 \quad (7.132)$$

и более сложную неявную, консервативную разностную схему бегущего счёта:

$$\frac{\hat{u}_j-u_j}{\tau}+\frac{\hat{u}_j^2-\hat{u}_{j-1}^2}{2h}=0\,, \quad (7.133)$$

вычисления по которой выполняются по формуле

$$\hat{u}_j=-\frac{h}{\tau}+\sqrt{\frac{h^2}{\tau^2}+\frac{2h}{\tau}u_j+\hat{u}_{j-1}^2}\,,\quad (j=1,2,3,\ldots)\,. \quad (7.134)$$

В задачах физики часто встречаются многомерные квазилинейные уравнения переноса. Некоторые методы численного решения таких уравнений рассмотрим на примере однородного двумерного уравнения

$$\frac{\partial u}{\partial t}+v\frac{\partial u}{\partial x}+w\frac{\partial u}{\partial v}=0\quad (0\le x\le l,\quad 0\le v\le v_{\max}) \quad (7.135)$$

для функции трёх переменных *u(x,v,t).* В общем случае функция *w* зависит от всех переменных:

$$w=w(x,v,u,t). \quad (7.136)$$

Вид (7.135) имеет, в частности, известное уравнение Власова, которое часто используется в физической электронике, физике плазмы, физике заряженных пучков.

При $w\equiv 0$ уравнение (7.135) имеет аналитическое решение вида

$$u(x,v,t)=ae^{ik(x-vt)}\quad (a=const)\,. \quad (7.137)$$

Согласно (7.137)

$$|u(x,v,t)|=|u(x,v,0)|=|a|=const\,, \quad (7,138)$$

тогда как

$$\left|\frac{\partial u}{\partial v}\right|=|-iktu(x,v,t)|=kt|a|\to\infty \quad (7,139)$$

при $t\to\infty$. Такое явление, когда решение ограничено, а его производная неограниченно возрастает со временем, называют «градиентной катастрофой». Отметим, что градиентная катастрофа – свойство аналитического решения и обусловлено физикой процесса, а не численным методом.

Для численного решения квазилинейного уравнения (7.135) конечно-разностным методом введём равномерную двумерную сетку по *x,v* с узлами x_j,v_i и шагами h_x,h_v соответственно. Согласно (7.137) длина

волны по v определяется равенством

$$kvt=2\pi, \qquad (7.140)$$

а поскольку минимальная длина волны составляет $2hv$, максимальное время счёта на выбранной сетке

$$t_{\max}=\frac{2\pi}{2khv}=\frac{\pi}{khv}. \qquad (7.141)$$

Как показывает детальный анализ, в течение интервала времени t_{max} спектр решения по v смещается в сторону всё более высоких частот ω_v вплоть до верхней границы

$$\omega_{v,\max}=\frac{\pi}{h_v}.$$

После этого возникает как бы отражение возмущения от верхней границы спектра , и в течение следующего интервала t_{max} спектр движется обратно в сторону низких частот, пока не приобретёт исходный характер. Далее возмущение отражается от нижней границы спектра $\omega_{v,\min}=\dfrac{\pi}{v_{\max}}$,

и весь процесс повторяется – возникают нефизические рекуррентные явления с периодом $2t_{max}$. Для устранения рекурсий следует не превышать длительность счёта t_{max} (7.141) или добавить в правую часть уравнения (7.135) член, имитирующий затухание волн в среде.

Обозначим теперь

$$\hat{u}_x\approx\frac{\hat{u}_{j+1,i}-\hat{u}_{j-1,i}}{2h_x}+O\left(h_x^2\right),\quad u_v\approx\frac{u_{j,i+1}-u_{j,i-1}}{2h_v}+O\left(h_v^2\right) \qquad (7.142)$$

$$\hat{u}_t\approx\frac{3\hat{u}_{ji}-4u_{ji}+\check{u}_{ji}}{2\tau}+O\left(\tau^2\right) \qquad (7.143)$$

и построим для уравнения (7.135) трёхслойную разностную схему

$$\hat{u}_t+v\hat{u}_x+wu_v=0. \qquad (7.144)$$

По t схема (7.144) явная, а по x на верхнем временном слое $t+\tau$ в схему входят три узла x_j, x_{j+1}, x_{j-1} и решение $\hat{u}$ на каждом временном шаге может быть найдено методом прогонки.

Записывая уравнение (7.135) в виде

$u_t=g,\ g=-vu_x-wu_v \qquad (7.145)$

и вводя неизвестные значения u_{ji} в узлах (x,v) - сетки можно также свести (7.145) к системе обыкновенных дифференциальных уравнений

$$\dot{u}_{ji}=g_{ji}, \qquad (7.146)$$

используя для ее численного решения описанные в главе 6 методы “предиктор-корректор” или интерполяционную и экстраполяционную формулы Адамса.

Помимо конечно-разностных методов для двумерных квазилинейных уравнений могут использоваться так называемые методы преобразований, аналогичные методам Галеркина (§ 7.5). В этих методах решение отыскивается в виде обобщенного двойного ряда Фурье

$$u(x,v,t)=\sum_{n,m}\bar{u}_{nm}(t)X_n(x)V_m(v), \qquad (7.147)$$

где $\{X_n(x)\}$, $\{V_m(v)\}$ — выбранные системы ортогональных функций, удовлетворяющие заданным краевым условиям, а $\bar{u}_{nm}(t)$ — коэффициенты ряда. Решение на шаге τ строится как ряд Тейлора

$$\hat{u}\approx\sum_{s=0}^{S}\frac{\tau^8}{s!}\frac{\partial^s u}{\partial t^s}, \qquad (7.148)$$

в котором производные высших порядков находятся путем точного (аналитического) дифференцирования уравнения (7.145). Порядок метода S обычно не выше 3. В качестве $\{X_n(x)\}$ часто принимаются тригонометрические функции, а в качестве $\{V_m(v)\}$ — полиномы Эрмита. Наибольшие трудности в методах преобразований составляет вычисление нелинейных членов вида uu_x, uu_v и т.п., хотя предложены некоторые способы преодоления этих трудностей путем усложнения вычислительного алгоритма и удлинения вычислений примерно вдвое.

В принципе, методы преобразований, как и методы Галеркина, могут обладать более высокой точностью по сравнению с разностными схемами благодаря отсутствию конечно-разностных производных и точному (аналитическому) дифференцированию рядов (7.147) по x и v, но для хорошего воспроизведения мелкомасштабных особенностей решения $u(x,v,t)$ в рядах вида (7.147)

приходится удерживать много членов, иногда много сотен. Тем не менее, вследствие неизбежного усечения рядов рекуррентные явления сохраняются, и для их подавления требуется вводить в алгоритм решения искусственное затухание.

Рассмотрим теперь довольно распространенный для уравнения (7.135) метод характеристик. Отметим, что система обыкновенных дифференциальных уравнений

$$\begin{cases} \dot{x} = v, \\ \dot{v} = w. \end{cases} \qquad (7.149)$$

называется уравнением характеристики, а его решение на плоскости (x,v) — характеристикой уравнения (7.135). Учитывая (7.149), находим, что вдоль характеристики

$$\frac{du}{dt} = 0, \qquad (7.150)$$

а следовательно

$$u(x,v,t) = u(x,v,0) = const, \qquad (7.151)$$

т.е. начальные значения u^0 не изменяясь распространяются вдоль характеристик.

Примем в момент t узлы (x_j, v_i) сетки (x,v) за начальные точки и построим выходящие из них отрезки характеристик на шаге τ, решая систему (7.149) одним из методов главы 6. Конечные точки $\hat{x}_{j'}$, $\hat{v}_{i'}$ отрезков характеристик назовем концевыми (рис. 37). Множество концевых точек образует подвижную, так называемую лагранжеву сетку. Для завершения временного шага надо пересчитать значения $\hat{u}$ с лагранжевой сетки в узлы неподвижной (эйлеровой) сетки по формуле

$$\hat{u}_{ji} = \sum_{j',i'} \hat{u}_{j'i'} \cdot S_x\left(x_j, \hat{x}_{j'}\right) \cdot S_v\left(v_i, \hat{v}_{i'}\right), \quad (7.152)$$

где S_x, S_v — весовые функции; j', i' — номера узлов лагранжевой сетки. Можно считать (7.152) интерполяционной формулой, связывающей узлы подвижной и неподвижной сеток. Весовые функции рекомендуется выбирать из условия сохранения моментов, вычисленных по узлам неподвижной и подвижной сеток:

$$\sum_{j,i} x_j^k v_i^n \hat{u}_{ji} = \sum_{j',i'} \hat{x}_{j'}^k \hat{v}_{i'}^n \hat{u}_{j'i'} \qquad (k,n = 0,1,\ldots). \qquad (7.153)$$

Сумма в левой части равенства (7.153) берется по узлам неподвижной сетки, а сумма в правой части — по концевым точкам. Обычно в (7.153) ограничиваются значениями $k = n = 0$; $k = 1, n = 0$ и $k = 0, n = 1$. При $k = n = 0$ равенство (7.153) означает сохранение на шаге среднего значения функции $u(x,v,t)$, а при $k = 1, n = 0$ и $k = 0, n = 1$ — сохранение первых моментов по x и v соответственно.

Метод характеристик можно распространить и на случай неоднородных квазилинейных уравнений с правой частью f, а также на трехмерные сетки. Хотя по точности он уступает методам преобразований и некоторым другим методам, но его достоинство — простой и быстрый вычислительный алгоритм.

Рассмотрим, наконец, оригинальную двумерную модель “водяной мешок”. Линию Γ на плоскости (x,v), описываемую уравнением

$$u(x,v,t) = C, \qquad (7.154)$$

где C — некоторая константа, назовем линией уровня. Придавая C различные значения $u_1, u_2, \ldots$ можно построить множество линий уровня $\Gamma_1, \Gamma_2, \ldots$, характеризующих “профиль” функции $u(x,v,t)$ в момент t, показанный на рис. 38,а. В силу единственности решения линии уровня нигде не пересекаются. Ограничимся случаем, когда функция w такова, что

$$\int u_v w_v dv = 0 \qquad (7.155)$$

по области, ограниченной какой-либо линией уровня Γ_i. Равенство (7.155) выполняется, например, если функция w не зависит от v. Но возможны и другие типы функций w, для которых выполняется (7.155). Можно показать, что при условии (7.155) для всех линий уровня Γ_i

$$\int_{V_i} u\, dx\, dv = const, \qquad (7.156)$$

где V_i — область, ограниченная линией Γ_i. При этом, естественно и для любой области V_{ik} между какими-либо двумя линиями уровня Γ_i и Γ_k

$$\int_{V_{ik}} u\, dx\, dv = const \qquad (7.157)$$

как разность двух констант.

Выбирая в момент t на каждой линии уровня Γ_i некоторое число точек, и построив выходящие из этих точек отрезки характеристик на шаге τ, можно найти форму линии Γ_i в следующий момент времени $t+\tau$, получая картину решения в виде динамики линий уровня. Поскольку все объемы V_{ik} между линиями уровня при движении сохраняются, их динамика напоминает изменение формы гибкого мешка с несжимаемой жидкостью — отсюда название “водяной мешок”. Достоинство метода — наглядность решения, но с течением времени форма линий уровня быстро усложняется (рис. 38,б) и на них приходится выбирать все больше точек, длительность моделирования оказывается ограниченной. Модель обычно применяется, когда можно ограничиться двумя-тремя линиями уровня, иногда даже одной, и дает скорее общую картину решения, чем его количественные результаты.

8.ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ ОПТИМИЗАЦИИ И ПОИСКА МИНИМУМА.

8.1. Постановка задач оптимизации и поиска минимума

Численными методами оптимизации называются методы построения алгоритмов численного нахождения минимумов или максимумов некоторых функций и точек (аргументов), в которых они достигаются. Заметим, что нахождение минимума некоторой функции $\Phi(x)$ эквивалентно нахождению максимума функции $-\Phi(x)$. Поэтому для определенности принято говорить о нахождении минимума. В прикладных задачах минимизируемую функцию часто называют целевой функцией. Целевая функция характеризует некоторое качество оптимизируемого объекта. Можно также называть эту функцию критерием (критерием качества или критерием оптимальности).
В более общем случае требуется найти такую точку x, в которой несколько целевых функций $\Phi_k(x)\ (k=1,2,...)$ совместно достигают минимума в том или ином смысле, который понимается по-разному в различных случаях. Такие задачи называются многокритериальными или задачами оптимизации по многим критериям.
Задачи поиска минимума тесно связаны с задачами решения систем уравнений, т .е . с задачами нахождения корней. Так, вместо отыскания минимума функции $\Phi(x)$ нескольких переменных $x=(x_1,x_2,...x_n)$ можно, в принципе, решать систему уравнений

$$\frac{\partial \Phi}{\partial x_i} = 0 \quad (i=1,2,...,n). \qquad (8.1)$$

Однако для этого функция $\Phi(x)$ должна быть дифференцируемой всюду в области поиска минимума, а дифференцируемость сложной функции иногда трудно установить аналитически. Кроме того, численное решение системы уравнений (8.1), в общем случае нелинейных, может быть сопряжено с большими трудностями. Поэтому иногда, наоборот, решение системы уравнений

$$f_k(x) = 0 \quad (k=1,2,...n) \qquad (8.2)$$

сводят к отысканию 0-го минимума некоторой функции, например,

$$\Phi(x) = \sum_{k=1}^{n} w_k f_k(x) \quad (w_k > 0), \qquad (8.3)$$

где $w_k\ (k=1,2,...n)$ – веса.

Функцию $\Phi(x)$ будем предполагать непрерывной или, по крайней мере, кусочно-непрерывной, и детерминированной. Последнее означает, что вычислительными и экспериментальными погрешностями можно пренебречь (вычислительными погрешностями, как уже говорилось во Введении, называются погрешности округления, обусловленные конечностью разрядной сетки компьютера). (Поиск минимума случайной функции рассматривается в 8.3). Множество X, на котором отыскивается минимум функции $\Phi(x)$, будем полагать компактным и замкнутым. (Напомним, что множество называется компактным, если из любого его бесконечного и ограниченного подмножества можно выделить сходящуюся последовательность. Множество называется замкнутым, если предел любой сходящейся последовательности его элементов принадлежит этому множеству).

Точку минимума, называемую также модой, будем обозначать $\bar{x}$.

Различают локальные и глобальные минимумы. Минимум $\Phi(\bar{x})$ называется локальным, если в некоторой его δ-окрестности

$$\Phi(\bar{x}) < \Phi(x), \ \|x - \bar{x}\| \le \delta. \qquad (8.4)$$

Минимум $\Phi(\bar{x})$ называется глобальным, если

$$\Phi(\bar{x}) = \inf_{\{X\}} \Phi(x) \qquad (8.5)$$

Локальных минимумов на множестве X может быть много. В принципе, если найти все локальные минимумы и взять из них наименьший, то получим глобальный минимум. Практически, однако, для сложных функций $\Phi(x)$ обычно не может быть уверенности в том, что все локальные минимумы уже найдены. Поэтому минимальный из локальных минимумов следует считать, строго говоря, лишь «кандидатом» на глобальный минимум, который может быть изменен при нахождении новых локальных минимумов.

Для решения задачи поиска минимума обычно стараются предварительно выделить такое множество (область) X, на котором функция $\Phi(x)$ имеет единственный минимум. В случае дифференцируемой функции нахождение такой области равносильно нахождению области, в которой система уравнений (8.1) имеет единственный корень. Для произвольной функции $\Phi(x)$ эта задача алгоритмически не разрешима и для ее решения необходимо привлекать какие-либо априорные или физические соображения. Функцию, имеющую в рассматриваемой области единственный минимум, называют унимодальной. В дальнейшем будем ограничиваться унимодальными функциями, если противоположное не будет специально указываться. Для моды $\bar{x}$ будем использовать обозначение $\bar{x} = arg\,min\Phi(x)$, понимая его следующим образом: $\bar{x}$ - аргумент (координаты) минимума функции $\Phi(x)$.

8.2. Поиск минимума функции одной переменной.

Будем рассматривать задачу поиска минимума функции $\Phi(x)$ одной переменной x на отрезке [a,b]. Простейшим методом является прямой поиск минимума путем перебора значений. В этом методе на отрезке [a,b] выбирается некоторое число точек $\{x_j;\ j=0,1,\ldots,n\}$, которые могут быть случайно расположенными или равноотстоящими с некоторым шагом *h:*

$$x_j = a + jh \quad (j=0,1,\ldots,n),\ h=(b-a)/n \qquad (8.6)$$

Далее последовательно вычисляются значения

$$\Phi_j = \Phi(x_j) \quad (j=0,1,2,..), \qquad (8.7)$$

пока не окажется, что

$$\Phi_j < \Phi_{j+1}. \qquad (8.8)$$

Выполнение неравенства (8.8) означает, что минимум $\bar{x}$ расположен на отрезке $[x_{j-1}, x_{j+1}]$ (рис.39). На этом отрезке снова выбирается некоторое число точек, и перебор повторяется до тех пор, пока длина очередного отрезка станет меньше 2ε, где ε - заданная погрешность.

Тогда в качестве $\bar{x}$ можно принять середину последнего отрезка.

Наибольшее распространение в задачах одномерного поиска минимума получили так называемые симметричные итерационные методы, итерации в которых строятся следующим образом:

1) на отрезке [*a,b*] выбирается точка $a<y<b$ (рис. 40);

2) далее берется симметричная ей точка *z*; симметрия означает, что

$y-a=b-z$, (8.9)

т.е.

$z=a+b-y$ (8.10)

3) если $\Phi(x)<\Phi(z)$, то для дальнейшего рассмотрения оставляется отрезок [*a*,*z*], иначе - отрезок [*y*,*b*], в результате длина отрезка, на котором отыскивается минимум, сокращается в отношении

$$q=\frac{z-a}{b-a}. \qquad (8.11)$$

Затем процесс повторяется.

Простейшим симметричным методом является метод дихотомии (половинного деления). Пусть на *k* - й итерации отрезок поиска минимума составляет

$[a_{k-1},b_{k-1}]\ \ (k=1,2,...)$, причем $a_0=a$, $b_0=b$. В качестве y_k, z_k примем точки

$$y_k=0{,}5(a_{k-1}+b_{k-1}-\varepsilon),\ \ z_k=0{,}5(a_{k-1}+b_{k-1}+\varepsilon), \qquad (8.12)$$

где ε - заданная погрешность (рис.41). При этом

$$q=0{,}5+\frac{\varepsilon}{2(b_{k-1}-a_{k-1})}. \qquad (8.13)$$

При $b_{k-1}-a_{k-1}>>\varepsilon$, т.е. на начальных итерациях отрезок уменьшается примерно вдвое на каждой итерации (отсюда – название метода).

Далее сравниваем $\Phi(y_k)$ с $\Phi(z_k)$. Если окажется, что

$$\Phi(y_k)<\Phi(z_k), \qquad (8.14)$$

то $\bar{x}\in[a_{k-1};z_k]$ и для следующей итерации полагаем

$a_k=a_{k-1}$, $b_k=z_k$. (8.15)

Если же выполняется неравенство $\Phi(y_k)>\Phi(z_k)$, (8.16)

то $\bar{x}\in[y_k,b_{k-1}]$ и полагаем $a_k=y_k$, $b_k=z_{k-1}$. (8.17)

Если $b_k-a_k<2\varepsilon$, то итерации прекращаются и в качестве приближенного значения $\bar{x}$ с погрешностью ε принимается середина последнего отрезка:

$$\bar{x}\approx\overline{x_k}=0{,}5(a_k+b_k). \quad (8.18)$$

Недостатками метода дихотомии являются необходимость дважды вычислять функцию $\Phi(x)$ на каждой итерации и замедление сходимости (т.е. увеличение q) по мере приближения длины отрезка (b_k-a_k) к значению 2ε.

Более экономичными являются симметричные методы, в которых требуется однократное вычисление функции $\Phi(x)$ на каждой итерации, кроме первой , на которой $\Phi(x)$ вычисляется дважды. Рассмотрим схему таких симметричных методов. Пусть на *k*-й итерации минимум отыскивается на отрезке $[a_{k-1},b_{k-1}]\ \ (k=1,2,...)$, причем $a_0=a$, $b_0=b$, и из предыдущей итерации (кроме первой) уже известны одна из двух внутренних точек, например y_k и соответствующее значение $\Phi(y_k)$.

Положим, что

$$y_k=a_{k-1}+c_{k-1}(b_{k-1}-a_{k-1})\ \ (k=1,2,..), \qquad (8.19)$$

где c_k - некоторый параметр:

$0<c_k<0{,}5\ \ (k=0,1,2,...)$. (8.20)

Симметричную точку z_k находим по формуле (8.10):

$$z_k=a_{k-1}+b_{k-1}-y_k=a_{k-1}+(1-c_{k-1})(b_{k-1}-a_{k-1}). \qquad (8.21)$$

Далее вычисляем $\Phi(z_k)$ и полагаем :

1) если $\Phi(z_k)\ge\Phi(y_k)$, то

$a_k=a_{k-1}$, $b_k=z_k$, $z_{k+1}=y_k$, $\Phi(z_{k+1})=\Phi(y_k)$; (8.22)

2) если $\Phi(z_k)<\Phi(y_k)$, то

$a_k=y_k$, $b_k=b_{k-1}$, $y_{k+1}=z_k$, $\Phi(y_{k+1})=\Phi(z_k)$ (8.23)

(рис.42). Для дальнейшего рассмотрения выбираем отрезок $[a_k,b_k]$, длина которого

$$b_k-a_k=(1-c_{k-1})(b_{k-1}-a_{k-1}). \qquad (8.24)$$

Как и в методе дихотомии итерации прекращаются, если

$$b_k-a_k<2\varepsilon, \quad (8.25)$$

где ε - заданная погрешность, и в качестве приближенного значения точки минимума $\bar{x}$ принимается $\overline{x_k}$ (8.18).

Различные симметричные методы отличаются выбором параметров c_k. Методы, в которых параметры c_k постоянны, т.е.

$c_k=c=const$, (8.26)

называются стационарными, а методы, в которых параметры c_k зависят от номера итерации k - нестационарными.

Наиболее быстрым из всех симметричных методов с однократным вычислением $\Phi(x)$ на итерации является нестационарный метод Фибоначчи. В этом методе длина *k-го* отрезка связана с длиной начального отрезка (b_0-a_0) равенством

$$b_k-a_k=\frac{b_0-a_0}{F_{k+2}}, \qquad (8.27)$$

где F_j - так называемые числа Фибоначчи, определяемые рекуррентной формулой

$F_j=F_{j-1}+F_{j-2}$, $(j=3,4,..)$, $F_1=F_2=1$. (8.28)

По этой формуле

$F_3=2$, $F_4=3$, $F_5=5$, $F_6=8$, $F_7=13$, $F_8=21$, $F_9=34$, $F_{10}=55$, $F_{11}=89$, $F_{12}=144$, $F_{13}=233$,..

Для чисел Фибоначчи имеются таблицы, но можно вычислять их и непосредственно по формулам (8.28).

Необходимое число итераций n выбирается из условия, что

$$b_n - a_n < 2\varepsilon \le b_{n-1} - a_{n-1}\,, \qquad (8.29)$$

где согласно (8.27)

$$b_n - a_n = \frac{b_0 - a_0}{F_{n+2}}; \quad b_{n-1} - a_{n-1} = \frac{b_0 - a_0}{F_{n+1}}. \qquad (8.30)$$

Можно показать, что в методе Фибоначчи

$$c_0 = \frac{F_n}{F_{n+2}}, \qquad (8.31)$$

а последующие значения c_k вычисляются по рекуррентной формуле

$$c_k = \frac{1-2c_{k-1}}{1-c_{k-1}} \quad (k = 1,2,...), \qquad (8.32)$$

причем точки y_k и z_k определяются формулами

$$\begin{cases} y_k = \dfrac{a_{k-1} + (b_0 - a_0)F_{n-k+1}}{F_{n+2}} \\ z_k = \dfrac{a_{k-1} + (b_0 - a_0)F_{n-k+2}}{F_{n+2}} \end{cases} \qquad (8.33)$$

При $k=n$ точки y_k и z_k совпадают, так как $F_1=F_2=1$, и итерационный процесс завершается.

Методу Фибоначчи незначительно (примерно на 17% при больших значениях n) уступает в скорости стационарный симметричный метод «золотого сечения», который получается, если в формулах (8.32), (8.19), (8.21) положить

$c_k=c_{k-1}=c=const$ (8.34)

Из уравнения (8.32) находим тогда, что

$$c = \frac{3-\sqrt{5}}{2} = \frac{2}{3+\sqrt{5}} = 0{,}381966 \qquad (8.35)$$

В этом методе выполняется равенство $\dfrac{z_k - a_k}{z_k - a_{k-1}} = \dfrac{y_k - a_{k-1}}{b_{k-1} - a_{k-1}}$, (8.36)

о котором говорят, что точка z_k делит отрезок $[a_{k-1}, b_{k-1}]$ «в среднем и крайнем отношении». Метод «золотого сечения» обеспечивает линейную сходимость к минимуму $\bar{x}$, т.е. сходимость со скоростью геометрической прогрессии, имеющей знаменатель

$$q = 1 - c = 0{,}618234 \approx 0{,}62\,. \qquad (8.37)$$

Такую сходимость называют линейной или сходимостью первого порядка.

Рассмотрим теперь некоторые методы, не относящиеся к группе симметричных. Предположим, что функция $\Phi(x)$ является выпуклой вниз. В математическом анализе действительную функцию $\Phi(x)$, определенную на некотором интервале $[a;b]$, называют выпуклой вниз, если для любых двух точек x_1, x_2 этого интервала выполняется условие

$$\Phi\left(\frac{x_1 + x_2}{2}\right) \le \frac{\Phi(x_1) + \Phi(x_2)}{2}. \qquad (8.38)$$

Условие (8.38) означает, что выпуклая вниз функция целиком располагается не выше любой хорды AB (рис.43). Если функция $\Phi(x)$ дифференцируема, то для ее выпуклости вниз необходимо и достаточно, чтобы производная $\Phi'(x)$ не убывала на $[a,b]$, а если существует и вторая производная $\Phi''(x)$, то, соответственно, необходимо и достаточно, чтобы $\Phi''(x) > 0$ на $[a,b]$.

Для выпуклых вниз функций применим метод касательных, основанный на постепенной замене графика непрерывно дифференцируемой, выпуклой вниз функции $\Phi(x)$ ломаной линией. Построим касательные к кривой $\Phi(x)$ в точках a и b. Точка пересечения этих касательных x_1 является точкой минимума образованной ими ломаной линии. В этой точке построим новую касательную к кривой $\Phi(x)$ и найдем точку минимума новой ломаной x_2, и т.д. (рис.44). Поскольку минимум ломаной всегда расположен в одной из ее вершин, он легко может быть найден.

Можно доказать, что

$$\lim_{k\to\infty} x_k = \bar{x}\,. \qquad (8.39)$$

Метод касательных можно распространить и на тот случай, когда производная $\Phi'(x)$ не является непрерывной.

Для дважды дифференцируемых функций $\Phi(x)$ применим метод

парабол. Пусть x_k - k-е приближение к минимуму $\bar{x}$. Представим $\Phi(x)$ в окрестности x_k в виде ряда Тейлора

$$\Phi(x) \approx \Phi(x_k) + (x - x_k)\Phi'(x_k) + \frac{1}{2}(x - x_k)^2 \Phi''(x_k) \qquad (8.40)$$

и примем минимум этой параболы в качестве следующего приближения

$$x_{k+1} = x_k - \frac{\Phi'(x_k)}{\Phi''(x_k)} \quad (k = 0,1,2,\ldots). \qquad (8.41)$$

Как видно из этой формулы, метод парабол равносилен решению уравнения

$$\Phi'(x) = 0 \qquad (8.42)$$

методом Ньютона (касательных) и, как и метод Ньютона, имеет сходимость второго порядка, т.е. $\left|x_{k+1} - \bar{x}\right| = O\left(\left|x_k - \bar{x}\right|^2\right)$ (8.43)

Чтобы не вычислять производные сложных функций $\Phi(x)$, в формуле (8.41) могут использоваться конечно-разностные производные:

$$x_{k+1} = x_k - \frac{h}{2} \cdot \frac{\Phi(x_k + h) - \Phi(x_k - h)}{\Phi(x_k + h) - 2\Phi(x_k) + \Phi(x_k + h)} \quad (k = 0,1,\ldots). \qquad (8.44)$$

Здесь h - некоторый шаг, подбираемый экспериментально. Начальное приближение x_0 в этом методе должно быть предварительно задано (выбрано).

8.3. Поиск минимума случайной функции одной переменной.

Положение минимума $\bar{x}$ случайной функции $\Phi(x)$ одной переменной x определим в вероятностном смысле условием:

$$P\left(\left|\Phi(\bar{x}) - \Phi(x)\right|\right) = 1 - \varepsilon \qquad (8.45)$$

при

$$\left\|x - \bar{x}\right\| < \delta. \qquad (8.46)$$

Здесь P означает «вероятность» (probability), а ε и δ - малые положительные числа. Условие (8.45), (8.46) означает, что в малой окрестности точки $\bar{x}$ значительные отклонения случайной функции $\Phi(x)$ от минимума $\Phi(\bar{x})$ возможны, но маловероятны.

Одним из методов поиска минимума случайной функции $\Phi(x)$ может служить, например, итерационный процесс

$$x^{(k+1)} = x^{(k)} - \frac{a_k}{b_k}\left(\Phi(x^{(k)} + b_k) - \Phi(x^{(k)} - b_k)\right) \quad (k = 1,2,\ldots), \qquad (8.47)$$

начинающийся с произвольного начального значения $x^{(1)}$. В (8.47) a_k и b_k - положительные числа, удовлетворяющие условиям:

$$\begin{cases} \lim\limits_{k\to\infty} a_k = 0, \ \lim\limits_{k\to\infty} b_k = 0, \\ \sum\limits_{k=1}^{\infty} a_k = \infty, \ \sum\limits_{k=1}^{\infty}\left(\frac{a_k}{b_k}\right)^2 < \infty. \end{cases} \qquad (8.48)$$

При выполнении условий (8.48) итерационный процесс (8.47) сходится к $\bar{x}$ с вероятностью 1, т.е. «почти всегда»:

$$\operatorname{plim}_{k\to\infty} x^{(k)} = \bar{x}. \qquad (8.49)$$

Здесь *plim* - «вероятностный предел», который означает, что при достаточно больших k вероятность отклонения $x^{(k)}$ от $\bar{x}$ сколь угодно мала.

Одним из вариантов выбора величин a_k , b_k может быть:

$$a_k = \frac{1}{k}, \quad b_k = k^{-\frac{1}{3}}. \qquad (8.50)$$

Сходимость описанного итерационного процесса оказывается очень медленной, так как за каждый шаг значение $x^{(k)}$ очень мало меняется.

8.4Поиск минимума функции нескольких переменных

Поверхность, описываемую уравнением

$$\Phi(x) = const\ ,\ x = (x_1, x_2, \ldots, x_n), \qquad (8.51)$$

называют поверхностью уровня функции $\Phi(x)$. Если функция $\Phi(x)$ однозначная, то через каждую точку $\tilde{x}$, выбранную в области определения функции V, проходит единственная поверхность уровня

$$\Phi(x) = \Phi(\tilde{x}). \qquad (8.52)$$

При $n = 2$ поверхность уровня вырождается в линию уровня, а при $\tilde{x} = \overline{x}$ – в точку $\overline{x}$. Если функция $\Phi(x)$ дифференцируема, то её градиент, который удобно обозначать штрихом,

$$\Phi' = grad\,\Phi(x) = \nabla\Phi(x) = \sum_{i=1}^{n} \frac{\partial \Phi}{\partial x_i} \vec{i}_i \quad (8.53)$$

направлен по нормали к поверхности уровня в точке *x* и указывает направление скорейшего нарастания функции $\Phi(x)$, а антиградиент

$$-\Phi' = -grad\,\Phi(x) = -\nabla\Phi(x) = -\sum_{i=1}^{n} \frac{\partial \Phi}{\partial x_i} \vec{i}_i \qquad (8.54)$$

— направление скорейшего убывания. В формулах (8.53), (8.54) $\vec{i}_i$ — единичный вектор (орт) в направлении x_i. Мы ограничимся случаем, когда все поверхности уровня выпуклые. Выпуклость означает, что тело, границей которого является поверхность уровня, вместе с любыми двумя точками содержит и все точки соединяющего их отрезка прямой.

Множество поверхностей уровня в области V характеризует рельеф функции $\Phi(x)$. Для поиска минимума наиболее благоприятен рельеф типа “котловина”(рис. 45), который могут иметь функции $\Phi(x)$, дифференцируемые всюду в области V, т.е. имеющие градиент (8.53) при любых значениях $x \subset V$. Значительно сложнее поиск минимума, если рельеф имеет тип, называемый оврагом, или “истинным оврагом”, и показанный на рис.46 для функции двух переменных ($n = 2$). В этом случае поверхности уровня имеют линии или точки излома, где не существует градиент $\Phi'(x)$. На рис.46 геометрическое место таких точек AB образует “дно оврага”. При итерационном поиске минимума очередное приближение к минимуму быстро оказывается на дне оврага. Дальнейшее убывание $\Phi(x)$ возможно лишь при движении по дну оврага, а его направление трудно найти из-за отсутствия градиента.

Промежуточным рельефом между котловиной и “оврагом” является так называемый “разрешимый овраг” (рис. 47). В этом рельефе на дне оврага градиент существует, но кривизна поверхности уровня очень велика, т.е. направление градиента быстро меняется. Практически обычно изломы поверхностей уровня слегка сглаживаются вследствие вычислительных погрешностей, и истинный овраг представляется разрешимым, но с очень большой кривизной поверхностей уровня вблизи сглаженных изломов на дне оврага, так что направление градиента плохо определено.

Рельефы, получаемые при изменении на противоположные знаков неравенств на рис.46, 47, называют, соответственно, “гребнем” и “разрешимым гребнем”. Наконец, функции, не являющиеся унимодальными, иногда могут иметь сложный, неупорядоченный рельеф со многими котловинами, “оврагами”, “гребнями” и т.п.(рис.48). В дальнейшем ограничимся лишь унимодальными дифференцируемыми функциями.

Если функция $\Phi(x)$ дважды дифференцируема, то вблизи минимума она может быть с погрешностью $O\left(\|\bar{x}-x\|^3\right)$ аппроксимирована рядом Тейлора:

$$\Phi(x)=\Phi(\bar{x})+\frac{1}{2}\times\sum_{i,j}\frac{\partial^2\Phi(\bar{x})}{\partial x_i\partial x_j}(\bar{x}_i-x_i)(\bar{x}_j-x_j)+O\left(\|x-\bar{x}\|^3\right),\quad(8.55)$$

т.е. квадратичной функцией. В этой формуле сумма представляет собой квадратичную форму, положительную всюду, кроме точки $\bar{x}$, в которой она равна 0, а следовательно матрица

$$\Phi''(\bar{X})=\left(\frac{\partial^2\Phi}{\partial x_i\partial x_j}\right)\Bigg|_{X=\bar{X}}\qquad(8.56)$$

является положительно определенной. Поверхности уровня квадратичной функции (8.55) — n-мерные эллипсоиды (в двухмерном случае — эллипсы) с центром в точке $\bar{x}$. Если при этом одна из главных осей эллипсоида во много раз длиннее остальных, то рельеф функции $\Phi(x)$ имеет тип “разрешимого оврага”.

Для поиска минимума функций нескольких переменных часто используются методы спуска. Простейшим из этих методов является координатный спуск, который строится следующим образом. Пусть $x^{(0)}$ — начальное приближение к моде $\bar{x}$, которое должно быть задано. Зафиксируем координаты $x_2^{(0)}, x_3^{(0)},\ldots,x_n^{(0)}$ и найдём значение $x_1^{(1)}$ как решение задачи

$$x_1^{(1)}=\arg\min\Phi\left(x_1,x_2^{(0)},x_3^{(0)},\ldots,x_n^{(0)}\right)\qquad(8.57)$$

каким-либо методом поиска минимума функции одной переменной, например, методом “золотого сечения”. Геометрически это означает, что минимум отыскивается вдоль прямой, проходящей через точку $x^{(0)}$ параллельно оси x_1, причём $x_1^{(1)}$ — координата точки , в которой эта прямая касается некоторой поверхности уровня (рис. 49).

Далее, аналогично находим $x_2^{(1)}, x_3^{(1)},\ldots,x_n^{(1)}$ как решения задач одномерной минимизации.

$$\begin{cases} x_2^{(1)} = \arg\min \Phi\left(x_1^{(1)}, x_2, x_3^{(0)}, \ldots, x_n^{(0)}\right), \\ x_3^{(1)} = \arg\min \Phi\left(x_1^{(1)}, x_2^{(1)}, x_3, x_4^{(0)} \ldots, x_n^{(0)}\right), \\ \ldots\ldots \\ x_n^{(1)} = \arg\min \Phi\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_{n-1}^{(1)}, x_n\right). \end{cases} \quad (8.58)$$

На этом первая итерация заканчивается. Как видно, итерационный процесс строится как последовательность решений одномерных задач минимизации по отдельным координатам. Далее, аналогично строятся следующие итерации.

Доказано, что в котловине метод координатного спуска сходится к минимуму со скоростью геометрической прогрессии, имеющей знаменатель $q<1$, т.е.

$$\left\|x^{(k)} - \bar{x}\right\| \le q\left\|x^{(k-1)} - \bar{x}\right\| \qquad (0<q<1) \qquad (8.59)$$

(сходимость первого порядка). Однако, во многих случаях (например, если рельеф близок к разрешимому оврагу) знаменатель прогрессии может быть близок к 1:

$$q \approx 1, \qquad (8.60)$$

сходимость оказывается очень медленной, и метод не эффективен.

Если уже выполнено не менее двух итераций, то некоторое ускорение сходимости дальнейших итераций можно получить, используя так называемый "δ^2 - процесс Эйткена". Предположим для этого, что уже известны $x^{(k-2)}, \Phi\left(x^{(k-2)}\right)$, $x^{(k-1)}, \Phi\left(x^{(k-1)}\right)$, $(k=2,3,\ldots)$ и будем считать полученное методом координатного спуска новое значение $\tilde{x}^{(k)}$ предварительным ("предсказанным"). Построим теперь для каждого аргумента x_i $(i=1,2,\ldots,n)$ функции $\Phi(x)$ интерполяционный полином Ньютона (параболу), проходящий через узлы $x^{(k-2)}, \Phi\left(x^{(k-2)}\right)$; $x^{(k-1)}, \Phi\left(x^{(k-1)}\right)$; $\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)$:

$$\begin{cases} \Phi_i(x) \simeq \Phi\left(\tilde{x}^{(k)}\right) + \left(x_i - \tilde{x}_i^{(k)}\right)\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}\right) + \\ \qquad + \left(x_i - \tilde{x}_i^{(k)}\right)\left(x_i - x_i^{(k-1)}\right)\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}, x^{(k-2)}\right). \end{cases} \quad (8.61)$$

Здесь $\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}\right)$, $\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}, x^{(k-2)}\right)$ — разделенные разности:

$$\begin{cases} \Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}\right) = \dfrac{\Phi\left(\tilde{x}^{(k)}\right) - \Phi\left(\tilde{x}^{(k-1)}\right)}{\tilde{x}_i^{(k)} - x_i^{(k-1)}}, \\ \Phi_i\left(x^{(k-1)}, x^{(k-2)}\right) = \dfrac{\Phi\left(x^{(k-1)}\right) - \Phi\left(x^{(k-2)}\right)}{x_i^{(k-1)} - x_i^{(k-2)}}, \\ \Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}, x^{(k-2)}\right) = \dfrac{\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}\right) - \Phi_i\left(x^{(k-1)}, x^{(k-2)}\right)}{\tilde{x}_i^{(k)} - x_i^{(k-2)}}. \end{cases} \quad (8.62)$$

После этого, в качестве уточненного ("исправленного") значения аргумента $x_i^{(k)}$ примем координату минимума параболы (8.61):

$$x_i^{(k)} = \frac{\tilde{x}_i^{(k)} + x_i^{(k-1)}}{2} - \frac{\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}\right)}{2\Phi_i\left(\tilde{x}^{(k)}, x^{(k-1)}, x^{(k-2)}\right)}. \quad (8.63)$$

Однако пользоваться процессом Эйткена следует с осторожностью ввиду некорректности постановки задачи численного дифференцирования. Некорректность в данном случае проявляется в том, что вблизи минимума $\bar{X}$ разделённые разности (8.62) приобретают характер неопределённости вида $0/0$ и вычислительная погрешность становится недопустимо большой.

Рассмотрим теперь метод градиентного спуска, называемый также методом скорейшего спуска. В этом методе очередное приближение к минимуму $x^{(k)}$ находится путём одномерной минимизации функции $\Phi(x)$ вдоль прямой, исходящей из точки $x^{(k-1)}$ в направлении антиградиента (8.54). Уравнение этой прямой в параметрической форме имеет вид

$$x = x^{(k-1)} - \tau \cdot \Phi'\left(x^{(k-1)}\right), \quad (8.64)$$

где τ — параметр. Итерации, следовательно, строятся по формулам

$$\tau^{(k)} = \arg\min \Phi\left(x^{(k-1)} - \tau \cdot \Phi'\left(x^{(k-1)}\right)\right), \quad (8.65)$$

$$x^{(k)} = {}^{(k-1)} - \tau^{(k)} \cdot \Phi'\left(x^{(k-1)}\right) \quad (8.66)$$

или, в компонентах,

$$x_i^{(k)} = x_i^{(k-1)} - \tau^{(k)} \cdot \frac{\partial \Phi\left(x^{(k-1)}\right)}{\partial x_i}. \quad (8.67)$$

Минимум $\Phi(x)$ по-прежнему достигается в точке $x^{(k)}$, где прямая (8.64) касается некоторой поверхности уровня. А поскольку на следующей итерации градиент $\nabla\Phi\left(x^{(k)}\right)$ нормален к поверхности уровня в точке $x^{(k)}$, то спуск в градиентном методе осуществляется по взаимно перпендикулярным направлениям (рис. 50).

Как видно, в методе градиентного спуска на каждой итерации одномерная минимизация выполняется однократно, тогда как в методе координатного спуска требуется n одномерных поисков минимума на каждой итерации. Однако для нахождения градиента требуется вычисление n первых производных функции $\Phi(x)$, которое обычно приходится выполнять с помощью методов численного дифференцирования, что значительно усложняет расчеты.

Метод градиентного спуска, как и метод координатного спуска, имеет в котловине линейную, т.е. первого порядка сходимость к минимуму, причем сходимость оказывается очень медленной, если поверхности уровня имеют большую кривизну.

Метод координатного или градиентного спуска даже для квадратичной функции вида (8.55) требует бесконечного числа итераций. Однако можно построить численный метод, который позволяет получить для квадратичной функции точное значение минимума не более чем за n одномерных минимизаций. Таким методом является метод сопряженный направлений. Основная идея метода — последовательный спуск по сопряженным направлениям, т.е. в направлении сопряженных векторов. Как известно, векторы x, y называются сопряженными относительно некоторой симметричной, положительно определенной матрицы A, если скалярное произведение

$$(x, Ay) = 0. \qquad (8.68)$$

В частности, попарно сопряженными являются собственные векторы x_i, x_j матрицы A, поскольку

$$(x_i, Ax_j) = 0.. \qquad (8.69)$$

Поэтому для квадратичной функции метод сопряженных направлений строится как последовательность спусков по направлениям главных осей эллипсоидов вида (8.51), как показано на рис. 51. В результате n спусков минимум $\bar{x}$ будет найден точно (т.е. с точностью до погрешностей округления). Если же выполнить менее n спусков, минимум будет найден приближенно, но с неплохой погрешностью. Однако поскольку все спуски выполняются приближенно, да и рельеф может не быть точной котловиной, для отыскания главных осей эллипсоидов необходимо неоднократно численно решать сложную полную проблему собственных значений (глава 5).

Практически вычисления могут строиться как последовательность итераций вида

$$x^{(k+1)} = x^{(k)} - r_k y^{(k)} \qquad (k = 0,1,2,\ldots), (8.70)$$

где $x^{(0)}$ — заранее выбранное начальное приближение, r_k — параметры, $y^{(0)}$ — градиент:

$$y^{(0)} = \Phi'\left(x^{(0)}\right). \qquad (8.71)$$

Следующие итерации находятся по формуле

$$y^{(k)} = \Phi'\left(x^{(k)}\right) - p_k y^{(k-1)} \qquad (k = 0,1,2,\ldots), (8.72)$$

а значения r_k — из условия:

$$r_k = \arg\min \Phi\left(x^{(k)} - r\, y^{(k)}\right). \quad (8.73)$$

Если все $p_k = 0$ $(k = 0,1,2,\ldots)$, получается метод наискорейшего спуска. Различные варианты метода сопряженных направлений отличаются выбором параметров p_k. Один из способов выбора:

$$p_k = \left\|\Phi'\left(x^{(k)}\right)\right\|^2 \Big/ \left\|\Phi'\left(x^{(k-1)}\right)\right\|^2. \qquad (8.74)$$

Поскольку на каждой итерации метода требуется вычисление градиента Φ', решение замедляется, иногда очень значительно.

Для овражного рельефа возможно применение "метода оврагов", иногда называемого также "блужданием по оврагам". В этом методе произвольно выбираются две близкие начальные точки $y^{(0)}$, $y^{(1)}$ и из каждой строятся последовательности координатных спусков, которые быстро заканчиваются, соответственно, в точках, $x^{(0)}$, $x^{(1)}$ вблизи дна оврага (рис. 52). Через точки $x^{(0)}, x^{(1)}$ проводится прямая, вдоль которой осуществляется продвижение приблизительно по дну оврага в сторону убывания функции $\Phi(x)$ на некоторое расстояние h_1 ("овражный шаг") до точки

$$y^{(2)} = x^{(1)} + sign\left(\Phi\left(x^{(0)}\right) - \Phi\left(x^{(1)}\right)\right)\left(x^{(1)} - x^{(0)}\right)h_1 \quad (8.75)$$

Из точки $y^{(2)}$ строиться координатный спуск до точки $x^{(2)}$ вблизи дна оврага. Затем вдоль прямой, проходящей через точки $x^{(1)}$ и $x^{(2)}$, выполняется аналогичное продвижение с новым шагом h_2 до точки $y^{(3)}$ и т.д. Если убывание $\Phi(x)$ прекращается, то достигнут минимум. Метод позволяет "блуждать" и по изогнутым оврагам, но шаги h_1, h_2, … часто приходится подбирать экспериментально в ходе счета. Иногда "блуждание" выводит к котловине, где могут использоваться более простые методы, например, координатный спуск.

Отметим еще "случайный координатный спуск" со случайным выбором осей координат, по которым осуществляется очередной спуск. Обычно мощные программы оптимизации (поиска минимума) автоматически меняют в ходе счета многие методы поиска для достижения наибольшей эффективности.

8.5 Оптимизация при наличии ограничений.

Часто встречаются задачи оптимизации (поиска минимума) при наличии ограничений типа равенств:

$$\varphi_i(x) = 0 \qquad (1 \le i \le m) \qquad (8.76)$$

и (или) неравенств

$$\psi_k(x) \ge 0. \qquad (1 \le k \le p) \qquad (8.77)$$

Ограничения типа равенств (8.76) задают некоторые дополнительные связи между переменными x_j $(j = 1, 2, \ldots, n)$, а неравенства (8.77) ограничивают область поиска минимума X. Теоретически можно с помощью равенств (8.76) исключить часть переменных и уменьшить размерность области X, но практически это обычно невозможно из-за сложного вида функций $\varphi_i(x)$.

Распространенным методом оптимизации с ограничениями является метод штрафных функций. В этом методе строится вспомогательная функция

$$\mathrm{F}(x)=\Phi(x)+\left[\mu\sum_{i=1}^{m}\varphi_i^2(x)+\lambda\sum_{k=1}^{p}\left(1-sign\psi_k(x)\right)\psi_k^2(x)\right] \quad (8.78)$$

где $\mu>0$, $\lambda>0$, и отыскивается минимум этой функции. Если все ограничения удовлетворяются, то в формуле (8.78) член в квадратных скобках обращается в 0 и задача сводится к минимизации исходной функции $\Phi(x)$. Если же ограничения не удовлетворяются, то $\mathrm{F}(x)>\Phi(x)$ на некоторую величину, которую можно трактовать как "штраф" за невыполнение ограничений.

Величина штрафа регулируется параметрами μ, λ, правильный выбор которых представляет собой достаточно сложную проблему. При малых значениях μ, λ штраф мал и ограничения плохо выполняются. При больших значениях μ, λ штраф велик, но найденный минимум функции $\mathrm{F}(x)$ может сильно отличаться от искомого минимума $\bar{x}$ функции $\Phi(x)$. Кроме того, при больших значениях μ, λ может сильно ухудшаться рельеф функции $\mathrm{F}(x)$, приобретая овражный или неупорядоченный характер. Поэтому для практического использования метода штрафных функций может потребоваться проведение большого объема тестовых численных экспериментов.

Особую задачу составляет поиск минимума линейной функции многих переменных

$$\Phi(x)=\sum_{i=1}^{n}c_i x_i=\min \quad (8.79)$$

с линейными ограничениями типа равенств и неравенств:

$$x_j\ge 0 \qquad (1\le j\le n), \quad (8.80)$$

$$\sum_{j=1}^{n}a_{ij}x_j=b_i \qquad (1\le i\le m), \quad (8.81)$$

$$\sum_{j=1}^{n}a_{ij}x_j\le b_i \qquad (m\le i\le M), \quad (8.82)$$

где c_i, a_{ij}, b_i — параметры задачи.

Решение задачи (8.79) — (8.82) составляет типичную проблему операционного раздела прикладной математики — линейного программирования. Область, ограниченная неравенствами (8.80), (8.82) — выпуклый многогранник в n - мерном пространстве (симплекс), одна из вершин которого расположена в начале координат. Минимум линейной функции $\Phi(x)$ обязательно находится в одной из вершин многогранника, число которых очень велико и составляет $O(2^N)$, где N — общее число неравенств в

формулах $(8.80),(8.82)$. Поскольку значение N может достигать 100÷1000 и более, поиск минимума путем простого сравнения значений $\Phi(x)$ во всех вершинах практически неосуществим, и поэтому организуется спуск от некоторой исходной ("опорной") вершины (которую также надо найти) по цепочке смежных ребер в сторону убывания функции $\Phi(x)$ до достижения ею минимума — так называемый "симплекс-метод". Метод программно реализован в большинстве современных математических пакетов прикладных программ.

9. ЧИСЛЕННЫЕ МЕТОДЫ МОНТЕ-КАРЛО.

9.1. Генерирование случайных чисел с заданным законом распределения.

Численными методами Монте-Карло называются методы, основанные на систематическом использовании случайных чисел. Различают "истинно случайные" числа, получаемые с помощью каких-либо физических датчиков, например, шумовых генераторов, и "псевдослучайные числа", которые вырабатываются специальными программами, т.е. фактически не являются случайными, но по своим статистическим характеристикам очень похожи на случайные. Достоинство физических датчиков — неограниченный запас случайных чисел и высокая скорость их генерации, но их серьезные недостатки — ненадежность (датчики нуждаются в регулярной проверке "случайности" генерируемых чисел) и невоспроизводимость результатов при повторных просчетах одного и того же варианта задачи (результаты совпадают не точно, а лишь в статистическом смысле). Поэтому общепринято использовать генераторы псевдослучайных чисел.

Основой всех генераторов случайных чисел являются генераторы чисел, равномерно распределенных (равновероятных) в интервале $(0,1)$. Такие числа будем обозначать R. Для начала работы генератора задается некоторое произвольное число C_0. Далее числа находятся по рекуррентным формулам. Последовательность генерируемых чисел имеет вид {A, B, B, …}, где A — отрезок апериодичности, а B — период. В расчетах рекомендуется использовать лишь отрезок {A, B}, который должен содержать много миллионов чисел. При необходимости можно получить новый отрезок {A, B}, обращаясь к генератору повторно с другим начальным числом C_0. Построение хорошего генератора псевдослучайных чисел, равновероятных в $(0,1)$ — серьезная вычислительная проблема. Практически все современные компьютеры, языки программирования и пакеты прикладных программ имеют в библиотеках стандартных программ хорошо проверенные генераторы таких чисел. Если же такой встроенный генератор отсутствует, можно воспользоваться одним из генераторов, опубликованных в литературе. Неплохие результаты дает, например, простой рекуррентный генератор, использующий формулу

$$R_j = \{13R_{j-1} + \pi\} \qquad (j = 1,2,\ldots), \qquad (9.1)$$

где j — номер случайного числа, фигурные скобки означают дробную часть числа, а в качестве R_0 можно принять произвольное число в интервале $(0,1)$.

Все остальные генераторы строятся с помощью генератора равномерно распределенных чисел. Существуют три метода генерирования псевдослучайных чисел (далее для краткости называемых случайными):

1) метод обратной функции,
2) метод отбора,
3) специальные методы.

В методе обратной функции случайные числа X, имеющие функцию распределения $F(x)$, вычисляются по формуле

$$X = F^{-1}(R), \tag{9.2}$$

где F^{-1} — обратная функция (рис. 53). Например, для экспоненциального распределения

$$F(x) = 1 - e^{-\lambda * x} \qquad (\lambda > 0, x \geq 0), \tag{9.3}$$

обратная функция

$$F^{-1}(F) = -\frac{1}{\lambda}\ln(1-F), \tag{9.4}$$

и, следовательно, случайное число

$$X = -\frac{1}{\lambda}\ln(1-R) \tag{9.5}$$

имеет распределение (9.3). Достоинство метода обратной функции — высокая скорость (по одному значению R можно вычислить одно значение X), но практически им удобно пользоваться, когда для F и F^{-1} можно получить достаточно простые формулы.

Метод отбора является практически универсальным. Для его использования необходимо знать лишь плотность распределения $f(x)$ случайной величины X и ее максимальное значение $f_{\max}$. Для типовых распределений, встречающихся в физике, всегда известно положение x_M максимума $f(x)$ (“мода”), а следовательно и $f_{\max} = f(x_M)$. Пусть $[x_1, x_2]$ — область определения $f(x)$; в качестве $[x_1, x_2]$ можно принять область, за пределами которой $f(x)$ достаточно мала, например, меньше (0.01÷0.001) $f_{\max}$. В методе отбора последовательно строятся случайные точки (X, Y), равномерно распределенные в прямоугольнике $(x_1 \leq X \leq x_2; 0 \leq Y \leq f_{\max})$, по формулам

$$X = x_1 + (x_2 - x_1)R_1, \tag{9.6}$$

$$Y = f_{\max} R_2, \tag{9.7}$$

где R_1, R_2 — случайные числа, равновероятные в интервале $(0,1)$. Если окажется, что

$$Y \leq f(X), \tag{9.8}$$

то это “полезная точка” и найденная величина X (9.6) имеет плотность распределения $f(x)$; иначе точка (X, Y) считается “бесполезной”, ее надо отбросить и перейти к генерации новой точки по формулам (9.6),(9.7) (рис. 54). При этом для отбора каждой “полезной точки” может потребоваться много испытаний (9.6)–(9.8).

Наконец, специальные методы генерирования случайных величин обычно основаны на использовании асимптотических формул теории вероятностей и преобразованиях (функциях) случайных величин. Отметим прежде всего, что если случайная величина X имеет нулевое математическое ожидание и единичную дисперсию, то случайная величина

$$Y = a + \sigma X \tag{9.9}$$

имеет математическое ожидание a и дисперсию σ^2. Для генерирования случайной величины X, имеющей стандартное нормальное распределение с плотностью вероятности

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \tag{9.10}$$

с нулевым математическим ожиданием и единичной дисперсией, обычно обозначаемое $N(0,1)$, можно, например, воспользоваться точной формулой

$$X = \sqrt{-2\ln(R_1)} * \cos(2\pi R_2), \tag{9.11}$$

а любое нормально распределенное число вычислить затем по формуле (9.9). Для вычисления случайной величины X, имеющей распределение $N(0,1)$ часто рекомендуется также пользоваться приближенной формулой

$$X = \sum_{j=1}^{12} R_j - 6, \qquad (9.12)$$

основанной на центральной предельной теореме теории вероятностей и имеющей очень малую погрешность. Недостаток генератора (9.12) — ограниченность диапазона случайных чисел: $-6 \le X \le 6$, что иногда нежелательно.

Используя выборки случайных величин с распределением $N(0,1)$ можно затем построить случайные величины, имеющие распределения χ ("хи"), χ^2 ("хи-квадрат"), F (Фишера), t (Стьюдента) и некоторые другие.

Существенно сложнее моделирование многомерных случайных величин $X = (X_1, X_2, \ldots, X_n)$. Некоторые возможности моделирования таких величин обсуждаются в литературе (см., например, [9]).

9.2. Вычисление многомерных интегралов методом Монте-Карло.

Рассмотрим задачу вычисления n-мерного интеграла

$$I = \int_{(V)} f(x)dx, \qquad X = (x_1, x_2, \ldots, x_n),$$

$$dx = dx_1 dx_2 \ldots dx_n \qquad (9.13)$$

в области V с границей Γ, вложенной в n-мерный параллелепипед

$$W = \{a_i \le x_i \le b_i; i = 1,2,\ldots,n\}, \qquad (9.14)$$

(рис. 55), имеющий объем

$$|W| = (b_1 - a_1)(b_2 - a_2)\ldots(b_n - a_n). \qquad (9.15)$$

По теореме о среднем в интегральном исчислении среднее значение $\bar{f}$ функции $f(x)$ по области V дается равенством

$$\bar{f} = \frac{1}{|V|}\int_{(V)} f(x)dx, \qquad (9.16)$$

где $|V|$ — объем области V. Для вычисления $\bar{f}$ с помощью генератора случайных чисел R, равновероятных на $(0,1)$, будем генерировать случайные точки $X^{(k)}$ с координатами

$$X_i^{(k)} = a_i + (b_i - a_i) * R_i^{(k)},$$
$$(i = 1,2,\ldots,n;\ k = 1,2,\ldots,n), \qquad (9.17)$$

равномерно распределенные в параллелепипеде W. Значение $\bar{f}$ находится усреднением $f(x)$ по точкам $X^{(k)}$, принадлежащим области V:

$$\bar{f} = \frac{1}{M}\sum_{X^{(k)} \in V} f\left(X^{(k)}\right), \qquad (9.18)$$

где M — число точек $X^{(k)}$, принадлежащих V.

Объем $|V|$ области V можно приближенно оценить по формуле

$$|V| \approx \frac{M}{N}|W|, \qquad (9.19)$$

где N — общее число случайных точек $X^{(k)}$ в W, а $|W|$ определяется формулой (9.15). Используя формулы (9.13), (9.15), (9.16), (9.18), (9.19) находим оценку интеграла I:

$$I = |V|\,\bar{f} \approx \frac{1}{N}|W| \sum_{X^{(k)} \in V} f\left(X^{(k)}\right). \qquad (9.20)$$

Принадлежность точки $X^{(k)}$ области V можно устанавливать по заданной границе Γ. Пусть, например, граница Γ имеет уравнение

$$\Gamma(x) = 0, \qquad (9.21)$$

и точки x, удовлетворяющие условию

$$\Gamma(x) \le 0, \qquad (9.22)$$

принадлежат области V. Тогда в сумму (9.20) следует включить все точки $X^{(k)}$, удовлетворяющие условию

$$\Gamma\left(X^{(k)}\right) \le 0. \qquad (9.23)$$

Недостаток методов Монте-Карло — невысокая точность $O\left(N^{-1/2}\right)$ и медленная сходимость по числу случайных точек N (для снижения погрешности на порядок следует увеличить число случайных точек N на два порядка). Тем не менее методы Монте-Карло экономичнее обычных методов численного интегрирования (например, методов с регулярно расположенными узлами интегрирования) уже при $n > 3$ и становятся единственно применимыми при $n > 6$.

Причина большой погрешности методов Монте-Карло — плохая равномерность расположения случайных точек $X^{(k)}$, которая особенно проявляется с ростом размерности области n. Показано, что точность и скорость сходимости методов Монте-Карло можно существенно, приблизительно до $O\left(N^{-1}\right)$, улучшить, используя неслучайные точки, генерируемые специальными программными датчиками [9] и с высокой степенью равномерности расположенные в многомерной области (так называемые $ËÏ_{\tau}$-точки, $ËÏ_{\tau}$-генераторы). Использование таких генераторов позволяет получать удовлетворительные результаты интегрирования даже в областях с размерностью n, составляющей несколько десятков.

9.3. Поиск минимума функции многих переменных.

В главе 8 рассматривались методы оптимизации (поиска минимума) функции нескольких переменных, наиболее распространенными из которых являются различные варианты методов спуска. Однако рельеф оптимизируемой функции, как правило, неизвестен, анализируется крайне сложно и зачастую оказывается очень неблагоприятным для применения этих методов. Ввиду недостатков, присущих методам спуска в областях со сложным рельефом для поиска минимума функции многих переменных часто более надежными оказываются методы прямого поиска, относящиеся к классу численных методов Монте-Карло.

В простейшей постановке метод прямого поиска минимума реализуется следующим образом. В области поиска минимума V выбирается некоторое число $N \gg 1$ случайных, равномерно распределенных точек $X^{(k)}$ $\left(k = 1, 2, \ldots, N\right)$. Отбор точек, принадлежащих V, можно производить так же, как в задаче численного интегрирования (§ 9.2). Во всех точках вычисляются значения оптимизируемой функции $\Phi\left(X^{(k)}\right)$, и в качестве оценки минимума $\tilde{x}$ принимается наименьшее из этих значений, т.е. полагается

$$\bar{x} \simeq \tilde{x} = \arg\min_k \Phi\left(X^{(k)}\right). \qquad (9.24)$$

Найденное таким образом значение $\bar{x}$ можно уточнить. Для этого будем полагать, что оно является первым приближением к минимуму:

$$\bar{x}^{(1)} = \arg\min_k \Phi\left(X^{(k)}\right). \qquad (9.25)$$

Далее в некоторой окрестности точки $\bar{x}^{(1)}$ снова выбирается N случайных точек, и процесс прямого поиска повторяется, приводя к значению $\bar{x}^{(2)}$ и т.д.

Достоинство методов прямого поиска — применимость к любым, сколь угодно сложным рельефам. Однако число точек N при этом должно быть очень велико. Например, для того, чтобы в длинном и узком овраге оказалось достаточное для поиска минимума число точек, общее число случайных точек N должно быть очень большим. Многое здесь зависит от качества генератора

равномерно распределенных случайных точек.

Доказано, что погрешность поиска минимума описанным методом пропорциональна $N^{-1/2}$:

$$\left\|\tilde{x}-\bar{x}\right\| \le const \cdot N^{-1/2}. \qquad (9.26)$$

Такая сходимость, как уже указывалось, является слишком медленной. Как и в методе численного интегрирования, существенного ускорения сходимости здесь можно достигнуть, используя неслучайные $ËÏ_\tau$-точки, генерируемые $ËÏ_\tau$-генераторами.

Отметим еще, что методы прямого поиска иногда комбинируются с методами спуска. В таких комбинированных методах начальное приближение к минимуму $\bar{x}^{(0)}$ находится методом прямого поиска, а последующие приближения — методом спуска.

10. МЕТОД КРУПНЫХ ЧАСТИЦ

10.1. Модель крупных частиц

Одним из наиболее распространенных методов решения задач механики сплошной среды, физики плазмы, физической электроники, ставшим в последнее время классическим, является рассматриваемый в этой главе метод крупных частиц (макрочастиц). Для описания движения сплошной среды, как известно, используются подход Эйлера или подход Лагранжа. В подходе Эйлера за сплошной средой наблюдают в определенных точках так называемого фазового пространства, которым может быть пространство координат, координат-скоростей, координат-импульсов, координат-скоростей-температур и т.п. В фазовом пространстве вводится неподвижная расчетная сетка (эйлерова сетка) и точки наблюдения выбираются в узлах этой сетки. Координаты точек наблюдения (узлов сетки) являются независимыми фазовыми переменными (эйлеровыми переменными) задачи. С течением времени через точки наблюдения проходят различные частицы среды. Примером подхода Эйлера является численное решение уравнения переноса конечно-разностными методами, описанное в §7.2.

В подходе Лагранжа следят за траекториями различных выбранных частиц в фазовом пространстве. Положения частиц в текущий момент времени рассматриваются как узлы подвижной (лагранжевой) сетки. Множество всех фазовых траекторий образует фазовый портрет ансамбля частиц и также дает исчерпывающую информацию о процессе. Независимыми переменными в лагранжевом подходе (лагранжевыми переменными) могут служить номера частиц или их начальные фазовые положения. При этом предполагается, что начальные и текущие фазовые положения частиц взаимно-однозначно связаны уравнением траекторий. Примером лагранжева подхода в физических задачах является описанная в этой главе модель (метод) крупных частиц.

Хотя оба подхода эквивалентны в отношении полноты описания процессов, для практических расчетов может оказаться предпочтительным тот или другой подход в зависимости от характера задачи и возможностей компьютера. Используя рассчитанные траектории частиц, всегда можно преобразовать любую величину от переменных Лагранжа к переменным Эйлера и наоборот. Оба подхода часто сочетаются в расчетах. Например, в методе крупных частиц применительно к задачам физики плазмы и физической электроники за движением частиц наблюдают в переменных Лагранжа, тогда как создаваемые частицами плотности заряда и тока, а также поля, в которых движутся частицы, определяют на неподвижной сетке, т.е. в переменных Эйлера.

Для получения достаточно полной информации о процессе приходится следить за большим числом частиц, статистически усредняя результаты наблюдения. Начальные положения частиц, т.е. исходные точки фазовых траекторий, во многих моделях выбираются случайным образом, в

соответствии с заданной статистикой. Поэтому часто говорят о статистическом моделировании, или о моделировании по методу Монте-Карло. В первых работах по статистическому моделированию заряженных потоков на компьютере, модельные частицы еще не были “крупными”, и для получения нормальной плотности пространственного заряда приходилось рассматривать области очень маленького объема. Это затрудняло получение практически полезных результатов.

В широко используемом в физике так называемом “дебаевском приближении” усредненное макроскопическое поле определяет движение сразу многих частиц. Это естественно приводит к идее наблюдать не за отдельными частицами ансамбля, а за большими группами близко расположенных в фазовом пространстве частиц, объединенных в “крупные частицы” (“макрочастицы”) – сгустки частиц определенной формы, зависящей от размерности и свойств симметрии исследуемой математической модели физической системы. Частицы системы могут быть нескольких сортов: электроны, ионы различной массы и зарядности, нейтральные атомы и молекулы разной массы.

Все множество частиц считается состоящим из нескольких компонент, которые могут быть и односортными, например, “медленные электроны” (электроны плазмы с тепловыми скоростями) и “быстрые электроны” (электроны ускоренного пучка). В начальный момент времени фазовый объем каждой компоненты ансамбля разбивается на некоторое количество непересекающихся элементарных объемов (ячеек) и движение каждого такого объема отождествляется с движением какой-либо одной его частицы с суммарными зарядом и массой. Получаемые таким образом модельные частицы, или макрочастицы будем в дальнейшем для краткости называть частицами (электронами, ионами), в отличие от физических частиц (электронов, протонов, ионов). Как видно из способа разбиения, частицы в общем случае могут иметь различный коэффициент укрупнения.

Первые модели частиц строились в значительной степени эмпирически. В дальнейшем были предложены некоторые общие подходы, позволяющие строить модели с нужными свойствами, например, оптимальные по некоторым критериям, с заданными законами сохранения, с уменьшенными нефизическими, счетными эффектами или с использованием вычислительных алгоритмов, допускающих наглядную физическую интерпретацию.

Начнем рассмотрение с некоторой математической модели системы объема V, в котором содержится $N>>>1$ частиц одного сорта с зарядом MZe (Z-зарядовое число, e-абсолютное значение заряда электрона), массой покоя Mm_0 и с тем же отношением заряда к массе покоя $\eta = Ze/m_0$, что и для физических частиц. Коэффициент укрупнения $M>>>1$ выбирается таким, что суммарный заряд всех частиц $MNZe$ и средняя плотность пространственного заряда

$$\bar{\rho} = MNZe/V \qquad (10.1)$$

в модели те же, что и в моделируемой физической системе. Благодаря сохранению значения η в модели сохраняются и уравнения движения частиц. Поэтому, с учетом сохранения плотности заряда, собственное поле и динамика потока в модели и в физической системе должны быть близки.

Число учитываемых в модели координат обозначим dr, а число учитываемых компонент импульса d_p. Размерностью модели часто называют

$$d=0{,}5(d_r+d_p), \qquad (10.2)$$

так что встречаются модели с размерностью от d=0,5 до d=3. В трехмерном пространстве (d_r=3) частицы представляют собой точки (рис. 56, а); в двухмерных (d_r=2) декартовых x,y или цилиндрических r,θ координат – отрезки прямых (“стержни”, рис. 56, б и 56, в); в цилиндрических координатах r,z – кольца переменного радиуса (рис. 56, г); в одномерном случае – плоскости (“модель плоских листов”, рис. 56, д), диски (“дисковая модель”, рис. 56, е), коаксиальные цилиндрические (рис. 56, ж) или концентрические

сферические (рис. 56, з) поверхности. Все частицы имеют нулевой собственный объем.

Некоторые физические характеристики ансамбля заряженных частиц при их укрупнении не меняются, например средняя скорость v_0, среднеквадратичная тепловая скорость $v_{\dot{O}}$, плазменная частота ω_p, циклотронная частота, дебаевское расстояние λ_D, тогда как другие величины изменяются. Так кинетическая температура T возрастает в M раз вследствие того, что "массивные" крупные частицы имеют тот же разброс скоростей, что и физические частицы. Длина свободного пробега $l_c \sim 1/M$, частота парных столкновений $\nu_c \sim M$, сечение столкновений $\sigma_{\tilde{n}} \sim M^2$, среднее расстояние между частицами $s_0 \sim M^{1/d_r}$, сечение парных столкновений $\sigma_c \sim M^2$, дебаевское число (число частиц в дебаевском объеме $V_D = \lambda_D^3$) $n_D \sim 1/M$. Сохранение динамики частиц, плазменной частоты и дебаевского расстояния указывает, что в моделях частиц могут правильно воспроизводиться коллективные явления (плазменные колебания, затухание Ландау, крупномасштабные возмущения). Однако сокращение длины свободного пробега l_c означает, что в модели резко искажаются парные взаимодействия. Соответственно, относительные среднеквадратичные отклонения (флуктуации) плотности частиц, плотности заряда и собственного поля, пропорциональные $1/n_D^{1/2}$, оказываются завышенными в моделях в отношении $M^{1/2}$. В отношении $M^{1/2}$ оказывается также завышенным дробовой шум – флуктуации тока эмиссии или тока частиц на электроды.

Снизить искажение парных взаимодействий можно, увеличивая число частиц N, но этому препятствует ограниченность ресурсов компьютера. Действительно, описание ограниченной системы частиц средствами статистической физики справедливо, если объем системы V существенно превышает дебаевский объем $V_D = \lambda_D^3$. Отсюда следует требование $N >> n_D >> 1$, весьма обременительное для компьютера. Поэтому путь увеличения N ведет лишь к временным успехам, хотя современные суперкомпьютеры допускают значения N до 10^9. При моделировании бесстолкновительных систем для сохранения в модели необходимого "запаса бесстолкновительности" следует резко уменьшить ближние (парные) взаимодействия. Способы, применяемые для этого, можно трактовать как переход к частицам конечного собственного объема. В этой трактовке каждая частица с номером α $(\alpha = 1,2,...,N)$ представляет собой некоторое облако объемного заряда с центром в точке $\boldsymbol{r}_\alpha$, где $\boldsymbol{r}$ – радиус-вектор, и тем или иным распределением плотности. Изучаемый ансамбль физических частиц воспроизводится, таким образом, с помощью модели "частиц-облаков".

Качественно ясно, что при сближении удаленных облаков сила взаимодействия между ними сначала возрастает, а затем, с началом взаимопроникновения облаков, слабеет. При полном совпадении облака не взаимодействуют. Это означает уменьшение ближних воздействий, т.е. столкновительных эффектов. С другой стороны, усечение ближних взаимодействий приводит к уменьшению высших пространственных гармоник, т.е. к сглаживанию исследуемых процессов. Следовательно, в модели будут неправильно воспроизводиться (сглаживаться) мелкомасштабные (коротковолновые) явления. Для ряда таких моделей облаков аналитически получены качественные частотные и пространственные характеристики, дисперсионные уравнения, оценены столкновительные эффекты, хорошо совпадающие с результатами численных экспериментов. В случае частиц конечного размера роль дебаевского числа играет число перекрытий облаков n_q. Естественно, чем больше протяженность облака, тем легче удовлетворить условию $n_q >> 1$ при том же количестве частиц N.

Сглаживание ближних взаимодействий обычно сводится к линейному интегральному преобразованию плотности частиц и плотности тока. Как правило, при этом используется симметричное или вырожденное ядро. Способ сглаживания (усреднения) составляет одну из важнейших характеристик модели. Многие модели частиц различаются именно способом этого усреднения. Метод частиц распространяется и на многокомпонентные ансамбли. Коэффициент укрупнения M_a при этом зависит от сорта частиц a. Иногда в некоторых моделях возникает необходимость изменения реального отношения масс частиц различных сортов (например, ионов и электронов) для ускорения вычислений (иначе ионы движутся слишком медленно). В таких моделях возможно, конечно, получение лишь качественных оценок.

Модели частиц, особенно многомерные ($d \geq 2$), требуют, как правило, предельной загрузки мощностей компьютеров из-за большого количества частиц и узлов сетки, используемой для вычисления собственного поля. Несмотря на это, в многомерных моделях достигаются относительно низкие (по сравнению с одномерными) плотности узлов и частиц в рассматриваемой области и небольшие значения дебаевского числа n_D. Успех метода в этих условиях обеспечивается применением быстрых и достаточно точных алгоритмов решения уравнений математической модели. Например, при использовании сеточных методов решения аналогичных задач требуется много узлов в фазовом пространстве координат-импульсов для достижения приемлемой погрешности, тогда как в методе частиц сетка вводится лишь в координатном пространстве, но требуется много частиц. Если, однако, не интересоваться деталями распределения частиц по скоростям и ограничиться моментами распределения не выше второго (т.е. оценивать лишь плотность частиц, среднюю скорость, дисперсию скоростей и температуру), то может оказаться достаточным приемлемое для компьютера количество частиц. Поэтому в настоящее время методы частиц позволяют изучать более широкий круг физических процессов, особенно многомерных, чем сеточные методы.

10.2. Методы моделирования

Типичная система уравнений математической модели частиц состоит из уравнений Максвелла, уравнений среды

$$\boldsymbol{D}=\varepsilon \boldsymbol{E},\ \boldsymbol{B}=\mu \boldsymbol{H}, \qquad (10.3)$$

связывающих напряженности электрического $\boldsymbol{E}$ и магнитного $\boldsymbol{H}$ поля с их индукциями $\boldsymbol{D}$, $\boldsymbol{B}$, и уравнений движения частиц. Эта система является квазилинейной, интегро-дифференциальной, поскольку действующее на частицы поле зависит от распределения плотностей заряда и тока, которые сами определяются интегральными зависимостями от управляемой полем динамики частиц.

При самосогласованном решении уравнений модели на каждом шаге в текущий момент времени t сначала находятся макроскопические (сглаженные) плотность заряда ρ и (или) плотность тока $\boldsymbol{j}$, входящие в уравнения Максвелла. Для частиц, поступающих на шаге моделирования τ в рассматриваемую область, необходимо предварительно воспроизвести их начальное распределение (точнее, распределение центра облаков) в фазовом пространстве в соответствии с известной (заданной) статистикой. При этом могут использоваться рассмотренные в §9.1 методы моделирования случайных величин с заданными законами распределения. После этого численно решаются уравнения Максвелла (Пуассона) одним из методов, данных в главе 7. Решение, как правило, находится в узлах некоторой пространственной сетки. Для численного интегрирования уравнений движения необходимо вычисление поля в промежуточных точках, где располагаются частицы (центры частиц). При этом используются различные методы интерполирования или численного дифференцирования сеточных функций, обычно

также со сглаживанием (главы 1, 4 и §10.3). Далее из дифференциальных уравнений движения численно находится (методами главы 6) расположение частиц в следующий момент времени t+τ и т.д.

Количество частиц в модели N(t) с течением времени меняется в следствие поступления (эмиссии, инжекции) и ухода частиц через границы области, а также действия источников, описывающих процессы ионизации м рекомбинации. Эти процессы необходимо, по возможности адекватно, воспроизводить. В ходе моделирования наблюдают за величиной N(t). Если на текущем шаге значение N(t) оказывается больше некоторого $N_{\max}$, определяемого емкостью памяти компьютера, вычисления приходится прекращать. (Их надо повторить сначала, увеличив коэффициент укрупнения M). При N(t), меньших некоторого $N_{\min}$, также целесообразно прекращение расчетов. Для оценки значения $N_{\min}$ может, например, служить неравенство $n_q >> 1$.

Результаты численного эксперимента методом частиц подвергаются такой же статистической обработке, как и результаты физического эксперимента. Получаемые результаты можно разделить на две группы: функции и поля. Под функциями понимаются процессы, зависящие только от времени: различные токи, мощности, суммарный заряд и др. К полям относятся процессы, зависящие от времени и координат: распределение в области потенциалов, плотностей, скоростей и т.п. Численный эксперимент позволяет исследовать не только функции, но и поля, хотя, конечно, в ограниченной степени из-за огромного объема информации, относящейся к полям.

Основными параметрами математической модели частиц являются количество частиц N (или, что то же самое, коэффициент укрупнения M), шаг по времени τ и шаг пространственной сетки h (или набор шагов). Эти параметры определяют дискретизацию заряда, пространства и времени и должны задаваться оптимальным, взаимосвязанным образом, обеспечивая возможность воспроизведения исследуемых характеристик процесса с заданной погрешностью при минимальных затратах машинного времени.

При использовании сеточных методов вычисления поля объемного заряда длительность счета практически линейно возрастает с увеличением числа временных шагов и числа частиц. Число частиц *N* должно обеспечить выполнение условия $n_q >> 1$ и ограничивается лишь ресурсами компьютера. Большое число частиц позволяет полнее исследовать многие важные характеристики процессов: динамику и конфигурацию пучков частиц, диффузию, токооседание, моменты функций распределения второго или даже более высокого порядка и др.

Оптимальный шаг по времени τ обычно выбирается из условий точности решения уравнений движения частиц и составляет некоторую часть, порядка 1/8 – 1/50 характерного временного периода, например плазменного, циклотронного или внешнего поля. Можно также выбирать τ из условий удовлетворительной воспроизводимости моделируемых статистических характеристик (первых и вторых моментов распределения). Так, если в численном эксперименте вычисляется корреляционная функция некоторого процесса, то можно с помощью тестов выбрать шаг τ, обеспечивающий приемлемую погрешность этих вычислений.

Шаг пространственной сетки *h* целесообразно выбирать существенно меньшим дебаевского расстояния λ_D и большим среднего расстояния между частицами s_0. При этом само дебаевское расстояние λ_D находится с помощью предварительных тестовых просчетов. (Точнее, из тестов находятся средняя плотность объемного заряда ρ_0, связанная с плазменной частотой ω_p известной формулой $\omega_p^2 = \eta\rho_0 / \varepsilon$, и среднеквадратичная тепловая скорость v_T, а затем

вычисляется $\lambda_D = v_T / \omega_p$). В многомерных моделях часто трудно удовлетворить условию h<< λ_D и шаг сетки выбирается порядка $h \sim \lambda_D$. В общем случае значения N, τ, h должны выбираться взаимосвязано.

В некоторых моделях коэффициент M_α может зависеть от номера частицы α. Например, в задачах, в которых важно хорошее воспроизведение относительно небольшой группы высокоскоростных частиц, им можно приписать меньший вес M_α, чтобы таких частиц было больше. При этом возможно изменение M_α в ходе моделирования, т.е. объединение или дробление частиц.

Из-за вычислительных погрешностей, вызванных конечностью *M*, τ, *h* в модели наблюдаются рассеяние частиц под действием многократных столкновений и нефизическое отклонение движения частиц от регулярного. Это проявляется в том, что в отсутствие каких-либо физических флуктуаций часть энергии регулярного движения частиц переходит в энергию хаотического теплового движения – происходит нефизический “разогрев”. Оптимально выбранные параметры модели должны обеспечивать максимальную постоянную времени этого разогрева при минимальной длительности счета. Как видно, правильный выбор параметров модели частиц требует большого объема тестовых расчетов. Но эти тесты необходимы, так как в конечном итоге они позволят с удовлетворительной точностью находить все характеристики моделируемых систем, интересующие исследователя.

10.3 Сглаживание концентраций, плотностей заряда, плотностей тока и действующей силы.

При рассмотрении методов сглаживания ограничимся моделями идентичных частиц с зарядом q=*MZe*, поскольку распространение получаемых формул на случай частиц разных зарядов тривиально. Макроскопическая плотность частиц – облаков $n_0(\boldsymbol{r}, t)$ получается сглаживанием плотности центров облаков $n_{0c}(\boldsymbol{r}, t)$ с помощью преобразования

$$n_0(\boldsymbol{r}, \mathrm{t}) = \int S_q(\boldsymbol{r}, \boldsymbol{r}') n_{0c}(\boldsymbol{r}', t) d\boldsymbol{r}' \qquad (10.4)$$

с ядром $S_q(\boldsymbol{r}, \boldsymbol{r}')$ -так называемым коэффициентом формы, имеющим размерность обратного объема $\left[i \right]^{-d_r}$. В (10.4) плотность

$$n_{0c} = \sum_\alpha \delta(\boldsymbol{r} - \boldsymbol{r}_\alpha(t)), \qquad (10.5)$$

где δ -функции также имеют размерность $\left[M\right]^{-dr}$. Подставляя (10.5) в (10.4) находим, что

$$\mathrm{n}_0(\boldsymbol{r}, \mathrm{t}) = \sum_\alpha n_\alpha(\boldsymbol{r}, t), \; n_\alpha(\boldsymbol{r}, t) = S_q(\boldsymbol{r}, \boldsymbol{r}_\alpha(t)). \qquad (10.6)$$

Здесь каждый член суммы $n_\alpha(\boldsymbol{r}, t)$ описывает плотность, создаваемую частицей-облаком конечного объема с центром в точке $\boldsymbol{r}_\alpha$(t); α =1,2,…,*N*. Коэффициент формы $S_q(\boldsymbol{r}, \boldsymbol{r}')$ нормируется на единицу, максимален при $\boldsymbol{r} = \boldsymbol{r}'$ и является достаточно плавным. Его можно рассматривать как некоторое «поле плотности», создаваемое частицей, или как вероятность для частицы, находящейся в точке с радиус-вектором $\boldsymbol{r}'$, быть обнаруженной в точке с радиус-вектором $\boldsymbol{r}$. Протяженность области, в которой коэффициент формы имеет заметную величину, называют конечным размером a_q частицы, причем, обычно a_q порядка шага сетки *h*, хотя в общем случае может задаваться независимо от *h*.

Коэффициент формы обычно выбирается симметричным:

$$S_q(\boldsymbol{r}, \boldsymbol{r}') = S_q(\boldsymbol{r} - \boldsymbol{r}') = S_q(\boldsymbol{r}' - \boldsymbol{r}). \qquad (10.7)$$

Это означает, что форма облака не зависит от его положения, и никакие направления заранее не выделены. Случай

$$S_q(\boldsymbol{r}, \boldsymbol{r}_\alpha) = \delta(\boldsymbol{r} - \boldsymbol{r}_\alpha) \qquad (10.8)$$

соответствует частицам нулевого объема (модель ZSP-Zero Size Particle). Плотность объемного заряда

$$\rho(\boldsymbol{r},t)=qn_0(\boldsymbol{r},t),\ \ \rho_\alpha(\boldsymbol{r},t)=qn_\alpha(\boldsymbol{r},t) \qquad (10.9)$$

В формулах (10.4)-(10.9) координата $\boldsymbol{r}$ произвольная. Практически, однако, в методах частиц плотность вычисляется не по формулам (10.4) и (10.6), а с помощью линейного интегрального преобразования вида:

$$n_0(\boldsymbol{r})=\sum_\alpha n_\alpha(\boldsymbol{r}),\ \ n_\alpha=S(\boldsymbol{r},\boldsymbol{r}_\alpha); \qquad (10.10)$$

$$S(\boldsymbol{r},\boldsymbol{r}_\alpha)=\int S_q(\boldsymbol{r}',\boldsymbol{r}_\alpha)\ \ w(\boldsymbol{r}',\boldsymbol{r})d\boldsymbol{r}' \qquad (10.11)$$

(сумма берется по всем *N*-частицам, зависимость от *t* для краткости опущена).

Здесь $w(\boldsymbol{r}_\alpha,\boldsymbol{r}')$ – весовая функция, которую также называют функцией усреднения или “размазывания”, поскольку она устанавливает способ усреднения (“размазывания”) облака по некоторой окрестности точки определения $\boldsymbol{r}$. Формула (10.9) сохраняется.

Так, в смешанном эйлерово-лагранжевом подходе, на основе которого строится большинство моделей частиц, плотности частиц, заряда и тока, а затем создаваемое ими электромагнитное поле вычисляется в узлах неподвижной пространственной сетки. Рассмотрим сначала неограниченную область, в которой введена прямоугольная равномерная сетка с шагами h_x,h_y,h_z и узлами $\boldsymbol{r}_i$. Пусть заряд *q* нулевого размера, расположенный в точке $\boldsymbol{r}_\alpha$, создает в *i*-ом узле плотность

$$\rho_\alpha(\boldsymbol{r}_i)=qw(\boldsymbol{r}_\alpha,\boldsymbol{r}_i) \qquad (10.12)$$

Здесь $w(\boldsymbol{r}_\alpha,\boldsymbol{r}_i)$ – интерполяционная формула, описывающая способ (правило) “раздачи заряда” в узлы сетки. Поскольку все узлы равноправны

$$w(\boldsymbol{r}_\alpha,\boldsymbol{r}_i)=w(\boldsymbol{r}_\alpha-\boldsymbol{r}_i)=w(\boldsymbol{r}_i-\boldsymbol{r}_\alpha), \qquad (10.13)$$

т.е. ядро преобразования (10.11) симметричное. На функцию $w(\boldsymbol{r}_\alpha,\boldsymbol{r}_i)$ целесообразно также наложить некоторое условие нормировки, например

$$\sum_i w(\boldsymbol{r}_\alpha,\boldsymbol{r}_i)=\sum_i w(\boldsymbol{r}_\alpha-\boldsymbol{r}_i)=\frac{1}{V_g} \qquad (10.14)$$

для любого значения $\boldsymbol{r}_\alpha$; V_g – объем сеточной ячейки. Следовательно,

$$V_g\sum_i\rho_\alpha(\boldsymbol{r}_i)=q\,. \qquad (10.15)$$

Если трактовать $V_g\rho_\alpha(\boldsymbol{r}_i)$ как заряд, внесенный зарядом *q* нулевого объема в узлы (ячейку) с номером i, то формула (10.15) представляет собой закон сохранения (и одновременно правило подсчета) заряда в сеточной области.

Применяя теперь формулу (10.12) к каждому элементу заряда,

$$dq=qS_q(\boldsymbol{r}'-\boldsymbol{r}_\alpha)d\boldsymbol{r}' \qquad (10.16)$$

частицы-облака с центром в точке $\boldsymbol{r}_\alpha$, находим плотность

$$\rho_\alpha(\boldsymbol{r}_i)=qS(\boldsymbol{r}_i-\boldsymbol{r}_\alpha),\ \ S(\boldsymbol{r})=\int S_q(\boldsymbol{r}'+\boldsymbol{r})w(\boldsymbol{r}')d\boldsymbol{r}'\,. \qquad (10.17)$$

В частности, если интерполяционная функция $w(\boldsymbol{r})=\delta(\boldsymbol{r})$, то весовая функция и плотность заряда

$$S(\boldsymbol{r})=S_q(\boldsymbol{r}),\ \ \rho_\alpha(\boldsymbol{r}_i)=qS_q(\boldsymbol{r}_i-\boldsymbol{r}_\alpha)\,. \qquad (10.18)$$

Если хотя бы одна из функций S_q или w в уравнении свертки (10.17) удовлетворяет условию нормировки вида (10.14), то

$$V_g\sum_i S(\boldsymbol{r}_i-\boldsymbol{r}_\alpha)=1 \qquad (10.19)$$

для любых $\boldsymbol{r}_\alpha$. Для ансамбля частиц-облаков с центрами в точках $\boldsymbol{r}_\alpha$ $(\alpha=1,2,...,N)$ плотность заряда находится суммированием:

$$\rho(\boldsymbol{r}_i)=\sum_\alpha\rho_\alpha(\boldsymbol{r}_i)=q\sum_\alpha S(\boldsymbol{r}_i-\boldsymbol{r}_\alpha)\,. \qquad (10.20)$$

Плотность заряда можно представить аналогично (10.4):

$$\rho(\boldsymbol{r}_i)=q\int S(\boldsymbol{r}_i-\boldsymbol{r})n_{0c}(\boldsymbol{r})d\boldsymbol{r}\,. \qquad (10.21)$$

Формулы (10.20) и (10.21) учитывают как форму облака, так и способ раздачи зарядов. Если подставить плотность

центров (10.5) в интеграл (10.21), получим сумму (10.20).

Очевидно, можно пользоваться лишь весовой функцией S, не интересуясь ее "происхождением" (10.17), и, вычисляя плотность заряда в узлах по формуле (10.20) или (10.21). На первый взгляд, количество арифметических операций , необходимых для вычисления плотности заряда (10.20), пропорционально *IN*, где I - число узлов. В действительности, однако, весовая функция $S(\boldsymbol{r}_i - \boldsymbol{r}_\alpha)$ отлична от нуля лишь в ограниченном числе узлов окрестности точки $\boldsymbol{r}_\alpha$, так что количество операций пропорционально *N* и не зависит от *I*.

Рассмотрим теперь типичные сглаживающие функции. Ограничимся одномерным случаем, так как многомерные сглаживающие функции можно представить в виде произведения соответствующих одномерных, причем, как правило, однотипных. В простейшем случае интерполяционная функция полагается кусочно-постоянной:

$$w(x) = \begin{cases} 1/h, |x| < h/2, \\ 0, |x| \geq h/2 \end{cases} \qquad (10.22)$$

(рис. 57, а), а облако нулевого размера, т.е. $S_q - \delta$-функция, $S(x) = w(x)$. Таким образом, частица вносит единичный вклад в плотность частиц в ближайшем к ней сеточном узле. В зарубежной литературе это способ называют NGP (Nearest Grid Point – ближайший узел сетки). Весовая функция NGP является разрывной, поэтому малые изменения координат частиц, приводящие к пересечению одной из границ сеточных ячеек – линий $x_{i\pm1/2} = (i \pm 1/2)h$ с полуцелыми индексами, - вызывают скачки плотности. Происходит как бы усиление вычислительных погрешностей – шумов счета.

В другом распространенном методе интерполяционная функция по-прежнему дается формулой (10.22), но облако имеет прямоугольную форму:

$$s_q(x) = \begin{cases} 1/a_q, |x| < 0{,}5a_q, \\ 0, |x| \geq 0{,}5a_q \end{cases} \qquad (10.23)$$

(рис.57, б). Свертка кусочно-постоянных функций (10.22) и (10.23):

$$S(x) = \begin{cases} \left[0{,}5(1 + h/a_q) - |x|/a_q\right]/h, \ 0{,}5(a_q - h) \leq |x| \leq 0{,}5(a_q + h), \\ 1/a_q, \ |x| \leq 0{,}5(a_q - h), \\ 0, \ |x| \geq 0{,}5(a_q + h) \end{cases} \qquad (10.24)$$

оказывается, кусочно-линейной (рис. 57, в).

В этом методе прямоугольное облако движется по сетке. Использование интерполяционной функции (10.22) означает, что доля объема облака, попадающая в ячейку сетки, которая окружает некоторый узел, составляет вклад в число частиц в этом узле. Поэтому описанный метод называют «облака в ячейках» - CIC (clouds in cells). В двумерном случае вклад частицы с номером α в число частиц в i-м узле, т.е. значение $V_q S(x_i - x_\alpha)$, представляет собой отношение заштрихованной площади (рис. 57, г) к площади частицы облака a_q^2. Поэтому такой способ называют еще взвешиванием по площадям (area weighting).

Если размер частицы $a_q > h$, то разрешение падает, так как протяженность области, в которой плотность частицы постоянна, превышает минимальный масштаб *h*, различимый на сетке. В случае $a_q < h$ существует отличный от нуля интервал расстояний между двумя облаками, на котором они не взаимодействуют. Обычно полагают $a_q = h$. При этом $S_q(x)$ совпадает с $w(x)$ (рис. 57, а), т.е. облако имеет форму сеточной ячейки; интервал, на котором два облака не взаимодействуют, стягивается в нуль; плоская вершина весовой функции исчезает:

$$S(x)=\begin{cases}(1-|x|/h)/h, |x|<h,\\ 0, |x|\ge h\end{cases} \qquad (10.25)$$

(рис. 57,д.). Функция-“крышка” (10.25) непрерывна, но ее первая производная разрывна. Малые изменения координат частиц в методе «облака в ячейках» вызывают малые изменения их плотности.

Можно также полагать, что весовая функция $S(x)$ (10.25) получена сверткой равной ей интерполяционной функции $w(x)=S(x)$, и коэффициента формы S_q в виде δ–функции. В такой интерпретации используемый метод называют «частицы в ячейках» - PIC (particles in cells). В вычислительном отношении методы PIC и CIC идентичны.

В случае “треугольного” облака

$$S_q(x)=\begin{cases}(1-|x|/a_q)/a_q, |x|<a_q,\\ 0, |x|\ge a_q\ge h\end{cases} \qquad (10.26)$$

(рис. 58, а) и кусочно-постоянной интерполяционной функции (20.22) весовая функция

$$S(x)=\begin{cases}0{,}5/h-x_1(1-0{,}5x_1/a_q)/(ha_q), a_q-h\le x_1\le a_q,\\ (1-|x|/a_q)/a_q, h/2\le|x|\le a_q-h/2,\\ (1-0{,}25h/a_q)/a_q-(x/a_q)^2/h, |x|\le h/2,\\ 0, |x_1|>a_q, (x_1\equiv|x|-h/2)\end{cases} \qquad (10.27)$$

составлена из отрезков трех парабол, сопряженных отрезками прямых (рис. 58,б).

В обычно используемом случае $a_q=h$ отрезки прямых исчезают, и весовая функция представляется квадратичным сплайном:

$$S(x)=\begin{cases}\left[3/4-(x/h)^2\right]/h, |x|\le h/2,\\ (3/2-|x|/h)^2/(2h), \quad h/2\le|x|\le 3h/2,\\ 0, \quad |x|\ge 3h/2,\end{cases} \qquad (10.28)$$

(см. рис. 58, в). Сплайн (10.28) имеет непрерывную первую, но разрывную вторую производную.

Используются также коэффициенты формы в виде конечного тригонометрического ряда Фурье

$$S_q(x)=\mathrm{Re}\sum_n \bar{S}_q(n)\exp(\mathrm{i}k_n x), \qquad (10.29)$$

где $\bar{S}_q(n)$ – коэффициент Фурье с номером n, i-мнимая единица, k_n-волновое число номера n, и в виде гауссовой кривой

$$S_q(x)=(2\pi a_q^2)^{-1/2}\exp(-0{,}5x^2/a_q^2). \qquad (10.30)$$

Модель с весовой функцией (10.28) в виде квадратичного сплайна из трех парабол иногда называют SSC (Spline shaped cloud), модель (10.29) – TSC (trigonometric shaped cloud), модель (10.30) – GSC (gaussian shaped cloud – облако гауссовой формы). С этими коэффициентами формы обычно используется интерполяционная функция $w(\boldsymbol{r})=\delta(\boldsymbol{r})$, так что $S=S_q$ (10.18).

Описана также модель, в которой гауссов коэффициент формы свертывается с кусочно-постоянной интерполяционной функцией (10.22), т.е. облако (10.30) взвешивается по площадям.

Функции (10.29) и (10.30) имеют непрерывные производные любого порядка. Как правило, однако, функции (10.29) и (10.30) обрываются при некоторых конечных x, так что функции разрывны на концах. Например, широко используется оборванная гауссова кривая

$$S_q(x)=\begin{cases}c_1\exp(-0{,}5x^2/a_q^2), |x|<x_{макс},\\ 0, |x|\ge x_{макс},\end{cases} \qquad (10.31)$$

где константы a_q и $x_{макс}$ определяются с помощью тестов из условия минимума погрешности и длительности счета, а c_1 - из условия нормировки.

Полученные формулы легко распространяются на ограниченные области. При наличии границ возможно обрывание коэффициента формы частицы-облака S_q и

некоторое нарушение закона сохранения числа частиц, когда часть облака оказывается за пределами области.

Для плотности тока сглаженные значения модели облаков

$$\boldsymbol{j}_\alpha(\boldsymbol{r}) = q\boldsymbol{v}_\alpha S_q(\boldsymbol{r}-\boldsymbol{r}_\alpha),\ \ \boldsymbol{j}(\boldsymbol{r}) = \sum_\alpha \boldsymbol{j}_\alpha(\boldsymbol{r}), \qquad (10.32)$$

а значения в узлах сетки

$$\boldsymbol{j}_\alpha(\boldsymbol{r}_i) = q\boldsymbol{v}_\alpha S(\boldsymbol{r}_i-\boldsymbol{r}_\alpha),\ \ \boldsymbol{j}(\boldsymbol{r}_i) = \sum_\alpha \boldsymbol{j}_\alpha(\boldsymbol{r}_i), \qquad (10.33)$$

где S_q и S - те же функции, которые применяются при сглаживании плотности заряда. Определение плотности заряда и тока по формулам (10.20) и (10.33) можно понимать как вычисление соответствующих интегралов методом Монте-Карло.

На погрешность аппроксимации плотности заряда и тока среди других факторов влияет конечность шага моделирования τ. Если за шаг $[t, t+\tau]$ центр облака переходит из i-й ячейки в ячейку i', то мгновенная картина расположения частиц в момент *t* даст завышенную плотность в ячейке i и заниженную в ячейке i'. Возникающий из-за этого счетный шум особенно велик в модели ZSP – NGP, но проявляется и в модели CIC. Шум объясняется тем, что во всех описанных методах при сглаживании используется лишь фазовое положение частиц, наблюдаемое в текущий момент времени t. Вычисляемая таким образом плотность дает поэтому мгновенную картину ансамбля, но не учитывает его динамику в течение шага τ.

Предполагая, что на шаге характер движения частиц ансамбля сохраняется, в качестве сглаженной весовой функции иногда принимают временное среднее несглаженной функции $S(\boldsymbol{r}_i, \boldsymbol{r}_\alpha(t))$ одного из указанных выше типов вдоль отрезка траектории проходимого частицей с номером α за интервал времени τ. Такой способ сглаживания приводит к увеличению эффективного объема частиц и числа перекрытия облаков n_q в отношении $a(a_q+l)/a$, где l – средний пролет частиц на шаге. В моделях пролет l часто в несколько раз превышает размер a_q, так что эффект сглаживания может быть существенным при незначительном удлинении счета.

При решения уравнений движения следует с помощью некоторой интерполяционной формулы вычислять напряженность поля $\boldsymbol{E}(\boldsymbol{r})$ в произвольной точке $\boldsymbol{r}$ по значениям $\boldsymbol{E}_i$ в сеточных узлах $\boldsymbol{r}_i$. Воспользуемся для этого той же формулой $w(\boldsymbol{r}, \boldsymbol{r}_i)$, которая была введена для раздачи заряда в узлы сетки. Тогда

$$\boldsymbol{E}(\boldsymbol{r}) = V_g \boldsymbol{E}_i w(\boldsymbol{r}, \boldsymbol{r}_i). \qquad (10.34)$$

Благодаря условию нормировки (10.14) в однородном поле, когда $\boldsymbol{E}_i = const$, $\boldsymbol{E}(\boldsymbol{r})$ также не зависит от $\boldsymbol{r}: \boldsymbol{E}(\boldsymbol{r}) \equiv \boldsymbol{E}_i = const$.

Напряженность поля $\boldsymbol{E}_\alpha$, действующая на частицу с номером α, получается усреднением напряженности (10.34) по объему облака с выбранным коэффициентом формы $S_q(\boldsymbol{r}, \boldsymbol{r}')$:

$$\boldsymbol{E}_\alpha = \int \boldsymbol{E}(\boldsymbol{r}') S_q(\boldsymbol{r}', \boldsymbol{r}_\alpha) d\boldsymbol{r}' = V_g \sum_i \boldsymbol{E}_i S(\boldsymbol{r}_i, \boldsymbol{r}_\alpha), \qquad (10.35)$$

где $S(\boldsymbol{r}, \boldsymbol{r}_\alpha)$ – та же весовая функция (10.11) (функция усреднения), которая использовалась для вычисления сеточной плотности объемного заряда. Действие формулы (10.35) является в некотором смысле обратным действию формул (10.10) и (10.11). Последние преобразует функции с континуума точек на сетку, тогда как формула (10.35) распространяет функцию с сетки на континуум. Также вычисляется магнитное поле $\boldsymbol{B}_\alpha$ и действующая сила

$$\boldsymbol{F}(\boldsymbol{r}_\alpha) \equiv \boldsymbol{F}_\alpha = q(\boldsymbol{E}_\alpha + \boldsymbol{v}_\alpha \times \boldsymbol{B}_\alpha). \qquad (10.36)$$

Выбор для напряженности поля (10.34) и плотности заряда (10.12) одинаковых интерполяционных формул, а, следовательно, одинаковых весовых функций, дает, по крайней мере, в электростатических задачах, два

преимущества. Во-первых, можно показать, что, если при этом конечно-разностное уравнение Пуассона, связывающее напряженность поля с плотностью заряда пространственно-симметрично, т.е. записано в центрально-разностных производных, то исключается *самодействие*: изолированная частица в неограниченной или периодической области, не испытывает действия сил.

Во-вторых, использование различных весовых функций S_1 и S_2 может привести к неустойчивости, которая называется гравитационной, и проявляется в том, что одинаковые частицы не отталкиваются, а притягиваются, как гравитационные. Если же $S_1 = S_2 = S$, гравитационная неустойчивость не возникает.

Во многих моделях напряженность электрического поля находится численным дифференцированием $\boldsymbol{E} = -grad\varphi$ сеточного потенциала, полученного в результате решения уравнения Пуассона. За значение напряженности поля в узлах обычно принимается центрально-разностная производная на целых или полуцелых шагах. Например, в двумерной области в узле (x_l, y_i) можно принять

$$E_x(l,i) = (\varphi(l+1,i) - \varphi(l-1,i))/(2h_x) \qquad (10.37)$$

или

$$E_x(l+1/2,i) = (\varphi(l+1,i) - \varphi(l,i))/h_x \qquad (10.38)$$

и т.д. Полученные таким образом напряженности затем используются в формуле взвешивания (10.35).

В некоторых важных случаях составляющие $\boldsymbol{E}_0$, $\boldsymbol{E}_w$ на каждом шаге моделирования являются известными аналитическими достаточно простыми функциями координат. Тогда целесообразно их уравнения Пуассона вычислять лишь кулоновский потенциал φ_ρ, сохраняя его в виде сеточной функции и численно дифференцируя для нахождения поля $\boldsymbol{E}_\rho$, например, по формулам (10.37), (10.38).

Помимо конечно-разностных формул для вычисления поля применяются сплайны. Сплайны позволяют получить более гладкое поле, с меньшими вычислительными погрешностями, но они замедляют расчеты и требуют дополнительной памяти для хранения коэффициентов сплайна. Эти трудности резко возрастают с увеличением размерности пространства. Поэтому в численном эксперименте сплайны используются, как правило, для интерполирования одномерных функций. Возможно, более радикальным является отыскание решения системы уравнений Максвелла или уравнения Пуассона непосредственно в виде сплайна, вместо определения решения на сетке с последующей сплайн-интерполяцией. В многомерных областях эта возможность пока не исследовалась.

Помимо исследования ансамблей заряженных частиц, метод частиц может успешно использоваться также для моделирования многих явлений газовой динамики, гидродинамики, пластичности, молекулярной физики, физики атмосферы, небесной механики, нейтронной физики, ядерной физики и других областей физики. Как теория метода, так и техника его практического использования активно развиваются и совершенствуются.

ЗАКЛЮЧЕНИЕ

Численные методы составляют сейчас важнейший раздел прикладной математики, интенсивно развивающийся одновременно с совершенствованием средств вычислительной техники и находящий все более широкое применение. Хотя современное состояние вычислительной техники и вычислительной математики в большинстве случаев не позволяет пока получать количественные результаты, полностью отвечающие потребностям экспериментаторов, проектировщиков и

конструкторов, однако численные результаты позволяют достигнуть нового, более высокого уровня понимания изучаемых процессов, приборов, установок.

Поэтому при выполнении расчетов на компьютере, численных (машинных) экспериментов целесообразно руководствоваться хорошим принципом, выдвинутым известным специалистом в области численных методов и их применения Р. В.Хеммингом: «Цель расчетов — не числа, а понимание». Но чтобы достигнуть этого понимания, надо хорошо ориентироваться в численных методах, отчетливо представлять себе, что могут и что не могут дать те или иные методы, каковы их особенности, характеристики, область применения. Без этих знаний важные физические явления могут остаться не обнаруженными, и наоборот, особенности методов можно ошибочно принять за особенности изучаемых объектов, например, неустойчивость трактовать как самовозбуждение, несохранение энергии вследствие погрешностей выбранного метода— как диссипацию и т.д.

Конечно, изложенные в этом небольшом учебном пособии численные методы следует рассматривать лишь как некоторую минимальную, общедоступную основу для понимания заложенных в них принципов и путей применения в инженерно-исследовательской практике. Для более полного изучения численных методов, как изложенных в учебном пособии, так и не вошедших в него, необходимо привлекать дополнительную литературу. Некоторые издания, наиболее подходящие, по нашему мнению, для студентов физических и технических специальностей, указаны в списке литературы [1—14].

Авторы надеются, что изложенный материал сможет способствовать пробуждению у читателей интереса к самостоятельному численному решению на компьютере разнообразных физических задач.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.Численные методы. М.: СПб, Физматлит,2000.
2. Волков Е. А. Численные методы. М.: Наука, 1987.
3. Калиткин Н. Н. Численные методы, М.: Наука, 1978.
4. Марчук Г. И. Методы вычислительной математики. М.: Наука, 1989.
5. Самарский А. А., Гулин А. В. Численные методы. М.: Нау ка, 1989.
6. Турчак Л. И. Основы численных методов. М.: Наука, 1987.
7. Хемминг Р. В. Численные методы для научных работников и инженеров. М.: Наука, 1968.
8. Самарский А. А., Гулин А. В. Численные методы математической физики. М.: Научный мир, 2000.
9. Соболь И.М. Численные методы Монте-Карло. М.:Наука, 1973.
10. Ильин В.П. Численные методы решения задач электрофизики. М.: Наука, 1985.
11. Карманов В.Г. Математическое программирование. М.: Наука, 1986.
12. Белоцерковский О.М., Давыдов Ю.М. Метод крупных частиц в газовой динамике. М.: Наука, 1982.
13. Хокни Р., Иствуд Дж. Численное моделирование методом частиц. М.: Мир, 1989.
14. Бэдсел Ч., Ленгдон А. Физика плазмы и численное моделирование. М.: Энергоатомиздат, 1989.

СОДЕРЖАНИЕ

Ращиков Владимир Иванович

Рошаль Анатолий Самуилович

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ФИЗИЧЕСКИХ ЗАДАЧ