Luvun e numeerinen approksimointi

Juha-Matti Vihtanen

20. toukokuuta 2019

Tiivistelmä

Sarjamuotoinen kokonaislukujen yhteen-, vähennys-, kerto- ja jakolaskuun perustuva esitys luvulle e ja virheanalyysi numeeriselle toteutukselle.

Eksponenttifunktio $f(z)=e^z$ on tunnetusti analyyttinen joukossa \mathbb{C} . Olkoon $z_0\in\mathbb{C}$, $R_1>0$, $R_2>0$ ja $R_2< R_1$. Eksponenttifunktiolla on silloin Taylorin lauseen (lause 7.4 kirjassa [1]) perusteella Taylorin sarja

$$f(z) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(z_0) (z - z_0)^k, \qquad (1)$$

joka suppenee joukossa $B(z_0,R_1)$. Edelleen saman lauseen perusteella Taylorin sarja (1) suppenee joukossa $B(z_0,R_2)$ tasaisesti, missä $B(z_0,r)=\{z\in\mathbb{C}\,|\,|z-z_0|< r\}$.

Tarkastellaan seuraavaksi sarjakehitelmää (1) reaalilukujen joukossa. Eksponenttifunktio toteuttaa reaaliakselilla kaavat

$$f^{(k)}(x) = \frac{d^k}{dx^k} f(x) = \frac{d^k}{dx^k} e^x = e^x$$
. (2)

Sijoittamalla x = 0 kaavaan (2) saadaan

$$f^{(k)}(0) = e^0.$$

Sijoittamalla $z=x,\,z_0=x_0$ ja $x_0=0$ Taylorin sarjan kaavaan saadaan tunnettu erikoistapaus, eksponenttifunktion $f(x)=e^x$ Maclaurinin sarja

$$e^{x} = f(x) = f(z) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(z_{0})(z - z_{0})^{k}$$
$$= \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(x_{0})(x - x_{0})^{k}$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(0) (x-0)^k = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(0) x^k$$
$$= \sum_{k=0}^{\infty} \frac{1}{k!} e^0 x^k = \sum_{k=0}^{\infty} \frac{1}{k!} 1 x^k = \sum_{k=0}^{\infty} \frac{1}{k!} x^k.$$

Sijoittamalla edelleen x=1 saadaan

$$e = e^{1} = e^{x} = \sum_{k=0}^{\infty} \frac{1}{k!} x^{k} = \sum_{k=0}^{\infty} \frac{1}{k!} 1^{k} = \sum_{k=0}^{\infty} \frac{1}{k!} 1$$
$$= \sum_{k=0}^{\infty} \frac{1}{k!}.$$

Tästä saadaan approksimaatiot

$$e \approx \sum_{k=0}^{n} \frac{1}{k!} \quad \forall n \in \mathbb{N}.$$

Olkoon $n \in \mathbb{N}$, $n \geq 2$ ja $p = \lceil \frac{n}{2} \rceil$. Todetaan ensin

$$n-p = n - \left\lceil \frac{n}{2} \right\rceil \ge n - \left(\frac{n}{2} + \frac{1}{2} \right) = n - \frac{n}{2} - \frac{1}{2}$$
$$= \frac{n}{2} - \frac{1}{2} \ge \frac{2}{2} - \frac{1}{2} = \frac{1}{2} > 0.$$

Todetaan myös $n-p \in \mathbb{N}$. Virhe

$$\begin{vmatrix} e - \sum_{k=0}^{n} \frac{1}{k!} \end{vmatrix} = \begin{vmatrix} \sum_{k=0}^{\infty} \frac{1}{k!} - \sum_{k=0}^{n} \frac{1}{k!} \end{vmatrix} = \begin{vmatrix} \sum_{k=n+1}^{\infty} \frac{1}{k!} \end{vmatrix} \le \sum_{k=n+1}^{\infty} \left| \frac{1}{k!} \right|$$

$$= \sum_{k=n+1}^{\infty} \frac{1}{k!} \le \sum_{k=n}^{\infty} \frac{1}{k!} = \sum_{k=n}^{\infty} \prod_{i=1}^{k} \frac{1}{i}$$

$$= \sum_{k=n}^{\infty} \prod_{i=1}^{k-p} \frac{1}{i} \prod_{i=k-p+1}^{k} \frac{1}{i} \le \sum_{k=n}^{\infty} \prod_{i=1}^{k-p} \frac{1}{i} \prod_{i=k-p+1}^{k-p+p} \frac{1}{p}$$

$$= \sum_{k=n}^{\infty} \left[\prod_{i=1}^{k-p} \frac{1}{i} \right] \left(\frac{1}{p} \right)^p = \sum_{k=n}^{\infty} \left(\frac{1}{p} \right)^p \prod_{i=1}^{k-p} \frac{1}{i}$$

$$= \left(\frac{1}{p} \right)^p \sum_{k=n}^{\infty} \prod_{i=1}^{k-p} \frac{1}{i} = \left(\frac{1}{p} \right)^p \sum_{k=n}^{\infty} \frac{1}{(k-p)!}$$

$$= \left(\frac{1}{p} \right)^p \sum_{k=n-p}^{\infty} \frac{1}{k!} \le \left(\frac{1}{p} \right)^p \sum_{k=0}^{\infty} \frac{1}{k!}$$

$$= \left(\frac{1}{\left\lceil \frac{n}{2} \right\rceil} \right)^{\left\lceil \frac{n}{2} \right\rceil} e \le \left(\frac{1}{\left\lceil \frac{n}{2} \right\rceil} \right)^{\frac{n}{2}} e \le \left(\frac{1}{\left\lceil \frac{n}{2} \right\rceil} \right)^{\frac{n}{2}} e$$

$$= e^{\ln \left(\frac{n}{2} \right)^{-1} \frac{n}{2}} e = e^{\left[-\ln \left(\frac{n}{2} \right)\right] \frac{n}{2}} e = e^{-\frac{n}{2} \ln \frac{n}{2} + 1}$$

$$= e^{\ln \left(\frac{n}{2} \right)^{-1} \frac{n}{2}} e = e^{\ln \left(10 - \frac{n}{2} \right) \ln \frac{n}{2} + 1} = e^{\ln \left(10 - \frac{n}{2} \right) \ln \frac{n}{2} + 1} = 10^{-\frac{n}{2} \ln \frac{n}{2} + 1} = 10^{-\frac{n}{2} \ln \frac{n}{2} + 1}$$

Kun valitaan n = 10000 saadaan

$$\begin{vmatrix} e - \sum_{k=0}^{n} \frac{1}{k!} \end{vmatrix} \leq 10^{\frac{-\frac{n}{2} \ln \frac{n}{2} + 1}{\ln 10}} = 10^{\frac{-\frac{10000}{2} \ln \frac{10000}{2} + 1}{\ln 10}}$$
$$= 10^{\frac{-5000 \ln 5000 + 1}{\ln 10}} < 10^{-18495}.$$

Yläraja on siis riittävän suuri lukuesitykselle, jossa on desimaaleja 18495. Voidaan siis käyttää esitystarkuutta, jossa on $2^{14}=16384$ desimaalia. Numeerisesti laskettaessa pitää ottaa huomioon myös katkaisuvirhe, joka muodostuu, kun desimaaliluku esitetään kokonaislukuna, jossa on 16384 merkitsevää numeroa. Kaavan muodossa voidaan todeta

$$e = \sum_{k=0}^{n} \frac{1}{k!} = \sum_{k=0}^{n} \frac{1}{10^{2^{14}}} 10^{2^{14}} \frac{1}{k!} = \frac{1}{10^{2^{14}}} \sum_{k=0}^{n} 10^{2^{14}} \frac{1}{k!}.$$

Sisimmäinen kertolasku toteutetaan rekursion alussa ja ulommainen pilkkua siirtämällä silmukan suoritettua. On huomioitavaa, että yhteenlaskussa muodostunut virhe on katkaisuvirheiden summa ja jakolaskussa algoritmin mukaan desimaalit ovat oikein esitystarkkuuteen asti, jonka jälkeen jakojäännöksestä määräytyy katkaisuvirhe.

Luku

$$a_k = 10^{2^{14}} \frac{1}{k!}$$

Silloin

$$a_k = 10^{2^{14}} \frac{1}{k!} = 10^{2^{14}} \frac{1}{(k-1)!} \frac{1}{k} = \frac{1}{k} 10^{2^{14}} \frac{1}{(k-1)!}$$

= $\frac{1}{k} a_{k-1}$.

Algoritmissa

$$s_0 = 0,$$
 $a_0 = 10^{2^{14}},$
 $s_k = s_{k-1} + a_{k-1},$
 $a_k = \frac{1}{k} a_{k-1}.$

Olkoon b_k algoritmissa lukua a_k vastaava rekrusiivisesti laskettu katkaistu desimaaliluku. Todetaan

$$b_0 = 10^{2^{14}} b_k = \left[\frac{1}{k} b_{k-1} \right] = \frac{1}{k} b_{k-1} - \delta_k.$$

Todetaan edelleen

$$\delta_k \geq 0 \quad \forall k \in \mathbb{N},$$

$$\delta_k < 1 \quad \forall k \in \mathbb{N}.$$

Todistetaan induktiolla, että numeerinen virhe

$$|a_n - b_n| < 2 \quad \forall n \in \mathbb{N}$$
.

Todetaan ensin, että tapaukselle n=0

$$|a_n - b_n| = |a_0 - b_0| = |10^{2^{14}} - 10^{2^{14}}| = 0 < 2.$$

Perusaskel n=1. Erotuksen itseisarvo

$$|a_n - b_n| = |a_1 - b_1| = \left| \frac{1}{1} a_0 - \left\lfloor \frac{1}{1} b_0 \right\rfloor \right| = \left| a_0 - \left\lfloor b_0 \right\rfloor \right|$$
$$= \left| 10^{2^{14}} - \left\lfloor 10^{2^{14}} \right\rfloor \right| = |10^{2^{14}} - 10^{2^{14}}|$$
$$= 0 < 2.$$

Induktioaskel $n=k+1,\; n\geq 2,\; k\geq 1.$ Erotuksen itseisarvo

$$|a_{n} - b_{n}| = |a_{k+1} - b_{k+1}| = \left| \frac{1}{k+1} a_{k} - \left(\frac{1}{k+1} b_{k} - \delta_{k+1} \right) \right|$$

$$= \left| \frac{1}{k+1} a_{k} - \frac{1}{k+1} b_{k} + \delta_{k+1} \right|$$

$$= \left| \frac{1}{k+1} (a_{k} - b_{k}) + \delta_{k+1} \right| \le \left| \frac{1}{k+1} (a_{k} - b_{k}) \right| + |\delta_{k+1}|$$

$$= \left| \frac{1}{k+1} |a_{k} - b_{k}| + |\delta_{k+1}| = \frac{1}{k+1} |a_{k} - b_{k}| + |\delta_{k+1}|$$

$$\le \frac{1}{1+1} |a_{k} - b_{k}| + |\delta_{k+1}| < \frac{1}{2} 2 + 1 = 2.$$

Tämä osoittaa väitteen kaikille $n \in \mathbb{N}$. \square

Kokonaisvirhe

$$= 10^{2^{14}} \left| e - \sum_{k=0}^{n} \frac{1}{k!} \right| + 2n \le 10^{2^{14}} 10^{-18495} + 2 \cdot 10000$$

$$= 10^{16384} 10^{-18495} + 20000 < 10^{18495} 10^{-18495} + 20000$$

$$= 10^{18495 - 18495} + 20000 = 10^{0} + 20000 = 1 + 20000$$

$$= 20001.$$

Saadun numeerisen tuloksen pilkun siirtämisen jälkeen virhe

$$\begin{vmatrix} e - 10^{-2^{14}} \sum_{k=0}^{n} b_k \end{vmatrix} = \begin{vmatrix} 10^{-2^{14}} 10^{2^{14}} e - 10^{-2^{14}} \sum_{k=0}^{n} b_k \end{vmatrix}$$
$$= 10^{-2^{14}} \begin{vmatrix} 10^{2^{14}} e - \sum_{k=0}^{n} b_k \end{vmatrix}$$
$$\leq 10^{-2^{14}} \cdot 20001$$
$$= 20001 \cdot 10^{-2^{14}}.$$

Numeerisen tuloksen viisi viimeistä desimaalia ovat 63091. Viimeisille desimaaleille ylärajana on siis 63091 + 20001 = 83092 ja alarajana 63091 – 20001 = 43090, joten kuudenneksi viimeinen desimaali ei muutu tarkkuusanalyysin rajoissa. Siten numeerinen tulos antaa luvun e katkaistusta desimaaliesityksestä $2^{14} - 5$ merkitsevää numeroa ja $2^{14} - 5 - 1 = 16379$ desimaalia oikein.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <math.h>
char * division(char * dominator, int denominat, char * ans);
char * addition(char * term1, char * term2, char * ans);
#define LENSTR 4
#define NSTR 4096
#define DATALENGTH NSTR*LENSTR
int main(void) {
  int n = 0;
  int k;
 FILE * fp;
  char * a = malloc(DATALENGTH);
  char * sn = malloc(DATALENGTH);
  a[0] = (char)(48+1);
  for (k=1;k<DATALENGTH;k++) {</pre>
   a[k] = (char)48;
 for(k=0;k<DATALENGTH;k++) {</pre>
   sn[k] = (char)48;
 n = 1;
 while(n<10000) {
   system("clear");
   printf("%d/10000\n", n);
    sn = addition(sn,a,sn);
   a = division(a,n,a);
   n++;
 }
 fp = fopen("e.dat", "w");
 fprintf(fp,"%c.",sn[0]);
  for(k=1;k<LENSTR*NSTR;k++) {</pre>
   fprintf(fp, "%c", sn[k]);
 return 0;
}
```

```
char * division(char * dominator, int denominat, char * ans) {
  int k;
  int j;
  int m = denominat;
  int k10;
  int intdata[NSTR];
  int intrmadata[NSTR];
  int divdata[NSTR];
  int rma;
  char tmpchar;
  char data[NSTR][LENSTR];
  char div[NSTR][LENSTR];
  char divstr[NSTR*LENSTR];
  for(k=0;k<NSTR;k++) {
    for(j=0;j<LENSTR;j++) {</pre>
      data[k][j] = dominator[LENSTR*k+j];
  }
  for(k=0;k<NSTR;k++) {
   k10=1000;
    intdata[k]=0;
    for(j=0;j<LENSTR;j++) {</pre>
      tmpchar = data[k][j];
      intdata[k] += k10 * ((int)tmpchar-48);
      k10 /= 10;
   }
  }
 rma = 0;
  for (k=0; k< NSTR; k++) {
    intrmadata[k] = 10000 * rma + intdata[k];
    rma = intrmadata[k] % m;
    divdata[k] = (intrmadata[k]-rma)/m;
  for (k=0; k< NSTR; k++) {
   k10 = 10;
    for(j=0;j<LENSTR;j++) {</pre>
      div[k][LENSTR-j-1] = (char)(10*(divdata[k] % k10)/k10+48);
      divdata[k] -= divdata[k] % k10;
      k10 *= 10;
  }
  for (k=0;k<NSTR;k++) {
    for(j=0;j<LENSTR;j++) {</pre>
      divstr[LENSTR*k+j] = div[k][j];
   }
  }
```

```
for(k=0;k<LENSTR*NSTR;k++) {</pre>
   ans[k] = divstr[k];
 return ans;
char * addition(char * term1, char * term2, char * ans) {
  int k;
  int j;
  int k10;
  int mmr;
  int intdata1[NSTR];
  int intdata2[NSTR];
  int intmmrdata[NSTR];
  char tmpchar;
  char chardata1[NSTR][LENSTR];
  char chardata2[NSTR][LENSTR];
  char sum[NSTR][LENSTR];
  for(k=0;k<NSTR;k++) {
    for(j=0;j<LENSTR;j++) {</pre>
      chardata1[k][j] = term1[LENSTR*k+j];
      chardata2[k][j] = term2[LENSTR*k+j];
  }
  for (k=0; k< NSTR; k++) {
   k10=1000;
    intdata1[k]=0;
    for(j=0;j<LENSTR;j++) {</pre>
      tmpchar = chardata1[k][j];
      intdata1[k] += k10 * ((int)tmpchar-48);
      k10 /= 10;
  for(k=0;k<NSTR;k++) {
   k10=1000;
    intdata2[k] = 0;
    for(j=0;j<LENSTR;j++) {</pre>
      tmpchar = chardata2[k][j];
      intdata2[k] += k10 * ((int)tmpchar-48);
      k10 /= 10;
  }
  mmr = 0;
  for(k=0;k<NSTR;k++) {
    intmmrdata[NSTR-k-1] = mmr + intdata1[NSTR-k-1] + intdata2[NSTR-k-1];
    if(intmmrdata[NSTR-k-1] >= 10000) {
```

```
intmmrdata[NSTR-k-1] -=10000;
      mmr = 1;
   } else {
      mmr = 0;
 }
  for (k=0; k< NSTR; k++) {
   k10 = 10;
   for(j=0;j<LENSTR;j++) {</pre>
      sum[k][LENSTR-j-1] = (char)(10*(intmmrdata[k] % k10)/k10+48);
      intmmrdata[k] -= intmmrdata[k] % k10;
      k10 *= 10;
   }
  }
  for (k=0;k<NSTR;k++) {
   for(j=0;j<LENSTR;j++) {</pre>
      ans[LENSTR*k+j] = sum[k][j];
   }
 }
 return ans;
}
```

Viitteet

[1] John W. Matthews, Russell W. Howell

Complex Analysis for Mathematics and Engineering
Jones and Barlett Publishers, Sudbury, USA, 1997.