1.2 Грешка. Източници на грешка. Представяне на числата в компютъра.

Както отбелязахме, повечето числени методи включват някаква апроксимация. Ето защо разбирането на идеята за грешка е от много голяма важност за ефективното им използване. Нека първо дадем следните дефиниции:

Дефиниция 1. ***Абсолютна грешка*** *наричаме разликата между точната и приближената стойност при дадена апроксимация:*

$$\varepsilon_a := exact\ value - approximation.$$

Дефиниция 2. ***Относителна грешка*** *дефинираме по следния начин:*

$$\varepsilon_r := \frac{exact\ value - approximation}{exact\ value} = \frac{\varepsilon_a}{exact\ value}.$$

Основните източници на грешка при решаването на една практическа задача са следните:

- Математическият модел – както казахме, математическият модел сам по себе си е една апроксимация на реалността, с други думи самото му съставяне въвежда грешка по отношение на реалния процес.

- Грешка от числения метод – обикновено числените методи се базират на някаква апроксимация, т.е. въвеждат някаква грешка. Тъй като ние на практика не знаем точното решение на съответната математическа задача, обикновено е невъзможно да намерим каква е грешката при въпросната апроксимация. От друга страна, за да разберем дали даден числен метод е приложим, или не, ние трябва да знаем с каква точност той ще реши съответната задача. Затова се налага да се правят оценки на грешката, например да се намери някаква стойност, която тя със сигурност не надминава, или да се определи нейният порядък. Така, при изучаването на различните числени методи в настоящия курс, ние най-често ще се спираме на два основни момента:

  - описание на самия метод;
  - начини за оценка на грешката.

- Грешки от закръгляване – те са свързани с начина, по който числата се представят в компютъра. Ще се спрем по-подробно на този вид грешка в настоящия параграф.

- Грешки от входните данни – математическите модели обикновено зависят от някакви параметри, стойностите на които се определят чрез провеждането на експерименти, правенето на измервания. Дори и най-съвършената техника позволява измерване с определена точност, т.е. стойностите на измерените величини, с които работим, също носят определена грешка.

Задача 1. Постройте в една координатна система графиките на функциите

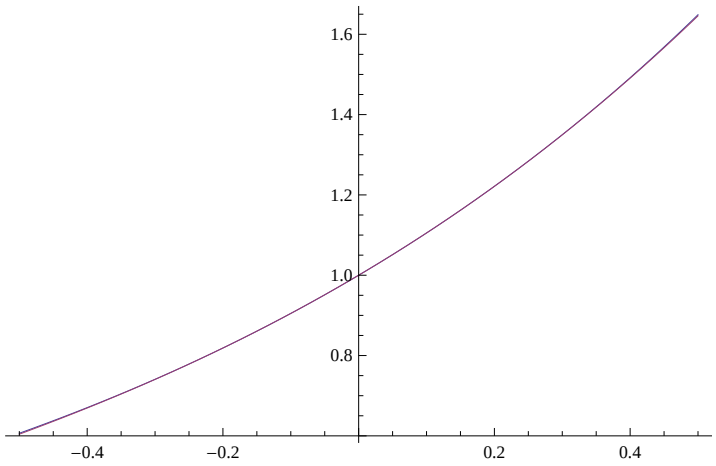
$$f(x) = e^x \text{ и } g(x) = 1 + x + 0.5x^2 + 0.1667x^3$$

в интервала [-0.5,0.5]. Постройте в същия интервал графиките на абсолютната и относителната грешка, които се получават при приближаването на $f(x)$ с $g(x)$, като функции на x.

Решение. Първо построяваме съответните графики, използвайки СКА Mathematica:

```
In[3]:= Plot[{E^x, 1 + x + 0.5 x^2 + 0.1667 x^3}, {x, -0.5, 0.5}]
```

Out[3]=



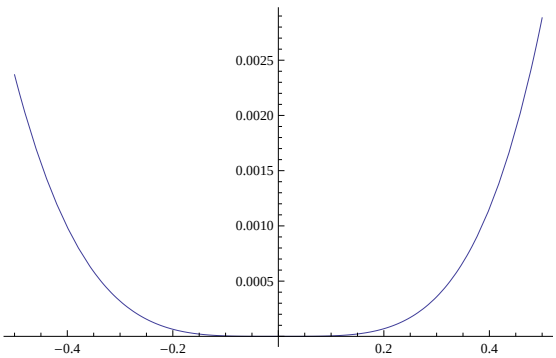
Визуално двете графики изглежда, че съвпадат. Това, което постигаме е, че експоненциалната функция, стойността на която не е лесно да се пресметне, сме приближили с алгебричен полином.

Абсолютната грешка, според Дефиниция 1, е

$$\varepsilon_a(x) = f(x) - g(x) = e^x - (1 + x + 0.5x^2 + 0.1667x^3).$$

```
In[1]:= Plot[E^x - (1 + x + 0.5 x^2 + 0.1667 x^3), {x, -0.5, 0.5}, PlotRange → All]
```

Out[1]=



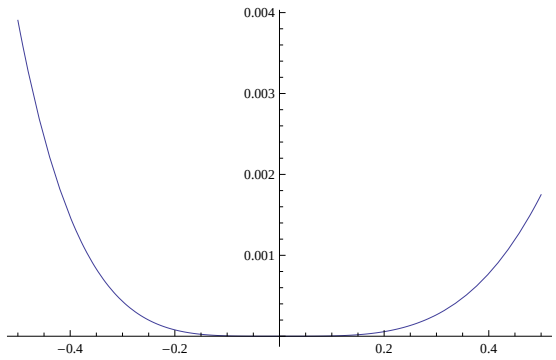
От графиката виждаме, че функцията $g(x)$ приближава сравнително добре функцията $f(x)$ в дадения интервал. Разбира се, дали точността на приближението е достатъчно добра, зависи от конкретния контекст, в който се разглежда задачата.

За относителната грешка, според Дефиниция 2, имаме

$$\varepsilon_r(x) = \frac{f(x) - g(x)}{f(x)} = \frac{e^x - (1 + x + 0.5x^2 + 0.1667x^3)}{e^x}.$$

In[2]:= `Plot[(E^x - (1 + x + 0.5 x^2 + 0.1667 x^3))/E^x, {x, -0.5, 0.5}, PlotRange → All]`

Out[2]=



От последната графика виждаме, че относителната грешка в разглеждания интервал не надминава 0.2%. □

Сега ще се спрем на грешката от закръгляване. Причината за нея, както казахме, е начинът, по който числата се представят в компютъра. По-точно, ще се занимаем с т.нар. числа с плаваща точка (floating-point). При този подход числото се представя чрез дробна част, наречена **мантиса**, и цяло число, което се нарича **експонента** или характеристика по следния начин:

$$m.b^e,$$

където m е мантисата, b е основата на бройната система, в която работим (в компютъра $b = 2$), e – експонентата. Например $156.78 = 0.15678 \times 10^3$ е представянето във вид на число с плаваща точка на числото 156.78 в десетична бройна система. Да обърнем внимание, че обикновено дробната част се нормализира, така че първият знак след десетичната точка да бъде различен от нула.

Предимството на числата с плаваща точка е, че те позволяват представянето както на дроби, така и на много големи числа. От друга страна обаче, се появява т.нар. грешка от закръгляване, тъй като мантисата може да съдържа само краен брой значещи цифри. В компютъра, с t-битова дума могат да се представят най-много 2^t различни реални числа. Очевидно има безброй много числа, които не могат да бъдат представени точно. За тяхното представяне се използва най-близкото число, което може да се представи точно. По този начин въвеждаме грешка от закръгляване. Нещо повече, тъй като има максимално (по абсолютна стойност) число, то при опит да запишем число, което има по-голяма стойност, получаваме т.нар грешка "overflow". Освен това, по аналогична причина, не можем да представяме много малки по абсолютна стойност числа (т.е. близки до нулата). Опитът за записването на такова число води до грешка "underflow". Нека отбележим, че някои компютри заместват "underflow" с нула.

За да илюстрираме ефектите от грешките от закръгляване, нека разгледаме един хипотетичен компютър, който използва десетична бройна система и представя числата с плаваща точка чрез 1-цифрена експонента със знак и 3-цифрена мантиса.

Най-малкото положително число, което можем да представим в този компютър, е 0.100×10^{-9}, а следващото по големина число е 0.101×10^{-9}. Всяко друго число между тези две трябва да бъде апроксимирано. Това ни дава минимална грешка от закръгляване 0.5×10^{-12}.

Най-голямото число, което можем да представим, е 0.999×10^9, докато следващото по-малко число е 0.998×10^9, което дава максимална грешка 0.5×10^6.

Вижда се, че грешката съществено зависи от големината на числата, които апроксимираме. Затова е по-смислено да говорим за относителната вместо за абсолютната грешка. Може да се покаже, че тя е под 5×10^{-3}, т.е. под 0.5%. Да разгледаме следния пример, който ще ни покаже защо относителната грешка е по-добрия показател за точността на приближението. Ясно е, че абсолютна грешка, равна на 1, при число от порядъка на 10^8 е, по принцип, много по-пренебрежима, отколкото грешка от 0.001 при число от порядъка на 10^{-2}. Относителните грешки в този случай са съответно 10^{-8} и 0.1.

Може да се покаже, че за относителната грешка е в сила

$$|\varepsilon_r| < 0.5 \times 10^{-p},$$

където p е броят значещи цифри в мантисата. За числа с двойна точност (double) $p \approx 16$, а с единична (float) – $p \approx 7$.

Като резултат от грешките от закръгляване, дори фундаменталните асоциативни и дистрибутивни закони на алгебрата може и да не са в сила при числени пресмятания. Да разгледаме следните примери:

- Асоциативност на събирането

  $$a + (b + c) = (a + b) + c.$$

  Нека $a = 0.456 \times 10^{-2}$, $b = 0.123 \times 10^{0}$, $c = -0.128 \times 10^{0}$. Тогава

  $$(a + b) + c = 0.128 \times 10^{0} - 0.128 \times 10^{0} = 0,$$
  $$a + (b + c) = 0.456 \times 10^{-2} - 0.500 \times 10^{-2} = -0.440 \times 10^{-3}.$$

  Очевидно първият резултат не е верен и причината за това е **събирането на голямо с малко число**. Можем да разгледаме и още по-показателен пример за този проблем – ако съберем 0.100×10^{0} с 0.100×10^{-3}, резултатът е 0.100×10^{0}, т.е. все едно не сме извършили събирането!

- Асоциативност на умножението

  $$a \times (b \times c) = (a \times b) \times c.$$

  При стойности $a = 10^{-6}$, $b = 10^{-6}$, $c = 10^{8}$ лявата страна на асоциативния закон дава верен резултат. При използване на дясната страна обаче, при изчисленията ще се получи "underflow". Виждаме, че дори при работата с числа, които могат да бъдат представени точно, не сме застраховани от наличието на тази грешка. Следователно **действието на един алгоритъм може да зависи съществено от реда, в който се извършват операциите в него.**

Горните примери ни показват, че всяка аритметична операция, която извършваме, би могла да въведе грешка. Както казахме, числените методи се базират на голям брой аритметични операции, така че това е нещо, което не можем да пренебрегнем при тяхното използване.

За да илюстрираме ефекта на грешките от закръгляване, нека разгледаме следния пример.

Задача 2. Даден е алгебричният полином

$$p(x) = (x-2)^9 = x^9 - 18x^8 + x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512.$$

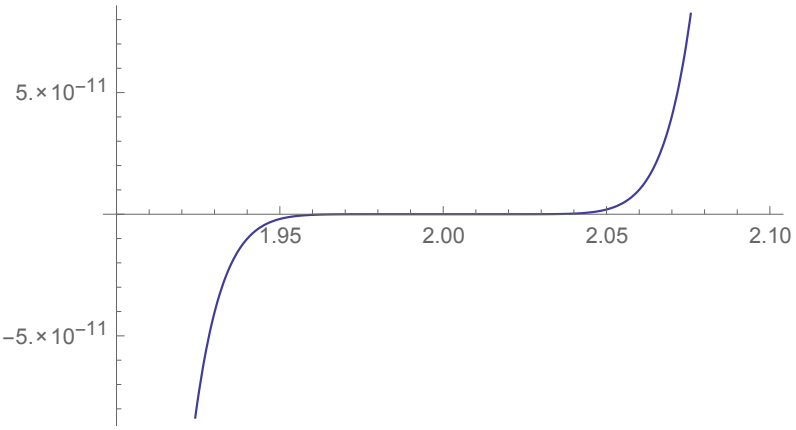
Да се построи неговата графика, като за пресмятане на стойностите му в точката x се използва

а) $p(x) = (x-2)^9$

б) $p(x) = x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512$.

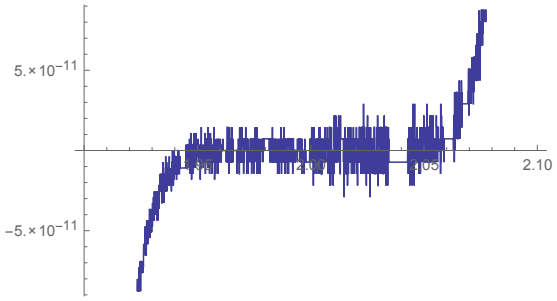
Решение. Решението на а) прилагаме по-долу.

```
f[x_]:=(x-2)^9
Plot[f[x],{x,1.9,2.1}]
```



За б) имаме.

```
p[x_]:=-512+2304 x-4608 x^2+5376 x^3-4032 x^4+2016 x^5-672 x^6+144 x^7-18 x^8+x^9
Plot[p[x],{x,1.9,2.1}]
```



Очевидно във втория случай резултатът е по-лош. Причината е в големия брой аритметични операции, които извършваме при него. Грешките от закръгляване водят до появилия се "шум". □

Вземайки предвид казаното дотук, **числените методи, които използваме трябва да са такива, че грешките от закръгляване да не водят до драстично изменение на резултата. Такива методи се наричат устойчиви.**

Библиография

[1] Боянов, Б.: Лекции по числени методи. Дарба, 2008

[2] Сборник по числени методи – http://www.fmi.uni-sofia.bg/econtent/nummeth

[3] Сендов, Бл., Попов, В.: Числени методи. Първа част. Университетско издателство „Св.Климент Охридски", 1996

[4] Kiusalaas, J.: Numerical Methods in Engineering. Cambridge University Press, 2010

[5] Chapra, S.: Applied Numerical Methods with Matlab for Engineers and Scientists. McGraw Hill, 2012

[6] Бахвалов, Н.С., Лапин, А.В., Чижонков, Е.В.: Численные методы в задачах и упражнениях. Высшая школа, 2000

[7] Hollis, S: Manual for Stewart's Single Variable Calculus. Brooks/Cole, 2008

[8] Antia, H. M.: Numerical Methods for Scientists and Engineers. McGraw-Hill Publishing Company Ltd., 1991