MA 214: Numerical Analysis Notes

Aryaman Maithani

2021-03-10 02:06:47+05:30
Until: Lecture 12

Disclaimer

This is just a collection of formulae/algorithms compiled together.
In the case of algorithms, I explain the procedure concisely. However, do not take this as a substitute for lecture slides as I don't go into the theory at all.
Also, I've modified some things compared to the lecture slides wherever I felt it was an error. It also is possible that I've made a typo of my own. So, be warned.
Sometimes, I also change the order in the notes compared to how it was taught in slides if I feel like that's more efficient.

---

1 Interpolation

1. **Lagrange Polynomials**
   Let $x_0,\ x_1,\ \ldots,\ x_n$ be $n+1$ distinct points in $[a,b]$. Let $f : [a,b] \to \mathbb{R}$ be a function whose value is known at those aforementioned points.
   We want to construct a polynomial $p(x)$ of degree $\leq n$ such that $p(x_i) = f(x_i)$ for all $i \in \{0,\ldots,n\}$.
   Towards this end, we define the polynomials $I_k(x)$ for $k \in \{0,\ldots,n\}$ in the following manner:
   $$I_k(x) := \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}.$$
   (Intuitive understanding: I_k is a degree n polynomial such that $I_k(x_j) = 0$ if $k \neq j$ and $I_k(x_k) = 1$.)
   Now, define $p(x)$ as follows:
   $$p(x) := \sum_{i=0}^{n} f(x_i) I_i(x)$$

2. **Newton's form**
   Let $x_0, x_1, \ldots, x_n$ be $n+1$ distinct points in $[a,b]$. Let $f : [a,b] \to \mathbb{R}$ be a function whose value is known at those aforementioned points.

We want to construct a polynomial $P_n(x)$ of degree $\leq n$ such that $p(x_i) = f(x_i)$ for all $i \in \{0, \ldots, n\}$.

We define the divided differences (recursively) as follows:

$$\begin{aligned} f[x_0] &:= f(x_0) \\ f[x_0, x_1, \ldots, x_k] &:= \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0} \qquad \text{for all } 1 < k \leq n \end{aligned}$$

With this in place, the desired polynomial $P_n(x)$ is (not so) simply:

$$\begin{aligned} P_n(x) := f[x_0] &+ f[x_0, x_1](x - x_0) \\ &+ f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &+ \cdots \\ &\vdots \\ &+ f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) \end{aligned}$$

Remarks. Note that $x - x_n$ does not appear in the last term.
Note that given $P_n(x)$, it is simple to construct $P_{n+1}(x)$.

3. **Osculatory Interpolation**
   This is essentially the same as the previous case.
   I'll state the problem in the form I think is the simplest. (Any other form can be reduced to this.)
   Suppose we are given $k+1$ distinct points $x_0, \ldots, x_k$ in $[a, b]$ and a function $f : [a, b] \to \mathbb{R}$ which is sufficiently differentiable.
   Suppose we are given the following values:

$$\begin{gathered} f^{(0)}(x_0), f^{(1)}(x_0), \ldots, f^{(m_1 - 1)}(x_0) \\ f^{(0)}(x_1), f^{(1)}(x_1), \ldots, f^{(m_2 - 1)}(x_1) \\ \vdots \\ f^{(0)}(x_k), f^{(1)}(x_k), \ldots, f^{(m_k - 1)}(x_k) \end{gathered}$$

(Notation: $f^{(0)}(x) = f(x)$ and $f^{(j)}(x)$ is the j^{th} derivative.)
Thus, we are given $m_1 + m_2 + \cdots m_k =: n + 1$ data. As usual, we now want to compute a polynomial $P_n(x)$ that agrees with f at all the data. (That is, all the given derivatives must also be same.) As it goes without saying, $P_n(x)$ must have degree $\leq n$.

To do this, we list the above points as follows:

$$\underbrace{x_0, x_0, \ldots x_0}_{m_1}, \underbrace{x_1, x_1, \ldots, x_1}_{m_2}, \ldots, \underbrace{x_k, x_k, \ldots, x_k}_{m_k}.$$

Now, we just apply the above (Newton's) formula with the following modification in the definition of the divided difference:

$$f[\underbrace{x_i, x_i, \ldots, x_i}_{p+1 \text{ times}}] := \frac{f^{(p)}(x_i)}{p!}.$$

4. **Richardson Extrapolation**
Suppose that for sufficiently small $h \neq 0$, we have the formula:

$$M = N_1(h) + k_1 h + k_2 h^2 + k_3 h^3 + \cdots,$$

for some constants $k_1, k_2, k_3, \ldots$.
Define the following:

$$N_j(h) := N_{j-1}(h/2) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{2^{j-1} - 1} \quad \text{for } j \geq 2.$$

Choose some h *sufficiently small* (whatever that means). Then, $N_j(h)$ keeps becoming a better approximation of M as j increases.
We create a table of h and $N_j(h)$ as follows:

h	$N_1(h)$	$N_2(h)$	$N_3(h)$	$N_4(h)$
h	$N_1(h)$			
$h/2$	$N_1(h/2)$	$N_2(h)$		
$h/4$	$N_1(h/4)$	$N_2(h/2)$	$N_3(h)$	
$h/8$	$N_1(h/8)$	$N_2(h/4)$	$N_3(h/2)$	$N_4(h)$

$N_4(h)$ will be a good approximation, then.
(Look at slide 15 of Lecture 7 for an example.)

Special case
Sometimes, we may have the following scenario:

$$M = N_1(h) + k_2 h^2 + k_4 h^4 + \cdots.$$

In this case, we define:

$$N_j(h) := N_{j-1}(h/2) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{4^{j-1} - 1} \quad \text{for } j \geq 2.$$

Then, we do the remaining stuff as before.
We define

$$R_h^k := \frac{N_k(h) - N_k(2h)}{N_k(h/2) - N_k(h)}.$$

The closer that R_h^k is to 4^k, the better is the approximation.

2 Numerical Integration

$$I = \int_a^b f(x)\mathrm{d}x$$

(Derivation: by approximating f in different ways using Newton's method.)

1. **Rectangle Rule**

$$I \approx (b-a)f(a)$$

$$E^R = f'(\eta)\frac{(b-a)^2}{2}, \text{ for some } \eta \in [a,b]$$

$x_0 = a$.

2. **Midpoint Rule**

$$I \approx (b-a)f\left(\frac{a+b}{2}\right)$$

$$E^M = \frac{f''(\eta)}{24}(b-a)^3, \text{ for some } \eta \in [a,b]$$

$x_0 = \frac{a+b}{2}$.

3. **Trapezoidal Rule**

$$I \approx \frac{1}{2}(b-a)[f(a)+f(b)]$$

$$E^T = -f''(\eta)\frac{(b-a)^3}{12}, \text{ for some } \eta \in [a,b]$$

$x_0 = a,\ x_1 = b$.

4. **Corrected Trapezoidal**

$$I \approx \frac{1}{2}(b-a)[f(a)+f(b)] + \frac{(b-a)^2}{12}(f'(a)-f'(b))$$

$$E^{CT} = f^{(4)}(\eta)\frac{(b-a)^5}{720}, \text{ for some } \eta \in [a,b]$$

$x_0 = x_1 = a,\ x_2 = x_3 = b$.

5. **Composite Trapezoidal**

$$I \approx \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{N-1} f(x_i) + f(x_N)\right]$$

$$E_C^T = -f''(\xi)\frac{h^2(b-a)}{12}, \text{ for some } \xi \in [a,b]$$

Here, $Nh = b-a$ and $x_i = a + ih$.

6. **Simpson's Rule**

$$I \approx \frac{b-a}{6}\left\{f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right\}$$

$$E^S = -\frac{1}{90}f^{(4)}(\eta)\left(\frac{b-a}{2}\right)^5, \text{ for some } \eta \in [a,b]$$

$x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b.$

7. **Composite Simpson's**

$$I \approx \frac{h}{6}\left[f(x_0) + 2\sum_{i=1}^{N-1} f(x_i) + 4\sum_{i=1}^{N} f\left(x_{i-1} + \frac{h}{2}\right) + f(x_N)\right]$$

$$E_C^S = -f^{(4)}(\xi)\frac{(h/2)^4(b-a)}{180}, \text{ for some } \xi \in [a,b]$$

Here, $Nh = b - a$ and $x_i = a + ih$.

8. **Gaussian Quadrature**

Let $Q_{n+1}(x)$ denote the $(n+1)^{\text{th}}$ Legendre polynomial.
Let $r_0, \ldots, r_{n+1}$ be its roots. (These will be distinct, symmetric about the origin and will lie in the interval $[-1, 1]$.)
For each $i \in \{0, \ldots, n\}$, we define c_i as follows:

$$c_i := \int_{-1}^{1}\left(\prod_{k=0, k\neq i}^{n}\frac{x - x_k}{x_i - x_k}\right)\mathrm{d}x.$$

Then, we have

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=0}^{n} f(r_i)c_i.$$

Moreover, if f is a polynomial of degree $\leq 2n+1$, then the above is "approximation" is exact.

Standard values:
$n = 0 : Q_1(x) = x$ and $r_0 = 0$. $c_0 = 2$.
$n = 1 : Q_2(x) = x^2 - \frac{1}{3}$ and $r_0 = -\frac{1}{\sqrt{3}},\ r_1 = \frac{1}{\sqrt{3}}$. $c_0 = c_1 = 1$.
$n = 2 : Q_3(x) = x^3 - \frac{3}{5}x$ and $r_0 = -\sqrt{\frac{3}{5}},\ r_1 = 0,\ r_2 = \sqrt{\frac{3}{5}}$. $c_0 = c_2 = 5/9,\ c_1 = 8/9$.

9. **Improper integrals using Taylor series method**

Suppose we have $f(x) = \frac{g(x)}{(x-a)^p}$ for some $0 < p < 1$ and are asked to calculate $I = \int_a^b f(x)\mathrm{d}x$.
For the sake of simplicity, I assume $a = 0$ and $b = 1$.

Let $P_4(x)$ denote the fourth Taylor polynomial of g around a. (In this case 0.)
Now, compute $I_1 = \displaystyle\int_0^1 \frac{P_4(x)}{x^p}\mathrm{d}x$. This can be integrated exactly. (Why?)
Now, we approximate $I - I_1$.
Define

$$G(x) := \begin{cases} \dfrac{g(x) - P_4(x)}{x^p} & \text{if } 0 < x \leq 1 \\ 0 & \text{if } x = 0 \end{cases}$$

Then, approximate $I_2 = \displaystyle\int_0^1 G(x)\mathrm{d}x$ using the composite Simpson's rule.
Then, $I = I_1 + I_2$.

For the case of $a = 0,\ b = 1$ and $N = 2$ for the composite Simpson's part, we get that
$I_2 \approx \frac{1}{12}[2G(0.5) + 4G(0.25) + 4G(0.75) + G(1)]$.

That is, finally:

$$I \approx \int_0^1 \frac{P_4(x)}{x^p}\mathrm{d}x + \frac{1}{12}[2G(0.5) + 4G(0.25) + 4G(0.75) + G(1)].$$

10. **Adaptive Quadrature**

Let $I = \displaystyle\int_a^b f(x)\mathrm{d}x$ be the integral that we want to approximate.
Suppose that ϵ is the accuracy to which we want I. That is, we want a number P such that $|P - I| < \epsilon$.

Here is what you do:
Subdivide $[a, b]$ into N intervals: $[x_0, x_1],\ [x_1, x_2],\ \ldots,\ [x_{n-1}, x_n]$.
(Naturally, $a = x_0 < x_1 < \ldots < x_n = b$.)
Now, for each subinterval, compute the following values:

$$S_i = \frac{h}{6}\left(f(x_i) + 4f\left(x_i + \frac{h}{2}\right) + f\left(x_{i+1}\right)\right), \text{ and}$$

$$\overline{S_i} = \frac{h}{12}\left(f(x_i) + 4f\left(x_i + \frac{h}{2}\right) + 2f\left(x_i + \frac{h}{2}\right) + 4f\left(x_i + \frac{3h}{4}\right) + f(x_{i+1})\right).$$

Now, calculate $E_i = \frac{1}{15}|\overline{S_i} - S_i|$.

Now, if $E_i \leq \dfrac{x_i - x_{i-1}}{b - a}\epsilon$, then move on to the next interval.

Otherwise, subdivide again to better approximate $\displaystyle\int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x$.

Finally, sum up all the $\overline{S_i}$s and that's the answer. That is,

$$I \approx P = \sum_{i=1}^{n} \overline{S_i}.$$

Check slide 25 of Lecture 6 for example.

11. **Romberg Integration**
Essentially the baby of composite Trapezoidal rule and Richardson extrapolation.
Suppose we want to calculate $\int_a^b f(x)\mathrm{d}x$.
Let N be a power of 2.
$T_N := \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{N-1} f(a+ih) + f(x_N)\right]$, where $Nh = b-a$.
Note that T_N can be computed using $T_{N/2}$ (assuming $N \neq 2^0$) as:

$$T_N = \frac{T_{N/2}}{2} + h\sum_{i=1}^{N/2} f\left(a + (2i-1)h\right).$$

In the above, we have $Nh = b-a$.

Now, for $m \geq 1$, we define:

$$T_N^m = T_N^{m-1} + \frac{T_N^{m-1} - T_{N/2}^{m-1}}{4^m - 1}.$$

(Where T_N^0 is just T_N.)
(Also, for some reason, T_N' has been used instead of T_N^1.)
Note that $\frac{N}{2^m}$ must be an integer for T_N^m to be defined. We create a table as follows:

N	T_N	T_N'	T_N^2	T_N^3
1	T_1			
2	T_2	T_2^1		
4	T_4	T_4^1	T_4^2	
8	T_8	T_8^1	T_8^2	T_8^3

T_8^3 will be a good approximation, then.
(Look at slide 25 of Lecture 7 for an example.)

Remark. It can be shown that $I = T_N + c_2h^2 + c_4h^4 + \cdots$. This is why we used the special case formula of 1. 4.

3 Numerical Differentiation

1.

$$f'(a) \approx \frac{f(a+h) - f(a)}{h}$$

$$E(f) = -\frac{1}{2}hf''(\eta) \quad \text{for some } \eta \in [a, a+h].$$

2. **Central Difference Formula**

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}$$

$$E(f) = -\frac{1}{6}h^2 f^{(3)}(\eta) \quad \text{for some } \eta \in [a-h, a+h].$$

Note that this is an $O(h^2)$ approximation. Thus, we can use the special case of §1. 4. for better accuracy.

3.

$$f'(a) \approx \frac{-3f(a) + 4f(a+h) - f(a+2h)}{2h}$$

$$E(f) = \frac{1}{3}h^2 f^{(3)}(\eta) \quad \text{for some } \eta \in [a, a+2h].$$

Formula 2 is always the better one whenever applicable. At end points, formula 3 is better than formula 1.

4. **Central difference for second derivative**

$$f''(x_0) = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi),$$

for some $\xi \in (x_0 - h, x_0 + h)$.

5. **Solving boundary-value problems in ODE**
Suppose that we want to solve the following (linear) ODE:

$$y''(x) + f(x)y'(x) + g(x)y = q(x)$$

in the interval $[a, b]$ such that we know $y(a) = \alpha$, and $y(b) = \beta$.

Set $h := \frac{b-a}{N}$ for some $N \in \mathbb{N}$ and $x_i = a + ih$ for $h \in \{0, 1, \ldots, N\}$.
Using central difference approximation, we set up $N - 1$ linear equations as follows:

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + f(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + g(x_i)(y_i) = q(x_i)$$

$$i = 1, 2, \ldots, N-1$$

The above equations can be rearranged as:

$$\left(1 - \frac{hf_i}{2}\right) y_{i-1} + (-2 + h^2 g_i)y_i + \left(1 + \frac{hf_i}{2}\right) y_{i+1} = h^2 q_i,$$

for $i = 1, \ldots, N-1$; where $f_i = f(x_i)$ and so on.

4 Solution of non-linear equations

Let f be a continuous function on $[a_0, b_0]$ such that $f(a_0)f(b_0) < 0$ in all these cases. We want to find a root of f in $[a_0, b_0]$. (Existence in implied.)

1. **Bisection Method**
   Set $n = 0$ to start with.
   Loop over the following:
   Set $m = \frac{a_n + b_n}{2}$.
   If $f(a_n)f(m) < 0$, then set $a_{n+1} = a_n$ and $b_{n+1} = m$.
   Else, set $a_{n+1} = m$ and $b_{n+1} = b_n$.
   Increase n by one.
   We still have a root in $[a_n, b_n]$.

2. **Regula-falsi or false-position method**
   Set $n = 0$ to start with.
   Loop over the following:
   Set $w = \dfrac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)}$.
   If $f(a_n)f(w) < 0$, then set $a_{n+1} = a_n$ and $b_{n+1} = w$.
   Else, set $a_{n+1} = w$ and $b_{n+1} = b_n$.
   Increase n by one.
   We still have a root in $[a_n, b_n]$.

3. **Modified regula-falsi**
   Set $n = 0$ and $w_0 = a_0$ to start with.
   Loop over the following:
   Set $F = f(a_n)$ and $G = f(b_n)$.
   Set $w_{n+1} = \dfrac{Ga_n - Fb_n}{G - F}$.
   If $f(a_n)f(w_{n+1}) \leq 0$, then set $a_{n+1} = a_n$ and $b_{n+1} = w_{n+1}$ and $G = f(w_{n+1})$.
   Furthermore, if we also have $f(w_n)f(w_{n+1}) > 0$, set $F = \frac{F}{2}$.
   Else, set $a_{n+1} = w_{n+1}$ and $b_{n+1} = b_n$ and $F = f(w_{n+1})$.
   Furthermore, if we also have $f(w_n)f(w_{n+1}) > 0$, set $G = \frac{G}{2}$.
   Increase n by one.
   We still have a root in $[a_n, b_n]$.

4. **Secant method**
   Set $x_0 = a$, $x_1 = b$ and until satisfied, keep computing x_n given by

   $$x_{n+1} = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} \qquad \text{for } n \geq 1.$$

   Remark. This process will be forced to stop if we arrive at $f(x_n) = f(x_{n-1})$ at some point.

5 Iterative methods

1. **Newton's Method**
   You are given a function f which is continuously differentiable and you want to find its root. You are also given some x_0.
   Compute the following sequence recursively until satisfied:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad \text{for } n \geq 0.$$

2. **Fixed point iteration**
   Let I be a closed interval in $\mathbb{R}$. Let $f : I \to I$ be a differentiable function such that there exists some $K \in [0, 1)$ such that $|f'(x)| \leq K$ for all $x \in I$.
   Then, there is a unique $\xi \in I$ such that $f(\xi) = \xi$. To find this fixed point, choose any $x_0 \in I$ and define the sequence

$$x_n := f(x_{n-1}) \quad n \geq 1.$$

   Then, $x_n \to \xi$.

3. **Aitken's Δ^2 Process**
   Definition. Given a sequence (x_n), let $\Delta x_n := x_{n+1} - x_n$.
   Then, $\Delta^2 x_n = x_{n+2} - 2x_{n+1} + x_n$.

   Given a sequence $x_0, x_1, \ldots$ converging to ξ, calculate $\widehat{x_1}, \widehat{x_2}, \ldots$ by

$$\widehat{x_n} := x_{n+1} - \frac{(\Delta x_n)^2}{\Delta^2 x_{n-1}}.$$

   Then, $\widehat{x_n} \to \xi$.

   If the sequence $x_0, x_1, \ldots$ converges linearly to ξ, that is, if

$$\xi - x_{n+1} = K(\xi - x_n) + \theta(\xi - x_n), \qquad \text{for some } K \neq 0$$

   then $\widehat{x_n} = \xi + O(\xi - x_n)$, that is, $\dfrac{\widehat{x_n} - \xi}{x_n - \xi} \to 0$.

4. **Steffensen iteration**
   Let $g(x)$ be the function whose fixed point is desired. Let y_0 be some given point.
   Set $n = 0$ to start with.
   Loop over the following:
   Set $x_0 = y_n$.
   Set $x_1 = g(x_0),\ x_2 = g(x_1)$.
   Calculate Δx_1 and $\Delta^2 x_0$.

Set $y_{n+1} = x_2 - \dfrac{(\Delta x_1)^2}{\Delta^2 x_0}$.

Increase n by 1.

Note that we get a sequence $y_0, y_1, y_2, \ldots$. However, we only ever have x_0, x_1 and x_2.

Definition 1. Let $x_0, x_1, x_2, \ldots$ be a sequence that converges to ξ and set $e_n = \xi - x_n$. If there exists a number P and a constant $C \neq 0$ such that

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^P} = C,$$

then P is called the **order of convergence** and C is called **asymptotic error constant**.

Examples.

1. **Fixed point iteration**
   ξ fixed point of $g : I \to I$ and $g'(\xi) \neq 0$.
   $P = 1$ and $C = |g'(\xi)|$.
2. **Newton's method**
   $\lim\limits_{n\to\infty} \dfrac{|e_{n+1}|}{|e_n|} = \dfrac{1}{2}\left|\dfrac{f''(\xi)}{f'(\xi)}\right|$.
   (If ξ is a double root, then $P = 1$.)
3. **Secant method**
   $$|e_{n+1}| = C|e_n||e_{n-1}|$$
   $P = \frac{1+\sqrt{5}}{2} = 1.618\ldots$.

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^P} = \left|\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right|^{1/P}, \quad \text{provided } f'(\xi) \neq 0.$$

Theorem 2. Let $f : [a, b] \to \mathbb{R}$ be in $C^2[a, b]$ and let the following conditions be satisfied:

1. $f(a)f(b) < 0$,
2. $f'(x) \neq 0$, for all $x \in [a, b]$,
3. $f''(x)$ doesn't change sign in $[a, b]$ (might be zero at some points),
4. $$\frac{|f(a)|}{|f'(a)|} \leq b - a \text{ and } \frac{|f(b)|}{|f'(b)|} \leq b - a.$$

Then, the Newton's method converges to the unique solution ξ of $f(x) = 0$ in $[a, b]$ for any choice $x_0 \in [a, b]$.

6 Solving systems of linear equations

6.1 LU Factorisation

We want solve $Ax = b$ where A is some known $n \times n$ matrix, b a known $n \times 1$ matrix and x is unknown.
Assumption: $Ax = b$ can be solved without any row interchange.
We define (finite) sequences of matrices $A^{(n)} = [a_{ij}^{(n)}]$ and $b^{(n)}$.

Define $A^{(1)} := A$. Let $m_{ji} := \dfrac{a_{ji}^{(i)}}{a_{ii}^{(i)}}$.

Define $M^{(1)}$ as

$$M^{(1)} := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -m_{21} & 1 & 0 & \cdots & 0 \\ -m_{31} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -m_{n-1,1} & 0 & 0 & \cdots & 0 \\ -m_{n1} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Thus, we can write $M^{(1)}A^{(1)}x = M^{(1)}b$.
Let $A^{(2)} := M^{(1)}A^{(1)}$ and $b^{(2)} = M^{(1)}b^{(1)}$.
Note that $A^{(2)}$'s first column will just have the top element non-zero and everything below will be zero.

We can similarly construct the later matrices that perform the row operations. In general, we have:

$$M^{(k)} := \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & -m_{k+1,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -m_{n,k} & \cdots & 1 \end{bmatrix},$$

along with

$$A^{(k+1)} = M^{(k)}A^{(k)} = M^{(k)} \cdots M^{(1)}A, \text{ and}$$
$$b^{(k+1)} = M^{(k)}b^{(k)} = M^{(k)} \cdots M^{(1)}b.$$

Finally, set $U = A^{(n)}$ and $L = [M^{(1)}]^{-1} \cdots \left[M^{(n-1)}\right]^{-1}$.
Then, we have

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n-1,1} & m_{n-1,2} & m_{n-1,3} & \cdots & 0 \\ m_{n1} & m_{n2} & m_{n3} & \cdots & 1 \end{bmatrix}.$$

Thus, we have $A = LU$. Now, set $y = Ux$. We solve $Ly = b$ for y. This is easy because L is lower triangular.
Then, we solve $Ux = y$ for x.

Check slide 27 of Lecture 11 for example.

LU factorisation requires $\theta(n^3/3)$ multiplication and division and $\theta(n^3/3)$ addition and subtraction. However, once the factorisation is done, it can be used again and again to solve $Ax = b$ for different values of b. The number of equations then taken to solve $Ax = b$ is $\theta(2n^2)$.

6.2 LPU Factorisation

In the previous case, we assumed that we aren't doing any row operations on A when doing the Gauss elimination. Now, we relax that condition.
Suppose, if possible, we have done some row exchanges while doing Gauss elimination on A. Construct the matrix P which is obtained upon doing those exact row exchanges on I. (Only the row exchanges.)
Suppose that the the final product of doing the Gauss elimination following the previous procedure gave us the upper triangular matrix as U and lower one as L.
Then, that means $PA = LU$. Thus, solving $Ax = b$ can first be reduced to $PAx = Pb = b'$. Now, converting PA to LU does not take any row exchange, so we're back in business. (Back to the previous case, that is.)
Check slide 15 of Lecture 12 for example.

6.3 Cholesky's Algorithm

Definition 3. A matrix A is positive definite if:

1. $A = A^t$, and
2. $x^tAx > 0$ for all $x \neq 0$.

Given any positive definite matrix A, we can write $A = LL^t$, where L is lower triangular. (And thus, L^T would be upper triangular.)
The algorithm to do so is as follows:
Let $L = (l_{ij})$. We now construct these entries.

`set` $l_{ii} := \sqrt{a_{11}}$.
`for` $j = 2, 3, \ldots, n$:
    `set` $l_{j1} := \frac{a_{j1}}{l_{11}}$.
`for` $i = 2, 3, \ldots, n-1$:

`set` $l_{ii} := \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$

`for` $j = i+1, \ldots, n$:

$$\texttt{set } l_{ji} := a_{ji} - \frac{\sum_{k=1}^{i-1} l_{jk}l_{ik}}{l_{ii}}.$$

$$\texttt{set } l_{nn} := \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2}.$$

Note that in the above, we've only defined l_{ij} for when $i \geq j$. The others are obviously 0.

Cholevsky's algorithm requires $\theta(n^3/6)$ multiplication and division and $\theta(n^3/6)$ addition and subtraction.

6.4 Scaled partial pivoting

Suppose we want to solve $Ax = b$. This could lead to round off errors (check slide 17 of Lecture 12). One possible way to combat this is to do the following:

First, define the scale factor s_i of row i as the maximum modulus of any element in the row. More precisely, $s_i := \max\limits_{1\leq j\leq n} |a_{ij}|$.
(Note that we assume that A is invertible and thus, every s_i will be nonzero.)

Now, along the first column, we find the row which has the greatest ratio $\frac{|a_{k1}|}{s_k}$. Suppose this is achieved for row p, that is,

$$\frac{|a_{p1}|}{s_p} = \max_{1\leq k\leq n} \frac{|a_{k1}|}{s_k}.$$

If $p \neq 1$, then we perform $R_1 \leftrightarrow R_p$.
(Now, when we refer to s_p, we will mean the scale factor of the new p, that is, the original s_1. Similarly, for s_1.)
Now, we make everything below a_{11} zero using the standard row operation.

Now, we repeat the same thing by going along column 2 and finding the row which has the greatest ratio $\frac{|a_{k2}|}{s_k}$. If the row is not 2, then we perform the interchange and go on.

Two things I would like to point out:
1. We never change the value of s_i except the interchanging that we do. In particular, after doing the row operations like $R_2 - m_{21}R_1$, I will **not** recalculate the value of s_2.
2. When considering the ratios $\frac{|a_{ki}|}{s_k}$ for column i, we will consider the a_{ki} that is currently present. That is, in this case, we will consider the values that we get after performing all the operations that have been performed until that point.

Consider the example given in slide 22 of Lecture 12. After doing $R_1 \leftrightarrow R_3, R_2 - \frac{4.01}{1.09}R_1$, and $R_3 - \frac{2.11}{1.09}R_1$, we *do not* recalculate s_i. (We do exchange the values.)

However, when considering the ratios, we consider the new matrix for taking a_{k2} obtained after the above subtractions.

7 Matrices

Definition 4 (Norm)**.** A norm is a function $a \mapsto \|a\|$ from $\mathbb{R}^n$ to $\mathbb{R}$ such that:

1. $\|a\| \geq 0$ for all $a \in \mathbb{R}^n$ and $\|a\| = 0 \iff a = 0$,
2. $\|ra\| = |r|\|a\|$ for all $r \in \mathbb{R}$ and $a \in \mathbb{R}^n$,
3. $\|a + b\| \leq \|a\| + \|b\|$ for all $a, b \in \mathbb{R}^n$.

Following are *some* examples of norms on $\mathbb{R}^n$:

1. $\|x\|_1 = |x_1| + \cdots |x_n|$,
2. $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$,
3. $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$.

(In the above, $x = (x_1, \ldots, x_n)$ as usual.)

Definition 5 (Matrix norm)**.** A norm is a function $A \mapsto \|A\|$ from $M_n(\mathbb{R})$ (set of all $n \times n$ real matrices) to $\mathbb{R}$ such that:

1. $\|A\| \geq 0$ for all $A \in M_n(\mathbb{R})$ and $\|A\| = 0 \iff A = 0$,
2. $\|rA\| = |r|\|A\|$ for all $r \in \mathbb{R}$ and $A \in M_n(\mathbb{R})$,
3. $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in M_n(\mathbb{R})$,
4. $\|AB\| \leq \|A\|\|B\|$ for all $A, B \in M_n(\mathbb{R})$.

Definition 6 (Induced norm)**.** Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, we get an induced matrix norm $\|\cdot\|$ on $M_n(\mathbb{R})$ defined as

$$\|A\| := \max_{\|x\|=1} \|Ax\|.$$

(The $\|\cdot\|$ on the right is the norm on $\mathbb{R}^n$ that we started with.)

Theorem 7. One has the equalities:

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1} |a_{ij}|$$

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1} |a_{ij}|.$$

That is, $\|A\|_\infty$ is the maximum of (modulus) row sums and $\|A\|_1$ of column.

Proposition 8 (Some properties)**.** Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let it also denote the corresponding induced norm. Fix a matrix $A \in M_n(\mathbb{R})$.

1. $\|I\| \geq 1$, where I is the identity matrix.

2. $\|Az\| \leq \|A\|\|z\|$ for all $z \in \mathbb{R}^n$.
3. If A is invertible and $z \in \mathbb{R}^n$, then
$$\frac{\|z\|}{\|A^{-1}\|} \leq \|Az\| \leq \|A\|\|z\|.$$
4. If A is invertible, then $\|A\|\|A^{-1}\| \geq 1$.

Definition 9 (Condition number)**.** For an invertible matrix A, $\operatorname{cond}(A) := \|A\|\|A^{-1}\|$.

Note that the condition is not intrinsic to A, it depends on the norm chosen as well.

Theorem 10. Let A be invertible. Then,
$$\frac{1}{\operatorname{cond} A} = \left\{ \frac{\|A - B\|}{\|B\|} : B \text{ is not invertible} \right\}.$$
In particular, if B is a matrix such that $\|A - B\| < \dfrac{1}{\|A^{-1}\|}$, then B is invertible.

The above is useful because one can pick any singular matrix B of choice and compute $\|A-B\|^{-1}$ to get a lower bound on $\|A^{-1}\|$.

Proposition 11. Let $A \in M_n(\mathbb{R})$ be invertible and **upper triangular**. Then,
$$\operatorname{cond} A \geq \frac{\|A\|_\infty}{\min_{1 \leq i \leq n} |a_{ii}|}.$$
(The $\operatorname{cond} A$ on the left is with respect to the induced $\|\cdot\|_\infty$ norm.)

7.1 Solving Linear Equations

Definition 12 (Error and residual error)**.** Suppose that we wish to solve $Ax = b$. Let $\bar{x}$ denote the exact solution and $\hat{x}$ the calculated solution. Then, we define the *error* e as
$$e := \bar{x} - \hat{x}$$
and the *residual error* r as
$$r := Ae = b - A\hat{x}.$$

Note that we always that the matrix A is invertible. In turn, we may assume $b \neq 0$ since $Ax = 0$ only has the trivial solution.

Proposition 13. We have the following inequalities
$$\begin{aligned}
\frac{\|r\|}{\|A\|} \leq \|e\| &\leq \|A^{-1}\|\|r\|, \\
\frac{\|b\|}{\|A\|} \leq \|\bar{x}\| &\leq \|A^{-1}\|\|b\|, \\
\frac{1}{\operatorname{cond} A}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|\bar{x}\|} &\leq \operatorname{cond} A \frac{\|r\|}{\|b\|}.
\end{aligned}$$

(The second actually follows from the first by taking $\hat{x} = 0$.)

Iterative improvement of solution

Let $Ax = b$ be the system we wish to solve.
Compute a first solution $\hat{x}^{(1)}$.
Consider the system $Ae = b - A\hat{x}^{(1)}$ for e unknown.
Let $\hat{e}^{(1)}$ be a calculated (approximate) solution of the above.
Then, $\hat{x}^{(2)} = \hat{x}^{(1)} + \hat{e}^{(1)}$ is a "better" solution.
Now solve $Ae = b - A\hat{x}^{(2)}$ for a new solution $\hat{e}^{(2)}$ and continue ad nauseam.

Iteration functions

Definition 14 (Contraction)**.** Let $S \subset \mathbb{R}^n$ be closed. A map $g : S \to S$ is said to be a *contraction* if there exists $k < 1$ such that

$$\|g(x) - g(y)\| \leq k\|x - y\|$$

for all $x, y \in S$.

It is okay if you do not recall what closed means. We will work with $S = \mathbb{R}^n$ anyway.

Proposition 15. Let $g : S \to S$ be a contraction. Then, g has a unique fixed point ξ in S. Moreover, starting with any $x_0 \in S$ and defining

$$x_{n+1} = g(x_n)$$

produces a sequence converging to ξ.

Definition 16 (Approximate inverse)**.** Let A be *any* real $n \times n$ matrix. $C \in M_n(\mathbb{R})$ is said to be an *approximate inverse* for A if there exists some matrix norm $\|\cdot\|$ such that

$$\|I - CA\| < 1.$$

Note that a priori, we have not required for A or C to be invertible.

Theorem 17. Let $A, C \in M_n(\mathbb{R})$ be such that C is an approximate inverse of A. Then, A and C are both invertible.

Let C now denote an approximate inverse of A. As usual, we wish to solve $Ax = b$.

Define the function $g : \mathbb{R}^n \to \mathbb{R}^n$ as

$$g(x) = Cb + (I - CA)x = x + C(b - Ax).$$

Then, $A\xi = b$ iff $g(\xi) = \xi$. Thus, solving the system is equivalent to finding a fixed point for g. Note that

$$\|g(x) - g(y)\| \leq \|I - CA\|\|x - y\|.$$

Since $\|I - CA\| < 1$, the map g is a contraction. Thus, any starting $x^{(0)}$ will yield a sequence converging to ξ. We have

$$x^{(m+1)} = x^{(m)} + C(b - Ax^{(m)})$$

for $m \geq 0$ that converges to ξ.

There are two common choices of C. Both follow by first defining $\hat{L}$, $\hat{D}$, and $\hat{U}$ as in:

$$\begin{aligned}
\hat{L} = (l_{ij}) \quad & l_{ij} = \begin{cases} a_{ij} & \text{if } i > j, \\ 0 & \text{if } i \leq j, \end{cases} \\
\hat{D} = (d_{ij}) \quad & d_{ij} = \begin{cases} a_{ij} & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \\
\hat{U} = (u_{ij}) \quad & u_{ij} = \begin{cases} a_{ij} & \text{if } i < j, \\ 0 & \text{if } i \geq j. \end{cases}
\end{aligned}$$

(Basically take A and decompose it into an upper-triangular, diagonal and lower-triangular matrix.)

We may assume that all diagonal entries of A are non-zero. (Otherwise we may do row operations to make that the case.) Thus, $\hat{D}$ is invertible.

Now, the two choices of C are:

1. (**Jacobi iteration**) $C = \hat{D}^{-1}$.
   Letting $x_i^{(m)}$ denote the i-th component of $x^{(m)}$, we get the recurrence

   $$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{i \neq j} a_{ij} x_j^{(m)} \right),$$

   for $i = 1, \ldots, n$.

2. (**Gauss-Seidel**) $C = (\hat{L} + \hat{D})^{-1}$.
   With notation as earlier, we have

   $$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(m+1)} - \sum_{j>i} a_{ij} x_j^{(m)} \right),$$

   for $i = 1, \ldots, n$.

Note that we are guessing that the above are actually approximate identities. For the second one, note that $I - CA = (\hat{L} + \hat{D})^{-1}\hat{U}$. Thus, the desired C is an approximate inverse iff

$$\|C\hat{U}\| < 1$$

for some matrix norm.

Lemma 18. Suppose that A is strictly diagonally dominant. Then, the Jacobi iteration converges.

Recall that A is said to be *strictly diagonally dominant* if the each diagonal entry has a greater absolute value than the sum of the absolute values of the other elements in its row. That is,

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \forall\, i = 1, \ldots, n.$$

Definition 19 (Spectral radius)**.** Given a matrix $B \in M_n(\mathbb{R})$, we define its spectral radius $\rho(B)$ as

$$\rho(B) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } B\}.$$

Theorem 20. C is an approximate inverse of $A \iff \rho(I - CA) < 1$.

8 Initial Value Problem

We now wish to solve an ODE of the form

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x, y) \\ y(x_0) &= y_0 \end{aligned}$$

in some neighbourhood of (x_0, y_0). We also assume that all partial derivatives of f exist.

The idea is to differentiate the equation with respect to x, keeping in mind that y is also a function of x. By abuse of notation, we use f' to denote the derivative of the function $x \mapsto f(x, y(x))$. (Here y is a solution of the DE.)

For example, if we have

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x, y) := y - x^2 + 1 \\ y(0) &= 0.5 \end{aligned}$$

on $0 \le x \le 2$. Then, we have

$$f'(x) = y'(x) - 2x = (y - x^2 + 1) - 2x.$$

Similarly, we may compute f'' by substituting y' as $y - x^2 + 1$ again and so on.

In general, we have

$$\begin{aligned} y' &= f(x, y), \\ y'' &= f' = f_x + f_y f, \\ y''' &= f'' = f_{xx} + f_{xy} f + f_{yx} f + f_{yy} f^2 + f_y f_x + f_y^2 f. \end{aligned}$$

Now, note that we know that value of y of x_0. Instead of solving it throughout, we will now approximate fix some *step* h and define the points $x_n = x + nh$ for $n = 0, 1, \ldots$ and let y_n denote the *approximate* solution at x_n. (The exact solution will always be denoted by $y(x_n)$ and approximate as y_n.)

In other words, we solve the problem *discretely.*

Define T_k so that we have

$$hT_k(x, y) = hf(x, y) + \frac{h^2}{2!} f'(x, y) + \frac{h^3}{3!} f''(x, y) + \cdots + \frac{h^k}{k!} f^{(k-1)}(x, y).$$

(Note that h is there on the right.)

8.1 Taylor algorithm of order k

We wish to find an approximate solution to

$$y' = f(x, y), \quad y(a) = y_0$$

over the interval $[a, b]$.

Algorithm:

1. Fix some N and let $h = \frac{b-a}{N}$. As mentioned, define $x_n := a + nh$ for $n = 0, \dots, N$.
2. Define y_{n+1} recursively as
$$y_{n+1} = y_n + hT_k(x_n, y_n)$$
for $n = 0, \dots, N$. (Note that y_0 is given already as $y(a)$.)

The local error is given as

$$E = \frac{h^{k+1}}{(k+1)!} y^{(k+1)}(\xi) = \frac{h^{k+1}}{(k+1)!} f^{(k)}(\xi, y(\xi))$$

for some $\xi \in [x_n, x_n + h]$.

The Taylor method of order $k = 1$ is called **Euler's method**. More explicitly, we have

$$y_{n+1} = y_n + hf(x_n, y_n),$$

nice and simple. The local error is

$$E = \frac{h^2}{2} y''(\xi), \quad x_n \le \xi \le x_{n+1}.$$

Note that it is $O(h^2)$.

Note that we have

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(\xi_n), \quad x_n \le \xi_n \le x_{n+1}.$$

(Note that the above have the exact values.)

Letting e_n denoting the error $y(x_n) - y_n$ gives

$$e_{n+1} = e_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{h^2}{2} y''(\xi_n).$$

Applying MVT, the above reduces to

$$e_{n+1} = e_n + hf_x(x_n, \bar{y}_n)e_n + \frac{h^2}{2} y''(\xi_n).$$

Theorem 21. Let y_n denote the approximate solution obtained via Euler's method to

$$y' = f(x, y), \quad y(a) = y_0.$$

If the exact solution y of the above DE has continuous second derivatives on $[a, b]$ and if we have the inequalities

$$|f_y(x, y)| \leq L, \quad |y''(x)| \leq Y$$

on the interval for fixed L and Y, then

$$|e_n| \leq \frac{hY}{2L}\left(e^{(x_n - x_0)L} - 1\right).$$

8.2 Runge-Kutta (R-K) Method

We continue with the same notations as before.

R-K Method of Order 2

We have the recurrence

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2),$$

where $K_1 = hf(x_n, y_n)$ and $K_2 = hf(x_n + h, y_n + K_1)$.

The local error is $O(h^3)$.

R-K Method of Order 4

We have the recurrence

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

where

$$\begin{aligned}
K_1 &= hf(x_n, y_n), \\
K_2 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_1\right), \\
K_3 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}K_2\right), \\
K_4 &= hf(x_n + h, y_n + K_3).
\end{aligned}$$

The local error is $O(h^5)$.