Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
Ярославский государственный университет
им. П.Г. Демидова
Кафедра математического моделирования

Базовые методы численного анализа динамических систем

Методические указания

Рекомендовано
Научно-методическим советом университета
для студентов, обучающихся по специальности
Прикладная математика и информатика

Ярославль 2009

ђddqyɐʎ pƎ dɐdƎbʎɜɐ b ɐɔɐdɔʎ Ɉddʎqʎè

УДК 519.6
ББК В 193я73
Б 17

Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2009 года

Рецензент
кафедра математического моделирования

Составитель Глызин Д.С.

Б 17 **Базовые методы численного анализа динамических систем: метод. указания** / сост. Д.С. Глызин; Яросл. гос. ун-т. – Ярославль: ЯрГУ, 2009. – 24 с.

В методических указаниях приведено описание лабораторных работ по курсу "Численные методы анализа динамических систем".

Предназначены для студентов четвертого курса, обучающихся по специальности 010501 Прикладная математика и информатика (дисциплина "Численные методы анализа динамических систем", блок ДС), очной формы обучения.

УДК 519.6
ББК В 193я73

ŋqdqɏeɎ bɆ dədɆbɎɜə d ɕɔɑbɔɏ Ɉqddɏqɏè

Содержание

1 Арифметика чисел с плавающей точкой

Главным источником информации для этого раздела стала работа [10], в которой на математическом уровне строгости излагаются основные проблемы компьютерной арифметики с нецелыми числами.

Наиболее распространенной формой представления действительных чисел в машинных вычислениях являются двоичные форматы с плавающей точкой. Эти форматы позволяют хранить точно (без ошибки округления) числа следующего вида:

$$\pm(d_0 + d_1 \cdot 2^{-1} + \cdots + d_{p-1} \cdot 2^{-(p-1)})2^e, \; d_i = 0, 1, \; e_{max} \le e \le e_{min}. \qquad (1)$$

Любое другое действительное число округляется до одного из представимых в форме (1).

Цифры $d_0 \ldots d_{p-1}$ образуют *мантиссу*, а целое число e называется показателем (экспонентой).

За точность представления числа отвечает параметр p, количество значащих цифр в мантиссе. Этот параметр определяется как программной, так и аппаратной частями платформы. Стандартом IEEE-854 предусмотрены следующие варианты высокоточных представлений:

- double, двойная точность;

- double extended, двойная расширенная.

Числа с плавающей точкой, первый знак которых отличен от нуля, называют нормализованными.

Упражнение. Найти количество нормализованных чисел с плавающей точкой в шарах единичного радиуса с центрами в точке 0 и точке 2^{10}.

Архитектура Intel x86 реализует в сопроцессоре вычисления с двойной расширенной точностью, таким образом, для большинства современных персональных компьютеров внутренним аппаратным форматом является 80-битовый double extended. Тем не менее такие числа поддерживаются не всеми компиляторами языков высокого уровня. Здесь мы рассмотрим язык C как самое универсальное средство создания консольных приложений под различными операционными системами, а также среду Turbo Delphi, основанную на языке Object Pascal, бесплатно распространяемую для некоммерческого использования, как одну из самых удобных сред быстрой разработки высокопроизводительных программ с графическим пользовательским интерфейсом.

В указанных языках высокоточные вычисления реализованы на основе следующих типов:

- в C: double (64 бита) и long double, реализация которого зависит от компилятора и его ключей, основные ограничения см. в [11].

- в Object Pascal: float (64 бита) и extended (80 бит).

Упражнение. Выяснить с помощью функции sizeof() размер типов double и long double в доступной вам реализации компилятора языка C с установками по умолчанию.

Упражнение. Написать программу на языке Object Pascal и выяснить, во сколько раз быстрее на вашем компьютере выполняются одни и те же действия над числами формата double, чем над числами в формате extended. *Это соотношение существенно меняется в зависимости от модели процессора!*

Указание: для точного замера начального и конечного моментов вычислений можно использовать следующую функцию:

```
function RdTsc: Int64;
asm
rdtsc
end;
```

Она возвращает текущее показание счетчика тактов процессора.

Выбор расширенной двойной точности в качестве основы вычислений должен быть осознанным, например, если он продиктован практической необходимостью в дополнительных значащих цифрах. Хотя использование расширенной точности выглядит естественным (именно в таком формате хранятся числа в сопроцессоре, а математические функции Delphi используют для передачи аргументов и значений тип extended), минусы такого подхода значительны: это не только более низкая (в разы) скорость вычислений, но и существенные (в случае C) проблемы портируемости.

В качестве важного примера оборудования, не поддерживающего double extended, можно привести графические процессоры, используемые для неграфических вычислений. Благодаря большому количеству ядер эти устройства демонстрируют высокую производительность на хорошо параллелизуемых задачах, но при этом лишь в последнем поколении (по состоянию на 2008 год) видеокарт появилась поддержка двойной (не расширенной) точности.

Даже две платформы, соответствующие стандарту IEEE-854, могут выполнять одну и ту же программу с разным результатом. Причины кроются в том, что некоторые действия компиляторов, влияющие на вычисления, никак не регламентируются стандартом. Одной из самых частых причин несогласованности является разница в формате, в котором хранятся промежуточные значения выражений.

Упражнения:

- вычислить, насколько будет отличаться от 1 сумма $\sum_{i=1}^{10^6} 10^{-6}$ при использовании чисел с двойной точностью;