Numerical Analysis Lecture Notes

Rebekah Dix

November 23, 2018

Contents

1 Results from Real Analysis

Theorem 1 (The Mean Value Theorem)**.** Suppose f is a real-valued function, defined and continuous on the closed interval $[a, b] \in \mathbb{R}$ and f differentiable on the open interval (a, b). Then there exists a number $\xi \in (a, b)$ such that

$$f(b) - f(a) = f'(\xi)(b - a) \tag{1}$$

Theorem 2 (Taylor's Theorem)**.** Suppose that n is a nonnegative integer, and f is a real-valued function, defined and continuous on the closed interval $[a, b]$ of $\mathbb{R}$, such that the derivatives of f of order up to and including n are defined and continuous on the closed interval $[a, b]$. Suppose further that $f^{(n)}$ is differentiable on the open interval (a, b). Then, for each value of $x \in [a, b]$, there exists a number $\xi = \xi(x)$ in the open interval (a, b) such that

$$f(x) = f(a) + (x - a)f'(a) + \cdots + \frac{(x - a)^n}{n!} f^{(n)}(a) + \frac{(x - a)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi) \tag{2}$$

2 Solution of equations by iteration

2.1 Simple Iteration

Theorem 3 (Existence of Root)**.** Let f be a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. Assume further, that $f(a)f(b) \leq 0$; then, there exists ξ in $[a, b]$ such that $f(\xi) = 0$.

Proof. The condition $f(a)f(b) \leq 0$ implies that $f(a)$ and $f(b)$ have opposite signs, or one of them is 0. If either $f(a)$ or $f(b)$ is 0, then we've found a root. Suppose that both endpoints are non-zero (in which case they have opposite signs). In this case, 0 must belong to the open interval whose endpoints are $f(a)$ and $f(b)$. The intermediate value theorem gives the existence of a root in the open interval (a, b). Thus, in both cases, a zero is guaranteed. $\square$

- The converse of Theorem 3 is clearly false.

Theorem 4 (Brouwer's Fixed Point Theorem)**.** Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Then, there exists $\xi \in [a, b]$ such that $\xi = g(\xi)$. ξ is called a fixed point of the function g.

Proof. Define a function $f(x) = x - g(x)$. If we find a root ξ of f, then ξ is a fixed point of g. Then,

$$f(a)f(b) = (a - g(a))(b - g(b)) \leq 0 \tag{3}$$

By assumption, $a \leq g(a), g(b) \leq b$. Therefore, the first term is negative and the second term is positive. Therefore, $f(a)f(b) \leq 0$. By Theorem 3, there exists a $\xi \in [a, b]$ such that $f(\xi) = 0$. Then, for this ξ, $g(\xi) = \xi$. $\square$

Definition 1 (Simple Iteration)**.** Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Given that $x_0 \in [a, b]$, the recursion defined by

$$x_{k+1} = g(x_k) \tag{4}$$

is called simple iteration; the numbers x_k, $k \geq 0$, are referred to as iterates.

- If this sequence converges, the limit must be a fixed of g, since g is continuous on a closed interval. Note that

$$\xi = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\left(\lim_{k\to\infty} x_k\right) = g(\xi) \tag{5}$$

Definition 2 (Contraction)**.** Let g be a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. Then, g is said to be a contraction on $[a, b]$ if there exists a constant L such that $0 < L < 1$ and

$$|g(x) - g(y)| \leq L|x - y| \quad \forall x, y \in [a, b] \tag{6}$$

Theorem 5 (Contraction Mapping Theorem)**.** Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Suppose g is a contraction on $[a, b]$. Then, g has a unique fixed point ξ in the interval $[a, b]$. Moreover, the sequence (x_k) defined by simple iteration converges to ξ as $k \to \infty$ for any starting value x_0 in $[a, b]$.

Let $\epsilon > 0$ be a certain tolerance, and let $k_0(\epsilon)$ denote the smallest positive integer such that x_k is no more than ϵ away from the fixed point ξ (i.e. $|x_k - \xi| \leq \epsilon$) for all $k \geq k_0(\epsilon)$. Then,

$$k_0(\epsilon) \leq \left\lfloor \frac{\ln|x_1 - x_0| - \ln(\epsilon(1-L))}{\ln(1/L)} \right\rfloor + 1 \tag{7}$$

Proof. Let $E_k = |x_k - \xi|$ be the error at k. Then

$$\begin{aligned} |x_{k+1} - \xi| &= |g(x_k) - g(\xi)| && \text{(definition of } g \text{ and } \xi \text{ a fixed point)} \\ &< L|x_k - \xi| && (g \text{ a contraction}) \end{aligned}$$

Therefore by induction

$$E_k \leq L^k E_0 \tag{8}$$

Since $L < 1$, $L^k \to 0$ as $k \to \infty$, so that $\lim_{k\to\infty} |x_k - \xi| = 0$. □

Theorem 6 (Contraction Mapping Theorem when Differentiable)**.** Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Let $\xi = g(\xi) \in [a, b]$ be a fixed point of g (the existence of this point is guaranteed by Brouwer's fixed point theorem). Assume g has a continuous derivative in some neighborhood of ξ with $|g'(\xi)| < 1$. Then the sequence

(x_k) defined by simple iteration $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ as $k \to \infty$, provided that x_0 is close to ξ.

Definition 3 (Stable, Unstable Fixed Point)**.** Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$, and let ξ denote a fixed point of g. ξ is a stable fixed point of g if the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ whenever the starting value x_0 is sufficiently close to ξ. Conversely, if no sequence (x_k) defined by this iteration converges to ξ for any starting value x_0 close to ξ, except for $x_0 = \xi$, then we say that ξ is an unstable fixed point of g.

- With this definition, a fixed point may be neither stable nor unstable.
- If $|g'(\xi)| < 1$, then ξ is a stable fixed point (provided g is continuous, differentiable etc.)

Theorem 7 (Unstable Fixed Points)**.** Suppose that $\xi = g(\xi)$, where the function g has a continuous derivative in some neighborhood of ξ, and let $|g'(\xi)| > 1$ (thus ξ is an unstable fixed point). Then the sequence (x_k) defined by simple iteration $x_{k+1} = g(x_k)$, $k \geq 0$, does not converge to ξ from any starting value x_0, $x_0 \neq \xi$.

Definition 4 (Rate of Convergence)**.** Suppose $\xi = \lim_{k\to\infty} x_k$. Define $E_k = |x_k - \xi|$.

- The sequence (x_k) converges to ξ linearly if there exists a number $\mu \in (0, 1)$ such that
$$\lim_{k\to\infty} \frac{E_{k+1}}{E_k} = \mu \tag{9}$$
- The sequence (x_k) converges to ξ superlineraly if $\mu = 0$. That is, the sequence of μ_k generated at each step $\to 0$ as $k \to \infty$.
- The sequence (x_k) converges to ξ with order q if there exists a $\mu > 0$ such that
$$\lim_{k\to\infty} \frac{E_{k+1}}{E_k^q} = \mu \tag{10}$$
In particular, if $q = 2$, then the sequence converges quadratically.

2.2 Newton's Method

Definition 5 (Newton's Method)**.** Newton's method for the solution of $f(x) = 0$ is defined by
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \tag{11}$$

Geometrically, $(x_{n+1}, 0)$ is the intersection of the x-axis and the tangent of the graph of f at $(x_n, f(x_n))$.

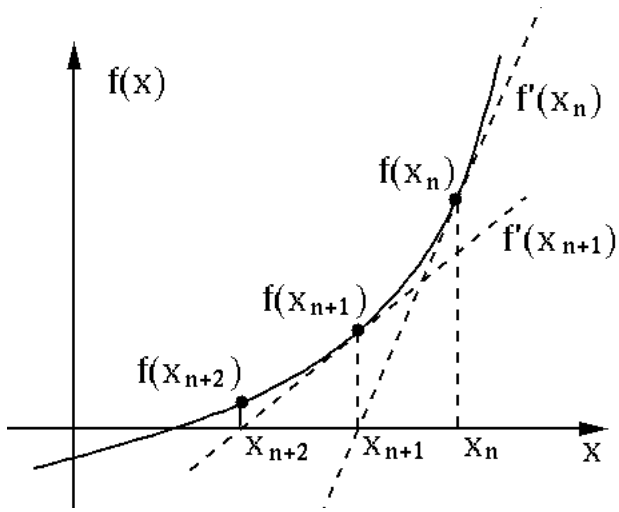

Figure 1: Geometric Interpretation of Newton's Method in $\mathbb{R}$

Intuitively, the fixed points of this iteration g will be stable.

We can show that $|g'(\xi)| < 1$.

$$
\begin{aligned}
g'(x) &= 1 - \frac{f' \cdot f' - f \cdot f''}{(f')^2} \\
&= 1 - \left(1 - \frac{f(x) \cdot f''(x)}{(f'(x))^2}\right) \\
&= \frac{f(x) \cdot f''(x)}{(f'(x))^2}
\end{aligned}
$$

Therefore

$$
|g'(\xi)| = \left|\frac{f(\xi) \cdot f''(\xi)}{(f'(\xi))^2}\right| = 0 < 1 \tag{12}
$$

Theorem 8 (Convergence of Newton's Method)**.** Suppose that f is a continuous real-valued function with continuous second derivative f'' defined on the closed interval $I_\delta = [\xi - \delta, \xi + \delta]$, $\delta > 0$, such that $f(\xi) = 0$ and $f''(\xi) \neq 0$. Additionally suppose that there exists a positive constant A such that

$$
\frac{|f''(x)|}{|f'(y)|} \leq A \quad \forall x, y, \in I_\delta \tag{13}
$$

If initially

$$
|\xi - x_0| \leq h = \min(\delta, \frac{1}{A}) \tag{14}
$$

then the sequence (x_k) defined by Newton's method converges quadratically to ξ.

Proof. We first compute the Taylor expansion of $f(\xi)$, expanding about the point $x_k \in I_\delta$, where $|\xi - x_k| \leq h = \min(\delta, \frac{1}{A})$. Thus

$$f(\xi) = f(x_k) + (\xi - x_k)f'(x_k) + \frac{(\xi - x_k)^2}{2} f''(\eta_k) \tag{15}$$

where η_k is between ξ and x_k. Recall that $f(\xi) = 0$. We can use this fact and the definition of Newton's iteration to rearrange the above expansion as

$$\xi - x_{k+1} = -\frac{(\xi - x_k)^2 f''(\eta_k)}{2f'(x_k)} \tag{16}$$

A small modification to this equation allows us to derive a relationship between adjacent errors

$$E_{k+1} = \frac{f''(\eta_k)}{2f'(x_k)} E_k^2 \tag{17}$$

Recall by assumption we have that $|\xi - x_k| \leq h = \min(\delta, \frac{1}{A})$ and $\frac{|f''(x)|}{|f'(y)|} \leq A \quad \forall x, y, \in I_\delta$. Therefore,

$$|E_{k+1}| = \frac{1}{2} \left| \frac{f''(\eta_k)}{f'(x_k)} \right| |E_k|^2 \leq \frac{1}{2} |E_k| \tag{18}$$

We are given that $|\xi - x_0| \leq h = \min(\delta, \frac{1}{A})$, so that induction gives that

$$|E_k| = |\xi - x_k| \leq \frac{1}{2^k} h \tag{19}$$

Therefore (x_k) converges to ξ as $k \to \infty$.

To show convergence is quadratic, notice that

$$\begin{aligned} \lim_{k\to\infty} \frac{|E_{k+1}|}{|E_k|} &= \lim_{k\to\infty} \frac{1}{2} \frac{|f''(\eta_k)|}{|f'(x_k)|} \\ &= \frac{1}{2} \frac{|f''(\xi)|}{|f'(\xi)|} = \mu \leq \frac{A}{2}. \end{aligned}$$

This shows that convergence is quadratic. □

2.3 Secant Method

Observe that Newton's method requires us to know the first derivative f' of f. In applications, we might not know f' or it could be expensive to calculate. This motivates approximating the $f'(x_k)$ in Newton's method with

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \tag{20}$$

Definition 6 (Secant Method)**.** The secant method is defined by

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{21}$$

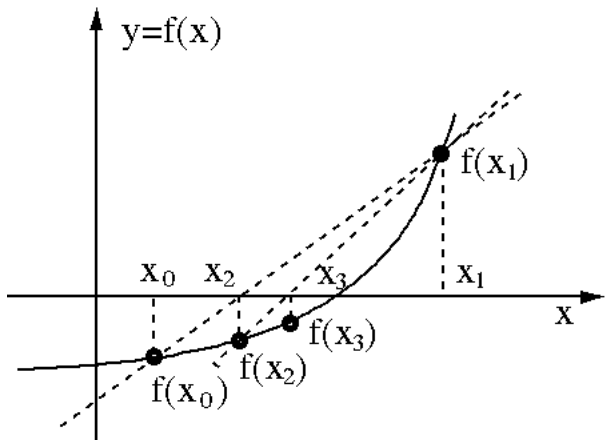

Figure 2: Geometric Interpretation of Secant Method in $\mathbb{R}$

Theorem 9 (Convergence of Secant Method)**.** Suppose that f is a real-valued function, defined and continuously differentiable on an interval $I = [\xi - h, \xi + h]$, $h > 0$, with center point ξ. Suppose further that $f(\xi) = 0$, $f'(\xi) \neq 0$. Then, the sequence (x_k) defined by the secant method converges at least linearly to ξ provided that x_0 and x_1 are sufficiently close to ξ.

Proof. Without loss of generality, assume that $\alpha = f'(\xi) > 0$ in a small neighborhood of ξ. We'll choose this neighborhood such that

$$0 < \frac{3}{4}\alpha < f'(x) < \frac{5}{4}\alpha \tag{22}$$

for all x in the interval.

Recall that the secant method is defined by

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{23}$$

We can repeatedly use the mean value theorem to approximate each of these terms. First observe that

$$\frac{f(x_k) - f(\xi)}{x_k - \xi} = f'(\eta_k) \tag{24}$$

for some η_k between x_k and ξ. Since $f(\xi) = 0$, this equation implies that

$$f'(x_k) = f'(\eta_k)(x_k - \xi) \tag{25}$$

Next observe that
$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} = f'(\theta_k) \tag{26}$$
for some θ_k between x_k and x_{k-1}. Therefore, we can put these pieces together to observe that,
$$_{k+1} = x_k - \frac{f'(\eta_k)(x_k - \xi)}{f'(\theta_k)} \tag{27}$$
To show convergence, we can compare successive error terms.

$$\begin{aligned} E_{k+1} &= x_{k+1} - \xi \\ &= E_k - \frac{f'(\eta_k)}{f'(\theta_k)} E_k \\ &= \left(1 - \frac{f'(\eta_k)}{f'(\theta_k)}\right) E_k \end{aligned}$$

Therefore

$$\begin{aligned} \frac{E_{k+1}}{E_k} &= 1 - \frac{f'(\eta_k)}{f'(\theta_k)} \\ &< 1 - \frac{5\alpha/4}{3\alpha/4} \\ &= \frac{2}{3} \\ &< 1 \end{aligned}$$

Therefore the secant method converges at least linearly. □

3 Solution of systems of linear equations

3.1 LU Decomposition

3.2 Least Squares

Given a system of equations $Ax = b$, the least squares problem is

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \tag{28}$$

We can expand the objective function out as

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T A x - 2b^T A x + b^T b \end{aligned}$$

To find the x that minimizes this expression we find the x that satisfies $\nabla_x F = 0$. That is

$$\nabla_x F = 0 = 2A^T Ax - 2A^T b \tag{29}$$

Therefore the minimizer is $x = (A^T A)^{-1} A^T b$. $(A^T A)^{-1} A^T$ is called the pseudo-inverse of A. If A is square and invertible, then the pseudo-inverse equals A^{-1}.

3.3 Gram-Schimdt Orthogonalization

Algorithm: Denote the columns of A by a_i.

1. $q_1 = a_1$. Then normalized by $q_1 = \frac{q_1}{\|q_1\|}$.
2. $q_2 = a_2 - \langle q_1, a_2 \rangle q_1$. Then normalize by $q_2 = \frac{q_2}{\|q_2\|}$. It's simple to verify that $q_2 \perp q_1$.
3. For an arbitrary k, $q_k = a_k - \langle a_k, q_1 \rangle q_1 - \langle a_k, q_2 \rangle q_2 - \ldots - \langle a_k, q_{k-1} \rangle q_{k-1}$. Then normalize by $q_k = \frac{q_k}{\|q_k\|}$.

We can observe the following properties:

1. $\|q_i\| = 1$ (this follows directly)
2. $q_i \perp q_j$ for all $i \neq j$
3. $q_k \in span(a_1, \ldots, a_k)$ and $a_k \in span(q_1, \ldots, q_k)$ so that $span(a_1, \ldots, a_k) = span(q_1, \ldots, q_k)$.

[[Write proof for 2]].

3.4 QR Factorization

Definition 7. (Unitary Matrix) A matrix $Q = [q_1 \ldots q_n] \in \mathbb{R}^{m \times n}$ is unitary if and only if $\langle q_i, q_j \rangle = \delta_{ij}$.

Observations about this definition:

1. $Q^T Q = I$
2. If Q is square, then $Q^T = Q^{-1}$.

To calculate the QR decomposition, we can find Q by using the Gram-Schmidt process. Then R can be found as

$$R = \begin{pmatrix} \langle e_1, a_1 \rangle & \langle e_1, a_2 \rangle & \langle e_1, a_3 \rangle & \ldots \\ 0 & \langle e_2, a_2 \rangle & \langle e_2, a_3 \rangle & \ldots \\ 0 & 0 & \langle e_3, a_3 \rangle & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{30}$$

3.4.1 Application to Least Squares

Suppose that we can write $A = QR$, where $A \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{m\times n}$ and unitary, and $R \in \mathbb{R}^{n\times n}$ and upper triangular. Then the least squares solution to $Ax = b$ is given by

$$\begin{aligned}
x &= (A^TA)^{-1}A^Tb \\
&= (R^TQ^TQR)^{-1}R^TQ^Tb \\
&= (R^TR)^{-1}R^TQ^Tb \\
\implies (R^TR)x &= R^TQ^Tb \\
Rx &= Q^Tb \qquad \text{(assume } R \text{ is invertible (i.e. no zeros on the diagonal))}
\end{aligned}$$

We can then solve for x using back substitution, which is $\mathcal{O}(n^2)$.

3.5 Norms and Condition Numbers

Definition 8. (Norm) Suppose that $\mathcal{V}$ is a linear space over the field $\mathbb{R}$. The *nonnegative* real-valued function $\|\cdot\|$ is a norm on $\mathcal{V}$ if the following axioms are satisfied: Fix $v \in \mathcal{V}$

1. Positivity: $\|v\| = 0$ if and only if $v = 0$
2. Scale Preservation: $\|\alpha v\| = |\alpha|\|v\|$ for all $\alpha \in \mathbb{R}$
3. Triangle Inequality: $\|v + w\| \leq \|v\| + \|w\|$.

Example 1 (Examples of Norms). 1. 1-norm:

$$\|v\|_1 = \sum_{i=1}^{n} |v_i| = |v_1| + \cdots + |v_n| \tag{31}$$

2. 2-norm:

$$\|v\|_2 = \left(\sum_{i=1}^{n} v_i^2\right)^{\frac{1}{2}} = \sqrt{v_1^2 + \cdots + v_n^2} = \sqrt{v^Tv} \tag{32}$$

3. ∞-norm

$$\|x\|_\infty = \max_{i=1,\ldots,n} |v_i| \tag{33}$$

4. p-norm

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}} \tag{34}$$

For the p-norm, proving the triangle inequality follows from the Minkowski's inequality.

Definition 9 (Operator Norm)**.** Let A be an $m \times n$ matrix. That is, A is a linear transformation form $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the operator norm (or subordinate matrix norm) of A is

$$\|A\|_{p,q} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_q}{\|x\|_p}. \tag{35}$$

Observations about this definition:

1. It's easy to check that this definition of the operator norm satisfies the properties of a norm given in Definition 8. For the triangle inequality, observe that

$$\begin{aligned} &\|(A+B)x\|_p \leq \|Ax\|_p + \|Bx\|_p && \text{(from Minkowski's inequality)} \\ \implies &\frac{\|(A+B)x\|_p}{\|x\|_p} \leq \frac{\|Ax\|_p}{\|x\|_p} + \frac{\|Bx\|_p}{\|x\|_p} \end{aligned}$$

   Taking the supremum of both sides over x shows that $\|A+B\|_p \leq \|A\|_p + \|B\|_p$.

2. The definition immediately implies that for an arbitrary $x \in \mathbb{R}^n$, $x \neq 0$,

$$\|Ax\|_q \leq \|A\|_{p,q} \|x\|_p \tag{36}$$

   We can generalize this inequality to claim that

$$\|AB\| \leq \|A\| \|B\| \tag{37}$$

   for conformable matrices A, B. Indeed, fix $0 \neq x \in R^n$. Then

$$\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\| \tag{38}$$

   We can divide all inequalities by $\|x\|$ to see that for all $x \neq 0$,

$$\frac{\|ABx\|}{\|x\|} \leq \|A\| \|B\| \tag{39}$$

   Taking the supremum over x on the left hand side shows that $\|AB\| \leq \|A\| \|B\|$.

Theorem 10 (The 1-norm of a matrix is the largest absolute-value column sum)**.** Let $A \in \mathbb{R}^{m \times n}$ and denote the columns of A by a_j, $j = 1, \ldots, n$. Then $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}| = \max_{j=1,\ldots,n} \|a_j\|$.

Proof. Fix $x \in \mathbb{R}^n$. Let $C = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}|$. First consider the product $A \cdot x$. The ith element is $\sum_{j=1}^{n} a_{ij}x_j$. Then

$$\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^{m} |(Ax)_i| = \sum_{i=1}^{m} |\sum_{j=1}^{n} a_{ij}x_j| \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}||x_j| && \text{(triangle inequality)} \\
&= \sum_{j=1}^{n} |x_j| \left(\sum_{i=1}^{m} |a_{ij}| \right) && \text{(interchange order of summation, assumed finite)} \\
&\leq C\|x\|_1
\end{aligned}$$

Therefore $\frac{\|Ax\|_1}{\|x\|_1} \leq C$ for all x. Next, we find an x such we achieve equality with C. Call index J the index such that $\|a_J\|_1 = C = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$. Then let e_J be the n-vector of zeros with a 1 in the Jth entry. Clearly $\|e_J\|_1 = 1$. But then

$$\|Ae_J\|_1 = \|a_J\|_1 = C \tag{40}$$

In sum, we first showed that for all $x \in \mathbb{R}^n$

$$\frac{\|Ax\|_1}{\|x\|_1} \leq C \tag{41}$$

We then found an $x \in \mathbb{R}^n$ such that $\frac{\|Ax\|_1}{\|x\|_1} = C$. Therefore

$$\|A\|_1 = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = C = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| = \max_{j=1,\ldots,n} \|a_j\| \tag{42}$$

□

Theorem 11 (The ∞-norm of a matrix is the largest absolute-value row sum)**.** Let $A \in \mathbb{R}^{m\times n}$ and denote the rows of A by b_i, $i = 1, \ldots, m$. Then $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1,\ldots,m} \|b_i\|$.

Proof. Fix $x \in \mathbb{R}^n$. Let $C = \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}|$.

$$\begin{aligned}
\|Ax\|_\infty &= \max_{i=1,\ldots,m} |\sum_{j=1}^n a_{ij}x_j| \\
&\leq \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}||x_j| && \text{(by the triangle inequality)} \\
&\leq \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}|\|x\|_\infty && \text{(since } |x_j| \leq \|x\|_\infty \text{ for all } j\text{)} \\
&= C\|x\|_\infty
\end{aligned}$$

Next, we find an x such we achieve equality with C. Call I the index for which $\|b_I\|_\infty = C$. Define

$$x_j = \begin{cases} 1 & a_{Ij} > 0 \\ -1 & a_{Ij} < 0 \end{cases} \tag{43}$$

Observe that $\|x\|_\infty = 1$. Then

$$\begin{aligned}
|A \cdot x|_I &= |b_I^T \cdot x| \\
&= |\sum_{j=1}^m a_{Ij}x_j| \\
&= |\sum_{j=1}^m a_{Ij}| \\
&= C
\end{aligned}$$

We then found an $x \in \mathbb{R}^n$ such that $\frac{\|Ax\|_\infty}{\|x\|_\infty} = C$. Therefore

$$\|A\|_\infty = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = C = \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}| = \max_{i=1,\ldots,m} \|b_i\| \tag{44}$$

□

Theorem 12 (The 2-norm of a symmetric positive definite matrix is the maximum absolute value of its eigenvalues)**.** Let A be a positive definite $n \times n$ matrix. Then

$$\|A\|_2 = \max_{i=1,\ldots,n} |\lambda_i| \tag{45}$$

Proof. Since A is positive definite, A has n distinct eigenvalues, which implies that it has n linearly independent eigenvectors. Therefore, for an arbitrary $x \in \mathbb{R}^n$, we can write x

as a linearly combination of the eigenvectors $x_1, \ldots, x_n$. Then

$$\begin{aligned} x &= c_1 x_1 + \cdots + c_n x_n \\ Ax &= c_1 A x_1 + \cdots + c_n A x_n \\ &= c_1 \lambda_1 x_2 + \cdots + c_n \lambda_n x_n \end{aligned}$$

We can normalize the eigenvectors of A so that $x_i^T x_i = 1$. Then $\|Ax\|_2 = \sqrt{\sum_{i=1}^n c_i^2 \lambda_i^2}$ and $\|x\|_2 = \sqrt{\sum_{i=1}^n c_i^2}$. Therefore

$$\frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\frac{\sum_{i=1}^n c_i^2 \lambda_i^2}{\sum_{i=1}^n c_i^2}} \leq \max_i |\lambda_i| = |\lambda_I| \tag{46}$$

Now we'll find an x such that we actually achieve equality. Call I the index of the maximum absolute value of an eigenvalue. Then, consider the eigenvector associated with this eigenvalue, called x_i. Then

$$\frac{\|Ax_I\|_2}{\|x_I\|_2} = \frac{|\lambda_I| \|x_I\|}{\|x_I\|} = |\lambda_I| \tag{47}$$

This shows that $\|A\|_2 = \max_i |\lambda_i|$. □

Theorem 13 (The 2-norm of a matrix $A_{m\times n}$ equals its largest singular value)**.** Let A be an $m \times n$ matrix and denote the eigenvalues of the matrix $B = A^T A$ by λ_i, $i = 1, \ldots, n$. Then

$$\|A\|_2 = \max_i \sqrt{\lambda_i} \tag{48}$$

The square roots of the (nonnegative) eigenvalues of $A^T A$ are referred to as the singular values of A.

3.5.1 Conditioning

Conditioning helps us quantify the sensitivity of the output to perturbations of the input. In what follows, let f be a mapping from a subset D of a normed linear space $\mathcal{V}$ to another normed linear space $\mathcal{W}$.

Definition 10 (Absolute Condition Number)**.**

$$Cond(f) = \sup_{x,y \in D, x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|} \tag{49}$$

Definition 11 (Absolute Local Condition Number)**.**

$$Cond_x(f) = \sup_{x+\delta x \in D, \delta x \neq 0} \frac{\|f(x+\delta x) - f(x)\|}{\|\delta x\|} \tag{50}$$

The previous two definitions depend on the magnitudes of $f(x)$ and x. In applications, it's often better to rescale as follows

Definition 12 (Relative Local Condition Number)**.**

$$cond_x(f) = \sup_{x+\delta x \in D, \delta x \neq 0} \frac{\|f(x+\delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \tag{51}$$

In these definitions, if f is differentiable then we can replace the differences with the appropriate derivatives.

Example 2 (Example of condition numbers)**.** Let D be a subinterval of $[0, \infty)$ and $f(x) = \sqrt{x}$. Then $f'(x) = \frac{1}{2\sqrt{x}}$.

1. If $D = [1, 2]$, then $Cond(f) = \frac{1}{2}$.
2. If $D = [0, 1]$, then $Cond(f) = \infty$.
3. If $D = (0, \infty)$, then the the absolute local condition number of f at $x \in D$ is

   $$Cond_x(f) = \frac{1}{2\sqrt{x}} \tag{52}$$

   Thus as $x \to$, $Cond_x(f) \to \infty$, and as $x \to \infty$, $Cond_x(f) \to 0$.
4. If $D = (0, \infty)$, then the relative local condition number of f is $cond_x(f) = 1/2$ for all $x \in D$.

Definition 13 (Condition Number of a Nonsingular Matrix)**.** The condition number of a nonsingular matrix A is defined by

$$\kappa(A) = \|A\| \|A^{-1}\| \tag{53}$$

If $\kappa(A) \gg 1$, the matrix is said to be ill-conditioned.

Observations about this definition:

1. $\kappa(A) = \kappa(A^{-1})$
2. For all A, $\kappa(A) \geq 1$. This follows because

   $$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\| \tag{54}$$

3. The condition number of a matrix is unaffected by scaling all its elements by multiplying by a nonzero constant.
4. There is a condition number for each norm, and the size of the condition number is strongly dependent on the choice of norm.

4 Special Matrices

4.1 Symmetric Positive Definite Matrices

Definition 14 (Symmetric, Positive Definite, spd)**.** The real matrix A is said to be symmetric if $A = A^T$. A square $n \times n$ matrix is called positive definite if

$$\boldsymbol{x}^T A \boldsymbol{x} > 0 \tag{55}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$.

Theorem 14 (Properties of spd matrices)**.** Let A be an $n \times n$ real, spd matrix. Then

1. $a_{ii} > 0$ for all $i = 1, \ldots, n$ (the diagonal elements of A are positive).
2. $Ax_i = \lambda_i x_i \implies \lambda_i \in \mathbb{R}_{>0}, \boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$ (the eigenvalues of A are real and positive, and the eigenvectors of A belong to $\mathbb{R}^n \setminus \{0\}$).
3. $x_i \perp x_j$ if $\lambda_i \neq \lambda_j$ (the eigenvectors of distinct eigenvalues of A are orthogonal)
4. $\det(A) > 0$ (the determinant of A is positive)
5. Every submatrix B of A obtained by deleting any set of rows and the corresponding set of columns from A is symmetric and positive definite (in particular, every principal submatrix is positive definite).

Proof. We prove each claim in the theorem as follows

1. Let e_i be the ith canonical basis vector in $\mathbb{R}^n$. Then

$$a_{ii} = e_i^T A e_i > 0 \tag{56}$$

   since A is pd. A few observations: this only relies on A being pd. $e_i^T A$ picks out the ith row of A. Ae_i picks out the ith column of A.

2. We'll first show that the eigenvalues of A are real. Suppose λ, $\boldsymbol{x}$ are an eigenvalue/vector pair of A. Thus $A\boldsymbol{x} = \lambda \boldsymbol{x}$. We can conjugate this equation to find that $\bar{A}\bar{\boldsymbol{x}} = A\bar{\boldsymbol{x}} = \bar{\lambda}\bar{x}$ (thus complex eigenvalues of real valued matrices come in conjugate pairs). Then

$$\boldsymbol{x}^T A \bar{\boldsymbol{x}} = \bar{\lambda} \boldsymbol{x}^T \bar{\boldsymbol{x}}$$
$$\boldsymbol{x}^T A^T \bar{\boldsymbol{x}} = (A\boldsymbol{x})^T \bar{\boldsymbol{x}} = \lambda \boldsymbol{x}^T \bar{\boldsymbol{x}}$$

   Since $A = A^T$, we know that $\lambda \boldsymbol{x}^T \bar{\boldsymbol{x}} = \lambda \boldsymbol{x}^T \bar{\boldsymbol{x}}$. As long as $\boldsymbol{x} \neq 0$, then $\boldsymbol{x}^T \bar{\boldsymbol{x}} \neq 0$. Therefore $\bar{\lambda} = \lambda$, which shows $\lambda \in \mathbb{R}$.

   The fact that the eigenvector associated with λ has real elements follows by noting that all elements of the singular matrix $A - \lambda I$ are real numbers. Therefore, the columns of $A - \lambda I$ are linearly dependent in $\mathbb{R}^n$. Hence there exists an $\boldsymbol{x} \in \mathbb{R}^n$ such that $(A - \lambda I)\boldsymbol{x} = 0$.

This proof only requires that A is symmetric – therefore any real, symmetric matrix has real eigenvalues/vectors.

Next we'll show the eigenvalues of A are positive. Suppose λ, $\boldsymbol{x}$ are an eigenvalue/vector pair of A. Then

$$0 < \boldsymbol{x}^T A \boldsymbol{x} = \lambda \boldsymbol{x}^T \boldsymbol{x} \tag{57}$$

Since $\boldsymbol{x} \neq 0$ and $\boldsymbol{x}^T\boldsymbol{x}$ is positive (it's actually the squared 2-norm of $\boldsymbol{x}$), then $\lambda > 0$. Note that this part of the proves requires A be pd.

3. Let λ_i, λ_j be distinct eigenvalues of A, and $\boldsymbol{x}_i, \boldsymbol{x}_j$ the corresponding eigenvectors. Then

$$\begin{aligned}\boldsymbol{x}_i^T A \boldsymbol{x}_j &= \lambda_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ \boldsymbol{x}_i^T A^T \boldsymbol{x}_j &= (A\boldsymbol{x}_i)^T x_j = \lambda_i \boldsymbol{x}_i^T \boldsymbol{x}_j\end{aligned}$$

Since A is symmetric, these two string of equalities must be equal. We can subtract them to find that

$$(\lambda_i - \lambda_j)\boldsymbol{x}_i^T \boldsymbol{x}_j = 0 \tag{58}$$

Since we assumed $\lambda_i \neq \lambda_j$, then it must be that $\boldsymbol{x}_i^T \boldsymbol{x}_j$. Therefore $x_i \perp x_j$. This proof again only relies on the symmetry of A. which is the product of the diagonal elements of the matrix (the eigenvalues).

4. This follows from the fact that the determinant of A is equal to the product of its eigenvalues. Or, we can write A in terms of its eigenvalue decomposition. Thus

$$A = X \Lambda X^{-1} \tag{59}$$

Therefore

$$\det(A) = \det(X) \det(\Lambda) \det(X)^{-1} = \det(\Lambda) \tag{60}$$

Or, we can write A in terms of its eigenvalue decomposition. Thus

$$A = X \Lambda X^{-1} \tag{61}$$

Therefore

$$\det(A) = \det(X) \det(\Lambda) \det(X)^{-1} = \det(\Lambda) \tag{62}$$

5. Let $I \subset \{1, 2, \ldots, n\}$ be a subset of indices and let $B = A_{II}$. A is symmetric, so that $A_{II} = A_{II}^T$. Therefore B is symmetric. Let $x \in \mathbb{R}^n$ and define a vector y that is 0 for the indices not included in I and follows the value of x for the indices included in I. Therefore, $x^T B x = y^T A y > 0$ since A is pd.

$\square$

4.2 Cholesky Factorization

Theorem 15 (Cholesky). If A is spd, then there exists a lower diagonal matrix L such that $A = LL^T$. This is called the Cholesky decomposition.

Algorithm 1 Cholesky Factorization

Require: $A \in \mathbb{R}^{n \times n}$, SPD
  $L_1 \leftarrow \sqrt{a_{11}}$
  for $k \leftarrow 2, 3, \ldots, n$ **do**
    Solve $L_{k-1} l_k = a_k$ for l_k
    $l_{kk} \leftarrow \sqrt{a_{kk} - l_k^T l_k}$
    $L_k \leftarrow \begin{pmatrix} L_{k-1} & 0 \\ l_k^T & l_{kk} \end{pmatrix}$
  end for

Notation:

- L_{k-1}: the first $k-1 \times k-1$ upper left corner of L
- a_k: the first $k-1$ entries in column k of A
- l_k: the first $k-1$ entries in column k of L^T [[?]]
- a_{kk}, l_{kk}: the kk entries of A and L, respectively

4.3 Banded Matrices and Differential Equations

Consider the two-point boundary value problem

$$u'' + 2u' = -1, \quad u(x=0) = 0, u(x=1) = 0 \tag{63}$$

where $x \in [0, 1]$.

Define a sequence of grid points $\{x_i\}_{i=0}^{N+1}$. We can approximate the derivative of u at each point on the grid as follows

$$\begin{aligned} u'(x_j) &= \lim_{\delta \to 0} \frac{u(x_j + \delta) - u(x_j - \delta)}{2\delta} \\ &\approx \frac{u(x_{j-1}) - u(x_{j+1})}{2\Delta x} \end{aligned}$$

where we use the centered difference quotient of order 2. Similarly, we can approximate the second derivative of u each each point in the domain as

$$\begin{aligned} u''(x_j) &= \lim_{\delta \to 0} \frac{u'(x_j+\delta) - u'(x_j-\delta)}{2\delta} \\ &\approx \frac{u'(x_{j+1}) - u'(x_{j-1})}{2\Delta x} \\ &= \frac{\frac{u(x_{j+2})-u(x_j)}{2\Delta x} - \frac{u(x_j)-u(x_{j-2})}{2\Delta x}}{2\Delta x} \\ &= \frac{u(x_{j+2}) - 2u(x_j) + u(x_{j-2})}{4\Delta x} \end{aligned}$$

Let's instead use the grid points adjacent to x_j:

$$u''(x_j) \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{\Delta x} \tag{64}$$

Then, going back to the initial differential equation, for x_j, we have

$$\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{\Delta x} + 2\frac{u(x_{j-1}) - u(x_{j+1})}{2\Delta x} = -1 \tag{65}$$

Let

$$U = \begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_N) \end{bmatrix} \tag{66}$$

and let $u_i = u(x_i)$. In this notation, the differential equation at x_j can be written as

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} + \frac{u_{j-1} - u_{j+1}}{\Delta x} = -1 \tag{67}$$

We can put these equations together into a matrix. Each row will only have 3 non-zero entries at $j-1$, j, and $j+1$. Thus the jth row is

$$\begin{pmatrix} 0 & 0 & \dots & \frac{1}{\Delta x^2} + \frac{1}{\Delta x} & \frac{-2}{\Delta x^2} & \frac{1}{\Delta x^2} - \frac{1}{\Delta x} & 0 & 0 & \dots \end{pmatrix} \tag{68}$$

Thus stacking these rows together will give a tridiagonal matrix. Call this matrix A. Then we have that

$$AU = -\mathbf{1} \tag{69}$$

5 Simultaneous nonlinear equations

5.1 Analysis Preliminaries

Definition 15 (Cauchy Sequence)**.** A sequence $(\boldsymbol{x}^{(k)}) \subset \mathbb{R}^n$ is called a Cauchy sequence in $\mathbb{R}^n$ if for any $\epsilon > 0$ there exists a positive integer $k_0 = k_0(\epsilon)$ such that

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(m)}\|_\infty < \epsilon \quad \forall k, m \geq k_0(\epsilon) \tag{70}$$

Remark 1. $\mathbb{R}^n$ is **complete** in the sense that every Cauchy sequence $(\boldsymbol{x}^{(k)})$ converges to some $\xi \in \mathbb{R}^n$.

Definition 16 (Continuous function)**.** Let $D \subset \mathbb{R}^n$ be nonempty and $f : D \to \mathbb{R}^n$. Given $\boldsymbol{\xi} \in D$, f is continuous at $\boldsymbol{\xi}$ if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that for every $\boldsymbol{x} \in B(\boldsymbol{\xi}; \delta) \cap D$

$$\|f(\boldsymbol{x}) - f(\boldsymbol{\xi})\|_\infty < \epsilon \tag{71}$$

Lemma 1. Let $D \subset \mathbb{R}^n$ be nonempty and $f : D \to \mathbb{R}^n$ be defined and continuous on D. If $(\boldsymbol{x}^{(k)}) \subset D$ converges in $\mathbb{R}^n$ to $\boldsymbol{\xi} \in D$, then $f(\boldsymbol{x}^{(k)})$ also converges to $f(\boldsymbol{\xi})$.

We want to find a vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $f(\boldsymbol{x}) = 0$.

Example 3. Consider the linear system

$$Ax = b \tag{72}$$

Then $A : \mathbb{R}^n \to \mathbb{R}^m$. Let $f(\boldsymbol{x}) = A\boldsymbol{x} - b$.

Example 4. Let

$$f = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ 5x_1^2 + 21x_2^2 - 9 \end{bmatrix} \tag{73}$$

Note that $x_1^2 + x_2^2 = 1$ is the 0 level set of f, and is the unit circle. $5x_1^2 + 21x_2^2 = 9$ is the 0 level set of f and is an ellipse.

This function has four zeros

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \pm\sqrt{3}/2 \\ \pm\frac{1}{2} \end{bmatrix} \tag{74}$$

5.2 Simultaneous iteration

Example 5.

Definition 17 (Lipschitz condition, constant, and contraction)**.** Let D be a closed subset of $\mathbb{R}^n$ and $g : D \to D$. If there exists a positive constant L such that

$$\|g(x) - g(y)\|_\infty \leq L\|x - y\|_\infty \tag{75}$$

for all $x, y \in D$, then g satisfies the Lipschitz condition on D in the ∞-norm. L is called the Lipschitz constant. If $L \in (0,1)$, then g is called a contraction on D in the ∞-norm.

Observations about this definition:

- Any function g that satisfies the Lipschitz condition on D is continuous on D (to see this, set $\delta = \frac{\epsilon}{L}$).
- If g satisfies the Lipschitz condition on D in the ∞-norm, then it also does in the p-norm for $p \in [1, \infty)$ and vice-versa. However the size of L depends on the choice of norm.

Theorem 16 (Contraction Mapping Theorem in $\mathbb{R}^n$)**.** Suppose D is a closed subset of $\mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ is defined on D, and $g(D) \subset D$. Suppose further that g is a contraction on D in the ∞-norm. Then,

1. g has a unique fixed point $\boldsymbol{\xi} \in D$
2. The sequence $(\boldsymbol{x}^{(k)})$ defined by $\boldsymbol{x}^{(k+1)} = g(\boldsymbol{x}^k)$ converges to $\boldsymbol{\xi}$ for any starting value $x^{(0)} \in D$.

Proof. The proof has three parts:

1. First prove uniqueness, assuming existence of a fixed point.
2. Prove the iteration generates a Cauchy sequence (then convergence to some $\boldsymbol{\xi}$ follows from the completeness of the space).
3. Show $\boldsymbol{\xi}$ is indeed the fixed point.

Uniqueness: Suppose $\boldsymbol{\xi}, \boldsymbol{\eta}$ are both fixed points of g in D. Then,

$$\begin{aligned} \|\boldsymbol{\xi} - \boldsymbol{\eta}\|_\infty &= \|g(\boldsymbol{\xi}) - g(\boldsymbol{\eta})\| && (\boldsymbol{\xi}, \boldsymbol{\eta} \text{ are fixed points}) \\ &\leq L\|\boldsymbol{\xi} - \boldsymbol{\eta}\|_\infty && (g \text{ is a contraction on } D) \end{aligned}$$

We can rearrange this to see that $(1 - L)\|\boldsymbol{\xi} - \boldsymbol{\eta}\|_\infty \leq 0$. By assumption, $L \in (0, 1)$, and the norm of a quantity is always weakly positive. Therefore, $\|\boldsymbol{\xi} - \boldsymbol{\eta}\|_\infty = 0$ which implies $\boldsymbol{\xi} = \boldsymbol{\eta}$.

Convergence: Assuming g has a fixed point $\boldsymbol{\xi} \in D$, the sequence $\boldsymbol{x}^{(k+1)} = g(\boldsymbol{x}^k)$ will converge to $\boldsymbol{\xi}$ for any $\boldsymbol{x}^{(0)} \in D$. This follows because

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{\xi}\|_\infty \leq L^k \frac{1}{1 - L} \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_\infty \tag{76}$$

Since $L \in (0, 1)$, $\lim_{k \to \infty} L^k = 0$, and therefore

$$\lim_{k \to \infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{\xi}\|_\infty = 0 \tag{77}$$

Existence: First observe that if $\boldsymbol{x}^{(0)}$ belongs to D, then $\boldsymbol{x}^{(k+1)} = g(\boldsymbol{x}^k) \in D$ for all $k \geq 1$ since $g(D) \subset D$ (this is important since the proof relies on g being a contraction on D). Next, consider the distance between two adjacent terms on the sequence $\boldsymbol{x}^{(k+1)} = g(\boldsymbol{x}^k)$

$$\begin{aligned}
\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|_\infty &= \|g(\boldsymbol{x}^{(k-1)}) - g(\boldsymbol{x}^{(k-2)})\|_\infty && \text{(definition of } g\text{)} \\
&\leq L\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k-2)}\|_\infty && (g \text{ is a contraction on } D) \\
&\leq L^{k-1}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_\infty && \text{(induction)}
\end{aligned}$$

Now, fix positive integers m, k such that $m > k$. Then

$$\begin{aligned}
\|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(k)}\|_\infty &= \|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)} + \boldsymbol{x}^{(m-1)} + \cdots + \boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\|_\infty \\
&\leq \|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m-1)}\|_\infty + \cdots + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_\infty && \text{(triangle inequality)} \\
&\leq (L^{m-1} + \cdots + L^k)\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_\infty && (g \text{ a contraction}) \\
&= L^k(L^{m-k-1} + \cdots + 1)\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_\infty \\
&\leq L^k \frac{1}{1-L}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_\infty && \text{(geometric series)}
\end{aligned}$$

Since $L \in (0, 1)$, $\lim_{k\to\infty} L^k = 0$. Therefore, $\boldsymbol{x}^{(k)}$ is a Cauchy sequence in $\mathbb{R}^n$, that is for all $\epsilon > 0$, there exists a k_0 such that

$$\|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(k)}\|_\infty < \epsilon \quad \forall m, k \geq k_0 \tag{78}$$

Any Cauchy sequence in $\mathbb{R}^n$ is convergent in $\mathbb{R}^n$. Thus, there exists some $\boldsymbol{\xi} \in \mathbb{R}^n$ such that $\boldsymbol{\xi} = \lim_{k\to\infty} \boldsymbol{x}^{(k)}$.

$\boldsymbol{\xi}$ is indeed the fixed point: Since g satisfies the Lipschitz condition on D, g is continuous on D. Therefore,

$$\boldsymbol{\xi} = \lim_{k\to\infty} \boldsymbol{x}^{(k+1)} = \lim_{k\to\infty} g(\boldsymbol{x}^{(k)}) = g\left(\lim_{k\to\infty} \boldsymbol{x}^{(k)}\right) = g(\boldsymbol{\xi}) \tag{79}$$

therefore $\boldsymbol{\xi}$ is a fixed point of g, and observe that $\boldsymbol{\xi} \in D$ since D is closed. □

Definition 18 (Jacobian)**.** Let $g = (g_1, \ldots, g_n)^T : \mathbb{R}^n \to \mathbb{R}^n$ be a function defined and continuous in an (open) neighborhood of $\boldsymbol{\xi} \in \mathbb{R}^n$. Suppose the first partial derivatives of each g_i exist at $\boldsymbol{\xi}$. The Jacobian matrix $J_g(\boldsymbol{\xi})$ of g at $\boldsymbol{\xi}$ is the $n \times n$ matrix with elements

$$J_g(\boldsymbol{\xi})_{ij} = \frac{\partial g_i}{\partial x_j}(\boldsymbol{\xi}) \tag{80}$$

Theorem 17 (Jacobian and Fixed Point Stability)**.** Let $g = (g_1, \ldots, g_n)^T : \mathbb{R}^n \to \mathbb{R}^n$ be a function defined and continuous on a closed set $D \subset \mathbb{R}^n$. Let $\boldsymbol{\xi} \in D$ be a fixed point of g. Suppose the first partial derivatives of each g_i are defined and continuous in some (open) neighborhood $N(\boldsymbol{\xi}) \in D$ of $\boldsymbol{\xi}$, with

$$\|J_g(\boldsymbol{\xi})\|_\infty < 1 \tag{81}$$

Then there exists $\epsilon > 0$ such that $g(\bar{B}_\epsilon(\boldsymbol{\xi})) \subset \bar{B}_\epsilon(\boldsymbol{\xi})$, and the sequence $\boldsymbol{x}^{(k+1)} = g(\boldsymbol{x}^k)$ converges to $\boldsymbol{\xi}$ for all $\boldsymbol{x}^{(0)} \in \bar{B}_\epsilon(\boldsymbol{\xi})$ (in other words, the sequence converges to $\boldsymbol{\xi}$ as long as $\boldsymbol{x}^{(0)}$ is close enough to $\boldsymbol{\xi}$).

Example 6.

Newton's Method

Definition 19 (Newton's Method)**.** The sequence defined by

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - [J_f(\boldsymbol{x}^{(k)})]^{-1} f(\boldsymbol{x}^{(k)}) \tag{82}$$

where $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, is called Newton's method.

Theorem 18. Suppose $f(\boldsymbol{\xi}) = 0$, that in some (open) neighborhood $N(\boldsymbol{\xi})$ of $\boldsymbol{\xi}$, where f is defined and continuous, all the second-order partial derivatives of f are defined and continuous, and that the Jacobian matrix $J_f(\boldsymbol{x}^{(k)})$ of f at the point $\boldsymbol{\xi}$ is nonsingular. Then the sequence defined by Newton's method converges to $\boldsymbol{\xi}$ provided that $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{\xi}$.

6 Eigenvalues of Eigenvectors of a symmetric matrix

Another matrix decomposition is

$$A = X\Lambda X^{-1} \tag{83}$$

where X is a matrix of the eigenvectors and Λ is a diagonal matrix with the eigenvalues.

6.1 Why we use iteration to calculate eigenvalues/eigenvectors

We call λ an eigenvalue and $x \neq 0$ an eigenvector of A if $Ax = \lambda x$. Thus, $(Ax - \lambda I) = 0$. Therefore, $x \in Null(A - \lambda I)$. Since $x \neq 0$, $A - \lambda I$ has a non-trivial nullspace, so we must have $\det(A - \lambda I) = 0$. This suggests a way to transform an eigenvalue finding problem to a root finding problem. Define

$$\rho(\lambda) = \det(A - \lambda I) \tag{84}$$

Recall that the determinant of a matrix is the product of its eigenvalues. If A is a $n \times n$ real, symmetric matrix, then $\rho(\lambda)$ is an n-th order polynomial in λ, whose roots are the eigenvalues of A.

Theorem 19 (Abel(-Ruffini) Theorem, or "No-go Theorem")**.** There is no algebraic solution (that is, a solution expressed in terms of radicals) to general polynomial equations of degree five or higher with arbitrary coefficients.

Therefore, there is no finite-number operation procedure that provides an eigenvalue decomposition.

6.2 Power Iteration

Find the biggest eigenvalue/vector.

Algorithm 2 Power Iteration

Require: $v^{(0)}$ = some vector with $\|v^{(0)}\| = 1$

```
for k ← 1, 2, ... do
    w ← A v^(k-1)                                        ▷ Apply A
    v^(k) ← w / ||w||                                    ▷ Normalize
    λ^(k) ← (v^(k))^T A v^(k) = ⟨v^(k), A v^(k)⟩         ▷ Rayleigh Quotient
end for
```

Theorem 20 (Convergence of Power Iteration)**.** Suppose $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$ and $q_1^T v^{(0)} \neq 0$. Then the iterates of power iteration satisfy

$$\|v^{(k)} - (\pm q_1)\| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \qquad \text{(error of eigenvector)}$$

$$|\lambda^{(k)} - \lambda_1| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \qquad \text{(error of eigenvalue)}$$

Proof. Convergence of eigenvector: Write $v^{(0)} = v$ as a linear combination of the orthonormal eigenvectors q_i:

$$v = c_1 q_1 + \cdots + c_n q_n \tag{85}$$

$v^{(k)}$ is a scalar multiple of $A^k v^{(0)}$. Therefore

$$\begin{aligned}
v^{(k)} &= \alpha_k A^k v^{(0)} && (\alpha_k \text{ a normalization constant}) \\
&= \alpha_k \left(\sum_{i=1}^{n} \lambda_i^k a_i q_i\right) \\
&= \alpha_k \lambda_1^k \left(c_1 q_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k q_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^k q_n\right)
\end{aligned}$$

We can choose α_k such that $\alpha_k \lambda_1^k$ is 1. Therefore, $c_1 q_1$ is dominating (as long as $c_1 \neq 0$). The other terms are of order $\mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$.

Convergence of eigenvalue: see proposition below.

□

Theorem 21 (Error of Rayleigh Quotient). Let x_1 be the eigenvector that corresponds to the largest (in absolute value) eigenvalue. If $\|x - x_1\| = \mathcal{O}(\epsilon)$, then

$$\left|\frac{\langle x, Ax\rangle}{\langle x, x\rangle} - \lambda_1\right| = \mathcal{O}(\epsilon^2) \tag{86}$$

Proof. TODO. □

6.3 Inverse Iteration

Find the smallest eigenvalue/vector.

6.4 Simultaneous Iteration

Obtain the full set of eigenvalues/vectors simultaneously.

Algorithm 3 Simultaneous Iteration

Require: $Q^{(0)} = V = I$, a list of vectors V, which we choose to be the identity

```
1: for k ← 1, 2, ... do
2:     Z ← A Q^(k-1)                          ▷ Apply A
3:     Z ← Q^(k) R^(k)                        ▷ QR factorization of Z
4:     A^(k) ← (Q^(k))^T A Q^(k)              ▷ A_ii^(k) = ⟨q_i^(k), A q_i^(k)⟩
5: end for
```

Intuitively,

$$A^K \cdot V = \left[\sum_i \lambda_i^k c_{1i} \tilde{q}_i \;\middle|\; \sum_i \lambda_i^k c_{2i} \tilde{q}_i \;\middle|\; \cdots\right] \tag{87}$$

The first column vector will converge to $\tilde{q}_1$. The second vector will converge to $\tilde{q}_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \tilde{q}_2$.

$\underline{Q}^{(k)}$ will converge to the matrix of eigenvectors:

$$X = \left[x_1 \;\middle|\; x_2 \;\middle|\; \cdots \;\middle|\; x_n\right] \tag{88}$$

$\underline{A}^{(k)}$ will converge to a diagonal matrix containing the eigenvalues.

6.5 Shifted Power Iteration

Find the eigenvalue close to a specific number.

6.6 QR Algorithm

The QR can be viewed as a stable procedure for computing QR factorizations of the matrix powers A, A^2, A^3, ...

Algorithm 4 QR Algorithm (without shifts)

Require: $A^{(0)} = A$

for $k \leftarrow 1, 2, \ldots$ **do**

$\quad Q^{(k)} R^{(k)} \leftarrow A^{(k-1)}$ ▷ QR factorization of $A^{(k-1)}$

$\quad A^{(k)} \leftarrow R^{(k)} Q^{(k)}$ ▷ Recombine factors in reverse order

end for

6.7 Simultaneous Iteration equivalent to QR Algorithm

The QR algorithm is equivalent to simultaneous iteration applied to a full set of initial vectors, namely, $\hat{Q}^{(0)} = I$. Summary of each algorithm:

Simultaneous Iteration

$$\begin{aligned}
\underline{Q}^{(0)} &= I && \text{(initial condition)} \\
Z &= A\underline{Q}^{(k-1)} && \text{(apply } A\text{)} \\
Z &= \underline{Q}^{(k)} R^{(k)} && \text{(resemblence of normalization, } QR \text{ factorization of } Z\text{)} \\
A^{(k)} &= (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} && \text{(resemblence of Rayleigh quotient)}
\end{aligned}$$

QR Algorithm

$$\begin{aligned}
A^{(0)} &= A && \text{(initial condition)} \\
A^{(k-1)} &= Q^{(k)} R^{(k)} && \text{(compute } QR \text{ factorization)} \\
A^{(k)} &= R^{(k)} Q^{(k)} && \text{(reverse order of factors)} \\
\underline{Q}^{(k)} &= Q^{(1)} Q^{(2)} \cdots Q^{(k)} && \text{(definition of } \underline{Q}^{(k)}\text{)}
\end{aligned}$$

and

$$\underline{R}^{(k)} = R^{(k)} R^{(k-1)} \cdots R^{(1)} \qquad \text{(definition of } \underline{R}^{(k)}\text{)}$$

[[WRONG]]

Theorem 22 (Equivalence of Simultaneous Iteration and the QR Algorithm)**.** Simultaneous Iteration and the QR Algorithm generate identical sequences of matrices $\underline{R}^{(k)}, \underline{Q}^{(k)}, A^{(k)}$. Both give

$$
\begin{aligned}
&(a) : A^{(k)} = \underline{Q}^{(k)}\underline{R}^{(k)} && (QR \text{ factorization of the } k\text{th power of } A)\\
&(b) : A^{(k)} = (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} && (\text{projection})
\end{aligned}
$$

Proof. By induction on k (number of iterations). The base case $k = 0$ is trivial.

1. QR gives (a): Assume $A^{(k-1)} = \underline{Q}^{(k-1)}\underline{R}^{(k-1)}$. The inductive hypothesis for (b) gives that $A^{(k-1)} = (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)}$ or that $\underline{Q}^{(k-1)} A^{(k-1)} = A\underline{Q}^{(k-1)}$. Then

$$
\begin{aligned}
A^{(k)} &= AA^{(k-1)} && (\text{decompose to use inductive hypothesis})\\
&= A\underline{Q}^{(k-1)}\underline{R}^{(k-1)} && (\text{inductive hypothesis})\\
&= \underline{Q}^{(k-1)}A^{(k-1)}\underline{R}^{(k-1)} && (\text{inductive hypothesis from } (b))\\
&= \underline{Q}^{(k-1)}R^{(k-1)}Q^{(k-1)}\underline{R}^{(k-1)} && (\text{from algorithm})\\
&= \underline{Q}^{(k)}\underline{R}^{(k)} && (\text{from definitions of } \underline{Q}^{(k)}, \underline{R}^{(k)})
\end{aligned}
$$

2. QR gives (b): Assume $A^{(k-1)} = (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)}$. From the relationship $A^{(k-1)} = Q^{(k)}R^{(k)}$ and the fact that $Q^{(k)}$ is orthogonal, we can apply $(Q^{(k)})^T$ to both sides (on the left) to get that $(Q^{(k)})^T A^{(k-1)} = R^{(k)}$. Then

$$
\begin{aligned}
A^{(k)} &= R^{(k)}Q^{(k)}\\
&= (Q^{(k)})^T A^{(k-1)} Q^{(k)}\\
&= (Q^{(k)})^T (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)} Q^{(k)} && (\text{inductive hypothesis})\\
&= (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} && (\text{definition of } \underline{Q}^{(k)})
\end{aligned}
$$

□

7 Polynomial Approximation

7.1 Polynomial Interpolation

Problem: Let $n \geq 1$, and suppose that $\{x_i\}_{i=0}^n$ are distinct real numbers and $\{y_i\}_{i=0}^n$ are real numbers. We wish to find $p_n \in \mathbb{P}_n$ such that $p_n(x_i) = y_i$ for $i = 0, 1, \ldots, n$.

7.1.1 Vandermonde Matrix

We'll consider a slightly more general version of the problem here:

Problem: Let $n \geq 1$, and suppose that $\{x_i\}_{i=0}^n$ are distinct real numbers and $\{f(x_i) = y_i\}_{i=0}^n$ are real numbers. We wish to find $p_k \in \mathbb{P}_k$ such that $p_k(x_i) = y_i$ for $i = 0, 1, \ldots, n$.

Let $\{a_i\}_{i=0}^k$ be the coefficients of the polynomial we're solving for. We place the data $\{x_i\}_{i=0}^n$ in a Vandermonde Matrix X and solve the following system

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^k \\ 1 & x_1 & x_1^2 & \ldots & x_1^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \tag{89}$$

There are three cases:

1. If $N = K + 1$, then we can uniquely determine the coefficients.
2. If $N > k + 1$, then we use least squares (or a similar method) to approximate a solution.
3. If $N < k + 1$, then there are infinitely many solutions.

Notes about Vandermonde matrix:

1. The Vandermonde matrix is non-singular (this is why we get a unique solution when $N = k + 1$) (of course the data $\{x_i\}_{i=0}^n$ need to be distinct).
2. The Vandermonde matrix has a large condition number. This means errors in the function data $\{f(x_i)\}_{i=0}^n$ will magnify the error in our approximations of the coefficients. This issue motivates the alternative method for interpolation discussed below.

7.1.2 Lagrange Interpolation

Definition 20 (Lagrange basis polynomial)**.** Given the data $\{x_i\}_{i=0}^n$, define

$$l_j(x) = \frac{\prod_{i \neq j}(x - x_i)}{\prod_{i \neq j}(x_j - x_i)} \tag{90}$$

which satisfies

$$l_j(x_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{91}$$

(note that $\prod_{i \neq j}(x - x_i)$ is an nth order polynomial (1 less degree than the number of data points) and $\prod_{i \neq j}(x_j - x_i)$ is a constant).

Definition 21 (Lagrange interpolation polynomial)**.** Given the data $\{x_i\}_{i=0}^n$ and corresponding function values $\{f(x_i)\}_{i=0}^n$ the Lagrange interpolation polynomial is

$$p(x) = \sum_{i=0}^{n} f(x_i) l_i(x) \tag{92}$$

Notice that $p(x)$ does indeed interpolate f at the data:

$$\begin{aligned} p(x_j) &= \sum_{i=0}^{n} f(x_i) l_i(x_j) \\ &= \sum_{i=0}^{n} f(x_i) \delta_{ij} \\ &= f(x_j) \end{aligned}$$

Theorem 23 (Error of Lagrange interpolation polynomial)**.** Suppose that $n \geq 0$ and the f is a real-valued function, defined and continuous on the closed real interval $[a, b]$, such that derivative of f or order $n+1$ exists and is continuous on $[a, b]$. Then, with $x \in [a, b]$, there exists $\xi = \xi(x)$ in (a, b) such that

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x - x_k) \tag{93}$$

is the interpolation error, where $p(x)$ is n-th order.

Proof. We denote the error as a function of x as:

$$E(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x - x_k) \tag{94}$$

Define an auxiliary function $G(x)$ (and a fixed t) as

$$G_t(x) = E(x) - \frac{\prod_{k=0}^{n}(x - x_k)}{\prod_{k=0}^{n}(t - x_k)} E(t) \tag{95}$$

Note that at a grid point the auxiliary function is 0:

$$0 = G_t(x_j) \tag{96}$$

Further, evaluated at t, we have that

$$G_t(t) = E(t) - E(t) = 0 \tag{97}$$

Thus $G_t(x)$ has $n+2$ zeros.

We then use the following lemma:

Lemma 2 (Rolle's Theorem)**.** If a function $f(x)$ has k zeros, then its derivative $f'(x)$ has $k-1$ zeros. Similarly, $f''(x)$ has $k-2$ zeros, an so on.

Proof. Between every two zeros of the original function, the derivative must have a 0. □

Thus, applying this lemma repeatedly, we have that $G^{(n+1)}$ has one zero, call it ξ, so that $G^{(n+1)}(\xi) = 0$. Further, by direct calculation,

$$0 = G^{(n+1)}(\xi) = E^{(n+1)}(\xi) - \frac{(n+1)!}{\prod_{k=0}^{n}(t-x_k)}E(t) \tag{98}$$

Then notice that

$$E^{(n+1)}(\xi) = f^{(n+1)}(\xi) \tag{99}$$

so that

$$f^{(n+1)}(\xi) - \frac{(n+1)!}{\prod_{k=0}^{n}(t-x_k)}E(t) = 0 \tag{100}$$

rearranging this equation gives the required expression for the error function. □

Observations:

1. If $f(x) \in \mathbb{P}_n(x)$, then $f^{(n+1)}(\xi) = 0$, so that $E(t) = 0$ for all t. In words, we can perfectly interpolate a polynomial of order n with $n+1$ grid points.

Remark 2. If we naively sample $x_i \in [a,b]$ evenly, then

$$\sup_{t\in[a,b]} \left| \prod_{k=0}^{n}(x-x_k) \right| \tag{101}$$

may be large. Further, we can encounter Runge's phenomenon of oscillation at the edges of an interval. This occurs when using polynomial interpolation with polynomials of high degree over a set of equispaced interpolation points.

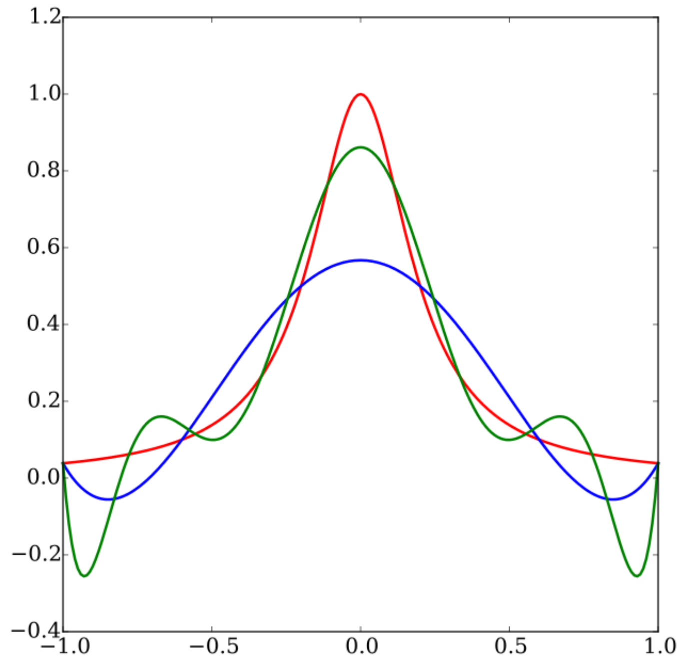

Figure 3: Runge's phenomenon. Red curve is the Runge function. Blue curve is 5-th order interpolating polynomial. Green curve is the 9th-order interpolating polynomial.

We can use Chebyshev grid points to minimize error.

Theorem 24 (Chebyshev grid to minimize polynomial interpolation error)**.** The solution to

$$\min_{\{x_i\}} \sup_{t\in[a,b]} \left| \prod_{k=0}^{n} (x - x_k) \right| \tag{102}$$

is given by a Chebyshev grid:

$$x_i = \cos(\theta_i), \quad \theta_i = \frac{i\pi}{n} \tag{103}$$

7.2 Polynomial Projection

Definition 22 (Orthogonal polynomials)**.** Given a domain $[a, b]$ and a weight function $w(x)$ on the domain, a set of orthogonal polynomials is a list of polynomials $\phi_0, \phi_1, \ldots, \phi_N, \ldots$ such that

$$\langle \phi_i, \phi_j \rangle = \int_a^b \phi_i(x)\phi_j(x)w(x)dx = \delta_{ij} \tag{104}$$

Theorem 25 (Orthogonal polynomials form a basis for the space of polynomials)**.**

$$\mathbb{P}_k = span(\phi_0, \ldots, \phi_k) \tag{105}$$

Example 7 (Examples of Orthogonal Polynomials)**.** The following are examples of Orthogonal Polynomials:

1. Legendre Polynomials
   (a) Domain: $[-1, 1]$
   (b) Weight: $w(x) = \frac{1}{2}$
   (c) Recurrence: $\phi_{n+1} = \frac{2n+1}{n+1} x\phi_n - \frac{n}{n+1}\phi_{n-1}$
2. Chebyshev Polynomials
   (a) Domain: $[-1, 1]$
   (b) Weight: $w(x) = \frac{1}{\sqrt{1-x^2}}$
   (c) Recurrence: $T_{n+1} = 2xT_n - T_{n-1}$
3. Hermite Polynomials
   (a) Domain: $[-\infty, \infty]$
   (b) Weight: $w(x) = e^{-x^2}$
   (c) Recurrence: $H_{n+1} = xH_n - nH_{n-1}$

7.2.1 Properties of Orthogonal Polynomials

1. Recurrence Relation: $\{\phi\}_{i=0}^{N}$ satisfies
   $$\phi_{n+1} = (\alpha_n x + \beta_n)\phi_n + \gamma_n \phi_{n-1} \tag{106}$$
   And these coefficients are **uniquely** determined by
   $$\begin{aligned} \langle \phi_{n+1}, \phi_{n+1} \rangle &= 1 \\ \langle \phi_{n+1}, \phi_n \rangle &= 0 \\ \langle \phi_{n+1}, \phi_{n-1} \rangle &= 0 \end{aligned}$$
2. ϕ_n has n zeros in the domain $[a, b]$.
3. The computation of the zeros ϕ_{n+1} follows from using the recurrence relation in matrix form.

Theorem 26 (OP Recurrence Relation)**.** A set of orthogonal polynomials $\{\phi\}_{i=0}^{\infty}$ satisfies

$$\phi_{n+1} = (\alpha_n x + \beta_n)\phi_n + \gamma_n \phi_{n-1} \tag{107}$$

Proof. Fix $i < n-1$. Let's first show that $\langle \phi_{n+1}, \phi_i \rangle = 0$. Then

$$\begin{aligned}
\langle \phi_{n+1}, \phi_i \rangle &= \alpha_n \langle x\phi_n, \phi_i \rangle + \langle \beta_n, \phi_i \rangle + \gamma_n \langle \phi_{n-1}, \phi_i \rangle \\
&= \alpha_n \langle x\phi_n, \phi_i \rangle + 0 + 0 && \text{(since } \phi_n, \phi_{n-1} \perp \phi_i, i < n-1) \\
&= \alpha_n \langle \phi_n, x\phi_i \rangle && \text{(move } x \text{ to second argument (from integral))}
\end{aligned}$$

Now ϕ_n is an nth order polynomial, and $x\phi_i$ is an $(i+1)$th order polynomial. Thus

$$x\phi_i \in span(\phi_0, \ldots, \phi_{i+1}) \in span(\phi_0, \ldots, \phi_{m-1}) \tag{108}$$

Therefore

$$\langle \phi_n, x\phi_i \rangle = 0 \tag{109}$$

Therefore the following $m-1$ conditions are automatically satisfied:

$$\langle \phi_n, \phi_i \rangle \tag{110}$$

where $i < m-1$, which leaves 3 conditions to find α, β, γ:

$$\begin{aligned}
&\langle \phi_{n+1}, \phi_{n+1} \rangle = 1 \\
&\langle \phi_{n+1}, \phi_n \rangle = 0 \\
&\langle \phi_{n+1}, \phi_{n-1} \rangle = 0
\end{aligned}$$

□

Theorem 27 (Roots of Orthogonal Polynomials)**.** If $\{\phi\}_{i=0}^{\infty}$, then $\phi_n(x)$ has n real roots, called Gaussian quadratures.

Proof. By induction. Clearly ϕ_0 = a constant, which has 0 roots. Next, assume, for the sake of contradiction, that ϕ_1 has no real roots in $[x_1, x_2]$. We know $\langle \phi_1, \phi_0 \rangle = 0$. Without loss of generality, assume that ϕ_1 is completely positive: $\phi_1(x) > 0$ for all $x \in [x_1, x_2]$. Then

$$\langle phi_1, \phi_0 \rangle = \int_{x_1}^{x_2} \phi_1 \phi_0 w(x) dx > 0 \tag{111}$$

since $\phi_1 > 0$, $\phi_0 > 0$ [[?]], $w(x) > 0$ by assumption; but this is a contradiction to orthogonality. Therefore the assumption that ϕ_1 has no real roots in $[x_1, x_2]$ is invalid, so there must be at least one root. But it can't have more than one, so it has exactly one.

Continuing, we can use the same argument to show that ϕ_2 has at least one root. Assume, for the sake of contradiction, that ϕ_2 has only one real root, call it ξ. Then $(x-\xi)\phi_2$ is either > 0 or < 0 [[?]]. Then

$$\langle (x-\xi)\phi_2, \phi_0 \rangle = \int_{x_1}^{x_2} (x-\xi)\phi_2 \phi_0 w(x) dx > 0 \tag{112}$$

However, we should have that

$$\langle \phi_2, (x-\xi)\phi_0 \rangle = 0 \tag{113}$$

since $(x-\xi)\phi_0 \in span(\phi_0, \phi_1)$. Thus have a reached a contradiction, so ϕ_2 has to have at least 2 real roots, and hence exactly 2 real roots. This argument extends to ϕ_n. □

Theorem 28 (Locations of Gaussian Quadratures from Recurrence Relation)**.** Give the recurrence relation

$$\phi_{n+1} = (\alpha_n x + \beta_n)\phi_n + \gamma_n \phi_{n-1} \tag{114}$$

we can rewrite this as

$$\alpha_n x \phi_n = \phi_{n+1} - \beta_n \phi_n - \gamma_n \phi_{n-1} \tag{115}$$

Thus for constants a_n, b_n, c_n we have that

$$x\phi_n = \phi_{n-1} + b_n \phi_n + c_n \phi_{n+1} \tag{116}$$

where this equality holds for all x in the domain. We can write this system in matrix form as follows

$$x \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \vdots \\ \phi_n(x) \end{pmatrix} = \begin{pmatrix} b_0 & c_0 & & & & \\ a_1 & b_1 & c_1 & & & \\ & a_2 & b_2 & \ddots & & \\ & & \ddots & \ddots & & \\ & & & & & c_{n-1} \\ & & & & a_n & b_n \end{pmatrix} \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \vdots \\ \phi_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ c_n \phi_{n+1} \end{pmatrix} \tag{117}$$

where A is the matrix of coefficients. We want to find the roots $\phi_{n+1}(x_i) = 0$, where $i = 1, \ldots, n+1$. Then the eigenvalues of A are the zeros of ϕ_{n+1}. In sum

$$\text{GQ of } \phi_{n+1} = eig(A) \tag{118}$$

7.3 Best Approximation in the 2-norm

Let $\{\phi_i\}_{i=0}^{\infty}$ be a set of orthogonal polynomials and $f \in \mathcal{C}^{\infty}$. Then we can write $f(x)$ as a linear combination of the orthogonal basis polynomials with projection coefficients $\{c_i\}_{i=0}^{\infty}$:

$$f(x) = \sum_{k=0}^{\infty} c_k \phi_k(x) \tag{119}$$

with coefficients

$$c_k = \langle f, \phi_k \rangle = \int_a^b f(x)\phi_k(x)w(x)dx \tag{120}$$

Thus

$$f(x) = \sum_{k=0}^{\infty} \langle f(x), \phi_k(x) \rangle \phi_k(x) \tag{121}$$

We define the projection

$$p_N(x) = \sum_{k=0}^{N} \alpha_k \phi_k(x) \tag{122}$$

where we approximate the coefficients

$$c_i \to \alpha_i = \sum_{k=0}^{N} f(x_k)\phi_k(x_k)w(x_k) \tag{123}$$

Theorem 29 (Exact integration of $f(x) \in \mathbb{P}_{2N+1}$ using $N+1$ grid points)**.** Suppose. $f(x) \in \mathbb{P}_{2N+1}$. Then

$$\int_a^b f(x)w(x)dx = \sum_{i=0}^{N} f(x_k)w_k \tag{124}$$

if $\{x_0, \ldots, x_N\}$ are the GQ (roots) of ϕ_{N+1}, where

$$w_k = \int_a^b l_k(x)w(x)dx \tag{125}$$

where $l_k(x)$ is a Lagrange polynomial.

Proof. We consider two cases. First suppose that $f \in \mathbb{P}_N$. Then

$$f(x) = \sum_{i=0}^{N} f(x_i)l_i(x) \tag{126}$$

Then

$$\begin{aligned}
\int_a^b f(x)w(x)dx &= \int_a^b \sum_{i=0}^{N} f(x_i)l_i(x)w(x)dx \\
&= \sum_{i=0}^{N} f(x_i) \int_a^b l_i(x)w(x)dx \\
&= \sum_{i=0}^{N} f(x_i)w_i
\end{aligned}$$

Thus the equality holds for $f \in \mathbb{P}_N$. Now suppose $f(x) \in \mathbb{P}_{2N+1} \setminus \mathbb{P}_N$. Then let

$$p(x) = \sum_{i=0}^{N} f(x_i)l_i(x) \tag{127}$$

and define the residual

$$r(x) = f(x) - p(x) \tag{128}$$

Notice that $r(x_i) = 0$ for $i = 0, 1, \ldots, N$, and that $r(x) \in \mathbb{P}_{2N+1}$, since $f(x) \in \mathbb{P}_{2N+1}$. Then, we can decompose $r(x)$ into an $N+1$th order polynomial and an Nth order polynomial $q(x)$ as follows

$$r(x) = (x - x_0)(x - x_1)(x - x_2) \cdots (x - x_N) \times q(x) \tag{129}$$

Then

$$\int_a^b r(x)w(x)dx = \int_a^b \prod_{i=0}^{N}(x - x_i)q(x)w(x)dx = 0 \tag{130}$$

This follows because $\prod_{i=0}^{N}(x - x_i)$ is actually just a constant multiple of the orthogonal polynomial $p_{N+1}(x)$ since the x_i are the Gaussian quadrature points. Further, since $q(x)$ has degree N, we know that $q(x) \in span\{p_0, \ldots, p_N\}$. Thus since $\{p_i\}$ are orthogonal polynomials, we know that the integral must evaluate to zero, since we are integrating p_{N+1} and a linear combination of lower order orthogonal polynomials. Therefore

$$\begin{aligned}
\int_a^b f(x)w(x)dx &= \int_a^b (p(x) + r(x))w(x)dx \\
&= \int_a^b p(x)w(x)dx && \left(\text{since } \int_a^b r(x)w(x)dx = 0\right) \\
&= \int_a^b \sum_{i=0}^{N} f(x_i)l_i(x)w(x)dx \\
&= \sum_{i=0}^{N} f(x_i) \int_a^b l_i(x)w(x)dx \\
&= \sum_{i=0}^{N} f(x_i)w_i
\end{aligned}$$

□

Theorem 30 (Projection Coefficients Equivalent to Numerical Representation)**.** Let $f(x) \in \mathbb{P}_{N+1}$. Then

$$\alpha_i = \langle f, \phi_i \rangle = \int_a^b f(x)\phi_i(x)w(x)dx = \sum_{k=0}^{N} f(x_k)\phi_i(x_k)w_k = c_i \tag{131}$$

That is the projection coefficients c_i are equal to the numerical representation α_i, where the grid points are the GQ of ϕ_{N+1}.

Proof. Notice that $f(x) \in \mathbb{P}_{N+1}$, $\phi_i \in \mathbb{P}_i \subsetneq \mathbb{P}_N$ so that $f(x)\phi_i(x) \in \mathbb{P}_{2N+1}$. □

Theorem 31 (Interpolation with Orthogonal Polynomials (Almost Unitary Matrix))**.** We interpolate f as follows:

$$p(x) = \sum_{n=0}^{N} c_n \phi_n(x) \tag{132}$$

such that $p(x_i) = f(x_i)$ where the x_i are the GQ of ϕ_{N+1}. Then

$$\begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_N(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_N(x_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_N(x_N) \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ \vdots \\ f(x_N) \end{bmatrix} \tag{133}$$

Then A, the matrix above, is almost unitary. In particular,

$$A^T \cdot W \cdot A = I \tag{134}$$

where W is a diagonal matrix with elements $w_0, w_1, \dots, w_N$.

Proof. We'll show that

$$(A^T WA)_{mm} = \delta_{mn} \tag{135}$$

We can write out the mnth entry of the matrix product as follows

$$\begin{aligned} (A^T WA)_{mm} &= \sum_{k=0}^{N} p_m(x_k) p_n(x_k) w_k \\ &= \int_a^b p_m(x) p_n(x) w(x) dx \\ &= \delta_{mn} \end{aligned}$$

where the second line follows from applying the above theorem. We can apply this theorem because $p_m(x) \in \mathbb{P}_{N+1}$ and $p_n \in \mathbb{P}_N$. The last line follows from the fact that p_m and p_n are orthogonal polynomials, so their product gives the Kronecker delta by definition. Thus, since $(A^T WA)_{mm} = \delta_{mn}$, $(A^T WA)_{mm}$ is the identity matrix.

Then the condition number of A is approximately $\frac{\max w_i}{\min w_i} \approx \mathcal{O}(1)$. □

Theorem 32 (Projection the best approximation in the L^2-norm:)**.** $p_N(x)$ is the best approximation in the L^2-norm:

$$\|f - p_N(x)\|_2^2 \leq \|f - q(x)\|_2^2 \tag{136}$$

for all $q \in \mathbb{P}_{\mathbb{N}}$.

Proof.

$$\begin{aligned}\langle f-q, f-q\rangle &= \langle f-p+p-q, f-p+p-q\rangle \\ &= \langle f-p, f-p\rangle + 2\langle f-p, p-q\rangle + \langle p-q, p-q\rangle\end{aligned}$$

Notice that $f-p \in span(\phi_{N+1}, \phi_{N+2}, \ldots)$. Further, $p-q \in span(\phi_0, \phi_1, \ldots, \phi_N)$. Thus, $\langle f-p, p-q\rangle = 0$. Therefore, we have that

$$\|f-q\|_2^2 = \|f-p\|_2^2 + \|p-q\|_2^2 \tag{137}$$

Since $\|p-q\|_2^2 \geq 0$, we have that

$$\|f-p_N(x)\|_2^2 \leq \|f-q(x)\|_2^2 \tag{138}$$

□

Theorem 33 (Error from Approximation by Projection)**.** Suppose $f \in \mathcal{C}^\infty$ and $\{\phi_i\}_{i=0}^{\infty}$ is a set orthogonal polynomials. We can write

$$f(x) = \sum_{k=0}^{\infty} c_k \phi_k(x) \tag{139}$$

and define the projection

$$p_N(x) = \sum_{k=0}^{N} \alpha_k \phi_k(x) \tag{140}$$

Then the error of this approximation is

$$error = \sum_{k=N+1}^{\infty} \alpha_k \phi_k(x) \tag{141}$$

which depends on $\{\alpha_{N+1}, \alpha_{N+2}, \ldots\}$. In particular, if $f(x) \in \mathcal{C}^\gamma$, then

$$\alpha_n = \mathcal{O}(n^{-\gamma}) \tag{142}$$

for $n > N$ and

$$\alpha_n = \mathcal{O}\left(\frac{1}{N^\gamma}\right) \tag{143}$$

for $n < N$.

Proof. Todo. □

8 Numerical Integration

In general, we take our domain $[a, b]$ and form an evenly spaced grid

$$\{x_0 = a, x_1, x_2, \ldots, x_k, \ldots, x_N = b\} \tag{144}$$

where

$$x_k = a + \frac{b-a}{N}k \tag{145}$$

Thus

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx \tag{146}$$

8.1 Trapezoidal Rule

We will approximate $f(x)$ on the interval $[x_k, x_{k+1}]$ by a first order polynomial $p_1(x)$ which interpolates $f(x)$ at x_k, x_{k+1}:

$$p_1(x) = f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(x - x_k) \tag{147}$$

Then

$$\int_{x_k}^{x_{k+1}} f(x)dx \to \int_{x_k}^{x_{k+1}} p_1(x)dx = f(x_k)\Delta x + \frac{f(x_{k+1}) - f(x_k)}{\Delta x} \int_{x_k}^{x_{k+1}} (x - x_k)dx \tag{148}$$

We can evaluate the final integral easily using the change of variable $x \equiv x - x_k$:

$$\int_{x_k}^{x_{k+1}} (x - x_k)dx = \int_0^{\Delta x} xdx = \frac{1}{2}x^2\Big|_0^{\Delta x} = \frac{1}{2}\Delta x^2 \tag{149}$$

Thus

$$\int_{x_k}^{x_{k+1}} p_1(x)dx = \frac{\Delta x}{2}(f(x_k) + f(x_{k+1})) \tag{150}$$

Putting these pieces together:

$$
\begin{aligned}
\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx \to \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} p_1(x)dx \\
&= \sum_{k=0}^{N-1} \left(\frac{\Delta x}{2} (f(x_k) + f(x_{k+1})) \right) \\
&= \frac{\Delta x}{2} (f(x_0) + f(x_1) + f(x_1) + f(x_2) + \ldots) \\
&= \frac{\Delta x}{2} (f(x) + 2 \sum_{k=1}^{N-1} f(x_k) + f(b))
\end{aligned}
$$

We now find the error in $[x_k, x_{k+1}]$. We can apply the exact expression found for error in polynomial interpolation found before (derived using Taylor's theorem). In this context, $N = 1$ (we're using a first order polynomial to interpolate). Thus

$$
f(x) - p_1(x) = f''(\xi_x)(x - x_k)(x - x_{k+1}) \tag{151}
$$

for some $\xi_x \in [x_k, x_{k+1}]$ (recall that the error depends on x). Then

$$
\begin{aligned}
\int_{x_k}^{x_{k+1}} f(x)dx - \int_{x_k}^{x_{k+1}} p_1(x)dx &= \int_{x_k}^{x_{k+1}} (f(x) - p_1(x)) \\
&= \int_{x_k}^{x_{k+1}} (f''(\xi_x)(x - x_k)(x - x_{k+1}))dx \\
&= f''(\eta) \int_{x_k}^{x_{k+1}} (x - x_k)(x - x_{k+1})dx
\end{aligned}
$$

Where the last inequality follows from an application of the Mean Value Theorem: For completeness we restate this theorem here.

Theorem 34 (First mean value theorem for definite integrals)**.** If $f : [a, b] \to \mathbb{R}$ is continuous and g is an integrable function that does not change sign on $[a, b]$, then there exists $c \in [a, b]$ such that

$$
\int_a^b f(x)g(x)dx = f(c) \int_a^b g(x)dx \tag{152}
$$

In this problem, notice that on the domain $[x_k, x_{k+1}]$, $(x - x_k)(x - x_{k+1})$ is a quadratic function that is always (weakly) negative and 0 at x_k and x_{k+1}. Next, we claim that

$$
f''(\eta) \int_{x_k}^{x_{k+1}} (x - x_k)(x - x_{k+1})dx = \mathcal{O}(\Delta x^3) \tag{153}
$$

This is because, using a simple change of variables $x \equiv x - x_k$,

$$
\int_{x_k}^{x_{k+1}} (x - x_k)(x - x_{k+1})dx = \int_0^{\Delta x} x(x - \Delta x)dx = \mathcal{O}(\Delta x^3) \tag{154}
$$

In sum, when $p(x)$ is the piecewise linear interpolation of $f(x)$ derived above,

$$\begin{aligned}
\int_a^b f(x)dx - \int_a^b p(x)dx &= \sum_{k=0}^{N-1} \mathcal{O}(\Delta x^3) \\
&\approx N\Delta x^3 \\
&= (b-a)\Delta x^2 && \text{(Recall } N\Delta x = b-a\text{)}
\end{aligned}$$

8.1.1 Richardson Extrapolation

We can define the error on the ith interval in the domain as

$$E_i = \int_{x_i}^{x_{i+1}} f(x)dx - \frac{\Delta x}{2}\left(f(x_i) + f(x_{i+1})\right) \tag{155}$$

which, by applying the MVT, equals

$$E_i = f''(\eta)\int_{x_i}^{x_{i+1}} (x-x_i)(x-x_{i+1})dx \tag{156}$$

for some $\eta \in [x_i, x_{i+1}]$. Then we define the total error $E^{(N)}$ as

$$\begin{aligned}
E^{(N)} &= \int_a^b f(x)dx - Tr(f;N) = \sum_{i=0}^{N-1} E_i \\
&= c\sum_{i=0}^{N-1} f''(\eta_i)\Delta x^3 \\
&= c\left[\sum_{i=0}^{N-1} f''(\eta_i)\Delta x\right]\Delta x^2
\end{aligned}$$

Then

$$\begin{aligned}
f''(\eta_i)\Delta x &\approx f'(x_{i+1}) - f'(x_i) \\
&\approx \int_{x_i}^{x_{i+1}} f''(x)dx && \text{(?)}
\end{aligned}$$

Continuing,

$$\begin{aligned}
E^{(N)} &= c\Delta x^2(f'(x_1) - f'(x_0) + f'(x_2) - f'(x_1) + \cdots) + \mathcal{O}(\Delta x^4) \\
&= c\Delta x^2(f'(x_N) - f'(x_0)) + \mathcal{O}(\Delta x^4) \\
&= c\Delta x^2(f'(b) - f'(a)) + \mathcal{O}(\Delta x^4)
\end{aligned}$$

8.2 Midpoint Rule

We will approximate $f(x)$ on the interval $[x_k, x_{k+1}]$ by a 0th order polynomial (i.e. a constant function), which interpolates $f(x)$ at $x_{k+\frac{1}{2}}$. Thus we use

$$f(x) \to p_0(x) = f\left(\frac{x_k + x_{k+1}}{2}\right) \tag{157}$$

Then

$$\int_{x_k}^{x_{k+1}} f(x)dx \to \int_{x_k}^{x_{k+1}} p_0(x)dx = f\left(\frac{x_k + x_{k+1}}{2}\right)\Delta x \tag{158}$$

Putting these pieces together gives

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)dx \to \Delta x \left(f(x_{\frac{1}{2}}) + f(x_{\frac{3}{2}}) + \ldots + f(x_{N-\frac{1}{2}})\right) \tag{159}$$

Using our exact expression for the error of polynomial interpolation gives that

$$f(x) - p_0(x) = f'(\xi_x)(x - x_{k+\frac{1}{2}}) \tag{160}$$

However, we *cannot* use the MVT theorem here (as before), since $(x - x_{k+\frac{1}{2}})$ is not necessarily always either strictly positive or strictly negative (i.e. does not change sign). Thus we will instead preform a Taylor expansion of $f(x)$ around the point $x_{k+\frac{1}{2}}$:

$$f(x) = f(x_{k+\frac{1}{2}}) + f'(x_{k+\frac{1}{2}})(x - x_{k+\frac{1}{2}}) + \frac{1}{2}f''(x_{k+\frac{1}{2}})(x - x_{k+\frac{1}{2}})^2 + \ldots \tag{161}$$

Then (using that $p_0(x) = f(x_{k+\frac{1}{2}})$)

$$\begin{aligned}
\int_{x_k}^{x_{k+1}} f(x)dx - \int_{x_k}^{x_{k+1}} p_0(x)dx &= \int_{x_k}^{x_{k+1}} f'(x_{k+\frac{1}{2}})(x - x_{k+\frac{1}{2}})dx \\
&\quad + \int_{x_k}^{x_{k+1}} \frac{1}{2}f''(x_{k+\frac{1}{2}})(x - x_{k+\frac{1}{2}})^2 dx \\
&\quad + \ldots \\
&= \mathcal{O}(\Delta x^3)
\end{aligned}$$

since

$$\int_{x_k}^{x_{k+1}} f'(x_{k+\frac{1}{2}})(x - x_{k+\frac{1}{2}})dx = f'(x_{k+\frac{1}{2}}) \int_{x_k}^{x_{k+1}} (x - x_{k+\frac{1}{2}})dx = 0 \tag{162}$$

8.3 Simpson's Rule

We will approximate $f(x)$ on the interval $[x_{2i}, x_{2i+2}]$ by a second order polynomial $p_2(x)$ which interpolates $f(x)$ at $x_{2i}, x_{2i+1}, x_{2i+2}$. Then

$$\int_{x_{2i}}^{x_{2i+2}} f(x)dx \to \int_{x_{2i}}^{x_{2i+2}} p_2(x)dx = \frac{\Delta x}{3}(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})) \tag{163}$$

8.4 Method of Undetermined Coefficients

9 Numerical ODE

9.1 Preliminaries

10 Initial Value Problems

Our model equation is

$$\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u}) \tag{164}$$

where

$$\boldsymbol{u}' = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \quad \boldsymbol{f}(\boldsymbol{u}) = \begin{pmatrix} f_1(u_1, \dots, u_n) \\ \vdots \\ f_n(u_1, \dots, u_n) \end{pmatrix} \tag{165}$$

10.1 Preparation

Theorem 35 (ODE reduction)**.** Any high order, non-autonomous ODE can be reduced to a 1st order, autonomous ODE (system).

Example 8 (Reduction of ODE)**.** Consider the 3rd order ODE

$$u''' = u'u - 2t(u')^2 \tag{166}$$

with the initial conditions

$$\begin{aligned} u(t=0) &= u_0 \\ u'(t=0) &= u_1 \\ u''(t=0) &= u_2 \end{aligned}$$

We can reduce this ODE to a 1st order system. Define the change of variables

$$\begin{aligned} y_0(t) &= u(t) \\ y_1(t) &= u'(t) \\ y_2(t) &= u''(t) \end{aligned}$$

which also gives that

$$y_0'(t) = y_1(t)$$
$$y_1'(t) = y_2(t)$$

We substitute these terms into the original ODE to get

$$y_2' = y1y_0 - 2t(y_1)^2 \tag{167}$$

with the initial condition

$$y_0(t=0) = u_0$$
$$y_1(t=0) = u_1$$
$$y_2(t=0) = u_2$$

Now we have a 1st order ODE system

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{y}, t) \tag{168}$$

Definition 23 (Autonomous)**.** If the force $\boldsymbol{f}$ has no explicit dependence on t, then we call the ODE (system) autonomous.

Example 9 (Autonomous ODE)**.** Continuing the above example. To reduce a non-autonomous ODE to and autonomous ODE, we can introduce another function $y_3(t) = t$. Notice that

$$y_0' = y_1$$
$$y_1' = y_2$$
$$y_2' = y_1y_002y_3(y_1)^2$$
$$y_3' = 1$$

with the respective initial conditions

$$\begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ 0 \end{pmatrix} \tag{169}$$

Thus, numerically, we only study 1st order autonomous ODEs since it's always possible to reduce a given problem to this context.

10.2 Well-posedness

We study the existence and uniqueness of 1st order ODEs.

Definition 24 (Lipshitz continuous)**.** If

$$|f(u) - f(u^*)| \leq L|u - u^*| \tag{170}$$

for u in a small neighborhood of u^*, then f is Lipshitz continuous at u^*. Note that if f' exists, then

$$L = |f'(u^*)| \tag{171}$$

Definition 25 (Uniformly Lipshitz continuous)**.** If L_u has an upper bound in the domain of f, then f is uniformly Lipshitz continuous.

Theorem 36 (Uniqueness)**.** If the force term $f(u)$ is uniformly Lipshitz, then the equation has a unique solution.

10.3 Difference Operator

Given a differential operator, we need to construct a difference operator to be able to compute numerical solutions to ODEs.

10.3.1 Second Order Forward Difference Approximation

We want to find a, b, c such that

$$f'(x_0) \approx af(x_0) + bf(x_0 + h) + cf(x_0 + 2h) \tag{172}$$

We will compute the Taylor expansion of f around $x_0 + 0h, x_0 + h, x_0 + 2h$. Thus

$$\begin{aligned}
x_0 &: f(x_0) \\
x_0 + h &: f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \frac{h^3}{6}f'''(x_0) + \dots \\
x_0 + 2h &: f(x_0 + 2h) = f(x_0) + 2hf'(x_0) + \frac{(2h)^2}{2}f''(x_0) + \frac{(2h)^3}{6}f'''(x_0) + \dots
\end{aligned}$$

We'll ignore the terms above order 2 (these will be our approximation error). Then we compute the following sum

$$f'(x_0) \approx af(x_0) + bf(x_0 + h) + cf(x_0 + 2h) \tag{173}$$

and match coefficients to determine a systems of equations to solve for a, b, c.

$$\begin{aligned}
f(x_0) &: 0 = a + b + c \\
f'(x_0) &: 1 = bh + 2hc \\
f''(x_0) &: 0 + b\frac{h^2}{2} + c\frac{(2h)^2}{2}
\end{aligned}$$

Solving this system gives

$$a = -\frac{3}{2h}, \quad b = \frac{2}{h}, \quad c = \frac{-1}{2h} \tag{174}$$

In sum, our second order forward difference operator is

$$f'(x_0) \approx -\frac{3}{2h} f(x_0) + \frac{2}{h} f(x_0 + h) - \frac{1}{2h} f(x_0 + 2h) \tag{175}$$

where the LHS and RHS differ by $\mathcal{O}(h^2)$.

10.3.2 First Order Forward Difference Approximation

Following the method above gives that

$$f'(x_0) \approx -\frac{1}{h} f(x_0) + \frac{1}{h} f(x_0 + h) \tag{176}$$

where the LHS and RHS differ by $\mathcal{O}(h)$.

10.3.3 Centered Difference Approximation to Second Derivative

We want to find a, b, c such that

$$f''(x_0) \approx af(x_0 - h) + bf(x_0) + cf(x_0 + h) \tag{177}$$

Following the method aboves results in the following system of equations:

$$\begin{aligned} f(x_0) &: 0 = a + b + c \\ f'(x_0) &: 1 = h(a - c) \\ f''(x_0) &: h^2 \left(\frac{1}{2} a + \frac{1}{2} c \right) \end{aligned}$$

which has solution

$$a = \frac{1}{h^2}, \quad b = -\frac{2}{h^2}, \quad c = \frac{1}{h^2} \tag{178}$$

However, this solution also satisfies the condition for the third derivative:

$$f'''(x_0) = \frac{h^3}{6}(a - c) \tag{179}$$

We did not explicitly impose/include 3rd order terms, but if we did, our solution would not change. Check. Thus, the error of approximation is

$$f'' - Df = \mathcal{O}(h^2) \tag{180}$$

10.4 Forward Euler (FE)

We continue studying the initial value problem

$$\begin{cases} u' = f(u) \\ u(t=0) = u_0 \end{cases} \tag{181}$$

Let $u_n = u(t_n)$ be the true solution. Let $\mathcal{U}_n$ be the numerical solution. Then the forward euler method is

$$u'(t_n) \to \frac{\mathcal{U}_n - \mathcal{U}_n}{\Delta t} \tag{182}$$

Thus at t_n, we have

$$\begin{cases} \frac{1}{\Delta t}(\mathcal{U}_{n+1} - \mathcal{U}_n) = f(\mathcal{U}_n) \\ \mathcal{U}_0 = u_0 \end{cases} \tag{183}$$

Rearranging gives that

$$\mathcal{U}_{n+1} = \mathcal{U}_n + \Delta t f(\mathcal{U}_n) \tag{184}$$

Example 10 (Linear ODE)**.**

$$\begin{cases} u' = \lambda u \\ u(t=0) = u_0 \end{cases} \tag{185}$$

We know that the analytical solution to this IVP is

$$u(t) = u_0 e^{\lambda t} \tag{186}$$

FE is

$$\frac{1}{\Delta t}(\mathcal{U}_{n+1} - \mathcal{U}_n) = \lambda \mathcal{U}_n \tag{187}$$

Gathering terms gives that

$$\frac{1}{\Delta t}\mathcal{U}_{n+1} - \left(\lambda + \frac{1}{\Delta t}\right)\mathcal{U}_n = 0 \tag{188}$$

Define

$$\boldsymbol{\mathcal{U}} = \begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \\ \vdots \\ \mathcal{U}_n \end{pmatrix} \tag{189}$$

We can combine these equations into matrix form

$$\frac{1}{\Delta t}\begin{pmatrix} 1 & 0 & & & \\ -(1+\lambda\Delta t) & 1 & & & \\ 0 & -(1+\lambda\Delta t) & 1 & & \\ & & \ddots & \ddots & \\ & & & -(1+\lambda\Delta t) & 1 \end{pmatrix}\begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \\ \vdots \\ \\ \vdots \\ \mathcal{U}_n \end{pmatrix} = \begin{pmatrix} \left(\lambda + \frac{1}{\Delta t}\right) u_0 \\ \vdots \\ \\ \vdots \\ 0 \end{pmatrix} \tag{190}$$

(notice this is a subdiagonal matrix, call it A). Then we have a linear system $AU = S$, $U = A^{-1}S$. For this particular example, we have an explicit formula for A^{-1}. Let $\mu = (1+\lambda\Delta t)$.

$$A^{-1} = \Delta t \begin{pmatrix} 1 & & & & & \\ \mu & 1 & & & & \\ \mu^2 & \mu & 1 & & & \\ \vdots & \mu^2 & \mu & 1 & & \\ \vdots & & \ddots & \ddots & \ddots & \\ \mu^{n-1} & \dots & & \mu^2 & \mu & 1 \end{pmatrix} \tag{191}$$

At the final time, n, we have that

$$\mathcal{U}_n = (A^{-1}\cdot S)_n = \Delta t\,(1+\lambda\Delta t)^{n-1}\cdot\left(\lambda + \frac{1}{\Delta t}\right)u_0 \tag{192}$$

which simplifies to

$$\mathcal{U}_n = (1+\lambda\Delta t)^n \cdot u_0 \tag{193}$$

Now assume that

$$T = n\Delta t \tag{194}$$

which means that

$$\mathcal{U}_n = \left(1 + \frac{\lambda T}{n}\right)^n \cdot u_0 \to e^{\lambda T}, \quad n \to \infty \tag{195}$$

Definition 26 (Local Truncation Error (LTE))**.** The local truncation error is by how much the true solution fails to satisfy the approximation scheme, which can be written as

$$\tau_n = \frac{u_{n+1} - u_n}{\Delta t} - f(u_n) \tag{196}$$

Definition 27 (Consistency)**.** We say a method is consistent is the LTE goes to 0 as $\Delta \to 0$.

Theorem 37 (Forward Euler (one-step) is consistent)**.**

Proof. We take the equation for LTE

$$\tau_n = \frac{u_{n+1} - u_n}{\Delta t} - f(u_n) \tag{197}$$

and substitute in the Taylor expansion of $u_{n+1} = u(t_{n+1})$ around $u_n = u(t_n)$. This Taylor expansion is

$$u_{n+1} = u_n + \tag{198}$$

Incomplete □

10.5 Stability

Example 11 (Stability of Linear ODE)**.** Using the Forward Euler method, our approximation to the solution of the linear ODE is $A\boldsymbol{\mathcal{U}} = \boldsymbol{S}$. Incomplete

«««< HEAD

11 Practice Problems

Exercise 1 (Hermite Polynomials)**.** Hermite polynomials are a set of polynomials that are orthonormal with respect to a Gaussian weight function $w(x) = \frac{1}{\sqrt{\pi}}e^{-x^2}$ on the domain $(-\infty, \infty)$. Thus, if H_m is the m-th order polynomials, then

$$\int_{-\infty}^{\infty} H_m(x)H_n(x)\frac{1}{\sqrt{\pi}}e^{-x^2}dx = \delta_{mn} \tag{199}$$

The first three Hermite polynomials are

$$H_0 = 1, \quad H_1 = \frac{2x}{\sqrt{2}}, \quad H_2 = \frac{4x^2 - 2}{\sqrt{8}}$$

Compute:

1. $\int_{-\infty}^{\infty} 2x^2 e^{-x^2} dx$
2. $\int_{-\infty}^{\infty} (4x^2 - 2x - 2)e^{-x^2} dx$

Solution 1. For the first integral, notice that

$$\begin{aligned}
\int_{-\infty}^{\infty} 2x^2 e^{-x^2} dx &= \sqrt{\pi}\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}\frac{2x}{\sqrt{2}}\frac{2x}{\sqrt{2}}e^{-x^2}dx \\
&= \sqrt{\pi}\delta_{11} \\
&= \sqrt{\pi}
\end{aligned}$$

For the second integral, notice that

$$
\begin{aligned}
\int_{-\infty}^{\infty}(4x^2-2x-2)e^{-x^2}dx &= \sqrt{8\pi}\int_{-\infty}^{\infty}\left(\frac{4x^2-2}{\sqrt{8}}-\frac{2x}{\sqrt{8}}\right)\frac{1}{\sqrt{\pi}}e^{-x^2}dx \\
&= \sqrt{8\pi}\int_{-\infty}^{\infty}H_2H_0w(x)dx - \sqrt{2\pi}\int_{-\infty}^{\infty}H_1H_0w(x)dx \\
&= 0
\end{aligned}
$$

Notice how for the second integral, the key observation is that we can separate the integral into two, and then make use of $H_0 = 1$.

Exercise 2 (Legendre Polynomials)**.** Legendre polynomials $\{p_n(x)\}$ are a set of orthogonal polynomials supported on $[-1,1]$ with weight $w(x) = 1$. It can be shown that

$$
Q[p_n] = \lambda_n p_n \tag{200}
$$

where $Q = \frac{d}{dx}\left((1-x^2)\frac{d}{dx}\right)$ is a second order differential operator, and $\lambda_n = -n(n+1)$. Show that

1. $\langle f, p_n\rangle = \frac{1}{\lambda_n}\langle Q[f], p_n\rangle$
2. Prove by induction, that if $f \in \mathcal{C}^{2\gamma}$, then $\langle f, p_n\rangle = \frac{1}{\lambda_n^{\gamma}}\langle Q^{\gamma}[f], p_n\rangle$.
3. Show that if $f \in \mathcal{C}^{2\gamma}$, then $\langle f, p_n\rangle = \mathcal{O}\left(\frac{1}{n^{2\gamma}}\right)$.

Solution 2. We show each item in turn. Beginning with (1): first notice that $p_n = \frac{Q[p_n]}{\lambda_n}$, so that we equivalently want to show that

$$
\frac{1}{\lambda_n}\langle f, Q[p_n]\rangle = \frac{1}{\lambda_n}\langle Q[f], p_n\rangle \tag{201}
$$

so we want to show that the two inner products are equal. We compute each inner product in turn to show equality:

$$
\begin{aligned}
\langle f, Q[p_n]\rangle &= \int_{-1}^{1} f[(1-x^2)p_n']'dx \\
&= \left(f(1-x^2)p_n'\right)\Big|_{-1}^{1} - \int_{-1}^{1} f'(1-x^2)p_n'dx \\
&= \int_{-1}^{1} f'(1-x^2)p_n dx \qquad \text{(the first term above evaluates to 0)}
\end{aligned}
$$

$$
\begin{aligned}
\langle Q[f], p_n \rangle &= \int_{-1}^{1} p_n((1-x^2)f')'dx \\
&= \Big(p_n(1-x^2)f'\Big)\Big|_{-1}^{1} - \int_{-1}^{1} f'(1-x^2)p_n'dx \\
&= \int_{-1}^{1} f'(1-x^2)p_n dx \qquad\qquad \text{(the first term above evaluates to 0)}
\end{aligned}
$$

Now, for (2), we have just proven the base case. Now assume that $\langle f, p_n \rangle = \frac{1}{\lambda_n^\gamma}\langle Q^\gamma[f], p_n \rangle$. Then

$$
\frac{1}{\lambda_n^{\gamma+1}}\langle Q^{\gamma+1}[f], p_n \rangle
$$