Уфа : УГАТУ, 2008 Вестник УГАТУ • Управление, ВТ и И Т. 11, № 1 (28). С. 181–187

УДК 519.6

Н. М. ШЕРЫХАЛИНА, С. С. ПОРЕЧНЫЙ

МОДЕЛИРОВАНИЕ ПОГРЕШНОСТИ И ЧИСЛЕННАЯ ФИЛЬТРАЦИЯ ПРИ РЕШЕНИИ СМЕШАННЫХ ЗАДАЧ ДЛЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Подходы и методы фильтрации численных результатов применяются при решении задач для дифференциальных уравнений с частными производными. Найдены и объяснены закономерности изменения погрешностей метода, исходных данных и округления от параметра дискретизации для явной и неявной разностных схем. *Численный эксперимент*; *фильтрация*; *разностные схемы*; *уравнения математической физики*

Ранее методы фильтрации результатов численного эксперимента применялись для увеличения точности, надежности и оценки погрешности решения задач численного дифференцирования [1] и интегрирования [1, 2]. С помощью фильтрации удалось практически полностью подавить погрешность численных методов, что позволило выявить и изучить поведение погрешностей исходных данных и округления. Исследование показало, что на эти погрешности влияют как фундаментальные математические закономерности, так и особенности выполнения арифметических операций с помощью конкретных устройств и программ. Для оценки этих погрешностей применялись разные модели. Исследование позволило найти диапазоны их применимости.

В данной работе методы фильтрации применяются к более сложной задаче: смешанной задаче для уравнения с частными производными с целью изучения специфики образования погрешностей и построения их моделей.

Идея численной фильтрации заключается в априорном представлении математической погрешности численного метода в виде суммы, например, степенных функций параметра дискретизации (числа узлов n или шага сетки h) и поочередном подавлении составляющих путем комбинации результатов, полученных при разных n и h. Здесь для фильтрации применяются методы, изложенные в [1, 2].

1. СМЕШАННАЯ ЗАДАЧА ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

В данной задаче искомая функция $u(x,t)$ должна удовлетворять уравнению параболического типа

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}\,, \tag{1}$$

начальному условию

$$u\,(x,0) = f\,(x) \tag{2}$$

и краевым условиям

$$\begin{aligned} &u\,(0,t) = \varphi_1\,(t)\,,\\ &u\,(1,t) = \varphi_2\,(t)\,,\\ &(t \geq \varsigma 0)\,. \end{aligned} \tag{3}$$

Задача решается методом конечных разностей. На плоскости (x,t) строится сетка с шагом h по переменной x ($x_i = ih$, $i = 0, \ldots, n$, $h = 1/n$) и с шагом τ по переменной t ($t_j = j\tau$, $j = 0, \ldots, m, \tau = b/m$). Вводятся обозначения $u\,(x_i, t_j) = u_{i,j}$.

При решении задачи по явной разностной схеме частные производные в уравнении (1) заменяют разностными аналогами следующим образом

$$\frac{u_{i+1,j} - 2u_{i,j} - u_{i-1,j}}{h^2} = \frac{u_{i,j+1} - u_{i,j}}{\tau}\,. \tag{4}$$

Обозначим $\lambda = \tau/h^2$ (число Куранта). Тогда разностное уравнение запишется в виде

$$\begin{aligned} &u_{i,j+1} = \lambda u_{i-1,j} + (1 - 2\lambda)\,u_{i,j} + \lambda u_{i+1,j}\,,\\ &i = 1, .., n-1; \quad j = 0, .., m-1\,. \end{aligned} \tag{5}$$

Контактная информация: (347)273-32-00

Таким образом, решение задачи сводится к повторному применению формулы (5) при последовательном возрастании j.

Однако явная схема является устойчивой только при $\lambda < 1/2$ [3, 4]. В противном случае развивается экспоненциальный рост погрешности. Поэтому при использовании явной схемы вычисления приходится вести с шагом по времени, зависящим от n.

При решении задачи по неявной схеме в (4) изменяется только вид частной разностной производной по времени

$$\frac{u_{i+1,j} - 2u_{i,j} - u_{i-1,j}}{h^2} = \frac{u_{i,j} - u_{i,j-1}}{\tau}, \qquad (6)$$

а разностное уравнение принимает вид

$$\lambda u_{i-1,j} - (1 + 2\lambda)\, u_{i,j} + \lambda u_{i+1,j} = -u_{i,j-1}\,,$$
$$i = 1, \ldots, n-1; \quad j = 0, \ldots, m\,. \qquad (7)$$

В соответствии с (2) и (3) значения $u_{i,0} = f(x_i)$, $u_{0,j} = \varphi_1(t_j)$, $u_{n,j} = \varphi_2(t_j)$, $i = 1, \ldots, n-1; j = 0, .., m$ являются известными. Тогда, подставляя в (7) $j = 1$, получим систему $n-1$ линейных алгебраических уравнений, решив которую можно определить $u_{i,1}$, $i = 1, \ldots, n-1$. При этом, поскольку $u_{0,1} = \varphi_1(t_1)$, $u_{n,1} = \varphi_2(t_1)$, известными оказываются все значения временного слоя $j = 1$, $(t = t_1)$. Затем, подставляя в (7) $j = 2$, решаем систему уравнений относительно $u_{i,2}$ и т. д. для всех $j = 2, \ldots, m$. Поскольку матрица системы уравнений (7) является трехдиагональной, эту систему можно решить методом прогонки.

Применение неявной схемы требует решения системы уравнений, но данная схема является безусловно устойчивой.

2. ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

В качестве примера на рис. 1 и 2 приводятся результаты решения задачи для $f(x) = \sin \pi x$; $\varphi_1(t) = f(0)$, $\varphi_2(t) = f(1)$. Время окончания счета состовляет $b = 0{,}05$ (при этом отличие решения от предельного при $t \to \infty$ составляет примерно 30%). Погрешности оценивались в точке $x = 0{,}2$.

Поскольку искомая функция $u(x, t)$ зависит от двух переменных, то необходима фильтрация двух видов последовательностей: сначала рассчитанных значений искомого параметра (например, $u_{n,m}$) при изменении m (и τ) для нескольких фиксированных $n(h)$, в результате чего получается последовательность отфильтрованных значений, соответствующих разным n, а затем фильтрации полученной последовательности по n.

Результаты фильтрации по m приведены на рис. 1 в виде зависимости десятичного логарифма относительной погрешности $-\lg \delta$ от $\lg m$ для двух значений n: 10 (рис. 1, *а*) и 320 (рис. 1, *б*). Точность, полученная при фильтрации, ограничивается накоплением погрешности округления (прямая $y = 16 - \frac{1}{2}\lg m$). Эта статистическая зависимость характерна также для методов численного интегрирования функций [1]. При $n = 320$ (рис. 1, *б*) ограничение по числу Куранта приводит к ограничению сверху шага по времени τ (ограничении снизу числа m). Это требует дополнительных затрат ресурсов. В результате максимальная полученная точность, оцененная с размытостью (отношением оценки погрешности оценки погрешности к оценке погрешности [6]) порядка 0,1 . . . 0,3, достигает 13 значащих цифр.

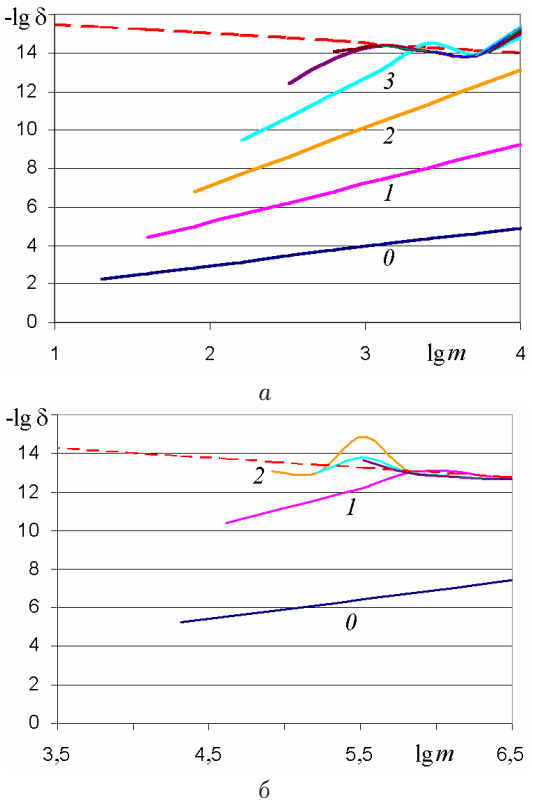


Рис. 1. Зависимость значений погрешностей решения смешанной задачи от шага по времени $\lg m = \lg b/\tau$: *а* — $n = 10$, *б* — $n = 320$. Явная схема. Пунктирная прямая $y = 16 - 0{,}5 \lg m$

Фильтрация полученной последовательности экстраполированных по m значений в сравнении с эталоном, найденным четвертой фильтрацией (рис. 2, *а*), показывает завышенные оценки (линия 3, рис. 2, *а*). Это следует из

рис. 2, б, где приведены результаты сравнения с точным решением

$$u(x,t) = e^{-\pi^2 t}\sin \pi x\,.$$

Этот результат не противоречит оценкам, сделанным выше, т. е. причиной этой погрешности является накопление погрешности округления при выполнении шага по времени (5).

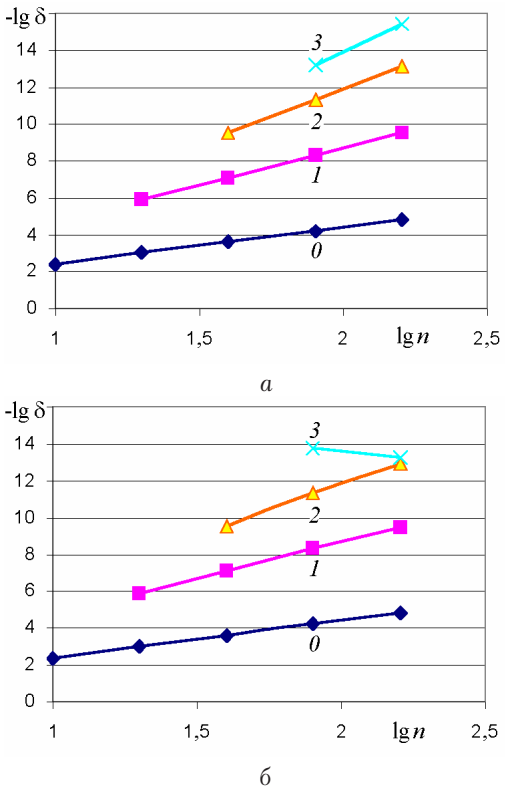


Рис. 2. Зависимость значений погрешностей решения смешанной задачи от шага по x: $а$ — сравнение с эталоном; $б$ — сравнение с точным значением. Явная схема

В особенности эффективно использование фильтрации при применении неявной схемы, поскольку она является безусловно устойчивой. На рис. 3 показаны результаты фильтрации по m для 4-х значений n.

В отличие от рис. 1, диапазон изменения m одинаков для всех n, поскольку явная схема безусловно устойчива.

Еще одним отличием является заметное ограничение точности (наличие порога) при возрастании n (рис. 3, $в$, $г$). Зависимость этого порогового значения точности от n (кривая E на рис. 4, $б$) аппроксимируется прямой $y = 16{,}5 - 2\cdot\lg n$, т. е. $\delta \sim n^2$. Квадратичная зависимость обусловлена погрешностью округления значений u_{ij} при вычислении второй разностной производной в (6), что согласуется с выводами [1]. Отметим, что для явной схемы этот вид погрешности не имеет места.

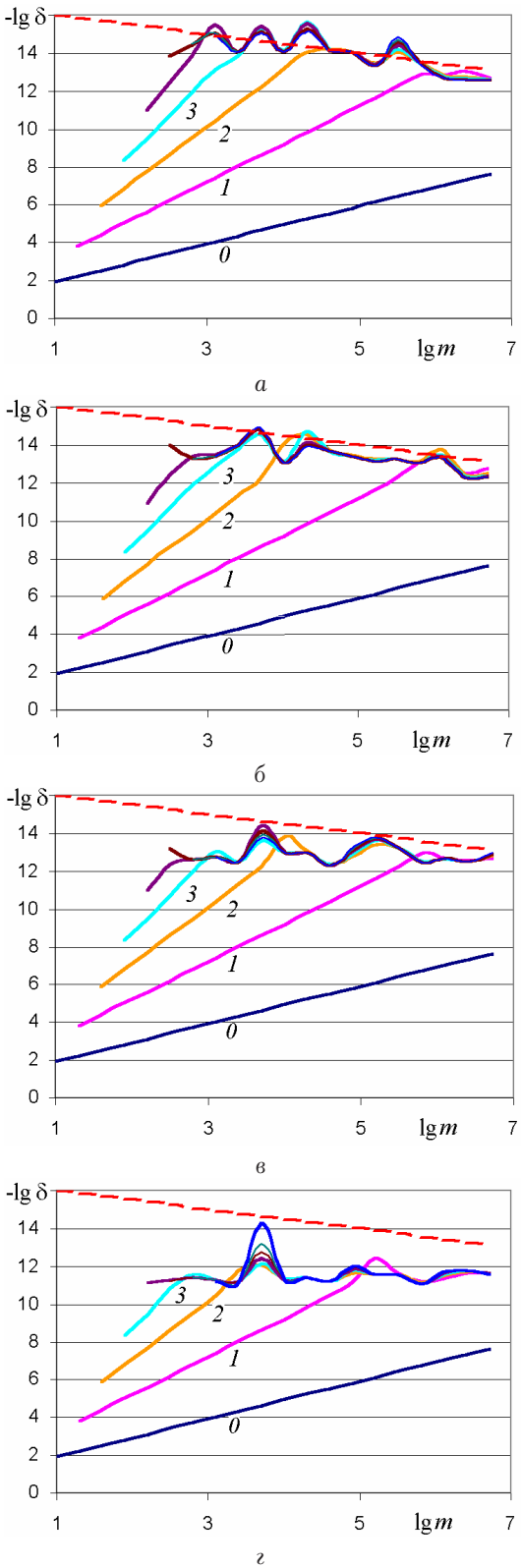


Рис. 3. Зависимость значений погрешностей решения смешанной задачи от шага по времени ($\lg m = \lg b/\tau$): $а$ — $n = 10$, $б$ — $n = 80$; $в$ — $n = 160$, $г$ — $n = 640$. Неявная схема. Пунктирная прямая $y = 16{,}5 - 0{,}5\lg m$

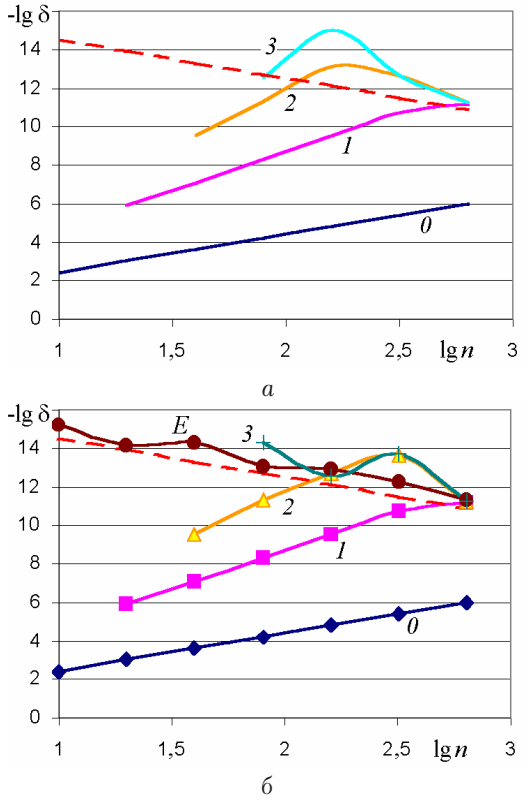


Рис. 4. Зависимость погрешностей решения задачи от шага по x: $а$ — сравнение с эталоном; $б$ — сравнение с точным значением. Неявная схема. Пунктирная прямая $y = 16{,}5 - 2 \cdot \lg n$

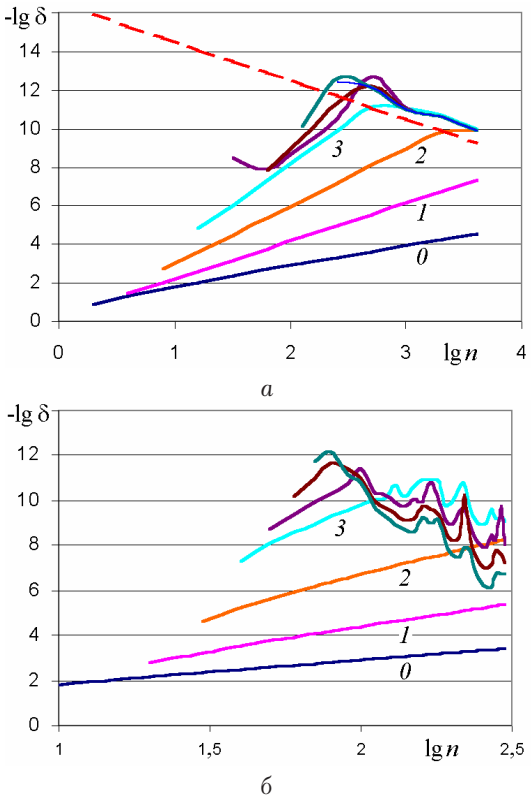


Рис. 5. Зависимость погрешностей решения задачи от шагов по x и t ($n = m$): $а$ — фильтрация методом Ромберга; $б$ — фильтрация методом Нэвилла. Неявная схема. Пунктирная прямая $y = 16{,}5 - 2 \cdot \lg n$

Для уменьшения объема вычислений по получению данных для фильтрации можно вместо поочередного изменения чисел m и n применить синхронное. На рис. 5 приведены результаты такого исследования. Сравнивая рис. 5, $а$ с рис. 4, $а$, $б$, следует отметить, что первый способ позволяет добиться несколько большей точности при меньших n. Уменьшить объем вычислений при синхронном изменении m и n можно, если применить метод Нэвилла [5], который требует увеличения числа узлов не в два раза, а на константу. Однако этот метод неустойчив к накоплению погрешности округления при проведении фильтраций. Этим и объясняется резкое увеличение погрешности при сравнительно небольших n (рис. 5, $б$).

Таким образом, применение неявной схемы связано с более быстрым нарастанием погрешности округления с увеличением числа узлов сетки по пространственной переменной. Скорость роста погрешности округления (угловой коэффициент k зависимости $y = 16{,}5 - k \cdot \lg n$) определяется номером старшей производной, входящей в уравнение (в данном случае $k = 2$).

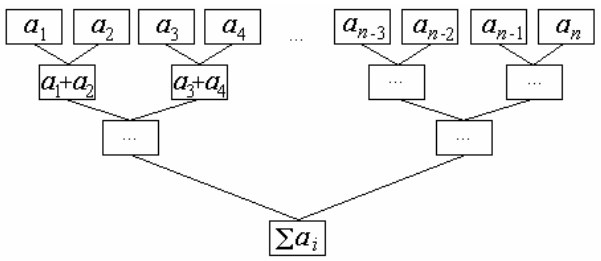
В явной схеме зависимость погрешности $y = 16{,}5 - 0{,}5 \cdot \lg m$ связана только с многократным повторением операции суммирования (5). Для уменьшения этой накапливающейся погрешности необходимо устранение причины: суммирования слагаемых разного порядка. Для этого можно предложить замену процедуры суммирования накоплением ($s := s + a_i$) суммированием пар слагаемых одного порядка. Суммирование организуется по схеме бинарного дерева: складываются пары слагаемых, затем пары пар и т. д. (рис. 6, $а$).

Отметим, что если число слагаемых в сумме не равно 2^k (для целого k), то по окончании попарного суммирования необходимо просуммировать все оставшиеся непарные слагаемые, начиная с первого по схеме уровня.

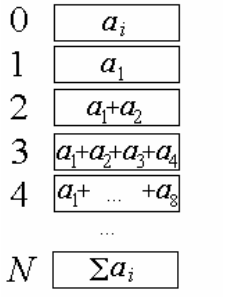
Для применения этого алгоритма при решении смешанной задачи введем новые переменные $v_{i,j} = u_{i,j} - u_{i,0}$ и преобразуем (5)

$$v_{i,j+1} = v_{i,j} + \\ + \lambda \left(v_{i-1,j} - 2v_{i,j} + v_{i+1,j} + u_{i-1,0} - 2u_{i,0} + u_{i+1,0} \right).$$

Результаты решения задачи этим способом приведены на рис. 7.

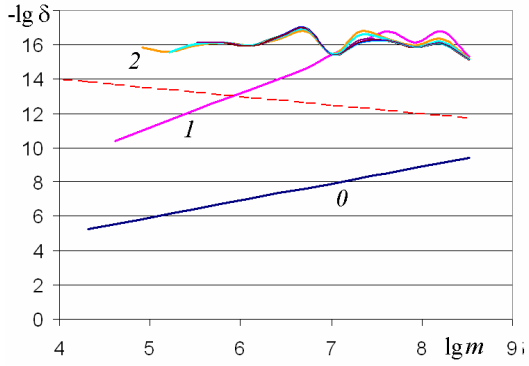


а

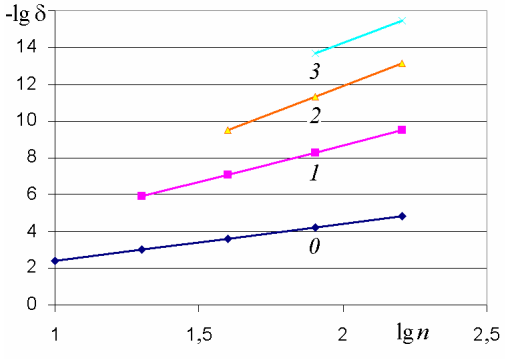


б

Рис. 6. Схема процесса попарного суммирования



а



б

Рис. 7. Результаты применения попарного суммирования для решения смешанной задачи: *а* — зависимость погрешности от шага по x; *б* — зависимость погрешности от шага по t (сравнение с точным значением). Явная схема

Видно, что накопления погрешности округления не происходит. Погрешность остается на уровне, характерном для погрешности исходных данных 10^{-15}.

Реализовать попарное суммирование весьма просто, если использовать идею стека (рис. 6, *б*). Вначале каждое слагаемое помещается на нулевой уровень. Далее слагаемое продвигается на следующий уровень. Если следующий уровень свободен (содержит нулевое значение), то данное слагаемое занимает свое место и операция добавления данного слагаемого на этом заканчивается. Если следующий уровень оказывается занятым (содержит отличное от нуля значение), то данное слагаемое складывается с содержимым данного уровня, а уровень, на котором слагаемое находилось, очищается. Теперь сумма должна быть продвинута дальше по уровню. Повторяется проверка следующего уровня на нуль и все описанные выше действия.

Программная реализация данного способа суммирования требует некоторого увеличения затрат ресурсов по времени и памяти: N машинных слов для 2^N слагаемых и максимум N операций сложения вместо одной на каждом шаге. Однако такие затраты не являются критическими, поскольку основное время затрачивается на вычисление значений функции.

3. АНАЛИЗ РЕЗУЛЬТАТОВ ЧИСЛЕННОГО ЭКСПЕРИМЕНТА

Алгоритм решения задачи по неявной схеме содержит дополнительный элемент — решение системы линейных алгебраических уравнений, что вызывает вопросы о возможном появлении дополнительной погрешности. Метод прогонки заключается в вычислении прогоночных коэффициентов и использовании рекуррентной зависимости для определения искомых величин. Рекуррентные формулы могут способствовать накоплению погрешности округления, аналогично численному интегрированию [1]. Для устранения этой возможности была предпринята попытка применения попарного суммирования. Однако результат такого видоизменения алгоритма практически не отличается от исходного.

Матрица системы уравнений (7), решаемой на каждом временном шаге, имеет меру обусловленности $\nu > 1$ [3, 4]. Но в этом случае может возникнуть экспоненциальный рост погрешности при повторных шагах по времени $\delta \sim e^{-M}\nu^m$. Этого не наблюдается в реальном вычислительном процессе (см. рис. 3, 4). Тем самым, теория может дать верхние оценки погрешности, которые невозможно использовать для практики.

Все сказанное выше заставляет искать причину погрешности в другом направлении. В [1] подобная квадратичная зависимость погрешности округления от n наблюдалась при фильтрации результатов вычисления второй разностной производной. При решении смешанной задачи также вычисляется вторая конечно-разностная производная. При этом погрешность округления определяется операцией вычитания близких значений функции u_{ij} $\Delta(n) \leq \frac{4\Delta_0}{h^2} = \frac{4\Delta_0}{(b-a)^2}n^2$ (где Δ_0 — величина порядка единицы последнего разряда числа в машинном слове), при вычислении второй производной по x, поэтому увеличивается как n^2 при увеличении n. Эта оценка и подтверждается численным экспериментом.

Величину Δ_0 в литературе часто называют неустранимой погрешностью. Однако влияние этой погрешности на результат в явной и неявной схемах отличается существенно.

В неявной схеме устранить рост погрешности, пропорциональный n^k, где k — порядок старшей производной, по-видимому, не удастся. При решении краевых задач теории прочности приходится иметь дело с производными четвертого порядка. Поэтому погрешность, связанная с округлением значений искомой функции, будет зависеть от числа разбиений n как n^4. При использовании двойной точности (мантисса около 16 десятичных разрядов) результат решения уже для $n = 10^4$, скорее всего, не будет иметь ни одной верной значащей цифры.

В явной схеме накопление происходит по другим законам и от порядка производной не зависит.

ВЫВОДЫ

Методы численной фильтрации позволили с помощью устранения погрешности численного метода изучить характер погрешности округления при решении смешанной задачи двумя методами.

Результаты эксперимента показали, что погрешность округления в ряде случаев оказывается существенно меньшей, чем это можно было ожидать, и отличается от общеизвестных (грубых) оценок. Это объясняется несоответствием модели предельной погрешности (предполагающей, как в интервальной математике, «наихудший случай») фактической реальности.

При решении по явной схеме погрешность определяется механизмом накопления погрешности округления при вычислении сумм [1] (зависит от числа разбиений m по статистическому закону как $\sqrt{m}$) и это накопление устраняется попарным суммированием.

При решении методом конечных разностей численная погрешность определяется погрешностью численного дифференцирования, связанной с округлением значений искомой функции, которая для второй производной оценивается как $10^{-M}n^2$ [1], где M — длина мантиссы в десятичных разрядах.

Отсутствие надежных оценок погрешности округления приводит либо к получению неверных результатов, либо к необходимости решения несуществующих проблем. В качестве примера можно привести работу [7], в которой авторы приходят к необходимости применения надстройки к комплексу OrCAD, которая довольно сложным способом помогает избежать отказов, связанных с точностью, устойчивостью вычислительных процессов, с одной стороны, и накоплением погрешности округления при численном интегрировании, с другой. На самом деле авторы комплекса OrCAD, не имея надежной информации о погрешности округления, использовали для ее оценки грубую линейную модель. Использование результатов, полученных в данной работе, позволит избежать таких ситуаций, а использование фильтрации и визуализации ее результатов в промышленных и исследовательских программных комплексах позволит уменьшить требования к ресурсам, с одной стороны, и существенно увеличить надежность получаемых результатов вычислений, с другой.

СПИСОК ЛИТЕРАТУРЫ

1. **Шерыхалина, Н. М.** Применение фильтрации для обработки результатов численного эксперимента / Н. М. Шерыхалина // Вестник УГАТУ (сер. «Управление, вычислительная техника и информатика»). 2007. Т. 9, № 7(25). С. 90–96.
2. **Житников, В. П.** Обоснование методов фильтрации результатов численного эксперимента / В. П. Житников, Н. М. Шерыхалина // Вестник УГАТУ (сер. «Фундаментальная и прикладная математика»). 2007. Т. 9, № 3(21). С. 71–79.
3. **Бахвалов, Н. С.** Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. М. : Наука, 2004. 636 с.
4. **Волков, Е. А.** Численные методы / Е. А. Волков. М. : Наука, 1982. 256 с.
5. **Шерыхалина, Н. М.** Методы обработки результатов численного эксперимента для увеличения их точности и надежности / Н. М. Шерыхалина // Вестник УГАТУ

(сер. «Управление, вычислительная техника и информатика»). 2007. Т. 9, № 2(20). С. 127–137.

6. **Шерыхалина, Н. М.** Численная фильтрация данных, искаженных нерегулярной погрешностью / Н. М. Шерыхалина, А. А. Ошмарин // Вестник УГАТУ. 2006. Т. 8, № 1(17). С. 138–141.
7. **Болотовский, Ю. И.** Опыт моделирования систем силовой электроники в среде OrCAD 9.2 / Ю. И. Болотовский, Г. И. Таназлы // Силовая электроника. 2004. № 1. С. 90–95.

ОБ АВТОРАХ

Шерыхалина Наталия Михайловна, доц. каф. комп. мат. Дипл. инж. (УГАТУ, 1993). Канд. физ.-мат. наук (БГУ, 1996). Иссл. в обл. волновых течений жидкости, уединенных волн, методов оценки погрешности численных результатов.

Поречный Сергей Сергеевич, асп. той же каф. Дипл. инж. и магистр по инфор. и прогр. обесп. САПР (УГАТУ, 2006). Готовит дис. в обл. мат. моделирования физ. процессов.