1 Jan 11, 2023

1.1 Trapezoid Rule

Definition 1: Trapezoid Rule

We do the following steps.

1. Dividing the interval $[a, b]$ into subintervals $[x_i, x_{i+1}]$, $i = 0, \cdots, n$, according to the partition $P = \{a = x_0 \leq x_1 \cdots \leq x_n = b\}$.
2. Estimated the total area $\int_a^b f(x)dx$ by $\frac{1}{2}\sum_{i=0}^{n-1}(x_{i+1} - x_i)[f(x_i) + f(x_{i+1})]$.

Example 1. Let's use the Trapezoid Method to estimate $f(x) = x^2$ on [0,2]. With the following partitions: $P_1 = [0, 1, 2]$, $P_2 = [0, 0.5, 1, 1.5, 2]$. We get for P_1:

$$\frac{8}{3} = \int_a^b f(x)dx \approx 0.5[(1-0)(f(1)+f(0)) + (2-1)(f(2)+f(1))] = 3$$

We get for P_2:

$$\frac{8}{3} = \int_a^b f(x)dx \approx 0.5[(0.5)(0.25) + (0.5)(1.25)) + (0.5)(3.25) + (0.5)(6.25)] = 2.75$$

The error when using P_1 is $\frac{1}{3}$. The error when using P_2 is about 0.083. For a simple function that we can graph, we can actually pick a non-uniform partition to improve the error. For more examples, please see the lecture notes and the reference book. □

1.2 How to control the error by making the partition fine enough

Indeed we can pick the partition as "fine" as possible to obtain the "most accurate" answer. However, to save time and resources, what is the minimum number of points required to achieve a given level of accuracy? This is answered by the following theorem.

Theorem 1: Theorem On Precision Of Trapezoid Rule

If f'' exists and is continuous on the interval $[a, b]$ and if the composite trapezoid rule T with uniform spacing h is used to estimate the integral $I = \int_a^b f(x)dx$, then for some ζ in (a, b),

$$I - T = -\frac{1}{12}(b-a)h^2 f''(\zeta)$$

Example 2. Compute $\int_0^1 e^{-x^2}dx$ with an error of at most $\frac{1}{2} \times 10^{-4}$. How many points should be used?

First, we check that f'' indeed exists and is continuous on $[0, 1]$. This is true since

$$f''(x) = (4x^2 - 2)e^{-x^2}.$$

Also, we estimate f'' by

$$|f''(x)| = |(4x^2 - 2)e^{-x^2}| \leq |(4x^2 - 2)| \leq 2.$$

(How to estimate f''?) Since f'' exists and is continuous. We can use the previous theorem. The error is given by

$$|I - T| = \left| -\frac{1-0}{12} h^2 f''(\zeta) \right| \overset{want}{\leq} \frac{1}{2} \times 10^{-4}$$

That is,

$$\left| \frac{1}{6} h^2 \right| \overset{want}{\leq} \frac{1}{2} \times 10^{-4}$$

So we can pick $h \leq 0.01732$. Since we are using uniform spacing, and $h = \frac{b-a}{\text{number of points}-1}$. So we need the number of points > 58. Thus we pick 59 or more points. □

2 Jan 18, 2023

2.1 More example on Error Analysis

- For a partition $P = \{a = x_0 \leq x_1 \cdots \leq x_n = b\}$, there are n intervals.
- There are totally $n + 1$ of points which includes x_0.
- The step size for uniform partition is $\frac{b-a}{n}$.

To make sure that $h = \frac{b-a}{n}$. Let me give a short proof here to convince you all.

Lemma 1: G

ven a uniform partition $P = \{a = x_0 \leq x_1 \cdots \leq x_n = b\}$, the step size h is given by,

$$h = \frac{b-a}{\text{number of intervals}} = \frac{b-a}{n} \tag{1}$$

Proof. Let's prove this by induction. (1) is obviously true when $n = 1$. Suppose (1) is true for $n = k$. When $n = k + 1$, we have added one more point. Then we have one more interval, thus the step size is $h = \frac{b-a}{k+1}$. □

Example 3. How many points are needed to estimate

$$\int_1^2 x \ln x \, dx \tag{2}$$

with an error of at most 10^{-3}?

We compute f'' with $f(x) = x \ln x$. We get

$$f'' = \frac{1}{x}. \tag{3}$$

So f'' exists and is continuous on $[1, 2]$. Next we get $|f''| \leq 1$. By the Theorem On Precision Of Trapezoid Rule, we have, for some $\zeta \in (1, 2)$,

$$|I - T| = |\frac{-1}{12}(b-a)h^2 f''(\zeta)| \leq \frac{1}{12}h^2 \overset{want}{\leq} 10^{-3} \tag{4}$$

Thus we get $h \leq 0.109545$.

The relationship between the number of points and the step size is,

$$h = \frac{b-a}{n}. \tag{5}$$

Thus $n \geq 9.1286$. So we need $n = 10$. But what is n? Looking at how we constructed the partition,

$$P = \{2 = x_0 \leq x_1 \cdots \leq x_{10} = 2\}. \tag{6}$$

So, for $n = 10$ we have in fact 11 points! Thus we need at least 10 points. I hope this explains better why we have calculated $n \geq 58$ but we need 59 points in the last discussion. The extra point is the x_0. □

2.2 Recursive Trapezoid Formula

Suppose you have two customers Anya and Loid. Anya might want to have an error of 10^{-3}. Loid might want to have an error of at most 10^{-10}. You might realize that if you use the above partition, you would need to recompute the values of the function that we are integrating specifically for Anya and Loid. It would be time-consuming to do that. But you are so smart and you notice that Anya and Loid did not state that they want to use a certain partition. Thus you can pick some partitions smartly that can minimize your work.

This is done by picking a uniform partition with $h = 2^N$. This can be seen by just drawing some pictures yourself, or you can look at Prof. Chen's Lecture notes. The set of points from the partition with $h = 2^N$ is a subset of the set of points from the partition with $h = 2^{N+1}$ for all N.

Let us denote the uniform partition with $h = 2^N$, then

$$R(n,0) = h\sum_{i=1}^{2^N-1} f(a+ih) + \frac{h}{2}[f(a)+f(b)]. \tag{7}$$

Actually, we have

$$R(n,0) - \frac{1}{2}R(n-1,0) = h\sum_{k=1}^{2^N-1} f[a+(2k-1)h]. \tag{8}$$

You can read more from the reference book to see how we get (8).

3 Jan 25, 2023

3.1 Deriving the Recursive Trapezoid Formula

If we have P with n intervals and a function f. The formula of the composite trapezoid rule gives

$$T(f;P) = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{n-1} f(x_i). \tag{9}$$

Since $x_i = x_0 + ih = a + ih$, where h is the step size. So we have

$$T(f;P) = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{n-1} f(a+ih). \tag{10}$$

In view of the uniform partition P of with 2^n equal intervals, $h = \frac{b-a}{2^n}$. The formula of the composite trapezoid rule becomes

$$T(f;P) = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{2^n-1} f(a+ih). \tag{11}$$

We have replaced n by 2^n.

Let us follow the notation used in the reference book and let

$$R(n,0) := T(f;P) = \frac{h}{2}[f(a)+f(b)] + h\sum_{i=1}^{2^n-1} f(a+ih). \tag{12}$$

The notation $A := B$ means A is defined to be B.

Next our goal is to do some computation and hopefully arrived at an expression like $R(n,0) =$ somethings that depends on $R(n-1,0)$. First, let's observe the obvious identity,

$$\begin{aligned} R(n,0) &= R(n,0) + \frac{1}{2}R(n-1,0) - \frac{1}{2}R(n-1,0) \\ &= \frac{1}{2}R(n-1,0) + \left[R(n,0) - \frac{1}{2}R(n-1,0)\right]. \end{aligned} \tag{13}$$

Let $C := \frac{h}{2}[f(a)+f(b)]$. From (12), we have

$$\begin{aligned} R(n,0) &= C + h\sum_{i=1}^{2^n-1} f(a+ih) \\ R(n-1,0) &= 2C + 2h\sum_{j=1}^{2^{n-1}-1} f(a+jh). \end{aligned} \tag{14}$$

Note that in (14), there are 2^n intervals in $R(n,0)$ and there are 2^{n-1} intervals in $R(n-1,0)$. Also, if the step size for $R(n,0)$ is h. Then the step size for $R(n-1,0)$ is $2h$.

From (14) we have

$$\begin{aligned} R(n,0) - \frac{1}{2}R(n-1,0) &= C + h\sum_{i=1}^{2^n-1} f(a+ih) - \left(C + h\sum_{j=1}^{2^{n-1}-1} f(a+jh)\right) \\ &= h\sum_{i=1}^{2^n-1} f(a+ih) - h\sum_{j=1}^{2^{n-1}-1} f(a+jh). \end{aligned} \tag{15}$$

We need to be careful when computing this sum. Notice that the indices are different in the first sum and the second sum. (The first sum depends on i, and the second sum depends on j.) Let us observe a fact about partition with 2^n steps here,

The points in the partition with 2^{N-1} steps are the even points in the partition with 2^N steps.

Proof. Consider any n. Let's us denote the points in the partition P_{N-1} with 2^{N-1} intervals by x_j, $1 \leq j \leq 2^{N-1}$. Denote the points in in the partition P_N with 2^N intervals by y_i, $1 \leq i \leq 2^N$. Picking any point x_p, then by the way that we construct the partition, it is obvious that x_p remains in P_N. Then for some integer k, there will be two points, y_k and y_{k+2}, both are not in P_{N-1}, such that

$$y_{k-1} < x_p = y_k < y_{k+1}. \tag{16}$$

Now I claim that k is even. If k is not even, then $k-1$ is even. By this argument, we see that all points in P_{N-1} will be the odd points in P_N. However x_0 is the even point that is both in P_{N-1} and P_N. Thus k must be even. □

With this observation, we can compute,

$$\begin{aligned} R(n,0) - \frac{1}{2}R(n-1,0) &= C + h\sum_{i=1}^{2^n-1} f(a+ih) - \left(C + h\sum_{j=1}^{2^{n-1}-1} f(a+jh)\right) \\ &= h\sum_{i=1}^{2^n-1} f(a+ih) - h\sum_{j=1}^{2^{n-1}-1} f(a+jh) \\ &= h\sum_{\substack{i=1 \\ \text{i is odd}}}^{2^n-1} f(a+ih). \end{aligned} \tag{17}$$

Becuase all the even points are canceled out by the second sum.

And this gives rise to the formula in the lectures and the book.

3.2 Romberg Algorithm

Romberg Algorithm is a combination of recursive trapezoid formula and extrapolation. We have seen the recursive trapezoid formula before. The extrapolation formula is

$$R(n,m) = R(n,m-1) + \frac{1}{4^m - 1}\left[R(n,m-1) - R(n-1,m-1)\right]. \tag{18}$$

Perhaps a picture shows this formula better,

$$\begin{matrix} R(n-1,m-1) & \\ R(n,m-1) & R(n,m) \end{matrix} \tag{19}$$

$R(n,m)$ is calculated with the values on the left shown above.

3.3 Simpson's Rules

Simpson's Rules is built on top of the observation that, we are using a straight line to estimate the curve between $f(a)$ and $f(b)$ define by the function f. We can add more points/ use a curve to do the estimation to improve our result.

Let's look at the Trap rule. For one interval, it has the form

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\left[f(a)+f(b)\right]. \tag{20}$$

This suggest that we can look for new integration rule of the following sense,

$$\int_a^b f(x)\,dx \approx Af(a)+Bf(b), \tag{21}$$

for some constant A and B. By this argument, for two intervals we have,

$$\int_a^b f(x)\,dx \approx Af(a)+Bf(\frac{a+b}{2})+Cf(b) \tag{22}$$

What is the general form for n-intervals like this?

How do we decide what are A, B, C? I am aware of two ways. First we can try to formulate finding A, B, C as an optimization problem that minimize some kind of error. Second way is to consider some kind of "toy problems" and make sure A, B, C gives the correct result for the "toy problems'.

4 Feb 1, 2023

4.1 Simpson's Rules

Definition 2: T

e basic Simpson's Rule is defined as follows

$$\int_a^b f(x)\,dx \approx \frac{1}{6}(b-a)\left[f(a)+4f(\frac{a+b}{2})+f(b)\right]. \tag{23}$$

Example 4. Find approximate values for the integral

$$\int_0^1 e^{-x^2}\,dx \tag{24}$$

using the basic Simpson's Rule. Carry five significant digits.

Solution: We can just apply the formula directly. From (23), we have

$$\int_0^1 e^{-x^2}\,dx \approx \frac{1}{6}\left[e^0+4f(e^{-0.25})+e^{-1}\right] \approx 0.7472. \tag{25}$$

The actual answer for this integral is approximately 0.74682. □

Example 5. Show that the basic Simpson's Rules integrate polynomial of degree at most two correctly on $[-1,1]$.

Solution: We first consider integrating on the interval $[-1,1]$ for $f(x) = 1/x/x^2$. The statement for other degree two polynomials will then follow. (WHY?) For $f(x) = 1$.

$$\int_{-1}^1 1\,dx = 2 = \frac{1}{3}\left[1+4+1\right]. \tag{26}$$

For $f(x) = x$.

$$\int_{-1}^1 x\,dx = 0 = \frac{1}{3}\left[-1+0+1\right]. \tag{27}$$

For $f(x) = x^2$.

$$\int_{-1}^1 x^2\,dx = \frac{2}{3} = \frac{1}{3}\left[1+0+1\right]. \tag{28}$$

□

Theorem 2: T

e error term for using the Simpson's Rule to integrate $f(x)$ on $[a,b]$ with partition $[a, a+h, a+2h=b]$is

$$\frac{-1}{90}\left(\frac{b-a}{5}\right)^5 f^{(4)}(\eta) \tag{29}$$

for some $\eta \in (a, b)$ if $f \in C^4$.

Example 6. If possible, find a formula

$$\int_{-1}^{1} f(x)\,dx \approx a_1 f(-1) + a_2 f(0) + a_3 f(1) \tag{30}$$

that gives the correct value for $f(x) = x, x^2, x^3$. Does it correctly integrate the function $f(x) = 1, x^4, x^5$?

Solution: Using (30) we obtain the following system of equations,

$$\begin{cases} 0 = \int_{-1}^{1} x\,dx = -a_1 + a_3 \\ \frac{2}{3} = \int_{-1}^{1} x^2\,dx = a_1 + a_3 \\ 0 = \int_{-1}^{1} x^3\,dx = -a_1 + a_3 \end{cases} \tag{31}$$

Solving this system we get, $a_1 = 1/3$ and $a_3 = 1/3$.

Next we check for $f(x) = 1, x^4, x^5$.

$$2 = \int_{-1}^{1} 1\,dx \stackrel{?}{=} 1/3 + a_2 + 1/3. \tag{32}$$

Now a_2 is arbitrary, we can pick it to be $4/3$.

For $f(x) = x^4$, we have

$$2/5 = \int_{-1}^{1} x^4\,dx \neq 1/3 + 1/3. \tag{33}$$

For $f(x) = x^5$, we have

$$0 = \int_{-1}^{1} x^5\,dx = -1/3 + 1/3. \tag{34}$$

□

4.2 Quadrature Formulas

In general, the Quadrature Formulas are of the form

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i). \tag{35}$$

Example 7. Determine the quadrature formula of the form (54) when the interval is $[-2, 2]$ and the nodes are -1, 0 and 1. With

$$A_i = \int_a^b L_i(x)\,dx, \quad L_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}. \tag{36}$$

Solution: For this case $a = -2, b = -2$ and $n = 2$. The nodes are $x_0 = -1$, $x_1 = 0$, $x_2 = 1$. We list L_i here,

$$L_0 = \prod_{j=1, j\neq 0}^{2} \frac{x - x_j}{-1 - x_j} = \frac{1}{2}x(x-1). \tag{37}$$

$$L_1 = \prod_{j=1, j\neq 1}^{2} \frac{x - x_j}{0 - x_j} = (x+1)(x-1). \tag{38}$$

$$L_2 = \prod_{j=1, j\neq 2}^{2} \frac{x - x_j}{1 - x_j} = x\frac{(x+1)}{2}. \tag{39}$$

Then $A_0 = 8/3$, $A_1 = -4/3$, $A_2 = 8/3$. So the required quadrature formula is,

$$\int_{-2}^{2} f(x)\,dx \approx \frac{8}{3}f(-1) - \frac{4}{3}f(0) + \frac{8}{3}f(1) \tag{40}$$

□

5 Feb 8, 2023

5.1 Gaussian Quadrature Formulas

In general, the Quadrature Formulas are of the form

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i). \tag{41}$$

Example 8. Determine the quadrature formula of the form (54) when the interval is $[-2, 2]$ and the nodes are -1, 0 and 1. With

$$A_i = \int_a^b L_i(x)\,dx, \quad L_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}. \tag{42}$$

Solution: For this case $a = -2, b = -2$ and $n = 2$. The nodes are $x_0 = -1$, $x_1 = 0$, $x_2 = 1$. We list L_i here,

$$L_0 = \prod_{j=1, j\neq 0}^{2} \frac{x - x_j}{-1 - x_j} = \frac{1}{2}x(x-1). \tag{43}$$

$$L_1 = \prod_{j=1, j\neq 1}^{2} \frac{x - x_j}{0 - x_j} = (x+1)(x-1). \tag{44}$$

$$L_2 = \prod_{j=1, j\neq 2}^{2} \frac{x - x_j}{1 - x_j} = x\frac{(x+1)}{2}. \tag{45}$$

Then $A_0 = 8/3$, $A_1 = -4/3$, $A_2 = 8/3$. So the required quadrature formula is,

$$\int_{-2}^{2} f(x)\,dx \approx \frac{8}{3}f(-1) - \frac{4}{3}f(0) + \frac{8}{3}f(1) \tag{46}$$

□

In (54) the points $x_0, x_1, x_2, \cdots, x_n$ are called the **nodes**. And the $A_0, A_1, A_2, \cdots, A_n$ are called the **weights**.

Example 9. Determine the Gaussian quadrature formula with three Gaussian nodes and three weights for the integral $\int_{-1}^{1} f(x)\,dx$. □

5.2 Spline Functions

A **spline function** is a function that consists of polynomial pieces joined together with certain smoothness conditions.

5.2.1 First Degree Spline

Definition 3

A function S is called a spline of degree 1 if:

1. The domain of S is an interval $[a, b]$.
2. S is continuous on $[a, b]$.
3. There is a partitioning of the interval $a = t_0 < t_1 < \cdots < t_n = b$ such that S is a linear polynomial on each subinterval $[t_i, t_{i+1}]$.

Example 10. Determine if

$$S(x) = \begin{cases} x, & x \in [-1, 0] \\ 1 - x, & x \in (-1, 1) \\ 2x - 2, & x \in [1, 2] \end{cases} \tag{47}$$

is a first degree spline function.

Solution: $S(x)$ is obviously defined on an interval and on each subinterval it is a linear function. However, it is discontinuous at $x = 0$. Thus this is not first degree spline function. □

5.2.2 Second Degree Splines

Definition 4

A function Q is called a spline of degree 2 (or called quadratic spline) if:

1. The domain of Q is an interval $[a, b]$.
2. Q and Q' are continuous on $[a, b]$.
3. There are points t_i (called knots) such that $a = t_0 < t_1 < \cdots < t_n = b$ such that Q is a linear polynomial on each subinterval $[t_i, t_{i+1}]$.

Note that unlike the First degree spline, which only requires continuity of itself. A second degree spline requires the continuity of itself and its first derivative.

Example 11. Check if

$$Q(x) \begin{cases} x^2, & x \in [10, 0] \\ -x^2, & x \in [0, 1] \\ 1 - 2x, & x \in [1, 202] \end{cases} \tag{48}$$

is a quadratic spline.

Solution: It is piecewise quadratic, we can also check that Q and Q' is continuous. Thus it is a quadratic spline. □

5.2.3 Spline of degree k

Definition 5

A function S is called a spline of degree k (or called quadratic spline) if:

1. The domain of Q is an interval $[a, b]$.
2. $S, S', S'', \cdots, S^{(k-1)}$ are continuous on $[a, b]$.
3. There are points t_i (called knots of S) such that $a = t_0 < t_1 < \cdots < t_n = b$ such that S is a linear polynomial on each subinterval $[t_i, t_{i+1}]$.

5.2.4 Natural cubic spline

Suppose we want are given some $(t_0, t_1, \cdots, t_n,$ and $y_0, y_1, \cdots, y_n$ where $f(t_i) = y_i$. We wish to interpolate f with $S(x)$ given by

$$S(x) = \begin{cases} S_0(x), & (t_0 \le x \le t_1) \\ \vdots & \\ S_{n-1}(x), & (t_{n-1} \le x \le t_n) \end{cases} \tag{49}$$

where each S_i is a cubic polynomial. To interpolate it we have to solve both the interpolation conditions and the continuity conditions. They are

$$S(t_i) = y_i, \quad \text{for } i = 0, 2, \cdots, n, \tag{50}$$

$$\lim_{x \to t_i^-} S^{(k)}(t_i) = \lim_{x \to t_i^+} S^{(k)}(t_i), \quad \text{for } i = 1, 2, \cdots, n-1. \tag{51}$$

Notice that we have not specified the endpoint conditions. One way to do that is called the **natural cubic spline**.

Definition 6: Natural cubic spline

The natural cubic spline is $S(x)$ constructed as above, with the additional conditions that

$$S''(t_0) = S''(t_n) = 0. \tag{52}$$

Example 12. Construct the natural cubic spline with t_i given by $(-1, 0, 1)$ and y_i given by $(1, 2, -1)$.

Solution: The interval here would be $[-1, 1]$. The function $S(x)$ is given by

$$S(x) = \begin{cases} S_0(x) = ax^3 + bx^2 + cx + d, & (-1 \le x \le 0) \\ S_{n-1} = ex^3 + fx^2 + gx + h), & (t_{n-1} \le x \le t_n) \end{cases} \tag{53}$$

after writing out continuity conditions and the interpolation conditions we see that $a = -1, b = -3, c = -1, d = 2, e = 1, f = -3, g = -1$, and $h = 2$. □

6 Feb 22, 2023

6.1 Gaussian Quadrature Formulas

In general, the Quadrature Formulas are of the form

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i). \tag{54}$$

One Quadrature Formula would be the *Gaussian Quadrature Formulas/ Gauss-Legendre quadrature formulas*. These formulas are given by the following theorem.

Theorem 3: Gaussian Quadrature Theorem

Let q be a nontrivial polynomial of degree $n+1$ such that

$$\int_a^b x^k q(x)\,dx = 0 \quad (0 \le k \le n). \tag{55}$$

Let $x_0, x_1, \cdots, x_n$ be the zeros(roots) of q. Then the formula

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i) \quad \text{where} \quad A_i = \int_a^b L_i\,dx \tag{56}$$

is exact for all polynomials of degree at most $2n+1$.

To summarize we have:

$$\text{Weights } A_i \text{ given by integration of Lagrange interpolation formula} \\ + \text{nodes } x_i \text{ given by zeros of } q = \text{Gaussian Quadrature formulas.} \tag{57}$$

Example 13. Determine the Gaussian quadrature formula with three Gaussian nodes and three weights for the integral $\int_{-1}^{1} f(x)\,dx$.

Solution. Let q be an arbitrary polynomial of degree 3, then $q(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$. From Theorem (3), we see that q must satisfies,

$$\int_{-1}^{1} q(x)\,dx = \int_{-1}^{1} xq(x)\,dx = \int_{-1}^{1} x^2 q(x)\,dx = 0. \tag{58}$$

Note that we only have three equations for four unknowns! Thus we must choose a variable on our own. In fact, we can pick

$$c_0 = c_2 = 0. \tag{59}$$

Note that this solves $\int_{-1}^{1} q(x)\,dx = \int_{-1}^{1} x^2 q(x)\,dx = 0$. Then we use this to find c_1 and c_3, we have

$$\int_{-1}^{1} xq(x)\,dx = \int_{-1}^{1} x(c_1 x + c_3 x^3)\,dx = 0. \tag{60}$$

We can pick $c_1 = -3$ and $c_3 = 5$. Hence,

$$q(x) = 5x^3 - 3x.$$

The roots are $-\sqrt{3/5}$, 0, and $\sqrt{3/5}$. To obtain A_i, we can either calculate it directly using (56). Or we can solve a system with A_i as unknowns. □

6.2 Numerical methods for Initial-Value Problem

In this section, we consider solving Ordinary Differential Equations (ODE) numerically.

A general ODE is given by,

$$\begin{cases} \dfrac{dx(t)}{dt} = x'(t) = f(t, x(t)) \\ \text{x(a) is given} \end{cases} \tag{61}$$

The problem is called an initial value because a is the initial time of x. Usually, a is denoted by t_0. Solving ODE is closely related to quadrature rules. WHY?

6.2.1 Taylor Series Method

Recall the Taylor series expansion of x about t is

$$x(t+h) = \sum_{i=0}^{\infty} \frac{h^i}{i!} x^{(i)}(t) \tag{62}$$

Example 14. Use the Taylor series method of degree 4 to solve the initial value problem

$$\begin{cases} \dfrac{dx}{dt} = x^4 \\ x(0) = 1 \end{cases} \tag{63}$$

Solution. We apply the Taylor series expansion directly, to obtain

$$x(h) = 1 + \frac{h}{1!}\, 4x^3 + \frac{h^2}{2!}\, 4 \times 3 \times x^2 + \frac{h^3}{3!}\, 4 \times 3 \times 2 \times x + \frac{h^4}{4!}\, 4 \times 3 \times 2.$$

□

Example 15. (Problem 15 in 10.1 in the book.) 15. Explain how to use the ODE method that is based on the Trapezoid Rule:

$$\begin{aligned} \hat{x}(t+h) &= x(t) + hf(t, x(t)), \\ x(t+h) &= x(t) + \frac{h}{2}[f(t, x(t)) + f(t+h, \hat{x}(t+h))]. \end{aligned} \tag{64}$$

This is called the *improved Euler's method or Heun's method.*

Solution. We can draw a picture to see how this work. But the idea is that we first do an approximation at $x(t+h)$ with the simple Euler's method. Then use this new information to improve the estimation $x(t+h)$ using the Trapezoid Rule. □

7 Mar 1, 2023

7.1 Runge-Kutta Method

The RK method imitates the Taylor series method without requiring analytic differentiation of the original differential equation. In the Taylor series method, we need to compute f', f'', etc. You may know very well the computation of derivative, but let's not assume everyone does! We hope to design a method that even a kid can use just by imputing numbers on the keypad.

The RK method of order 2 is given here.

Definition 7: Runge-Kutta method of order 2

Define

$$K_1 = hf(t,x), \quad K_2 := hf(t+h, x+K_1). \tag{65}$$

The second-order Runge-Kutta method is given by

$$\begin{aligned} x(t+h) &= x(t) + \frac{1}{2}(K_1 + K_2) \\ &= x(t) + \frac{h}{2}f(t,x) + \frac{h}{2}f(t+h, x+hf(t,x)). \end{aligned} \tag{66}$$

Example 16. Solve the differential equation

$$\begin{cases} \dfrac{dx}{dt} = -tx^2 \\ x(0) = 2 \end{cases} \tag{67}$$

with $h = 0.2$ using one step of RK method of order 2.

Solution: We can compute the following

$$\begin{aligned} K_1 &= 0.2 \times -1 \times 0 \times 2^2, \\ K_2 &= 0.2 \times -1 \times (0+0.2) \times (2+0)^2, \\ x(0.2) = x(0+0.2) &= 2 + \frac{1}{2}(K_1 + K_2). \end{aligned} \tag{68}$$

□

7.2 Matrix Factorizations

We will use bold font to denote matrices and vectors, i.e., $\mathbf{A}$ is a matrix and $\mathbf{b}$ is a vector.

We will study the lower triangular matrix $\mathbf{L}$ and the upper triangular matrix $\mathbf{U}$. An LU decomposition of a matrix $\mathbf{A}$ is the existence of $\mathbf{L}, \mathbf{U}$ such that $\mathbf{A} = \mathbf{LU}$. (Is this always possible?) We can use Gaussian Elimination to find the LU decomposition. Before doing an example, let's explain the idea behind this. As we know row operations in Gaussian Elimination are the same as multiplying the so called *elementary matrices* on the left. The

correspondence is, for a 3×3 matrix,

$$\text{Row-switching transformations}: \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{What does this matrix do?} \tag{69}$$

$$\text{Row-multiplying transformations}: \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{What does this matrix do?} \tag{70}$$

$$\text{Row-multiplying transformations}: \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \text{What does this matrix do?} \tag{71}$$

Note that **Elementary matrices for reducing a matrix into a upper triangular form are lower triangular matrices!** That means during the Gaussian Elimination we would have done something like this

$$\mathbf{M_1 M_2 \cdots M_k A = U}$$

where $\mathbf{M_i}$ are all lower triangular and $\mathbf{U}$ is upper triangular (that's what we should get after doing Gaussian Elimination). Then we can easily recover the LU factorization. (WHY?)

Example 17. Find the LU factorization for

$$\begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix} \tag{72}$$

using Gaussian Elimination.

Solution:

$$\begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{2}{3} & 1 \end{pmatrix} \begin{pmatrix} 3 & 3 \\ 0 & -1 \end{pmatrix} \tag{73}$$

□

8 Mar 7, 2023

8.1 Matrix Multiplication

Previously we learned that we can use three types of row operations to transform a matrix into its upper triangular. We want to find out a way to write any elementary matrices. We want to find a way to easily write any elementary matrices.

Suppose we are given a matrix $\mathbf{A}$. We just need to focus on one column of $\mathbf{A}$. (WHY?). Let the first column of $\mathbf{A}$ be

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}. \tag{74}$$

Suppose we have an arbitrary matrix $\mathbf{M} = (\mathbf{m_1}|\mathbf{m_2}|\mathbf{m_1})$, where is $\mathbf{m_i}$ is the column i of $\mathbf{M}$. Then

$$\begin{pmatrix} \mathbf{m_1} & \mathbf{m_2} & \mathbf{m_1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \mathbf{m_1}a_1 + \mathbf{m_2}a_2 + \mathbf{m_3}a_3 = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \end{pmatrix}. \tag{75}$$

In this way, it is easy to write out all elementary matrices.

Keep the matrix $\mathbf{A}$ *the same (Idenity):*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{76}$$

Multiply row 2 *of* $\mathbf{A}$ *by* k*:*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{77}$$

Add k *times row* 2 *to row* 3*:*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & k & 1 \end{pmatrix}. \tag{78}$$

All elementary matrices that are used to transform $\mathbf{A}$ to upper triangular form are lower triangular. For the first two cases, it is obvious. For the last case, note that we will have no need to add row i to row j when $j > i$.

We can use the same method for a 4×4 matrix. Now let $\mathbf{A}$ be a 4×4 matrix. *Keep the matrix* $\mathbf{A}$ *the same (Identity):*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{79}$$

Multiply row 4 *of* $\mathbf{A}$ *by* k*:*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & k \end{pmatrix}. \tag{80}$$

Add k times row 1 to row 3:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ k & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{81}$$

Add k times row 3 to row 4:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & k & 1 \end{pmatrix}. \tag{82}$$

8.2 Inverse of elementary matrices

The inverse of elementary matrices is the inverse process of the row operation. For example,

1. The inverse of Multiplying k to row i is dividing row i by k.
2. Adding k row i to row j is adding $-k$ row i to row j.

8.3 Example of LU

Example 18. Find the QR factorization of

$$\begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix}. \tag{83}$$

Solution:

Step 1: Multiply on left,

$$\underbrace{\begin{pmatrix} 1 & 0 \\ \frac{4}{3} & 1 \end{pmatrix}}_{\mathbf{L}^{-1}} \begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 3 & 2 \\ 0 & \frac{11}{3} \end{pmatrix}}_{U}. \tag{84}$$

Step 2: The inverse of adding $\frac{4}{3}$ of row 2 to row 1 is adding $\frac{-4}{3}$ of row 2 to row 1. Thus

$$\mathbf{L} = \begin{pmatrix} 1 & 0 \\ -\frac{4}{3} & 1 \end{pmatrix}. \tag{85}$$

Step 3: Thus we have found the LU factorization,

$$\begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{4}{3} & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 0 & \frac{11}{3} \end{pmatrix}. \tag{86}$$

□

Example 19. Find the LU factorization of

$$\begin{pmatrix} 5 & 6 & 4 \\ -4 & 1 & 5 \\ 8 & 1 & 5 \end{pmatrix}. \tag{87}$$

Solution:

Step 1:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ \frac{4}{5} & 1 & 0 \\ -\frac{8}{5} & 0 & 1 \end{pmatrix}}_{\mathbf{L}_1} \begin{pmatrix} 5 & 6 & 4 \\ -4 & 1 & 5 \\ 8 & 1 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 6 & 4 \\ 0 & \frac{29}{5} & \frac{41}{5} \\ 0 & \frac{-43}{5} & \frac{-7}{5} \end{pmatrix} \tag{88}$$

Step 2:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{43}{5}\frac{5}{29} & 1 \end{pmatrix}}_{\mathbf{L}_2} \begin{pmatrix} 5 & 6 & 4 \\ 0 & \frac{29}{5} & \frac{41}{5} \\ 0 & \frac{-43}{5} & \frac{-7}{5} \end{pmatrix} = \underbrace{\begin{pmatrix} 5 & 6 & 4 \\ 0 & \frac{29}{5} & \frac{41}{5} \\ 0 & 0 & \frac{312}{29} \end{pmatrix}}_{\mathbf{U}} \tag{89}$$

Step 3: We need to find $\mathbf{L}_1^{-1}$ and $\mathbf{L}_2^{-1}$. Using the same reasoning, we have

$$\mathbf{L}_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{8}{5} & 0 & 1 \end{pmatrix}, \quad \mathbf{L}_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{43}{5}\frac{5}{29} & 1 \end{pmatrix} \tag{90}$$

Step 4: Thus the LU factorization is

$$\begin{aligned} &\begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{8}{5} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{43}{5}\frac{5}{29} & 1 \end{pmatrix} \begin{pmatrix} 5 & 6 & 4 \\ 0 & \frac{29}{5} & \frac{41}{5} \\ 0 & 0 & \frac{312}{29} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \frac{-4}{5} & 1 & 0 \\ \frac{8}{5} & -\frac{43}{29} & 1 \end{pmatrix} \begin{pmatrix} 5 & 6 & 4 \\ 0 & \frac{29}{5} & \frac{41}{5} \\ 0 & 0 & \frac{312}{29} \end{pmatrix} \\ &= \begin{pmatrix} 5 & 6 & 4 \\ -4 & 1 & 5 \\ 8 & 1 & 5 \end{pmatrix}. \end{aligned} \tag{91}$$

□

9 Mar 15, 2023

9.1 Singular Value Decomposition

We write a SVD as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Steps to find the SVD are as follows:

Step 1: Given $\mathbf{A}_{m\times n}$, $m \geq n$. Compute $\mathbf{A}^T\mathbf{A}$. Find the eigenvalues λ_i and the orthonormal eigenvectors v_i of $\mathbf{A}^T\mathbf{A}$.

Step 2: Compute $\sigma_i = \sqrt{\lambda_i}$. Then $\mathbf{\Sigma}$ is given by a square matrix with diagonal element $\sigma_1 > \sigma_2 > \cdots \sigma_n$. Next we add some zeros rows if $m > n$ to make the matrix multiplication well-defined.

Step 3: For those non-zero σ_i, compute $u_i = \frac{Av_i}{\sigma_i}$. And then add some orthonormal vectors to make U a square matrix.

Step 4: We form the SVD as:

$$\mathbf{A} = \left(\begin{array}{c|c|c|c} u_1 & u_2 & \cdots & u_m \end{array}\right) \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \\ 0 & \cdots & 0 & \\ \vdots & & \vdots & \\ 0 & \cdots & 0 & \end{pmatrix} \left(\begin{array}{c|c|c|c} v_1 & v_2 & \cdots & v_m \end{array}\right)^T \tag{92}$$

Example 20. Compute the SVD for

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \\ 2 & 1 \end{pmatrix}.$$

Solution:
Step 1: First compute $\mathbf{A}^T\mathbf{A}$. We get

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 9 & 8 \\ 9 & 9 \end{pmatrix}.$$

The eigenvalues can be found by solving

$$\det(\lambda\mathbf{I} - \mathbf{A}^T\mathbf{A}) = 0.$$

We will get $\lambda_1 = 17$ and $\lambda_2 = 1$. Thus $\sigma_1 = \sqrt{17}, \sigma_2 = 1$. The corresponding eigenvector is

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Step 2: Hence

$$\mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Step 3: We also have

$$u_1 = \frac{1}{\sqrt{17}}\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 2\\ 2 & 2\\ 2 & 1\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix}$$

and

$$u_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 2\\ 2 & 2\\ 2 & 1\end{pmatrix}\begin{pmatrix}1\\ -1\end{pmatrix}.$$

We can pick

$$u_3 = \begin{pmatrix}\frac{2}{\sqrt{17}}\\ \frac{-3}{\sqrt{17}}\\ \frac{2}{\sqrt{17}}\end{pmatrix}\begin{pmatrix}1\\ -1\end{pmatrix}.$$

Hence

$$\mathbf{A} = \begin{pmatrix}\frac{3}{\sqrt{34}} & \frac{-1}{\sqrt{2}} & \frac{2}{\sqrt{17}}\\ \frac{4}{\sqrt{34}} & 0 & \frac{-3}{\sqrt{17}}\\ \frac{3}{\sqrt{34}} & \frac{1}{\sqrt{2}} & \frac{2}{\sqrt{17}}\end{pmatrix}\begin{pmatrix}\sqrt{17} & 0\\ 0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}^T.$$

□

9.2 Power Method

If you keeping apply a matrix $\mathbf{A}$ to some vector x then x will converges to the eigenvector that correspond to a largest eigenvalue! In particular, keep applying $\mathbf{B} := (\mathbf{A} - c\mathbf{I})^{-1}$ to x, then x will converge to the eigenvector that correspond to the eigenvalue closest to c.

Example 21. Let

$$\mathbf{A} = \begin{pmatrix}4 & -5 & 1\\ -5 & 3 & 0\\ 1 & 0 & -3\end{pmatrix}.$$

Carry out the some iterations of the power method to find the eigenvalue that is closet to 7. (The eigenvalue closest to 7 should be around 8.57236.)

Solution: We need to apply $\mathbf{B} := (\mathbf{A} - 7\mathbf{I})^{-1}$ to a random vector, let's use $x_0 = (1, 1, 1)^T$. Notice that we are actually finding x_1 such that,

$$\mathbf{B}x_0 = x_1. \tag{93}$$

That is equivalent to find x_1 such that,

$$(\mathbf{A} - 7\mathbf{I})x_1 = x_0. \tag{94}$$

(WHY?). Then in the next step we will find x_2 such that

$$\mathbf{B}x_1 = x_2. \tag{95}$$

Again this is equivalent to finding x_2 such that

$$(\mathbf{A} - 7\mathbf{I})x_2 = x_1. \tag{96}$$

This can be generalize to any step.

Let's do one iteration to test out the idea. First,

$$\mathbf{A} - 7\mathbf{I} = \begin{pmatrix} -3 & -5 & 1 \\ -5 & -4 & 0 \\ 1 & 0 & -10 \end{pmatrix}.$$

To apply B to x_0 is the same as find x_1 such that

$$\begin{pmatrix} -3 & -5 & 1 \\ -5 & -4 & 0 \\ 1 & 0 & -10 \end{pmatrix} x_1 = x_0 \tag{97}$$

This can be solve by Gaussian elimination of any other method you like. After solving (97), we have

$$x_1 = (-3/67, -13/67, -7/67). \tag{98}$$

Now we can find the first approximation of the eigenvalue, we compute

$$\lambda_1 = \frac{x_1(1)}{x_0(1)} = \frac{-3}{67}. \tag{99}$$

□