ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1982

Seria: INFORMATYKA z. 3 Nr kol. 708

Marta ŚWIERNIAK

Zakład Problemów Organizacji i Zarządzania
PAN - Bytom

NIEKTÓRE PROBLEMY KONSTRUKCJI BIBLIOTEK STANDARDOWYCH FUNKCJI MATEMATYCZNYCH DLA MINIKOMPUTERA

Streszczenie: W pracy przedstawiono pewne zagadnienia związane z algorytmizacją metod przybliżania wybranych funkcji matematycznych w bibliotekach minikomputerów oraz wykonaniem programów tych bibliotek. Wykorzystano przy tym doświadczenia uzyskane przy konstruowaniu biblioteki standardowych funkcji matematycznych dla minikomputera P.

WPROWADZENIE

Biblioteki standardowych funkcji matematycznych stanowią ważną część oprogramowania każdej maszyny cyfrowej.
Konstrukcja biblioteki musi uwzględniać specyficzne własności maszyny, dla której jest przeznaczona, tzn. jej wykorzystanie, możliwości, istniejące oprogramowanie.

Realizacja biblioteki wymaga rozwiązania dwóch grup zagadnień:

1) numerycznych - związanych z wyborem i algorytmizacją metod przybliżania wybranych funkcji matematycznych,
2) programowych - związanych z wykonaniem programów bibliotecznych w języku assemblera.

Wyboru zestawu funkcji elementarnych realizowanych w bibliotece można dokonać kierując się potrzebami korzystania z tych funkcji, możliwościami ich realizacji za pomocą prostych operacji arytmetycznych oraz możliwościami zastąpienia wyznaczania wartości pewnych funkcji przez wykonywanie prostych operacji arytmetycznych na innych funkcjach lub ich argumentach.

Szczególnie istotne jest to zagadnienie w przypadku bibliotek minikomputerów, z uwagi na stosunkowo małą pamięć operacyjną tych maszyn. W skonstruowanej dla minikomputera P bibliotece LIB-P zaproponowano następujący, najbardziej typowy zestaw funkcji: wartość bezwzględna (ABS), zmiana znaku (ZMZN), pierwiastek kwadratowy (SQRT), sinus (SIN), cosinus (COS),

tangens (TAN), arcus tangens (ATAN), logarytm naturalny (LOGAR), funkcja wykładnicza o podstawie e (EXP).

Indywidualnym problemem dla każdej maszyny cyfrowej jest wybór metody przybliżania funkcji. Wynika on bowiem zarówno z wielkości pamięci operacyjnej, długości słowa maszynowego, sposobu reprezentacji liczby, zestawu operacji arytmetycznych, czasu ich trwania, sposobu ich realizacji, jak również przeznaczenia maszyny cyfrowej.

Minikomputer P przeznaczony jest do sterowania procesami w czasie rzeczywistym. Wszystkie działania wykonywane są w arytmetyce "uzupełnienia do 2". Mnożenie i dzielenie realizowane są programowo. Słowo maszynowe jest 16-bitowe, liczba zmiennoprzecinkowa zajmuje dwa słowa, z czego 8 starszych bitów przypada na cechę, dziewiąty jest bitem znakowym, a pozostałe służą do zapisu mantysy. Przy zapisie liczby stałoprzecinkowej wykorzystywane są również dwa słowa, znak liczb przechowywany jest na najstarszym bicie.

Podstawowym kryterium dla minikomputerów przeznaczonych do sterowania w czasie rzeczywistym jest szybkość obliczeń. Z kolei programowa realizacja mnożenia i dzielenia stałoprzecinkowego, wpływają na relatywnie długi czas wykonywania tych operacji, determinuje konieczność minimalizacji ich liczby w algorytmach.

METODY PRZYBLIŻANIA FUNKCJI Z ZADANĄ DOKŁADNOŚCIĄ

Przybliżenie funkcji z zadaną dokładnością sprowadza się do rozwiązania zadania aproksymacji jednostajnej przy określonym błędzie tej aproksymacji. Obliczanie funkcji aproksymującej w maszynie realizowane jest przez wykonanie pewnej liczby operacji elementarnych przy wprowadzeniu stosunkowo niedużej informacji wejściowej. Możliwości takie dają wielomiany, funkcje wymierne, a także proste procesy iteracyjne.

Najbardziej elementarne rozwinięcie funkcji w szereg potęgowy ma niekorzystne własności. O wiele lepsze efekty daje zastąpienie potęg argumentu wielomianami, które jako przybliżenie dawałoby równomierny rozkład błędu. Z teorii aproksymacji wynika bowiem, że optymalny wielomian w sensie aproksymacji jednostajnej daje równomierny rozkład błędu. Szczególnie przydatne są do tego celu szeregi Czebyszewa [1], [2], charakteryzujące się z reguły dobrą zbieżnością.

Ponadto ortogonalność wielomianów Czebyszewa w przestrzeni L_2 $(-1,1;$ $(1-x^2)^{-1/2})$ umożliwia niezależne od siebie wyznaczanie współczynników szeregów Czebyszewa istotne przy zmianie rzędu wielomianu aproksymującego. Rekurencyjne własności wielomianów Czybyszewa umożliwiają wyznaczanie wartości odpowiedniej sumy częściowej szeregu w ilości kroków nie większej niż w przypadku zwykłego przybliżenia potęgowego.

W bibliotece LIB-P szeregów Czebyszewa użyto do wyznaczania wartości funkcji SIN, COS, TAN, ATAN, wykorzystując przy tym istniejące tablice [3] współczynników sum częściowych. Stabilność wielomianów Czebyszewa zagwarantowana jest jedynie w przedziale <-1, 1>.

Zapewnienie dobrej zbieżności szeregu wymaga często dalszego zawężenia przedziału argumentów, tak że ostateczny przedział jest kompromisem między wymaganiem szybkiej zbieżności szeregu a potrzebą aproksymowania funkcji w możliwie szerokim przedziale. Konieczność sprowadzania całego zakresu zmienności argumentów do określonego przedziału wiąże się bowiem z wykonaniem pewnego ciągu operacji, a ponadto może wnosić dodatkowy błąd do obliczeń.

Zbieżność rozwinięć w szeregi Czebyszewa nie dla wszystkich funkcji jest zadowalająca. Przykładowo szeregi Czebyszewa dla funkcji tangens i logarytm naturalny są słabo zbieżne. W przypadku funkcji TAN skorzystano więc z rozwinięcia funkcji x.ctgx i związków między tangensem a cotangensem ($tgx = \frac{x}{x\ ctgx}$).

W przypadku funkcji LOGAR posłużono się przybliżeniem wymiernym, będącym sumą pewnego nieskończonego szeregu geometrycznego.

Zastosowanie funkcji wymiernych w niewielkim stopniu komplikuje proces obliczeniowy w stosunku do przybliżeń wielomianowych. Szczególnie widoczną poprawę uzyskuje się przez zastosowanie aproksymacji wymiernej dla funkcji e^x. Dokładność może być lepsza nawet o rząd [4].

Jedną z najprostszych aproksymacji wymiernych jest aproksymacja Pade, dająca błąd przybliżenia mały w pobliżu zera, a rosnący w miarę oddalania się od niego. W przedziale <-0.75, 0.75> najlepszym przybliżeniem Pade jest postać o równych stopniach licznika i mianownika, której zaletą jest równość bezwzględnych wartości współczynników, przy jednakowych potęgach argumentu w liczniku i mianowniku. Upraszcza to znacznie proces obliczeniowy.

Zakładając pożądaną dokładność przybliżania funkcji należy liczyć się głównie z 2 typami błędów:

1) błędem metody wynikającym z założonego algorytmu przybliżania funkcji,
2) błędami zaokrągleń wynikającymi z propagacji błędu reprezentacji liczby w maszynie cyfrowej.

Przy wyborze funkcji przybliżającej trzeba zatem zdecydować się na rozwiązanie kompromisowe wynikające z faktu, że zmniejszenie błędu metody wymaga zwiększenia ilości działań, która z kolei mogą być przyczyną komulacji błędów zaokrągleń. Przykładowo w bibliotece LIB-P w przypadku zastosowanych przybliżeń wielomianowych przyjęto jako dopuszczalny maksymalny błąd metody 10^{-6}, co odpowiada w przybliżeniu dokładnym 20 bitom. Błąd zaokrąglenia przy zastosowanych w tym przypadku sumach częściowych szeregu nie powinien bowiem przekroczyć tych wartości.

W przypadku działań stałoprzecinkowych operuje się na dłuższym słowie, co gwarantuje z reguły lepszą dokładność.

Rozwinięcie zastosowane do przybliżania funkcji arcus tangens miało postać:

$$\operatorname{arctg}\left(x \cdot \operatorname{tg}\frac{\pi}{8}\right) = x \cdot \sum_{k=0}^{\infty} a_k T_k\,(2x^2-1) \qquad |x| \leq 1 \tag{4}$$

Dzieląc y określone związkiem (3) przez $\operatorname{tg}\frac{\pi}{8}$ otrzymujemy żądany we wzorze (4) zakres argumentu. Działanie algorytmu przedstawia schemat blokowy 1.

Do realizacji funkcji wykładniczej EXP zastosowano, jak już wspomniano, aproksymację Pade 3 stopnia o równych stopniach licznika i mianownika.

Sprowadzenia argumentu do odpowiedniego przedziału dokonano mając na celu:

- uzyskanie zmienności argumentu, w którym aproksymacja o równych stopniach licznika i mianownika daje najmniejszy błąd spośród przybliżeń Pade;
- przesunięcie argumentu w zakres wartości ujemnych (błąd jest nieco mniejszy niż dla argumentów dodatnich);
- zapewnienie odpowiedniej dokładności reprezentacji stałoprzecinkowej argumentu (ograniczenie górne);
- dalsze zmniejszenie przedziału zmienności argumentu (błąd aproksymacji wówczas maleje).

Dla dowolnego argumentu $-\infty < x < \infty$ liczbę $x_1 = x \log_2 e$ przedstawić można w postaci $x_1 = y+z$, gdzie y jest częścią całkowitą liczby x_1, z częścią ułamkową $(0 \leq z < 1)$.
Korzystając z tożsamości

$$e^x = 2^y \cdot e^{z \ln 2}$$

sprowadzamy argument przybliżonej funkcji do przedziału $\langle 0, \ln 2)$. Następnie odejmujemy od z wartość $\frac{3}{4}$ dla $z \leq \frac{1}{2}$ oraz $1\frac{1}{4}$ dla $z > \frac{1}{2}$, otrzymując

$$e^x = \begin{cases} 2^{y+1} \cdot 2^{\frac{1}{4}} \cdot e^{z_2} & z_1 = (z - 1\frac{1}{4}) \ln 2, \; z > \frac{1}{2} \\ 2^y \cdot 2^{\frac{3}{4}} \cdot e^{z_1} & z_1 = (z - \frac{3}{4}) \ln 2, \quad z \leq \frac{1}{2} \end{cases}$$

gdzie $\frac{3}{4} \cdot \ln 2 < z_1 \leq \frac{1}{4} \ln 2$.
Funkcja przybliżająca ma postać [3]:

$$e^{z_1} = E = \frac{120+60z_1+12z_1^2+z_1^3}{120-60z_1+12z_1^2-z_1^3} \tag{5}$$

Realizacji funkcji (5) dokonuje się po odpowiednim przeskalowaniu. Przyjmując jako współczynnik skalujący 2^{-8}, otrzymuje się

$$E = \frac{L(z_1)}{M(z_1)} \quad \text{gdzie} \quad L(z_1) = 0{,}46875 + (\frac{1}{4} - \frac{1}{64})z_1 + (\frac{1}{16} - \frac{1}{64})z_1^2 + \frac{1}{256} z_1^3$$

$$M(z_1) = L(z_1) - 2(\frac{1}{4} - \frac{1}{64})z_1 - \frac{1}{128} z_1^3$$

Obecnie realizacja wszystkich mnożeń przy wyznaczaniu wartości funkcji E odbywa się przez przesunięcia.

UWAGI KOŃCOWE

W pracy skoncentrowano się na jednym z zagadnień występujących przy konstrukcji bibliotek standardowych funkcji matematycznych dla minikomputera, a mianowicie na problemie wyboru i algorytmizacji metod przybliżania funkcji.

Przedstawiono przy tym przykładowo nietypowe, zaproponowane przez autora, rozwiązania. Drugie z zagadnień, tzn. realizacja programowa algorytmów i organizacja biblioteki, jest bowiem tak ściśle powiązana z budową i oprogramowaniem minikomputera, że nie wydaje się celowe omawianie go w pracy.

Należy podkreślić, że programy biblioteki zostały przetestowane na licznych przykładach. We wszystkich przypadkach uzyskano dokładność lepszą od założonej, wynoszącej 10^{-6}.

LITERATURA

[1] RALSTON A.: Wstęp do analizy numerycznej. PWN, Warszawa 1971.

[2] PASZKOWSKI S.: Zastosowanie numeryczne wielomianów i szeregów Czebyszewa. PWN, Warszawa 1975.

[3] LUSTERNIK L.A. i inni: Obliczanie funkcji elementarnych. PWN, Warszawa 1975.

[4] MEINARDUS G.: Aproksymacja funkcji i jej metody numeryczne. PWN, Warszawa 1968.

[5] ŚWIERNIAK M.: Praca dyplomowa. Gliwice 1980.

Wpłynęło do Redakcji: 17.12.1980 r.
W ostatecznej formie przyjęto: 20.05.1981 r.

Recenzent: Dr inż. Maciej Bargielski

НЕКОТОРЫЕ ПРОБЛЕМЫ КОНСТРУКЦИИ БИБЛИОТЕК МАТЕМАТИЧЕСКИХ ФУНКЦИЙ ДЛЯ МИНИКОМПЬЮТЕРА

Р е з ю м е

В статье представлены некоторые вопросы, связанные с апроксимацией выбранных математических функций в библиотеках миникомпьютеров, а также реализацией программ этих библиотек. Использованы опыты по конструкции библиотеки для миникомпьютера П.

SOME PROBLEMS OF DESIGNING OF MATHEMATICAL FUNCTIONS LIBRARIES

S u m m a r y

In the paper some problems connected with approximation algorithms for mathematical functions in minicomputer libraries are presented. Experiences in designing of library for P minicomputer have been used.