Федеральное государственное автономное образовательное учреждение
высшего образования «Санкт-Петербургский политехнический университет
Петра Великого»
Институт компьютерных наук и технологий
Программная инженерия

ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

Отчёт по лабораторным работам

по дисциплине «Вычислительная математика»

Студент гр. 5130904/30008 Ребдев П.А
Руководитель: Скуднева Е.В

Санкт-Петербург
2025

Оглавление

Лабораторная работа №1

1. Задание

Вариант №16

Для функции f(x) = 1 - exp(-x) по узлам x_k= 0.3k (k=0,1, …10) построить полином Лагранжа L(x) 10-й степени и сплайн-функцию S(x). Вычислить значения всех трех функций в точках y_k = 0.15 +0.3k (k=0,1, …9). Результаты отобразить графически.

Используя программу QUANC8, вычислить интегралы:

$\int_{0.5}^{1} |\sin(x) - 0.6|^m \, dx$, для m = -1 и для m = -0.5

2. Решение

2.1 Ход решения

1. Построение вектора из 10 x координат по заранее заданному правилу (x[i] = (0.3d * i))

2. Вычисление вектора значений функции в точках, необходимого для построения интерполяционного полинома Лагранжа и SPLINE

3. Построение интерполяционного полинома Лагранжа с использованием вычисленного на прошлом шаге вектора значений функций

4. Построение SPLINE с использованием вычисленного на втором шаге вектора значений функций

5. Смещение изначального вектора x координат на +0.15

6. Вычисление значений функции, интерполяционного полинома Лагранжа и SPLINE в новых x

7. Вычисление погрешности интерполяционного полинома Лагранжа и SPLINE

2.2 Основная программа

main.cpp

```
#include <iostream>
#include <vector>
#include <cmath>

#include "myFuncs.hpp"
#include "paint.hpp"
#include "quanc8.hpp"

int main()
{
 const unsigned pointsNum = 10;

 std::cout <<  std::fixed << std::setprecision(8);
 std::cout << "\033[1;32m";
 std::cout << "Base points = 0,3 * k (k = 0, 1, ...10)\nf(x) = 1 - exp(x)\nPoints to calculate = 0,15 + 0,3k (k = 0, 1, ...9)\n";
 std::cout << "\033[0m\n";

 std::vector< double > x(pointsNum);

 for (size_t i = 0; i <= pointsNum; ++i)
 {
  x[i] = (0.3d * i);
 }
```

```
 std::vector< double > orig(pointsNum);
 std::vector< double > Lagrange(pointsNum);
 std::vector< double > spline(pointsNum);

 for (size_t i = 0; i < pointsNum; ++i)
 {
  double point = 0.15d + 0.3d * i;
  std::cout << "x = " << point;

  orig[i] = baseFunc(point);
  std::cout << ";\t f(x) = " << orig[i];

  Lagrange[i] = LagrangePolynomial(pointsNum, x, point);
  std::cout << ";\t L(x) = " << Lagrange[i];

  spline[i] = splineNum(pointsNum, x, point);
  std::cout << ";\t S(x) = " << spline[i] << '\n';
 }
 std::cout << '\n';
 paint(x, orig, Lagrange, spline);

 return 0;
}
```

2.3 Интерполяционный полином Лагранжа

myFuncs.cpp

```
double LagrangePolynomial(const unsigned pointsNum, const std::vector< double > & x, double point)
{
 double result = 0;
 for (size_t i = 0; i < pointsNum; ++i)
 {
  double localResult = baseFunc(x[i]);
  for (size_t j = 0; j < pointsNum; ++j)
  {
   if (j != i)
   {
    localResult *= (point - x[j]);
    localResult /= (x[i] - x[j]);
   }
  }
  result += localResult;
 }
 return result;
}
```

2.4 Программа spline

myFuncs.cpp

```
double splineNum(const unsigned pointsNum, const std::vector< double > & x, double point)
{
 std::vector< double > func(pointsNum);
 for (size_t i = 0; i < pointsNum; ++i)
 {
```

```
    func[i] = baseFunc(x[i]);
  }
  tk::spline s(x, func);
  return s(point);
}
```

2.5 Программа QUANC8

main.cpp

```
  double * result = new double;
  double * errest = new double;
  int * nofunR = new int;
  double * posnR = new double;
  int * flag = new int;

  std::cout << "\033[1;32m";
  std::cout << "Find integral from 0.5 to 1 abs(sin(x) -0.6)^m dx. With m = -1 and m = -0.5";
  std::cout << "\033[0m\n";
  quanc8(quanc1, 0.5, 1, std::pow(10, -6), std::pow(10, -7), result, errest, nofunR, posnR, flag);
  std::cout << "I(-1) = " << *result << '\n';
  quanc8(quanc2, 0.5, 1, std::pow(10, -6), std::pow(10, -7), result, errest, nofunR, posnR, flag);
  std::cout << "I(-0.5) = " << *result << "\n\n";

  delete result;
  delete errest;
  delete nofunR;
  delete posnR;
  delete flag;
```

myFuncs.cpp

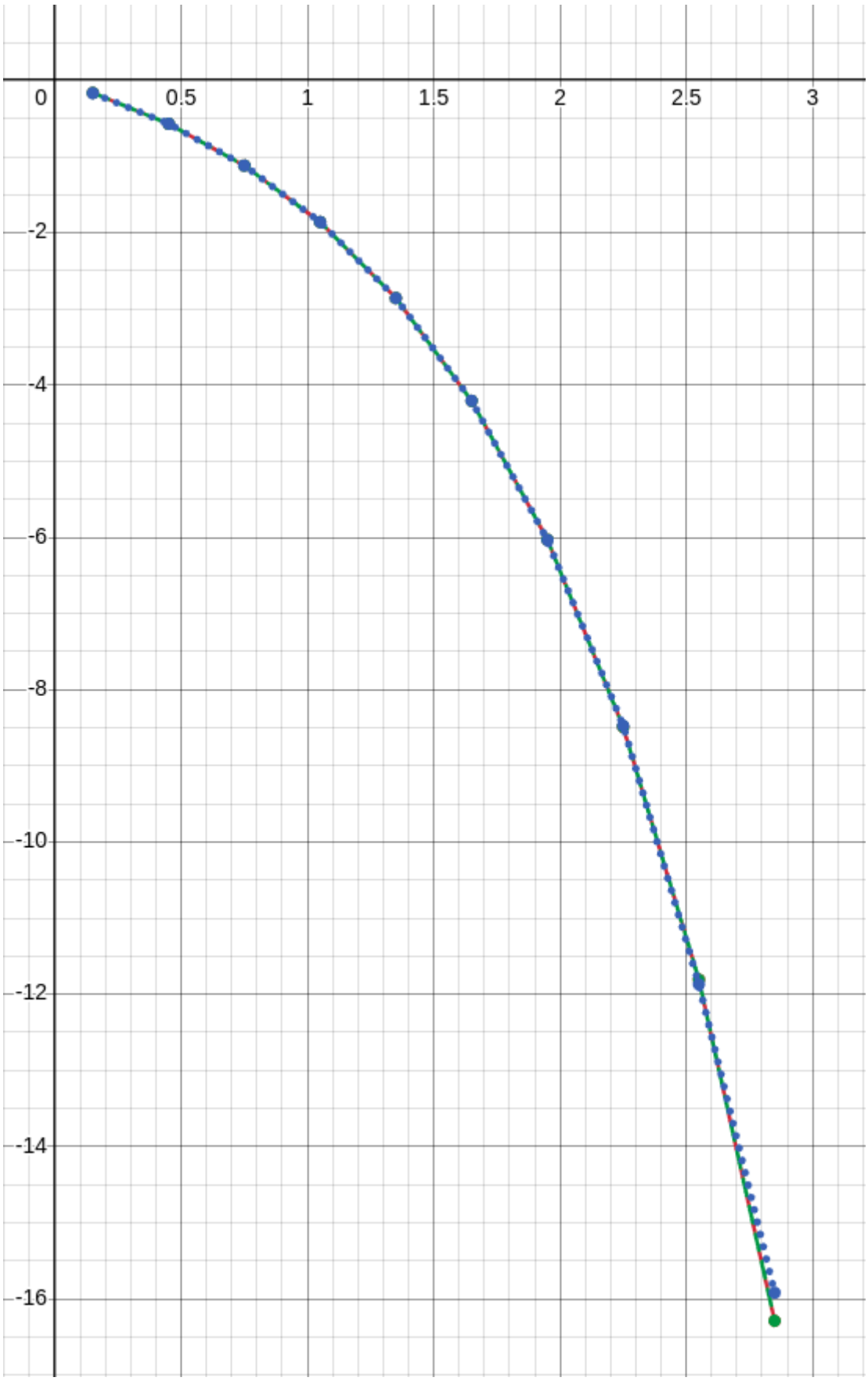
```
double quanc1(double x)
{
  return std::pow(std::abs(std::sin(x) - 0.6), -1);
}
double quanc2(double x)
{
  return std::pow(std::abs(std::sin(x) - 0.6), -0.5);
}
```

3. Результаты

3.1 Нахождение значения функции в точках

В результате вычислений была получена следующая таблица:

x	f(x)	L(x)	|L(x) - f(x)|	S(x)	|S(x) - f(x)|
0.15	-0.161834	-0.161834	0	-0.165899	0,004065
0.45	-0.568312	-0.568312	0	-0.567179	0,001133
0.75	-1.117000	-1.117000	0	-1.117272	0.000272
1.05	-1.857651	-1.857651	0	-1.857430	0,000221
1.35	-2.857426	-2.857426	0	-2.857681	0,000255
1.65	-4.206980	-4.206980	0	-4.205697	0,001283
1.95	-6.028688	-6.028688	0	-6.032910	0,004222
2.25	-8.487736	-8.487736	0	-8.471245	0,016491
2.55	-11.807104	-11.807104	0	-11.867653	0,060549
2.85	-16.287782	-16.287777	$5 * 10^{-6}$	-15.919744	0,368038



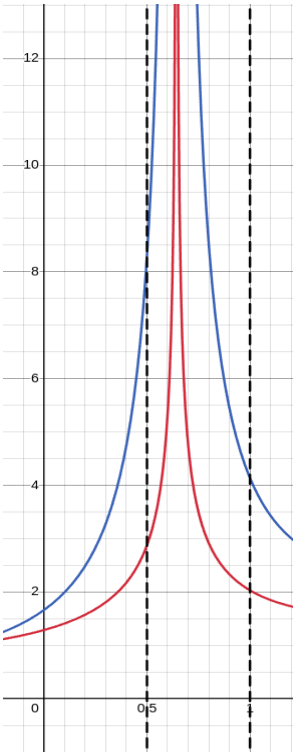
На графике красным обозначена оригинальная функция, зелёным интерполяционный полином Лагранжа, синим сплайн функция

3.2 Вычисления интеграла

При вычисления интеграла с помощью программы QUANC8 были получены следующие результаты:

m = -1; I = 60.089301

m = -0.5; I = 2.2098554



На графики чёрным пунктиром обозначены границы интегрирования, красным m = -0.5, синим m = -1

4. Выводы

В ходе выполнения лабораторной работы с помощью интерполяционного полинома Лагранжа и программы SPLINE были найдены значения функции в точках. При использование типа данных double (16 чисел после мантиссы), и округление до 6го знака после запятой были получены следующие погрешности:

	полином Лагранжа	программа SPLINE
Максимальная разница	0,000005	0,368038
	0,00003%	2,25%
Минимальная разница	0	0,000272
	0	0,0002%

По данным из таблицы можно сделать следующие выводы:

1. Полином Лагранжа отлично подходит для аппроксимации функций, погрешность в 6м разряде начинает появляться при вычисление чисел с двумя целыми разрядами, что является довольно хорошим результатом

2. Высокая погрешность программы SPLINE скорее всего свидетельствует о не качественности конкретной реализации SPLINE алгоритма.

С помощью программы QUANC8 были вычислены значения интеграла для m = -0.5 и m = -1 с абсолютной и относительной погрешностями равными 10^-7. Погрешность вычисления при m = -0.5 составила 0.00000003, а при m = 0.01433744. Можно сделать вывод что программа QUANC8 вычисляет значение интеграла с высокой точностью.

5. Дополнения

5.1 Полный код всех программ

Описание файлов:

- main.cpp отвечает за хранение данных и их вывод в консоль
- myFunc.* отвечают за реализацию интерполяционного полинома Лагранжа, вычисление значения функции и является «прослойкой» между main.cpp, spline.h и quanc8.cpp
- paint.* отвечает за визуализацию вычисленных значений функции, интерполяционного полинома Лагранжа и программы SPLINE

main.cpp

```
#include <iostream>
#include <vector>
#include <cmath>
#include <iomanip>

#include "myFuncs.hpp"
#include "paint.hpp"
#include "quanc8.hpp"

int main()
{
 const unsigned pointsNum = 10;

 std::cout <<  std::fixed << std::setprecision(8);
 std::cout << "\033[1;32m";
 std::cout << "Base points = 0,3 * k (k = 0, 1, ...10)\nf(x) = 1 - exp(x)\nPoints to calculate = 0,15 + 0,3k (k = 0, 1, ...9)\n";
 std::cout << "\033[0m\n";

 std::vector< double > x(pointsNum);

 for (size_t i = 0; i <= pointsNum; ++i)
 {
  x[i] = (0.3d * i);
 }

 std::vector< double > orig(pointsNum);
 std::vector< double > Lagrange(pointsNum);
 std::vector< double > spline(pointsNum);

 for (size_t i = 0; i < pointsNum; ++i)
 {
  double point = 0.15d + 0.3d * i;
  std::cout << "x = " << point;

  orig[i] = baseFunc(point);
```

```
    std::cout << ";\t f(x) = " << orig[i];

    Lagrange[i] = LagrangePolynomial(pointsNum, x, point);
    std::cout << ";\t L(x) = " << Lagrange[i];

    spline[i] = splineNum(pointsNum, x, point);
    std::cout << ";\t S(x) = " << spline[i] << '\n';
  }
  std::cout << '\n';
  paint(x, orig, Lagrange, spline);


  double * result = new double;
  double * errest = new double;
  int * nofunR = new int;
  double * posnR = new double;
  int * flag = new int;

  std::cout << "\033[1;32m";
  std::cout << "Find integral from 0.5 to 1 abs(sin(x) -0.6)^m dx. With m = -1 and m = -0.5";
  std::cout << "\033[0m\n";
  quanc8(quanc1, 0.5, 1, std::pow(10, -7), std::pow(10, -7), result, errest, nofunR, posnR, flag);
  std::cout << "I(-1) = " << *result << '\n';
  quanc8(quanc2, 0.5, 1, std::pow(10, -7), std::pow(10, -7), result, errest, nofunR, posnR, flag);
  std::cout << "I(-0.5) = " << *result << "\n\n";

  delete result;
  delete errest;
  delete nofunR;
  delete posnR;
  delete flag;

  return 0;
}
```

myFuncs.hpp

```
#ifndef MYFUNCS_HPP
#define MYFUNCS_HPP

#include <vector>

double baseFunc(double point);
double LagrangePolynomial(const unsigned pointsNum, const std::vector< double > & x, double point);
double splineNum(const unsigned pointsNum, const std::vector< double > & x, double point);
double quanc1(double x);
double quanc2(double x);

#endif
```

myFuncs.cpp

```
#include "myFuncs.hpp"

#include <cmath>

#include "spline.h"
```

```
double baseFunc(double point)
{
 return (1 - std::exp(point));
}
double LagrangePolynomial(const unsigned pointsNum, const std::vector< double > & x, double point)
{
 double result = 0;
 for (size_t i = 0; i < pointsNum; ++i)
 {
  double localResult = baseFunc(x[i]);
  for (size_t j = 0; j < pointsNum; ++j)
  {
   if (j != i)
   {
    localResult *= (point - x[j]);
    localResult /= (x[i] - x[j]);
   }
  }
  result += localResult;
 }
 return result;
}
double splineNum(const unsigned pointsNum, const std::vector< double > & x, double point)
{
 std::vector< double > func(pointsNum);
 for (size_t i = 0; i < pointsNum; ++i)
 {
  func[i] = baseFunc(x[i]);
 }
 tk::spline s(x, func);
 return s(point);
}
double quanc1(double x)
{
 return std::pow(std::abs(std::sin(x) - 0.6), -1);
}
double quanc2(double x)
{
 return std::pow(std::abs(std::sin(x) - 0.6), -0.5);
}
```

paint.hpp

```
#ifndef PAINT_HPP
#define PAINT_HPP

#include <vector>

using vec = std::vector< double >;

void paint(const vec & x, const vec & orig, const vec & Lagrange, const vec & spline);

#endif
```

paint.cpp

```
#include "paint.hpp"
```

```
#include <SFML/Graphics.hpp>

void paint(const vec & x, const vec & orig, const vec & Lagrange, const vec & spline)
{
 const double scale = 100;
 const double w = 1920;
 const double h = 1080;

 sf::RenderWindow window(sf::VideoMode(w, h), "UwU");

 while (window.isOpen())
 {
  sf::Event event;
  while (window.pollEvent(event))
  {
   if (event.type == sf::Event::Closed)
   {
    window.close();
   }
  }
  window.clear(sf::Color::White);

  for (size_t i = 0; i < x.size(); ++i)
  {
   sf::CircleShape circle(5);
   circle.setPosition((w / 20) - orig[i] * scale, (h / 3) + x[i] * scale);
   circle.setFillColor(sf::Color(255, 0, 0, 150));
   window.draw(circle);

   circle.setPosition((w / 20) - Lagrange[i] * scale, (h / 3) + x[i] * scale);
   circle.setFillColor(sf::Color(0, 255, 0, 150));
   window.draw(circle);

   circle.setPosition((w / 20) - spline[i] * scale, (h / 3) + x[i] * scale);
   circle.setFillColor(sf::Color(0, 0, 255, 150));
   window.draw(circle);
  }
  for (size_t i = 0; i < orig.size() - 1; ++i)
  {
   sf::Vertex line[] = {
    sf::Vertex(sf::Vector2f((w / 20) - orig[i] * scale, (h / 3) + x[i] * scale), sf::Color(255, 0, 0, 150)),
    sf::Vertex(sf::Vector2f((w / 20) - orig[i + 1] * scale, (h / 3) + x[i + 1] * scale), sf::Color(255, 0, 0, 150))
   };
   window.draw(line, 5, sf::Lines);

   line[0] = sf::Vertex(sf::Vector2f((w / 20) - Lagrange[i] * scale, (h / 3) + x[i] * scale), sf::Color(0, 255, 0, 150));
   line[1] = sf::Vertex(sf::Vector2f((w / 20) - Lagrange[i + 1] * scale, (h / 3) + x[i + 1] * scale), sf::Color(0, 255, 0, 150));
   window.draw(line, 5, sf::Lines);

   line[0] = sf::Vertex(sf::Vector2f((w / 20) - spline[i] * scale, (h / 3) + x[i] * scale), sf::Color(0, 0, 255, 150));
   line[1] = sf::Vertex(sf::Vector2f((w / 20) - spline[i + 1] * scale, (h / 3) + x[i + 1] * scale), sf::Color(0, 0, 255, 150));
   window.draw(line, 5, sf::Lines);
  }
  window.display();
 }
}
```

quanc8.hpp

```
#pragma once

int quanc8(double(*fun)(double x), double a, double b,
           double abserr, double relerr,
           double *resultR, double *errestR,
           int *nofunR,
           double *posnR, int *flag);
```

quanc8.cpp

```
#include <math.h>


/*
 fun       : The name of the user defined function f(x).
 a        : The lower limit of integration.
 b        : The upper limit of integration.
 abserr    : The absolute error tolerance. Should be positive.
 relerr    : The relative error tolerance. Should be positive.

 result    : An approximation to the integral I.
 errest    : An estimate of the absolute error in I.
 nofun      : The number of function evaluations used in the
                              calculation of result.
 posn      : If flag < 0, then posn is the point reached when
                              the limit on nfe was approached.
 flag     : Error indicator.
                              = 0, normal return.
                              = 1, invalid user input, relerr < 0, abserr < 0
                              < 0, If flag = -n then n subintervals did not
                                        converge.  A small number (say 10) of
                                        unconverged subintervals may be acceptable.
                                        Check posn for further information.
*/


int quanc8(double(*fun)(double x), double a, double b,
           double abserr, double relerr,
           double *resultR, double *errestR,
           int *nofunR,
           double *posnR, int *flag)

{  /* begin function quanc8 */

  /* this code has been translated from fortran, hence use
  elements 1 .. n */
           double w0, w1, w2, w3, w4, area, x0, f0, stone, step, cor11;
           double qprev, qnow, qdiff, qleft, esterr, tolerr;
           static double qright[32], f[17], x[17], fsave[9][31];
           static double xsave[9][31];
           double posn, result, errest;
           double temp, temp1;
           int nofun;
           int levmin, levmax, levout, nomax, nofin, lev, nim, i, j, ii;
```

```
/* check user input */
if (abserr < 0.0 || relerr < 0.0)
{
        *flag = 1;
        *resultR = 0.0;
        *errestR = 0.0;
        *posnR = 0.0;
        *nofunR = 0;
        return (0);
}

/*  ***  stage 1 ***  general initialization
set constants.  */
levmin = 1;
levmax = 30;
levout = 6;
nomax = 5000;

{
        ii = 1;
        for (i = 0; i <= levout; i++) ii *= 2;
} /* ---> ii = 2 ** (levout+1) */
nofin = nomax - 8 * (levmax - levout + ii);
/* note that there will be trouble when nofun reaches nofin */

temp = 14175.0;
w0 = 3956.0 / temp;
w1 = 23552.0 / temp;
w2 = -3712.0 / temp;
w3 = 41984.0 / temp;
w4 = -18160.0 / temp;

/* initialize running sums to zero. */
*flag = 0;
posn = 0.0;
result = 0.0;
cor11 = 0.0;
errest = 0.0;
area = 0.0;
nofun = 0;
if (a == b) goto ExitQuanc8;

/* ***  stage 2 ***  initialization for first interval */
lev = 0;
nim = 1;
x0 = a;
x[16] = b;
qprev = 0.0;

/* set up evenly spaced panels */
f0 = (*fun) (x0);
stone = (b - a) / 16.0;
x[8] = 0.5 * (x0 + x[16]);
x[4] = 0.5 * (x0 + x[8]);
x[12] = 0.5 * (x[8] + x[16]);
x[2] = 0.5 * (x0 + x[4]);
```

```
        x[6] = 0.5 * (x[4] + x[8]);
        x[10] = 0.5 * (x[8] + x[12]);
        x[14] = 0.5 * (x[12] + x[16]);
        for (j = 2; j <= 16; j = j + 2)  f[j] = (*fun) (x[j]);
        nofun = 9;

        /* ***  stage 3 ***  central calculation
        requires qprev,x0,x2,x4,...,x16,f0,f2,f4,...,f16.
        calculates x1,x3,...x15, f1,f3,...f15,qleft,qright,qnow,qdiff,area.
        */
Stage3:
        x[1] = 0.5 * (x0 + x[2]);
        f[1] = (*fun) (x[1]);
        for (j = 3; j <= 15; j = j + 2)
        {
                x[j] = 0.5 * (x[j - 1] + x[j + 1]);
                f[j] = (*fun) (x[j]);
        }
        nofun += 8;
        step = (x[16] - x0) / 16.0;
        qleft = (w0 * (f0 + f[8]) + w1 * (f[1] + f[7]) + w2 * (f[2] + f[6])
                + w3 * (f[3] + f[5]) + w4 * f[4]) * step;
        qright[lev + 1] = (w0 * (f[8] + f[16]) + w1 * (f[9] + f[15]) + w2 * (f[10] + f[14])
                + w3 * (f[11] + f[13]) + w4 * f[12]) * step;
        qnow = qleft + qright[lev + 1];
        qdiff = qnow - qprev;
        area += qdiff;

        /* ***  stage 4 *** interval convergence test */
        esterr = fabs(qdiff) / 1023.0;
        {
                tolerr = abserr;
                temp = relerr * fabs(area);
                if (temp > tolerr) tolerr = temp;
        }  /* ---> tolerr = max(abserr, relerr * abs(area)) */
        tolerr *= (step / stone);

        if (lev < levmin)     goto Stage5;  /* keep subdividing              */
        if (lev >= levmax)    goto Stage6B; /* too many nested levels        */
        if (nofun > nofin)    goto Stage6;  /* close to limit on fn calls    */
        if (esterr <= tolerr) goto Stage7;  /* this interval has converged   */

                                                                        /* ***  stage 5  ***  no
convergence
                                                                        locate next interval.          */
Stage5:
        nim *= 2;
        ++lev;

        /* store right hand elements for future use. */
        for (i = 1; i <= 8; i++)
        {
                fsave[i][lev] = f[i + 8];
                xsave[i][lev] = x[i + 8];
        }
```

```
        /* assemble left hand elements for immediate use. */
        qprev = qleft;
        for (i = 1; i <= 8; i++)
        {
                j = (-i);
                f[2 * j + 18] = f[j + 9];
                x[2 * j + 18] = x[j + 9];
        }

        goto Stage3;

        /* ***  stage 6  ***  trouble section
        number of function values is about to exceed limit.  */
Stage6:
        nofin *= 2;
        levmax = levout;
        posn = x0;   /* this is our trouble spot */
        goto Stage7;

        /* current level is levmax. */
Stage6B:
        --(*flag);   /* another interval has not converged */

                                /* ***  stage 7  ***  interval converged
                                add contributions into running sums.  */
Stage7:
        result += qnow;
        errest += esterr;
        cor11 += qdiff / 1023.0;

        /* locate next interval. */
        while (nim != (2 * (nim / 2)))
        {
                nim /= 2;
                --lev;
        }
        ++nim;
        if (lev <= 0) goto Stage8;

        /* assemble elements required for the next interval. */
        qprev = qright[lev];
        x0 = x[16];
        f0 = f[16];
        for (i = 1; i <= 8; i++)
        {
                f[2 * i] = fsave[i][lev];
                x[2 * i] = xsave[i][lev];
        }
        goto Stage3;

        /* ***  stage 8  ***  finalize and return */
Stage8:
        result += cor11;

        /* make sure errest not less than roundoff level. */
        if (errest == 0.0) goto ExitQuanc8;
```

```
        temp1 = fabs(result);
        temp = temp1 + errest;
        while (temp == temp1)
        {
                errest *= 2.0;
                temp = temp1 + errest;
        }

ExitQuanc8:
        /* --- return variables --- */
        *resultR = result;
        *errestR = errest;
        *posnR = posn;
        *nofunR = nofun;
        return (0);
}  /* end of quanc8() */
```

spline.h

```
/*
 * spline.h
 *
 * simple cubic spline interpolation library without external
 * dependencies
 *
 * ---------------------------------------------------------------------
 * Copyright (C) 2011, 2014, 2016, 2021 Tino Kluge (ttk448 at gmail.com)
 *
 *  This program is free software; you can redistribute it and/or
 *  modify it under the terms of the GNU General Public License
 *  as published by the Free Software Foundation; either version 2
 *  of the License, or (at your option) any later version.
 *
 *  This program is distributed in the hope that it will be useful,
 *  but WITHOUT ANY WARRANTY; without even the implied warranty of
 *  MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
 *  GNU General Public License for more details.
 *
 *  You should have received a copy of the GNU General Public License
 *  along with this program.  If not, see <http://www.gnu.org/licenses/>.
 * ---------------------------------------------------------------------
 *
 */


#ifndef TK_SPLINE_H
#define TK_SPLINE_H

#include <cstdio>
#include <cassert>
#include <cmath>
#include <vector>
#include <algorithm>
#ifdef HAVE_SSTREAM
#include <sstream>
#include <string>
#endif // HAVE_SSTREAM
```

```
// not ideal but disable unused-function warnings
// (we get them because we have implementations in the header file,
// and this is because we want to be able to quickly separate them
// into a cpp file if necessary)
#pragma GCC diagnostic push
#pragma GCC diagnostic ignored "-Wunused-function"

// unnamed namespace only because the implementation is in this
// header file and we don't want to export symbols to the obj files
namespace
{

namespace tk
{

// spline interpolation
class spline
{
public:
   // spline types
   enum spline_type {
      linear = 10,          // linear interpolation
      cspline = 30,         // cubic splines (classical C^2)
      cspline_hermite = 31    // cubic hermite splines (local, only C^1)
   };

   // boundary condition type for the spline end-points
   enum bd_type {
      first_deriv = 1,
      second_deriv = 2
   };

protected:
   std::vector<double> m_x,m_y;          // x,y coordinates of points
   // interpolation parameters
   // f(x) = a_i + b_i*(x-x_i) + c_i*(x-x_i)^2 + d_i*(x-x_i)^3
   // where a_i = y_i, or else it won't go through grid points
   std::vector<double> m_b,m_c,m_d;       // spline coefficients
   double m_c0;                          // for left extrapolation
   spline_type m_type;
   bd_type m_left, m_right;
   double  m_left_value, m_right_value;
   bool m_made_monotonic;
   void set_coeffs_from_b();             // calculate c_i, d_i from b_i
   size_t find_closest(double x) const;    // closest idx so that m_x[idx]<=x

public:
   // default constructor: set boundary condition to be zero curvature
   // at both ends, i.e. natural splines
   spline(): m_type(cspline),
      m_left(second_deriv), m_right(second_deriv),
      m_left_value(0.0), m_right_value(0.0), m_made_monotonic(false)
   {
      ;
   }
```

```
  spline(const std::vector<double>& X, const std::vector<double>& Y,
         spline_type type = cspline,
         bool make_monotonic = false,
         bd_type left  = second_deriv, double left_value  = 0.0,
         bd_type right = second_deriv, double right_value = 0.0
        ):
      m_type(type),
      m_left(left), m_right(right),
      m_left_value(left_value), m_right_value(right_value),
      m_made_monotonic(false) // false correct here: make_monotonic() sets it
  {
      this->set_points(X,Y,m_type);
      if(make_monotonic) {
          this->make_monotonic();
      }
  }


  // modify boundary conditions: if called it must be before set_points()
  void set_boundary(bd_type left, double left_value,
                    bd_type right, double right_value);

  // set all data points (cubic_spline=false means linear interpolation)
  void set_points(const std::vector<double>& x,
                  const std::vector<double>& y,
                  spline_type type=cspline);

  // adjust coefficients so that the spline becomes piecewise monotonic
  // where possible
  //   this is done by adjusting slopes at grid points by a non-negative
  //   factor and this will break C^2
  //   this can also break boundary conditions if adjustments need to
  //   be made at the boundary points
  // returns false if no adjustments have been made, true otherwise
  bool make_monotonic();

  // evaluates the spline at point x
  double operator() (double x) const;
  double deriv(int order, double x) const;

  // returns the input data points
  std::vector<double> get_x() const { return m_x; }
  std::vector<double> get_y() const { return m_y; }
  double get_x_min() const { assert(!m_x.empty()); return m_x.front(); }
  double get_x_max() const { assert(!m_x.empty()); return m_x.back(); }

#ifdef HAVE_SSTREAM
  // spline info string, i.e. spline type, boundary conditions etc.
  std::string info() const;
#endif // HAVE_SSTREAM

};



namespace internal
```

```
{

// band matrix solver
class band_matrix
{
private:
   std::vector< std::vector<double> > m_upper;  // upper band
   std::vector< std::vector<double> > m_lower;  // lower band
public:
   band_matrix() {};                            // constructor
   band_matrix(int dim, int n_u, int n_l);      // constructor
   ~band_matrix() {};                           // destructor
   void resize(int dim, int n_u, int n_l);      // init with dim,n_u,n_l
   int dim() const;                             // matrix dimension
   int num_upper() const
   {
      return (int)m_upper.size()-1;
   }
   int num_lower() const
   {
      return (int)m_lower.size()-1;
   }
   // access operator
   double & operator () (int i, int j);           // write
   double   operator () (int i, int j) const;     // read
   // we can store an additional diagonal (in m_lower)
   double& saved_diag(int i);
   double  saved_diag(int i) const;
   void lu_decompose();
   std::vector<double> r_solve(const std::vector<double>& b) const;
   std::vector<double> l_solve(const std::vector<double>& b) const;
   std::vector<double> lu_solve(const std::vector<double>& b,
                                bool is_lu_decomposed=false);

};

} // namespace internal




// ---------------------------------------------------------------------
// implementation part, which could be separated into a cpp file
// ---------------------------------------------------------------------

// spline implementation
// -----------------------

void spline::set_boundary(spline::bd_type left, double left_value,
                          spline::bd_type right, double right_value)
{
   assert(m_x.size()==0);          // set_points() must not have happened yet
   m_left=left;
   m_right=right;
   m_left_value=left_value;
   m_right_value=right_value;
```

```
}


void spline::set_coeffs_from_b()
{
   assert(m_x.size()==m_y.size());
   assert(m_x.size()==m_b.size());
   assert(m_x.size()>2);
   size_t n=m_b.size();
   if(m_c.size()!=n)
      m_c.resize(n);
   if(m_d.size()!=n)
      m_d.resize(n);

   for(size_t i=0; i<n-1; i++) {
      const double h  = m_x[i+1]-m_x[i];
      // from continuity and differentiability condition
      m_c[i] = ( 3.0*(m_y[i+1]-m_y[i])/h - (2.0*m_b[i]+m_b[i+1]) ) / h;
      // from differentiability condition
      m_d[i] = ( (m_b[i+1]-m_b[i])/(3.0*h) - 2.0/3.0*m_c[i] ) / h;
   }

   // for left extrapolation coefficients
   m_c0 = (m_left==first_deriv) ? 0.0 : m_c[0];
}

void spline::set_points(const std::vector<double>& x,
                  const std::vector<double>& y,
                  spline_type type)
{
   assert(x.size()==y.size());
   assert(x.size()>2);
   m_type=type;
   m_made_monotonic=false;
   m_x=x;
   m_y=y;
   int n = (int) x.size();
   // check strict monotonicity of input vector x
   for(int i=0; i<n-1; i++) {
      assert(m_x[i]<m_x[i+1]);
   }


   if(type==linear) {
      // linear interpolation
      m_d.resize(n);
      m_c.resize(n);
      m_b.resize(n);
      for(int i=0; i<n-1; i++) {
         m_d[i]=0.0;
         m_c[i]=0.0;
         m_b[i]=(m_y[i+1]-m_y[i])/(m_x[i+1]-m_x[i]);
      }
      // ignore boundary conditions, set slope equal to the last segment
      m_b[n-1]=m_b[n-2];
      m_c[n-1]=0.0;
```

```
    m_d[n-1]=0.0;
  } else if(type==cspline) {
    // classical cubic splines which are C^2 (twice cont differentiable)
    // this requires solving an equation system

    // setting up the matrix and right hand side of the equation system
    // for the parameters b[]
    internal::band_matrix A(n,1,1);
    std::vector<double>  rhs(n);
    for(int i=1; i<n-1; i++) {
      A(i,i-1)=1.0/3.0*(x[i]-x[i-1]);
      A(i,i)=2.0/3.0*(x[i+1]-x[i-1]);
      A(i,i+1)=1.0/3.0*(x[i+1]-x[i]);
      rhs[i]=(y[i+1]-y[i])/(x[i+1]-x[i]) - (y[i]-y[i-1])/(x[i]-x[i-1]);
    }
    // boundary conditions
    if(m_left == spline::second_deriv) {
      // 2*c[0] = f''
      A(0,0)=2.0;
      A(0,1)=0.0;
      rhs[0]=m_left_value;
    } else if(m_left == spline::first_deriv) {
      // b[0] = f', needs to be re-expressed in terms of c:
      // (2c[0]+c[1])(x[1]-x[0]) = 3 ((y[1]-y[0])/(x[1]-x[0]) - f')
      A(0,0)=2.0*(x[1]-x[0]);
      A(0,1)=1.0*(x[1]-x[0]);
      rhs[0]=3.0*((y[1]-y[0])/(x[1]-x[0])-m_left_value);
    } else {
      assert(false);
    }
    if(m_right == spline::second_deriv) {
      // 2*c[n-1] = f''
      A(n-1,n-1)=2.0;
      A(n-1,n-2)=0.0;
      rhs[n-1]=m_right_value;
    } else if(m_right == spline::first_deriv) {
      // b[n-1] = f', needs to be re-expressed in terms of c:
      // (c[n-2]+2c[n-1])(x[n-1]-x[n-2])
      // = 3 (f' - (y[n-1]-y[n-2])/(x[n-1]-x[n-2]))
      A(n-1,n-1)=2.0*(x[n-1]-x[n-2]);
      A(n-1,n-2)=1.0*(x[n-1]-x[n-2]);
      rhs[n-1]=3.0*(m_right_value-(y[n-1]-y[n-2])/(x[n-1]-x[n-2]));
    } else {
      assert(false);
    }

    // solve the equation system to obtain the parameters c[]
    m_c=A.lu_solve(rhs);

    // calculate parameters b[] and d[] based on c[]
    m_d.resize(n);
    m_b.resize(n);
    for(int i=0; i<n-1; i++) {
      m_d[i]=1.0/3.0*(m_c[i+1]-m_c[i])/(x[i+1]-x[i]);
      m_b[i]=(y[i+1]-y[i])/(x[i+1]-x[i])
          - 1.0/3.0*(2.0*m_c[i]+m_c[i+1])*(x[i+1]-x[i]);
```

```
    }
    // for the right extrapolation coefficients (zero cubic term)
    // f_{n-1}(x) = y_{n-1} + b*(x-x_{n-1}) + c*(x-x_{n-1})^2
    double h=x[n-1]-x[n-2];
    // m_c[n-1] is determined by the boundary condition
    m_d[n-1]=0.0;
    m_b[n-1]=3.0*m_d[n-2]*h*h+2.0*m_c[n-2]*h+m_b[n-2];   // = f'_{n-2}(x_{n-1})
    if(m_right==first_deriv)
      m_c[n-1]=0.0;   // force linear extrapolation

  } else if(type==cspline_hermite) {
    // hermite cubic splines which are C^1 (cont. differentiable)
    // and derivatives are specified on each grid point
    // (here we use 3-point finite differences)
    m_b.resize(n);
    m_c.resize(n);
    m_d.resize(n);
    // set b to match 1st order derivative finite difference
    for(int i=1; i<n-1; i++) {
      const double h  = m_x[i+1]-m_x[i];
      const double hl = m_x[i]-m_x[i-1];
      m_b[i] = -h/(hl*(hl+h))*m_y[i-1] + (h-hl)/(hl*h)*m_y[i]
             +  hl/(h*(hl+h))*m_y[i+1];
    }
    // boundary conditions determine b[0] and b[n-1]
    if(m_left==first_deriv) {
      m_b[0]=m_left_value;
    } else if(m_left==second_deriv) {
      const double h = m_x[1]-m_x[0];
      m_b[0]=0.5*(-m_b[1]-0.5*m_left_value*h+3.0*(m_y[1]-m_y[0])/h);
    } else {
      assert(false);
    }
    if(m_right==first_deriv) {
      m_b[n-1]=m_right_value;
      m_c[n-1]=0.0;
    } else if(m_right==second_deriv) {
      const double h = m_x[n-1]-m_x[n-2];
      m_b[n-1]=0.5*(-m_b[n-2]+0.5*m_right_value*h+3.0*(m_y[n-1]-m_y[n-2])/h);
      m_c[n-1]=0.5*m_right_value;
    } else {
      assert(false);
    }
    m_d[n-1]=0.0;

    // parameters c and d are determined by continuity and differentiability
    set_coeffs_from_b();

  } else {
    assert(false);
  }

  // for left extrapolation coefficients
  m_c0 = (m_left==first_deriv) ? 0.0 : m_c[0];
}
```

```
bool spline::make_monotonic()
{
   assert(m_x.size()==m_y.size());
   assert(m_x.size()==m_b.size());
   assert(m_x.size()>2);
   bool modified = false;
   const int n=(int)m_x.size();
   // make sure: input data monotonic increasing --> b_i>=0
   //           input data monotonic decreasing --> b_i<=0
   for(int i=0; i<n; i++) {
      int im1 = std::max(i-1, 0);
      int ip1 = std::min(i+1, n-1);
      if( ((m_y[im1]<=m_y[i]) && (m_y[i]<=m_y[ip1]) && m_b[i]<0.0) ||
         ((m_y[im1]>=m_y[i]) && (m_y[i]>=m_y[ip1]) && m_b[i]>0.0) ) {
         modified=true;
         m_b[i]=0.0;
      }
   }
   // if input data is monotonic (b[i], b[i+1], avg have all the same sign)
   // ensure a sufficient criteria for monotonicity is satisfied:
   //    sqrt(b[i]^2+b[i+1]^2) <= 3 |avg|, with avg=(y[i+1]-y[i])/h,
   for(int i=0; i<n-1; i++) {
      double h = m_x[i+1]-m_x[i];
      double avg = (m_y[i+1]-m_y[i])/h;
      if( avg==0.0 && (m_b[i]!=0.0 || m_b[i+1]!=0.0) ) {
         modified=true;
         m_b[i]=0.0;
         m_b[i+1]=0.0;
      } else if( (m_b[i]>=0.0 && m_b[i+1]>=0.0 && avg>0.0) ||
               (m_b[i]<=0.0 && m_b[i+1]<=0.0 && avg<0.0) ) {
         // input data is monotonic
         double r = sqrt(m_b[i]*m_b[i]+m_b[i+1]*m_b[i+1])/std::fabs(avg);
         if(r>3.0) {
            // sufficient criteria for monotonicity: r<=3
            // adjust b[i] and b[i+1]
            modified=true;
            m_b[i]   *= (3.0/r);
            m_b[i+1] *= (3.0/r);
         }
      }
   }

   if(modified==true) {
      set_coeffs_from_b();
      m_made_monotonic=true;
   }

   return modified;
}

// return the closest idx so that m_x[idx] <= x (return 0 if x<m_x[0])
size_t spline::find_closest(double x) const
{
   std::vector<double>::const_iterator it;
   it=std::upper_bound(m_x.begin(),m_x.end(),x);      // *it > x
   size_t idx = std::max( int(it-m_x.begin())-1, 0);   // m_x[idx] <= x
```

```
  return idx;
}

double spline::operator() (double x) const
{
  // polynomial evaluation using Horner's scheme
  // TODO: consider more numerically accurate algorithms, e.g.:
  //   - Clenshaw
  //   - Even-Odd method by A.C.R. Newbery
  //   - Compensated Horner Scheme
  size_t n=m_x.size();
  size_t idx=find_closest(x);

  double h=x-m_x[idx];
  double interpol;
  if(x<m_x[0]) {
    // extrapolation to the left
    interpol=(m_c0*h + m_b[0])*h + m_y[0];
  } else if(x>m_x[n-1]) {
    // extrapolation to the right
    interpol=(m_c[n-1]*h + m_b[n-1])*h + m_y[n-1];
  } else {
    // interpolation
    interpol=((m_d[idx]*h + m_c[idx])*h + m_b[idx])*h + m_y[idx];
  }
  return interpol;
}

double spline::deriv(int order, double x) const
{
  assert(order>0);
  size_t n=m_x.size();
  size_t idx = find_closest(x);

  double h=x-m_x[idx];
  double interpol;
  if(x<m_x[0]) {
    // extrapolation to the left
    switch(order) {
    case 1:
      interpol=2.0*m_c0*h + m_b[0];
      break;
    case 2:
      interpol=2.0*m_c0;
      break;
    default:
      interpol=0.0;
      break;
    }
  } else if(x>m_x[n-1]) {
    // extrapolation to the right
    switch(order) {
    case 1:
      interpol=2.0*m_c[n-1]*h + m_b[n-1];
      break;
    case 2:
```

```
        interpol=2.0*m_c[n-1];
        break;
      default:
        interpol=0.0;
        break;
      }
    } else {
      // interpolation
      switch(order) {
      case 1:
        interpol=(3.0*m_d[idx]*h + 2.0*m_c[idx])*h + m_b[idx];
        break;
      case 2:
        interpol=6.0*m_d[idx]*h + 2.0*m_c[idx];
        break;
      case 3:
        interpol=6.0*m_d[idx];
        break;
      default:
        interpol=0.0;
        break;
      }
    }
    return interpol;
}

#ifdef HAVE_SSTREAM
std::string spline::info() const
{
    std::stringstream ss;
    ss << "type " << m_type << ", left boundary deriv " << m_left << " = ";
    ss << m_left_value << ", right boundary deriv " << m_right << " = ";
    ss << m_right_value << std::endl;
    if(m_made_monotonic) {
      ss << "(spline has been adjusted for piece-wise monotonicity)";
    }
    return ss.str();
}
#endif // HAVE_SSTREAM


namespace internal
{

// band_matrix implementation
// -------------------------

band_matrix::band_matrix(int dim, int n_u, int n_l)
{
    resize(dim, n_u, n_l);
}
void band_matrix::resize(int dim, int n_u, int n_l)
{
    assert(dim>0);
    assert(n_u>=0);
    assert(n_l>=0);
```

```
   m_upper.resize(n_u+1);
   m_lower.resize(n_l+1);
   for(size_t i=0; i<m_upper.size(); i++) {
      m_upper[i].resize(dim);
   }
   for(size_t i=0; i<m_lower.size(); i++) {
      m_lower[i].resize(dim);
   }
}
int band_matrix::dim() const
{
   if(m_upper.size()>0) {
      return m_upper[0].size();
   } else {
      return 0;
   }
}


// defines the new operator (), so that we can access the elements
// by A(i,j), index going from i=0,...,dim()-1
double & band_matrix::operator () (int i, int j)
{
   int k=j-i;       // what band is the entry
   assert( (i>=0) && (i<dim()) && (j>=0) && (j<dim()) );
   assert( (-num_lower()<=k) && (k<=num_upper()) );
   // k=0 -> diagonal, k<0 lower left part, k>0 upper right part
   if(k>=0)    return m_upper[k][i];
   else        return m_lower[-k][i];
}
double band_matrix::operator () (int i, int j) const
{
   int k=j-i;       // what band is the entry
   assert( (i>=0) && (i<dim()) && (j>=0) && (j<dim()) );
   assert( (-num_lower()<=k) && (k<=num_upper()) );
   // k=0 -> diagonal, k<0 lower left part, k>0 upper right part
   if(k>=0)    return m_upper[k][i];
   else        return m_lower[-k][i];
}
// second diag (used in LU decomposition), saved in m_lower
double band_matrix::saved_diag(int i) const
{
   assert( (i>=0) && (i<dim()) );
   return m_lower[0][i];
}
double & band_matrix::saved_diag(int i)
{
   assert( (i>=0) && (i<dim()) );
   return m_lower[0][i];
}

// LR-Decomposition of a band matrix
void band_matrix::lu_decompose()
{
   int  i_max,j_max;
   int  j_min;
```

```
   double x;

   // preconditioning
   // normalize column i so that a_ii=1
   for(int i=0; i<this->dim(); i++) {
      assert(this->operator()(i,i)!=0.0);
      this->saved_diag(i)=1.0/this->operator()(i,i);
      j_min=std::max(0,i-this->num_lower());
      j_max=std::min(this->dim()-1,i+this->num_upper());
      for(int j=j_min; j<=j_max; j++) {
         this->operator()(i,j) *= this->saved_diag(i);
      }
      this->operator()(i,i)=1.0;         // prevents rounding errors
   }

   // Gauss LR-Decomposition
   for(int k=0; k<this->dim(); k++) {
      i_max=std::min(this->dim()-1,k+this->num_lower());  // num_lower not a mistake!
      for(int i=k+1; i<=i_max; i++) {
         assert(this->operator()(k,k)!=0.0);
         x=-this->operator()(i,k)/this->operator()(k,k);
         this->operator()(i,k)=-x;                       // assembly part of L
         j_max=std::min(this->dim()-1,k+this->num_upper());
         for(int j=k+1; j<=j_max; j++) {
            // assembly part of R
            this->operator()(i,j)=this->operator()(i,j)+x*this->operator()(k,j);
         }
      }
   }
}
// solves Ly=b
std::vector<double> band_matrix::l_solve(const std::vector<double>& b) const
{
   assert( this->dim()==(int)b.size() );
   std::vector<double> x(this->dim());
   int j_start;
   double sum;
   for(int i=0; i<this->dim(); i++) {
      sum=0;
      j_start=std::max(0,i-this->num_lower());
      for(int j=j_start; j<i; j++) sum += this->operator()(i,j)*x[j];
      x[i]=(b[i]*this->saved_diag(i)) - sum;
   }
   return x;
}
// solves Rx=y
std::vector<double> band_matrix::r_solve(const std::vector<double>& b) const
{
   assert( this->dim()==(int)b.size() );
   std::vector<double> x(this->dim());
   int j_stop;
   double sum;
   for(int i=this->dim()-1; i>=0; i--) {
      sum=0;
      j_stop=std::min(this->dim()-1,i+this->num_upper());
      for(int j=i+1; j<=j_stop; j++) sum += this->operator()(i,j)*x[j];
```

```
      x[i]=( b[i] - sum ) / this->operator()(i,i);
   }
   return x;
}

std::vector<double> band_matrix::lu_solve(const std::vector<double>& b,
      bool is_lu_decomposed)
{
   assert( this->dim()==(int)b.size() );
   std::vector<double>  x,y;
   if(is_lu_decomposed==false) {
      this->lu_decompose();
   }
   y=this->l_solve(b);
   x=this->r_solve(y);
   return x;
}

} // namespace internal


} // namespace tk


} // namespace

#pragma GCC diagnostic pop

#endif /* TK_SPLINE_H */
```

Лабораторная работа №2

1. Задание

Вариант №16

Написать процедуру формирования матрицы B по формулам:

$$B_{ik}=\begin{cases}\dfrac{0.01}{(N-i+k)(i+1)} & для \quad i=k \\ 0 & для \quad i<k \\ i(N-K) & для \quad i>k\end{cases}$$

Вычислить матрицу B^{-1}, используя процедуры DECOMP и SOLVE, и найти норму матрицы $R = BB^{-1} - E$ для N = 3, 6, 9 . Объяснить результаты.

$$\|R\|=\sqrt{\sum_{i}^{N}\sum_{k}^{N}R_{ik}^{2}}$$

2. Решение

2.1 Ход решения

1. Сформировать матрицы по заданному правилу

2. Вычислить обратные матрицы с помощью процедур DECOMP и SOLVE

3. Сформировать R матрицы, которые равны разнице произведений оригинальных и обратных матриц и единичных матриц

4. Вычислить нормы R матриц

2.2 Основная программа

main.cpp

```
#include <iostream>
#include <iomanip>
#include <vector>

#include "matrixFunc.hpp"

int main()
{
  size_t n[] = {3, 6, 9};

  for (size_t q = 0; q < 3; ++q)
  {
    std::vector< double > b(n[q] * n[q]);
    fillMatrix(b);
    std::cout << "\033[1;32mStandart matrix " << n[q] << 'x' << n[q] << ":\033[1;0m\n";
    printMatrix(b, std::cout);

    std::vector< double > antiB(antiMatrix(b));
    std::cout << "\033[1;33mAnti matrix:\033[1;0m\n";
    printMatrix(antiB, std::cout);

    std::vector< double > r(multiMatrix(b, antiB));
    for (int i = 0; i < n[q]; ++i)
    {
      r[i * n[q] + i] -= 1;
    }
    std::cout << "\033[1;34mR matrix:\033[1;0m\n";
    printMatrix(r, std::cout);
    std::cout << "\033[1;35mNorm of R: " << calcNorm(r) << "\033[1;0m\n\n";
  }
  return 0;
}
```

2.3 Вычисление обратной матрицы

matrixFunc.cpp

```
vec antiMatrix(const vec & m)
{
  int dim = std::sqrt(m.size());
  vec copy(m);
  double cond;
  std::vector< int > ipvt(dim);
  vec work(dim);

  decomp_(&dim, &dim, copy.data(), &cond, ipvt.data(), work.data());

  vec result(m.size());
  for (int i = 0; i < dim; ++i)
  {
    vec b(dim, 0.0);
    b[i] = 1.0;
    solve_(&dim, &dim, copy.data(), b.data(), ipvt.data());
    for (int j = 0; j < dim; ++j)
    {
      result[i * dim + j] = b[j];
    }
  }
  return result;
}
```

2.4 Перемножение матриц

matrixFunc.cpp

```
vec multiMatrix(const vec & first, const vec & second) noexcept
{
  vec result(first.size());
  int n = std::sqrt(first.size());
  for (size_t i = 0; i < n; ++i)
  {
    for (size_t j = 0; j < n; ++j)
    {
      result[i * n + j] = 0;
      for (size_t q = 0; q < n; ++q)
      {
        result[i * n + j] += first[i * n + q] * second[q * n + j];
      }
    }
  }
  return result;
}
```

2.5 Вычисление нормы матрицы

matrixFunc.cpp

```
double calcNorm(const vec & m)
{
  int n = std::sqrt(m.size());
  double norm = 0;
  for (size_t i = 0; i < n; ++i)
  {
    for (size_t j = 0; j < n; ++j)
```

```
  {
    norm += std::pow(m[i * n + j], 2);
  }
 }
 norm = std::sqrt(norm);
 return norm;
}
```

3. Результаты

3.1 Оригинальные матрицы

N = 3

3,33E-03	0	0
3	1,67E-03	0
6	4	1,11E-03

N = 6

1,67E-03	0	0	0	0	0
6	8,33E-04	0	0	0	0
12	10	5,56E-04	0	0	0
18	15	12	4,17E-04	0	0
24	20	16	12	3,33E-04	0
30	25	20	15	10	2,78E-04

N = 9

1,11E-03	0	0	0	0	0	0	0	0
9	5,56E-04	0	0	0	0	0	0	0
18	16	3,70E-04	0	0	0	0	0	0
27	24	21	2,78E-04	0	0	0	0	0
36	32	28	24	2,22E-04	0	0	0	0
45	40	35	30	25	1,85E-04	0	0	0
54	48	42	36	30	24	1,59E-04	0	0
63	56	49	42	35	28	21	1,39E-04	0
72	64	56	48	40	32	24	16	1,23E-04

3.2 Обратные матрицы

N = 3

300	0	0
-5,40E+05	600	0
1,94E+09	-2,16E+06	900

N = 6

600	0	0	0	0	0
-4,32E+06	1200	0	0	0	0
7,77E+10	-2,16E07	1800	0	0	0
-2,24E+15	6,22E+11	-5,18E+07	2400	0	0
8,06E+19	-2,24E+16	1,87E+12	-8,64E+07	3000	0
-2,90E+24	8,06E+20	-6,72E+16	3,11E+12	-1,08E+08	3600

N = 9

900	0	0	0	0	0	0	0	0
-1,46E+07	1800	0	0	0	0	0	0	0
6,30E+11	-7,78E+07	2700	0	0	0	0	0	0
-4,76E+16	5,88E+12	-2,04E+08	3600	0	0	0	0	0
5,14E+21	-6,35E+17	2,20E+13	-3,89E+08	4500	0	0	0	0
-6,94E+26	8,57E+22	-2,98E+18	5,25E+13	-6,08E+08	5400	0	0	0
1,05E+32	-1,230E+28	4,50E+23	-7,97E+18	9,19E+13	-8,16E+08	6300	0	0
-1,59E+37	1,96E+33	-6,80E+28	1,20E+24	-1,39E+19	1,23E+14	-9,53E+08	7200	0
2,06E+42	-2,54E+38	8,82E+33	-1,56E+29	1,80E+24	-1,60E+19	1,23E+14	-9,33E+08	8100

3.3 R матрицы

N = 3

0	0	0
-1,14E-13	0	0
-4,66E-10	0	0

N = 6

0	0	0	0	0	0
-4,55E-13	0	0	0	0	0
0	2,98E-08	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1,31E+05	0	0	0	0	0

N = 9

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
-128	0	0	0	0	0	0	0	0
-5,03E+07	2048	0	0	0	0	0	0	0
-4,40E+12	0	0	0	0	0	0	0	0
0	-3,52E+13	0	0,25	0	0	0	0	0
7,56E+22	-9,22E+18	0	-2,15E+09	3,27E+04	-0,25	0	0	0

3.4 Нормы

N = 3; ||R|| = 4,66E-10

N = 6; ||R|| = 1,31E+05

N = 9; ||R|| = 7,56E+22

4. Выводы

В ходе выполнения лабораторной работы был изучен метод нахождения обратной матрицы с помощью процедур DECOMP и SOLVE. Было обнаружено, что с увеличением стороны квадратной матрицы в 2 раза, среднеквадратичная норма R вырастает на десятки порядков, что может свидетельствует о том, что с ростом размера матрицы значительно возрастают погрешности при вычисление обратной матрицы, а так же погрешность при перемножение матриц. В ходе выполнения лабораторный работы для хранения элементов матрицы использовался тип данных double, что может являться причиной резкого роста погрешности, так как double поддерживает лишь 16 чисел после мантиссы. Можно с уверенностью заключить, что с ростом размера матрицы значительно вырастают и погрешности при вычислениях

5. Дополнения

5.1 Полный код всех программ

Описание файлов:

- main.cpp является связующим звеном между функциями реализованными в matrixFunc.*
- matrixFunc.* содержит функции заполнения матрицы, вывода матрицы, нахождения нормы матрицы, перемножения матриц и нахождения обратной матрицы с использованием процедур DECOMP и SOLVE
- *.f файлы содержащие оригинальные Фортран процедуры DECOMP и SOLVE

main.cpp

```
#include <iostream>
#include <iomanip>
#include <vector>

#include "matrixFunc.hpp"

int main()
{
  size_t n[] = {3, 6, 9};

  for (size_t q = 0; q < 3; ++q)
  {
    std::vector< double > b(n[q] * n[q]);
    fillMatrix(b);
    std::cout << "\033[1;32mStandart matrix " << n[q] << 'x' << n[q] << ":\033[1;0m\n";
    printMatrix(b, std::cout);

    std::vector< double > antiB(antiMatrix(b));
    std::cout << "\033[1;33mAnti matrix:\033[1;0m\n";
    printMatrix(antiB, std::cout);

    std::vector< double > r(multiMatrix(b, antiB));
    for (int i = 0; i < n[q]; ++i)
    {
      r[i * n[q] + i] -= 1;
    }
    std::cout << "\033[1;34mR matrix:\033[1;0m\n";
    printMatrix(r, std::cout);
    std::cout << "\033[1;35mNorm of R: " << calcNorm(r) << "\033[1;0m\n\n";
  }
  return 0;
}
```

matrixFunc.hpp

```
#ifndef MATRIXFUNC_HPP
#define MATRIXFUNC_HPP

#include <iostream>
#include <vector>

using vec = std::vector< double >;

void fillMatrix(vec & m) noexcept;
void printMatrix(const vec & m, std::ostream & out);
vec antiMatrix(const vec & m);
vec multiMatrix(const vec & first, const vec & second) noexcept;
double calcNorm(const vec & m);

#endif
```

matrixFunc.cpp

```
#include "matrixFunc.hpp"

#include <cmath>
#include <iomanip>

extern "C" {
  int decomp_(int *ndim, int *n, double *a, double *cond, int *ipvt, double *work);
  int solve_(int *ndim, int *n, double *a, double *b, int *ipvt);
}

void fillMatrix(vec & m) noexcept
{
  int n = std::sqrt(m.size());
  for (int i = 0; i < n; ++i)
  {
    for (int j = 0; j < n; ++j)
    {
      if (i == j)
      {
        m[i * n + j] = 0.01 / ((n - i + j) * (i + 1));
      }
      else if (i < j)
      {
        m[i * n + j] = 0;
      }
      else
      {
        m[i * n + j] = i * (n - j);
      }
    }
  }
}
void printMatrix(const vec & m, std::ostream & out)
{
  int n = std::sqrt(m.size());
  for (size_t i = 0; i < n; ++i)
  {
    for (size_t j = 0; j < n; ++j)
    {
```

```
      out << m[i * n + j] << '\t';
    }
    out << '\n';
  }
}
vec antiMatrix(const vec & m)
{
  int dim = std::sqrt(m.size());
  vec copy(m);
  double cond;
  std::vector< int > ipvt(dim);
  vec work(dim);

  decomp_(&dim, &dim, copy.data(), &cond, ipvt.data(), work.data());

  vec result(m.size());
  for (int i = 0; i < dim; ++i)
  {
    vec b(dim, 0.0);
    b[i] = 1.0;
    solve_(&dim, &dim, copy.data(), b.data(), ipvt.data());
    for (int j = 0; j < dim; ++j)
    {
      result[i * dim + j] = b[j];
    }
  }
  return result;
}
vec multiMatrix(const vec & first, const vec & second) noexcept
{
  vec result(first.size());
  int n = std::sqrt(first.size());
  for (size_t i = 0; i < n; ++i)
  {
    for (size_t j = 0; j < n; ++j)
    {
      result[i * n + j] = 0;
      for (size_t q = 0; q < n; ++q)
      {
        result[i * n + j] += first[i * n + q] * second[q * n + j];
      }
    }
  }
  return result;
}
double calcNorm(const vec & m)
{
  int n = std::sqrt(m.size());
  double norm = 0;
  for (size_t i = 0; i < n; ++i)
  {
    for (size_t j = 0; j < n; ++j)
    {
      norm += std::pow(m[i * n + j], 2);
    }
  }
```

```
 norm = std::sqrt(norm);
 return norm;
}
```

decomp.f

```
c  Code from http://www.netlib.org/fmm/
      subroutine decomp(ndim,n,a,cond,ipvt,work)
c
      integer ndim,n
      double precision a(ndim,n),cond,work(n)
      integer ipvt(n)
c
c     decomposes a double precision matrix by gaussian elimination
c     and estimates the condition of the matrix.
c
c     use solve to compute solutions to linear systems.
c
c     input..
c
c        ndim = declared row dimension of the array containing  a.
c
c        n = order of the matrix.
c
c        a = matrix to be triangularized.
c
c     output..
c
c        a  contains an upper triangular matrix  u  and a permuted
c          version of a lower triangular matrix  i-l  so that
c          (permutation matrix)*a = l*u .
c
c        cond = an estimate of the condition of  a .
c          for the linear system  a*x = b, changes in  a  and  b
c          may cause changes  cond  times as large in  x .
c          if  cond+1.0 .eq. cond , a is singular to working
c          precision.  cond  is set to  1.0d+32  if exact
c          singularity is detected.
c
c        ipvt = the pivot vector.
c          ipvt(k) = the index of the k-th pivot row
c          ipvt(n) = (-1)**(number of interchanges)
c
c     work space..  the vector  work  must be declared and included
c                   in the call.  its input contents are ignored.
c                   its output contents are usually unimportant.
c
c     the determinant of a can be obtained on output by
c        det(a) = ipvt(n) * a(1,1) * a(2,2) * ... * a(n,n).
c
      double precision ek, t, anorm, ynorm, znorm
      integer nm1, i, j, k, kp1, kb, km1, m
      double precision dabs, dsign
c
      ipvt(n) = 1
      if (n .eq. 1) go to 80
      nm1 = n - 1
```

```
c
c     compute 1-norm of a
c
      anorm = 0.0d0
      do 10 j = 1, n
         t = 0.0d0
         do 5 i = 1, n
            t = t + dabs(a(i,j))
    5    continue
         if (t .gt. anorm) anorm = t
   10 continue
c
c     gaussian elimination with partial pivoting
c
      do 35 k = 1,nm1
         kp1= k+1
c
c        find pivot
c
         m = k
         do 15 i = kp1,n
            if (dabs(a(i,k)) .gt. dabs(a(m,k))) m = i
   15    continue
         ipvt(k) = m
         if (m .ne. k) ipvt(n) = -ipvt(n)
         t = a(m,k)
         a(m,k) = a(k,k)
         a(k,k) = t
c
c        skip step if pivot is zero
c
         if (t .eq. 0.0d0) go to 35
c
c        compute multipliers
c
         do 20 i = kp1,n
            a(i,k) = -a(i,k)/t
   20    continue
c
c        interchange and eliminate by columns
c
         do 30 j = kp1,n
            t = a(m,j)
            a(m,j) = a(k,j)
            a(k,j) = t
            if (t .eq. 0.0d0) go to 30
            do 25 i = kp1,n
               a(i,j) = a(i,j) + a(i,k)*t
   25       continue
   30    continue
   35 continue
c
c     cond = (1-norm of a)*(an estimate of 1-norm of a-inverse)
c     estimate obtained by one step of inverse iteration for the
c     small singular vector.  this involves solving two systems
c     of equations, (a-transpose)*y = e  and  a*z = y  where  e
```

```
c     is a vector of +1 or -1 chosen to cause growth in y.
c     estimate = (1-norm of z)/(1-norm of y)
c
c     solve (a-transpose)*y = e
c
      do 50 k = 1, n
        t = 0.0d0
        if (k .eq. 1) go to 45
        km1 = k-1
        do 40 i = 1, km1
          t = t + a(i,k)*work(i)
   40   continue
   45   ek = 1.0d0
        if (t .lt. 0.0d0) ek = -1.0d0
        if (a(k,k) .eq. 0.0d0) go to 90
        work(k) = -(ek + t)/a(k,k)
   50 continue
      do 60 kb = 1, nm1
        k = n - kb
        t = 0.0d0
        kp1 = k+1
        do 55 i = kp1, n
          t = t + a(i,k)*work(i)
   55   continue
        work(k) = t + work(k)
        m = ipvt(k)
        if (m .eq. k) go to 60
        t = work(m)
        work(m) = work(k)
        work(k) = t
   60 continue
c
      ynorm = 0.0d0
      do 65 i = 1, n
        ynorm = ynorm + dabs(work(i))
   65 continue
c
c     solve a*z = y
c
      call solve(ndim, n, a, work, ipvt)
c
      znorm = 0.0d0
      do 70 i = 1, n
        znorm = znorm + dabs(work(i))
   70 continue
c
c     estimate condition
c
      cond = anorm*znorm/ynorm
      if (cond .lt. 1.0d0) cond = 1.0d0
      return
c
c     1-by-1
c
   80 cond = 1.0d0
      if (a(1,1) .ne. 0.0d0) return
```

```
c
c     exact singularity
c
   90 cond = 1.0d+32
      return
      end
```

solve.f

```
c  Code from http://www.netlib.org/fmm/
      subroutine solve(ndim, n, a, b, ipvt)
c
      integer ndim, n, ipvt(n)
      double precision a(ndim,n),b(n)
c
c   solution of linear system, a*x = b .
c   do not use if decomp has detected singularity.
c
c   input..
c
c     ndim = declared row dimension of array containing a .
c
c     n = order of matrix.
c
c     a = triangularized matrix obtained from decomp .
c
c     b = right hand side vector.
c
c     ipvt = pivot vector obtained from decomp .
c
c   output..
c
c     b = solution vector, x .
c
      integer kb, km1, nm1, kp1, i, k, m
      double precision t
c
c     forward elimination
c
      if (n .eq. 1) go to 50
      nm1 = n-1
      do 20 k = 1, nm1
        kp1 = k+1
        m = ipvt(k)
        t = b(m)
        b(m) = b(k)
        b(k) = t
        do 10 i = kp1, n
           b(i) = b(i) + a(i,k)*t
   10   continue
   20 continue
c
c     back substitution
c
      do 40 kb = 1,nm1
        km1 = n-kb
        k = km1+1
```

```
      b(k) = b(k)/a(k,k)
      t = -b(k)
      do 30 i = 1, km1
         b(i) = b(i) + a(i,k)*t
30    continue
40 continue
50 b(1) = b(1)/a(1,1)
   return
   end
```