Part I
Lectures by Glas

1 Basics

1.1 Normed spaces

Definition (Norm over a vector space)**.** Let X be (the set underlying) an $\mathbb{R}$ or $\mathbb{C}$-vector space. A map $\|\cdot\| : X \to [0, \infty)$ is called a **norm** over X if

1. $\forall \alpha \in F : \forall x \in X : \|\alpha x\| = |\alpha| \|x\|$
2. $\forall x, y \in X : \|x + y\| \leq \|x\| + \|y\|$
3. $\forall x \in X : \|x\| = 0 \iff x = 0_X$

The ordered pair $(X, \|\cdot\|)$ is called a **normed** (vector) **space**.

Example. Recall the p-norm on a $\mathbb{R}^n$ vector space with pointwise scaling and addition in the canonical basis ($n \in \mathbb{N}$, $\mathbb{N} \ni p \geq 1$). For all $x \in \mathbb{R}^n$, we define

$$\|x\|_p := \left(\sum_{i=1}^{n} |x^i|^p \right)^{1/p}$$

where x^i are the components of $x \in \mathbb{R}^n$ in the canonical basis, $i = 1 \ldots, n$. Then $(\mathbb{R}^n, \|\cdot\|_p)$ constitutes a normed vector space. A special case is when "$p = \infty$", where we define

$$\|p\|_\infty := \max_{i=1,\ldots,n} |x^i|$$

Definition. Idiotic definition incoming. A matrix norm is simply a norm on linear mappings $\mathbb{R}^n \to \mathbb{R}^m$, $n, m \in \mathbb{N}$.

Instead of leaving it at that, the lecture notes decide to define a matrix norm instead as follows: A **matrix norm** is a map $\|\cdot\| : \mathbb{R}^{n \times m} \to [0, \infty)$ if

1. $\forall \alpha \in \mathbb{R} : \forall A \in \mathbb{R}^{n \times m} : \|\alpha A\| = |\alpha| \|A\|$
2. $\forall A, B \in \mathbb{R}^{n \times m} : \|A + B\| \leq \|A\| + \|B\|$
3. $\forall A \in \mathbb{R}^{n \times m} : \|A\| = 0 \iff A = 0$

In other words, a matrix norm is a norm over $\mathbb{R}^{n \times m}$ (whatever that set even is).

Example. The **Frobenius norm** of a linear map $\phi : V \to W$ can be given by choosing first a basis on V and W defining components ϕ^i_j in that basis, $i = 1, \ldots, \dim V$, $j = 1, \ldots, \dim W$, then defining

$$\|\phi\| := \sqrt{\sum_{i=1}^{\dim V} \sum_{j=1}^{\dim W} |\phi^i_j|^2}$$

Notice that this definition is independent of choice of basis, being the trace of $\phi \circ \phi^*$, which can all be defined independent of a basis, given V and W are finite-dimensional.

Definition. Let V and W be finite-dimensional vector spaces and let $\|\cdot\|_M$ be a 'matrix norm'/norm on linear maps $V \to W$, and let $\|\cdot\|_V$, $\|\cdot\|_W$ be norms on V and W, respectively.

- $\|\cdot\|_M$ is said to be **consistent** with $\|\cdot\|_V$ and $\|\cdot\|_W$ if for all $\phi : V \to W$ homo and $v \in V$, we have

$$\|\phi(v)\|_W \leq \|\phi\|_M \|v\|_V$$

- $\|\cdot\|_M$ is said to be **sub-multiplicative** if for all $\phi, \psi : V \to W$ homo, we have

$$\|\phi \circ \psi\|_M \leq \|\phi\|_M \|\psi\|_M$$

Definition. Let $\|\cdot\|_V$, $\|\cdot\|_W$ be norms on a finite-dimensional vector spaces V and W, respectively. Let $\phi : V \to W$ be a homo. We then define the **induced** norm on $\mathrm{Hom}(V, W)$ **given** $\|\cdot\|_V$ by

$$\|\phi\| := \sup_{v \neq 0_V} \frac{\|\phi(v)\|_W}{\|v\|_V}$$

Theorem. *The induced norm is indeed a norm. Note that we will not prove this as this has been proven twice before in previous courses. Note that the following properties also hold:*

- *The induced norm is consistent with the original norm.*
- $\|\mathrm{id}_V\| = 1$
- *The induced norm is sub-multiplicative.*

Proof. Almost all of the properties are trivial, we shall only prove the last property.

$$\|(\phi \circ \psi)(v)\| \leq \|\phi\| \|\psi(v)\| \leq \|\phi\| \|\psi\| \|v\|$$

This trivially implies the desired result. □

1.2 Big $\mathcal{O}$ notation

Definition. Let $\Omega \subseteq \mathbb{R}$. Let $g : \Omega \to R$, where $\forall x \in \Omega : g(x) \neq 0$, e.g. g is zero nowhere. We define the following sets:

- If there exists an $a \in \Omega$ such that also $(a, \infty) \subseteq \Omega$, we define the set $\underset{x\to\infty}{\mathcal{O}}(g(x))$ by specifying its elements: let $f : \Omega \to \mathbb{R}$, then

$$f \in \underset{x\to\infty}{\mathcal{O}}(g(x)) \iff \exists C > 0 : \exists x_0 \in \Omega : \forall x \in \Omega : x \geq x_0 \implies |f(x)| \leq C|g(x)|$$

- Let $a \in \Omega$. We define the set $\underset{x\to a}{\mathcal{O}}(g(x))$ by specifying its elements: let $f : \Omega \to \mathbb{R}$. $f \in \underset{x\to a}{\mathcal{O}}(g(x))$ if and only if there exists an open interval $a \in I \subseteq \Omega$ and some $C > 0$ such that for all $x \in I$, $x \neq a \implies |f(x)| \leq C|g(x)|$.

Example.

- $\sin \in \mathcal{O}_{x\to\infty}(1)$
- Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial of order at most $n \in \mathbb{N}$, e.g. there exists $a_0, \ldots, a_n$ such that for all $x \in \mathbb{R}$

$$f(x) = \sum_{i=0}^{n} a_i x^i$$

  Then $f \in \mathcal{O}_{x\to\infty}(x^n)$. Proof is trivial for constant polynomials. Then a polynomial either approaches $-\infty$ or ∞. Proof is again trivial for polynomials approaching $-\infty$. Suppose it approaches ∞ then. Consider that

$$\frac{f(x)}{x^n} \overset{x\to\infty}{\longrightarrow} a_n$$

  Also, for sufficiently large x, we have that f is positive. f is also eventually increasing. Therefore, we have that sufficiently large x

$$\frac{|f(x)|}{x^n} = \frac{f(x)}{x^n} < a_n$$

- Informally, instead of writing $f \in \mathcal{O}_{x\to\cdot}(g(x))$, we write $f(x) = \mathcal{O}_{x\to\cdot}(g(x))$. We then also write for some function $h : \Omega \to \mathbb{R}$ that $f(x) = h(x) + \mathcal{O}_{x\to\cdot}(g(x))$ if $f - h \in \mathcal{O}_{x\to\cdot}(g(x))$.
- We then can for example write that

$$\exp(x) = 1 + x + \frac{x^2}{2} + \mathcal{O}(x^3) \qquad (x \to 0)$$

  This statement really means that

$$\sum_{k=3}^{\infty} \frac{x^k}{k!} = \mathcal{O}(x^3) \qquad (x \to 0)$$

  Do note that proving this is trivial for finite partial sums, but for the infinite sum this is a little more tricky. However, we can restrict ourselves to the interval $(0, 1)$ and simply divide both sides by x^3, and following approximately the same reasoning as earlier.

2 Numerical representation

Definition. Let $L, U \in \mathbb{Z}$ (the **lower and upper bound**) with $L < U$. Let $b \in \mathbb{N}$ with also $b \geq 2$ (the **base**). Let $t \in \mathbb{N}^*$ (the **number of digits**). We then define the **set of floating point numbers** with base b and 'mantissa length' t and exponent bounds L and U, by

$$\mathbb{F}(b,t,L,U) := \{0\} \cup \left\{ x \in \mathbb{R} \;\middle|\; (*)\, x = (-1)^s b^e \sum_{i=1}^{t} d_i b^{-i} \right\}$$

$$(*) : \exists s \in \{0,1\} : \exists \underset{\cap}{L} \leq \underset{\mathbb{Z}}{e} \leq U : \exists d_1, \ldots, d_t \in \{0, \ldots, b-1\} : d_1 \neq 0 \implies \ldots$$

Remark. Observe that for $x \in \mathbb{F}(b,t,L,U)$ then

$$\begin{aligned} b^{L-1} \leq |x| = b^e \sum_{i=1}^{t} d_i b^{-i} &\leq b^U \sum_{i=1}^{t} (b-1) b^{-i} \\ &= b^U \left(1 - b + \sum_{i=0}^{t} (b-1) b^{-i} \right) \\ &= (b-1) b^U \left(\frac{1 - b^{-t-1}}{1 - b^{-1}} - 1 \right) \\ &= (b-1) b^U \left(\frac{1 - b^{-t}}{b-1} \right) \\ &= b^U \left(1 - b^{-t} \right) \end{aligned}$$

Furthermore, trivially, if $x \in \mathbb{F}(b,t,L,U)$ then $-x \in \mathbb{F}(b,t,L,U)$, but also if $y \in \mathbb{F}(b,t,L,U)$ then not necessarily $x + y \in \mathbb{F}(b,t,L,U)$.

Definition. We define a 'rounding mapping' $\mathrm{fl} : \mathbb{R} \to \mathbb{F}(b,t,L,U)$ (that is by the way surjective) for all $x \in \mathbb{R}$ by $\mathrm{fl}(0) = 0$ and otherwise by first defining the exponent $e := \lceil \log_b |x| \rceil$. Then clearly the number $y := |x| \cdot b^{-e} < 1$. We then define for all $i \in \{1, \ldots, t+1\}$

$$d_i := \mathrm{mod}(\lfloor y b^i \rfloor, b) \qquad \text{(e.g. the } i\text{-th digit of } y \text{ in base } b)$$

and then for all $i \in \{1, \ldots, t-1\}$, $x_i := d_i$ and

$$x_t := \begin{cases} d_t & \text{if } d_{t+1} < b/2 \\ d_t + 1 & \text{if } d_{t+1} \geq b/2 \end{cases}$$

Now we define recursively $y_t := x_t$, and for all $i \in \{1, \ldots, t-1\}$

$$y_i := \begin{cases} x_i + 1 & \text{if } y_{i+1} \geq b \\ x_i & \text{otherwise} \end{cases}$$

Finally we define for all $i \in \{1, \ldots, t\}$

$$\tilde{d}_i := \begin{cases} 0 & \text{if } y_i \geq b \\ y_i & \text{otherwise} \end{cases}$$

Then

$$\mathrm{fl}(x) := \mathrm{sgn}(x) b^e \sum_{i=1}^{t} \tilde{d}_i b^{-i} \in \mathbb{F}(b,t,L,U)$$

More simply put (but less explicitly),

$$\mathrm{fl}(x) := \underset{y \in \mathbb{F}(b,t,L,U)}{\mathrm{argmin}} |y - x|$$

Note that this might not be unique, ties are broken by picking the smallest.

Definition. The number $\varepsilon_M := b^{1-t}$ is called the **machine epsilon**.

Remark. Observe that if $b^{L-1} \leq |x| \leq b^U(1-b^{-t})$ then there exists $\delta \in \mathbb{R}$ such that $|\delta| \leq \varepsilon_M/2$ such that $\mathrm{fl}(x) = x(1+\delta)$. But then also for very large x, $\delta \approx \varepsilon_M/2$ is possible, and then $\mathrm{fl}(x)$ scales approximately linearly in x, which means that the error for larger floating point numbers will be increasingly larger.

Remark. A clear consequence is that

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} < \varepsilon_M/2$$

Definition. Let $x \in \mathbb{R}$. If $|x| > \sup_{y \in \mathbb{F}(b,t,L,U)} |y|$ we say that x is an **overflowing value**. If $|x| < \inf_{y \in \mathbb{F}(b,t,L,U)} |y|$ then clearly $\mathrm{fl}(x) = 0$ and we say that x is an **underflowing value**.

2.1 Condition of problems

Let $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$ be normed spaces. Quick definition, a map $f : X \to Y$ is said to be well-posed if

- f is bijective
- f^{-1} is continuous

Quick notation, for $x \in X$ and $\Delta X \in X$ we define y and Δy implicitely by

$$f(x) = y \qquad\qquad f(x + \Delta x) = y + \Delta y$$

Using this notation, we define the **relative condition number** for all $x \in X$, $\kappa_f(x)$, by

$$\kappa_f(x) := \sup\left\{ \Delta x \in X \setminus \{0_X\} \;\middle|\; \frac{\|\Delta y\|/\|y\|}{\|\Delta x\|/\|x\|} \right\}$$

If f is additionally totally differentiable, we can consider $df : X \to Y$ and then of course we can induce a (submultiplicative) norm on linear maps $X \to Y$ and then use a higher-dimensional Taylor expansion to approximate

$$\kappa_f(x) \approx \|df\| \frac{\|x\|}{\|y\|}$$

We now heuristically say that f is well-conditioned if κ_f is bounded by a (small) strictly positive real number.

Example. If $f : X \to Y$ is linear, and X and Y are finite-dimensional and over the same field, then we choose a basis $e_1, \ldots, e_d$ and observe that

$$\frac{\partial f}{\partial e_i}(x) = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} = f(e_i)$$

The partial derivative being continuous trivially implies continuity, hence (where $n := \dim(Y)$)

$$df(x) = \sum_{i=1}^{n} x^i f(e_i) = f(x)$$

Anyway, observe that (if we assume f is invertible and thus well-posed)

$$\begin{aligned}
\kappa_f(x) &\approx \|df\| \frac{\|x\|}{\|y\|} \\
&= \|f\| \frac{\left\|f^{-1}(y)\right\|}{\|y\|} \\
&\leq \|f\| \sup_{y \in Y \setminus \{0_Y\}} \frac{\left\|f^{-1}(y)\right\|}{\|y\|} \\
&= \|f\| \left\|f^{-1}\right\|
\end{aligned}$$

We can then also prove well-conditionedness by studying that last expression, which we aptly then identify with its own definition:

$$\kappa(f) := \kappa_f := \|f\| \left\|f^{-1}\right\|$$

3 Direct methods for solving linear systems

Theorem. *Let $\phi : V \to V$ be an endomorphism between over a finite-dimensional vector space V such that $\det(\phi) \neq 0$. Then f is an isomorphism.*

Proof. We know that if $\det(\phi) \neq 0$, then for any basis $e_1, \ldots, e_n$, $n := \dim(V)$, we have that $\phi(e_1), \ldots, \phi(e_n)$ are linearly independent. The proof for this is relatively straightforward and thus omitted. It suffices now to show bijectivity of ϕ. By the fundamental theorem on homomorphisms, proving ϕ has a trivial kernel is sufficient. So suppose $v \in \ker \phi$ and observe that

$$0 = \phi(v) = \phi(v^i e_i) = v^i \phi(e_i)$$

By linear independence, this implies $v^i = 0$ for all $i = 1 \ldots, n$. But then trivially, $v = 0_V$ and then also $\ker \phi = \{0_V\}$, e.g. trivial and we are done.

Note that the converse is also true, since if ϕ is an isomorphism, we already trivially have linear independence of $\phi(e_1), \ldots, \phi(e_n)$ because of having a trivial kernel again, so $\det \phi \neq 0$. □

Remark. Because I am so nice I will prove the fact that we glossed over. Suppose $\det \phi \neq 0$ and pick then a basis, we know for any nonzero top form Ω on V that

$$\det \phi = \frac{\Omega(\phi(e_1), \ldots, \phi(e_n))}{\Omega_{1\ldots n}}$$

We get

$$\Omega(\phi(e_1), \ldots, \phi(e_n)) \neq 0$$

We shall prove that this gives linear independence by contradiction. Suppose $\phi(e_1), \ldots, \phi(e_n)$ are linearly dependent, we can then relabel such that

$$\phi(e_1) = \sum_{i=2}^{n} a^i \phi(e_i)$$

for some $a^2, \ldots, a^n \in F$. Observe that

$$\det \phi = \frac{\Omega\left(\sum_{i=2}^{n} a^i \phi(e_i), \ldots, \phi(e_n)\right)}{\Omega_{1\ldots n}} = \sum_{i=2}^{n} a^i \frac{\Omega(\phi(e_i), \ldots, \phi(e_i), \ldots, \phi(e_n))}{\Omega_{1\ldots n}} = 0$$

which is a contradiction (given 2 is invertible / nonzero in the field / the field has order not equal to 2).

3.1 LU factorization

Recall the following definitions:

Definition. Let $\phi : V \to V$ be an endomorphism over a finite-dimensional vector space V. ϕ is said to have a lower-triangular representation if there exists a basis such that the components of ϕ, ϕ^i_j, are lower-triangular, that is,

$$j > i \implies \phi^i_j = 0$$

Mutatis mutandis for upper-triangular, such that

$$j < i \implies \phi^i_j = 0$$

Recall that we can also prove that if we have such a triangular representation,

$$\det \phi = \prod_{i=1}^{\dim V} \phi^i_i$$

So trivially if the diagonal components are all nonzero, ϕ is invertible (the converse also holds). Also, lets suppose ϕ has a lower-triangular representation, is invertible and lets then work in the basis for

which that is the case. Let $w \in V$ and we now want to find $v \in V$ such that $\phi(v) = w$. Well, considering each component separately, clearly $v^j \phi^i_j = w^i$ must hold for all $i = 1, \dots, \dim V$. This then by lower-triangularity implies

$$\sum_{j=1}^{i} v^j \phi^i_j = w^i \implies v^i \phi^i_i = w^i - \sum_{j=1}^{i-1} v^j \phi^i_j$$

This implicitely allows us to recursively find the components of v, thereby uniquely determining it. Observe that if we instead assume upper-triangularity, we get

$$\sum_{j=i}^{\dim V} v^j \phi^i_j = w^i \implies v^i \phi^i_i = w^i - \sum_{j=i+1}^{\dim V} v^j \phi^i_j$$

Letting w take basis vector values, we can fully determine ϕ^{-1} as well. Now, since we can write a linear algorithm (so to speak) of solving these 'inverse problems' easily for triangular matrices, we are motivated to look at the case when an endomorphism ϕ, not necessarily triangular, can be 'decomposed' into morphisms $\Lambda, \Upsilon : V \to V$ such that $\phi = \Lambda \circ \Upsilon$, Λ is lower-triangular and Υ is upper-triangular (or vice-versa).

Definition. Let ϕ be an automorphism on a finite-dimensional vector space V. An upper-triangular endomorphism U and lower-triangular endomorphism L on V are said to be an **LU-factorization** of ϕ if $\phi = L \circ U$.
Note that triangularity here is taken to be in a certain basis. L and U must be triangular in the same basis.

Remark. Because ϕ is assumed to be an automorphism, if we additionally assume the field of V is not finite, we find that L and U must also be invertible, due to

$$0 \neq \det(\phi) = \det(L) \det(U)$$

Definition. Let ϕ be an endomorphism on a finite-dimensional vector space V. Let $d := \dim V$. Let $i \in \{1, \dots, d\}$. Let $e_1, \dots, e_d$ be a basis for V. Let $W := \operatorname{span}\{e_1, \dots, e_i\}$. A map $\psi : W \to W$ is said to be the i-th leading principal subendomorphism if for all $j, k = 1, \dots, i$, we have $\psi^j_k = \phi^j_k$.

Remark. Note the above definition simplifies to

$$\forall w \in W : \psi(w) = \phi(w) \in W$$

The equivalence in one direction is easy, the other is not. Assume the condition for the original definition is true. Let $w \in W$.

$$\phi(w) = \sum_{i=1}^{d} w^i \phi(e_i) = w^d \phi(e_d) + \sum_{i=1}^{d-1} w^i \psi(e_i) = \psi(w)$$

Theorem. *Let ϕ be an automorphism on a finite-dimensional vector space V over an infinite field. Let $d := \dim V$. Let $e_1, \dots, e_d$ be a basis for V. Then there exist unique endomorphisms L, U on V with L lower-triangular and its main diagonal entries in the basis equal to 1, U upper-triangular, $\phi = L \circ U$ if and only if all leading principal subendomorphisms are invertible.*

Proof. Suppose first the reverse direction, namely, invertible subendomorphisms. We proceed by induction on d. The base case is trivial. Let ψ be the $(d-1)$-th subendomorphism of ϕ. Suppose there exists now a unique factorization as specified above, such that $\psi = L \circ U$. We proceed by specifying uniquely endomorphisms L^*, U^* on $\operatorname{span}\{e_1, \dots, e_d\}$ by specifying its components in the given basis. Let $(L^*)_{d-1} := L$ (that is, in components) and similarly for U. Let $(L^*)^d_d := 1$. Let $1 \leq i, j < d$ and define

$$(L^*)^i_d := 0 =: (U^*)^d_j \qquad (U^*)^i_d := \sum_{k=1}^{d-1} \phi^k_d (L^{-1})^i_k$$

$$(U^*)^i_d := \sum_{k=1}^{d-1} \phi^k_d (L^{-1})^i_k \qquad (L^*)^d_j := \sum_{k=1}^{d-1} \phi^d_k (U^{-1})^j_k$$

$$\phi_d^d := \sum_{k=1}^{d} (U^*)_d^k (L^*)_d^i$$

Therefore, we have fully specified U^* and L^* by components. We shall verify now that $\phi_j^i = (L^* \circ U^*)_j^i$ for all $i, j = 1, \ldots, d$. Observe that generally

$$\begin{aligned}(L^* \circ U^*)_j^i &= \sum_{m=1}^{d} (L^*)_m^i (U^*)_j^m \\ &= \sum_{m=1}^{\min\{i,j\}} (L^*)_m^i (U^*)_j^m && \text{(triangularity)}\end{aligned}$$

If we let $i, j < d$ then

$$(L^* \circ U^*)_j^i = \sum_{m=1}^{d-1} L_m^i U_j^m = (L \circ U)_j^i = \phi_j^i$$

If we let $i < j$ and $j = d$ then

$$\begin{aligned}(L^* \circ U^*)_d^i &= \sum_{m=1}^{i} \sum_{k=1}^{d-1} \phi_d^k L_m^i (L^{-1})_k^m \\ &= \sum_{m=1}^{d-1} \sum_{k=1}^{d-1} \phi_d^k L_m^i (L^{-1})_k^m - \sum_{m=i+1}^{d-1} \sum_{k=1}^{d-1} \phi_d^k L_m^i (L^{-1})_k^m \\ &= \sum_{k=1}^{d-1} \phi_d^k (L \circ L^{-1})_k^i && \text{(triangularity)} \\ &= \sum_{k=1}^{d-1} \phi_d^k \delta_k^i = \phi_d^i\end{aligned}$$

A similar argument can be made for $j < i$ and $i = d$. This finishes the proof in the backwards direction. Now for the converse. Suppose we have a unique LU-factorization of ϕ. We know that

$$\det(\phi) = \det(L) \det(U) = \prod_{i=1}^{d} L_i^i U_i^i \neq 0$$

Under the assumption that the vector space is over an infinite field, we get that all diagonal entries are nonzero. Observe that we can also trivially find an LU-factorization of the i-th leading principle subendomorphism ψ, like in the above proof, which has the same diagonal entries as L and U. Hence $\det \psi \neq 0$ and is thus invertible, which additionally holds for all i. □

Remark. Do also note that uniqueness can be shown more simply. Let L, U, L^*, U^* be appropriate triangular automorphisms on V satisfying the above conditions and $L \circ U = L^* \circ U^*$. Then also

$$(L^*)^{-1} \circ L = U^* \circ U^{-1}$$

Notice that the composition of triangular maps is triangular in the same fashion, and additionally, the components of $(L^*)^{-1} \circ L$ should be equal to 1 on the diagonal, making it the identity, which implies $U = U^*$ and $L = L^*$.

Remark. Uniqueness does not hold in general. Suppose an LU-factorization, $\phi = L \circ U$, then for some diagonalizable isomorphism ι on V, we have that $\tilde{L} := L \circ \iota$ and $\tilde{U} := \iota^{-1} \circ U$ gives $\phi = \tilde{L} \circ \tilde{U}$.

Example. This is more of a non-example actually. Let T be an endomorphism on V such that $T(e_1) = e_2$ and $T(e_2) = e_1$, in some basis e_1, e_2 for V. T is clearly invertible. However, consider that its only

leading principal subendomorphism is not invertible, it being the zero map. We will now observe that an LU-factorization does not exist. Suppose it does in some basis, and we have $T = L \circ U$. Then

$$0 = \phi_1^1 = (L \circ U)_1^1 = L_1^1 U_1^1 \implies U_1^1 = 0$$

$$1 = \phi_1^2 = L_1^2 U_1^1 = 0$$

which is a clear contradiction.

Definition. Let V be a d-dimensional vector space, let $e_1, \ldots, e_d$ be a basis for V. Consider a permutation $\pi : \{1, \ldots, d\} \to \{1, \ldots, d\}$ (e.g. π is bijective). Then the unique isomorphism $P_\pi : V \to V$ such that for all $i = 1, \ldots, d$ we have $P_\pi(e_i) = e_{\pi(i)}$, is said to be a **permutation automorphism** on V.

Remark. The map P_π is fully determined by its action on basis vectors, so it suffices to store the information of π. You can do this by storing a tuple $(\pi(1), \ldots, \pi(d))$, on a computer for example.

Definition. Let ϕ be an automorphism on a finite-dimensional vector space V. Let P be a permutation automorphism on V. Let then L, U be automorphisms such that $P \circ \phi$ has LU-factorization $P \circ \phi = L \circ U$. Such factorization is said to be a **pivoted LU-factorization** of ϕ.

Theorem. *Let ϕ be an automorphism on a finite-dimensional vector space V. Then there exists a permutation automorphism P on V such that, with $\psi := P \circ \phi$, all leading principal subendomorphisms of ψ are invertible.*

Corollary. *There exists P such that $P \circ \phi$ has an LU-factorization.*

Corollary. *Every automorphism ϕ has a pivoted LU-factorization.*

3.2 Gaussian elimination with partial pivoting (GEPP)

An explicit algorithm exists to compute P, L and U. For every column $i = 1, \ldots, n-1$, where $n := \dim V$, we do the following:

- Let the morphism of the previous step be ψ. At the first step, $\psi = \phi$.
- Find j such that ψ_i^j is the largest component in absolute value, in column i, with $j \geq i$.
- Let P be the permutation that swaps e_j and e_i, and consider $\tilde{\psi} := P \circ \psi$: now $\tilde{\psi}_i^i = \psi_i^j$
- Consider the unique isomorphism L such that for all $k > i$ we have $L_i^k = -\tilde{\psi}_i^k / \tilde{\psi}_i^i$, every diagonal entry equal to 1 and 0 everywhere else. L is clearly lower-triangular. By the way, this is well-defined since $\tilde{\psi}_i^i$ is never 0, if it was then $\phi(e_i) = 0_V$, which contradicts invertibility of ϕ.
- The morphism for the next step is now $L \circ \tilde{\psi}$.
- Observe that for all $k > i$ we get the following components:

$$\begin{aligned}(L \circ \tilde{\psi})_i^k &= \sum_{m=1}^{k} L_m^k \tilde{\psi}_i^m \\ &= \tilde{\psi}_i^k + \sum_{m=1}^{k-1} L_m^k \tilde{\psi}_i^m \\ &= \tilde{\psi}_i^k + L_i^k \tilde{\psi}_i^i = 0\end{aligned}$$

We call the final morphism U, since it is upper-triangular in the considered basis, which we can see by considering that for all $i = 1, \ldots, n-1$, if $k > i$, then $U_i^k = 0$. We now have permutations $P_1, \ldots, P_{n-1}$ and lower-triangular morphisms $L_1, \ldots, L_{n-1}$, such that

$$U = L_{n-1} \circ P_{n-1} \circ \ldots \circ L_1 \circ P_1 \circ \phi$$

Let now $\tilde{L}_{n-1} := L_{n-1}$ and for all $k = 1, \ldots, n-2$

$$\tilde{L}_k := P_{n-1} \circ \ldots \circ P_{k+1} \circ L_k \circ (P_{k+1})^{-1} \circ \ldots \circ (P_{n-1})^{-1}$$

(e.g. we swap rows and columns of L_k appropriately). We additionally set

$$P := P_{n-1} \circ \ldots \circ P_1, \qquad L^{-1} := \tilde{L}_{n-1} \circ \ldots \circ \tilde{L}_1$$

Then clearly

$$U = L^{-1} \circ P \circ \phi \iff P \circ \phi = L \circ U$$

Since also the diagonal entries of L are all 1, we have found an LU-factorization like in an earlier theorem for $P \circ \phi$, proving existence of P for any isomorphism ϕ by explicit construction.

Remark.

- You might think, hold on, don't we need to compute the inverses of the L_k-morphisms? Well, you can just take the additive inverse of the off-diagonal elements.
- You might also think, hold on, who says that $\tilde{L}_k$ are lower-triangular and have entries in a single column? This can also be checked.
- One can play some memory tricks in a computer, by computing U entirely within the original memory of ϕ, then using the places where zero's get introduced to fit in nicely the nontrivial components of the L maps. Then all of the information of L and U fits in the matrix of ϕ.
- The time complexity for computing P, U and L is $\mathcal{O}_{n\to\infty}(n^3)$, then to solve a linear system the complexity is $\mathcal{O}_{n\to\infty}(n^2)$.

4 Iterative methods

Let ξ be an automorphism on a finite-dimensional vector space V. Let $v \in V$ and $b := \xi(v)$, e.g. we consider a linear system $\xi(v) = b$ with solution v. Consider any automorphism ψ on V. We trivially get that

$$\psi(v) + (\xi - \psi)(v) = b$$

Consider that if we then have a guess $x_k \in V$, a next guess $x_{k+1} \in V$ better satisfies

$$x_{k+1} = x_k - \psi^{-1}(\phi(x_k) - b)$$

In other words, we define a map $\phi : V \to V$ by

$$\phi(x) := x - \psi^{-1}(\xi(x) - b) = (\mathrm{id}_V - \psi^{-1} \circ \xi)(x) + \psi^{-1}(b)$$

that has a fixed point $\phi(v) = v$. Now instead of ψ^{-1} we can still pick any automorphism. Consider that ϕ now maps in such a way that we only really care about what $\mathrm{id}_V - \psi^{-1} \circ \xi$ does. Now, if we pick a norm on V in order to discuss convergence, we want to see if ϕ is a contraction. Consider two guesses $x, y \in V$. Observe that (if we also pick a norm on $\mathrm{End}(V)$ that is consistent)

$$\begin{aligned} \|\phi(x) - \phi(y)\| &= \left\|(\mathrm{id}_V - \psi^{-1} \circ \xi)(x - y)\right\| \\ &\leq \left\|\mathrm{id}_V - \psi^{-1} \circ \xi\right\| \|x - y\| \end{aligned}$$

So indeed, ϕ is a contraction if and only if $\left\|\mathrm{id}_V - \psi^{-1} \circ \xi\right\| < 1$, which means that any sequence defined by iterating ϕ on an initial guess converges to a solution due to the Banach fixed point theorem. The speed of convergence then also depends on said norm. For convenience, write $B := \mathrm{id}_V - \psi^{-1} \circ \xi$. Consider that for any eigenvector w of B with eigenvalue λ, we have

$$|\lambda| \|w\| = \|\lambda w\| = \|B(w)\| \leq \|B\| \|w\|$$

E.g. we convergence implies every eigenvalue λ satisfies $|\lambda| < 1$.
For now though, we consider different choices for ψ. Because we really like to keep things numerical / less abstract, we first reformulate the problem in terms of matrices: we have a nonsingular $n \times n$ matrix A, some $b \in \mathbb{R}^n$ in the canonical basis, and we look for $x \in \mathbb{R}^n$ such that $Ax = b$. The above derivation then corresponds to finding a nonsingular $n \times n$ matrix G, and we define our map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ through

$$\phi(x) := x - G^{-1}(Ax - b) = (I - G^{-1}A)x + G^{-1}b$$

We observe that, analogous to our previous derivation, $I - G^{-1}A$ is an 'iteration matrix'. Observe that for any A, there exist uniquely $n \times n$ matrices L, D, U, where L is stricly lower-triangular (diagonal entries are zero), D is diagonal and U is strictly upper-triangular, such that

$$A = L + D + U$$

Possible choices for G are:

- For some $\gamma > 0$, we take $G := \gamma \cdot I$, this is called the **Richardson method**. Its inverse is then $\gamma^{-1} \cdot I$.
- $G := D$, which leads to the **Jacobi method**
- $G := L + D$ or $G := D + U$, e.g. you take A and remove either the upper half or the lower half, which clearly gives you a triangular matrix, which is invertible if and only if A is. This leads to the **Gauss-Siedel methods**.

4.1 Discussing convergence

We consider in the above context a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ such that indeed $x_{k+1} = \phi(x_k)$ for all $k \in \mathbb{N}$ and additionally $x_0 \in \mathbb{R}^n$ can be any arbitrary initial guess. We define a new sequence $e_k := x - x_k$, e.g. the signed error from the true solution to $Ax = b$. By induction, you can prove that

$$e_k = x - x_k = (I - G^{-1}A)^k(x - x_0)$$

Then clearly if $\left\|I - G^{-1}A\right\| < 1$, we have that $e_k \to 0$ as $k \to \infty$.

Definition. Let ϕ be an automorphism on a finite-dimensional $\mathbb{C}$-vector space V. Let $E \subseteq \mathbb{C}$ be the set of eigenvalues of ϕ. Then we define the **spectral radius**, $\rho(\phi)$, by

$$\rho(\phi) := \max_{\lambda\in E} |\lambda|$$

Observe that this definition is basis-independent, and thus we can say that for any representation A of ϕ in any basis, $\rho(A) := \rho(\phi)$.

Theorem. *Let $A, G \in \mathbb{C}^{n\times n}$ be invertible, let $x_0, x, b \in \mathbb{C}^n$, and suppose $Ax = b$. Define the map*

$$\phi(x) := (I - G^{-1}A)x + G^{-1}b$$

Let $x_{k+1} := \phi(x_k)$ for all $k \in \mathbb{N}$. Then $(x_k)_{k\in\mathbb{N}}$ converges to x if and only if $\rho(I - G^{-1}A) < 1$.

This we do not prove in this course, but instead we prove what we had already seen: if $\|B\| < 1$ then we also have convergence.

4.2 Convergence criteria per method

Definition. Let $A \in \mathbb{R}^{n\times n}$. We say that A is **symmetric positive definite** if A is symmetric and all eigenvalues are strictly positive.

Theorem. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite with respective minimal and maximal eigenvalues $\lambda^-, \lambda^+ > 0$. Consider the Richardson method, e.g. let $G := \gamma \cdot I$ for some $\gamma > 0$, then $G^{-1} = \gamma^{-1} \cdot I$ and we define the iteration matrix $B := I - \gamma^{-1}A$. Then*

- $\rho(B) = \max\{|1 - \gamma^{-1}\lambda^-|, |1 - \gamma^{-1}\lambda^+|\}$
- *The Richardson method converges if and only if $\gamma > \lambda^+/2$.*

Proof. Let v^+ and v^- be the eigenvectors corresponding to λ^+, λ^-. Observe that

$$B(v^+) = v^+ - \gamma^{-1}\lambda^+ v^+ = (1 - \gamma^{-1}\lambda^+)v^+$$

and analogous for $B(v^-)$. E.g. all eigenvectors of A are eigenvectors of B. Similarly, if λ is an eigenvalue of B with eigenvector v, then $B(v) = \lambda v = v - \gamma^{-1}Av$, e.g. $Av = \gamma(1-\lambda)v$, therefore to consider the spectral radius $\rho(B)$ we already know that (if E is the set of eigenvalues of A)

$$\rho(B) = \max_{\lambda\in E} |1 - \gamma^{-1}\lambda|$$

Now to prove convergence it suffices to show $\rho(B) < 1$, e.g. $|1 - \gamma^{-1}\lambda^-| < 1$ and also for λ^+. Therefore we have convergence if and only if

$$0 < \gamma^{-1}\lambda^- < 2 \wedge 0 < \gamma^{-1}\lambda^+ < 2$$

□

Definition. Let $A \in \mathbb{C}^{n\times n}$. We say that A is **strictly diagonally dominant** if for all $i = 1, \ldots, n$ we have

$$\sum_{j=1\wedge j\neq i}^{n} |a_{ij}| < |a_{ii}|$$

Theorem. *Let $A \in \mathbb{C}^{n\times n}$ be invertible and strictly diagonally dominant. Then the Jacobi method converges.*

Proof. It suffices to show that for the decomposition $A = L+D+U$, that $\|B\| < 1$, with $B := I - D^{-1}A$. We consider first that $(D^{-1})^i_j = (D^i_j)^{-1} = (A^i_j)^{-1}\delta^i_j$ and then

$$(D^{-1} \circ A)^i_j = (D^{-1})^i_k A^k_j = (A^i_k)^{-1}\delta^i_k A^k_j = (A^i_i)^{-1}A^i_j$$

Therefore

$$B^i_j = \delta^i_j - (A^i_i)^{-1}A^i_j$$

In particular,

$$B^i_i = 1 - 1 = 0$$

We can consider the ∞-norm for example, and observe that

$$\begin{aligned}
\|B\|_\infty &= \max_{1\le i\le n} \sum_{j=1}^{n} |B^i_j| \\
&= \max_{1\le i\le n} \sum_{j=1}^{n} |\delta^i_j - (A^i_i)^{-1}A^i_j| \\
&= \max_{1\le i\le n} \sum_{j=1\wedge i\neq j}^{n} |(A^i_i)^{-1}A^i_j| \\
&= \max_{1\le i\le n} |A^i_i|^{-1} \sum_{j=1\wedge i\neq j}^{n} |A^i_j| \\
&< \max_{1\le i\le n} |A^i_i|^{-1}|A^i_i| = 1
\end{aligned}$$

Therefore, $\|B\| < 1$ and we have convergence. □

Remark.

- Without proof, the same theorem holds for the Gauss-Seidel methods under the same conditions as above.
- We have not discussed stopping criteria just yet.

5 Least squares problems

We would like to perform linear fits on datasets, where we assume linear dependence, e.g. given an independent variable x and dependent variable y, there exists an affine map (or, equivalently, linear when considering differences) $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = y$. In practice, this means there exist $a, b \in \mathbb{R}$ such that $f(x) = ax + b$, as we all know. Despite f being an **affine map**, we consider it **linear** due to it looking like a line :)
We consider f to be the 'best fit' if it minimizes the expression

$$\sum_{i=1}^{m} (f(x_i) - y_i)^2$$

for data points (x_i, y_i). We can reformulate by defining in the canonical basis a map $A : \mathbb{R}^2 \to \mathbb{R}^m$ such that $A(e_1) = 1$, the one vector, and $A(e_2) = (x_1, \ldots, x_m)$, the vector $x := (b, a)$ for $a, b \in \mathbb{R}$ and finally $b := (y_1, \ldots, y_m)$. Then clearly we wish to minimize in the 2-norm $\|A(x) - b\|$. Observe that $A(x)$ must then be an orthogonal projection on the span of b. This projection is unique, and we can find x, e.g. our parameters, by inverting A. However, A is not going to be invertible, so we need to look for something else.

5.1 QR-factorization

Definition. Let $A \in \mathbb{R}^{n \times n}$. We say A is **orthogonal** if $A^T = A^{-1}$. Equivalently, the operator that in the canonical basis has representation A should be Hermitian.

Remark. Recall that for an orthogonal matrix A and $x \in \mathbb{R}^n$ we have in the 2-norm that

$$\|A(x)\|^2 = \langle A(x), A(x) \rangle = \langle x, A^*(A(x)) \rangle = \langle x, x \rangle = \|x\|^2$$

Furthermore recall that

$$\operatorname{rank}(A) := \dim \operatorname{im} A$$

Definition. Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ such that $\operatorname{rank}(A) = n$. Let $Q \in \mathbb{R}^{m \times m}$ be orthogonal and let $R \in \mathbb{R}^{m \times n}$ be upper triangular. Suppose $A = QR$. This is called a **QR-factorization** of A. Additionally, define $\tilde{Q} \in \mathbb{R}^{m \times n}$ to have the first n columns of Q. Suppose there exists then a $\tilde{R} \in \mathbb{R}^{n \times n}$ such that $A = \tilde{Q}\tilde{R}$. This is called a **reduced QR-factorization** of A.

Remark. Suppose A admits a QR-factorization $A = QR$ and we define $\tilde{Q}$ as above. We can then define $\tilde{R}$ by components: $\tilde{R}^i_j = R^i_j$ for $i, j = 1, \ldots, n$. Observe that for $i = 1, \ldots, m$ and $j = 1, \ldots, n$ we have

$$(\tilde{Q}\tilde{R})^i_j = \sum_{k=1}^{n} \tilde{Q}^i_k \tilde{R}^k_j = \sum_{k=1}^{n} Q^i_k R^k_j = \sum_{k=1}^{m} Q^i_k R^k_j = (QR)^i_j$$

Hence we can always, given a QR-factorization, find a reduced factorization.

Remark. Consider that if we have $A = QR$ and $A = \tilde{Q}\tilde{R}$ then

$$\begin{aligned}
\|Ax - b\|^2 &= \|QRx - b\|^2 \\
&= \left\|Q(Rx - Q^{-1}b)\right\|^2 \\
&= \|Rx - Q^*b\|^2 \\
&= \left\|\tilde{R}x - \tilde{Q}^*b + Bb\right\|^2 \\
&= \left\|\tilde{R}x - \tilde{Q}^*b\right\|^2 + 2\left\langle \tilde{R}x - \tilde{Q}^*b, Bb \right\rangle + \|Bb\|^2 \\
&= \left\|\tilde{R}x - \tilde{Q}^*b\right\|^2 + \|Bb\|^2
\end{aligned}$$

where B is some mapping that represents the unmapped part of Q. The last step is justified by the following two observations:

- Q is a vector space isomorphism, so the image of $\tilde{Q}$ and B are disjoint, therefore because they map to an orthogonal basis, map to orthogonal sets.
- R is upper-triangular, so the components with respect to the basis of the image of B are 0, so in the standard inner product we get 0.

Now because $\tilde{R}$ is invertible we get that there is a unique minimizer, namely $x = \tilde{R}^{-1}\tilde{Q}^* b$. We conclude that if there exists a QR-factorization, then there is a unique best fit. Now $\tilde{Q}^*$ can be found by removing columns from Q and transposing, and $\tilde{R}$ can be inverted using backsubstitution (it is triangular).

5.2 Computing a QR-factorization

Consider that if $\operatorname{rank}(A) = n$ for $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, then equivalently there is a linear map $\phi : \mathbb{R}^n \to \mathbb{R}^m$ with $\dim \operatorname{im} \phi = n$, so $\phi(e_1), \ldots, \phi(e_n)$ is a basis for $\operatorname{im} \phi$, therefore, the columns of A are linearly independent. We can obtain an orthonormal basis for $\mathbb{R}^n$ now using the Gramm-Schmidt orthogonalization process. Let Q be the orthogonalization of A. Of course, R must be upper-triangular. Furthermore, Q is invertible. Therefore:

$$A(e_i) = Q(R(e_i)) = \sum_{j=1}^{i} R_i^j Q(e_j) = \sum_{j=1}^{i} R_i^j \tilde{e}_j$$

Therefore, the components of R can be uniquely determined by finding the components of columns of A in the basis $\tilde{e}_j := Q(e_j)$.

Theorem. *Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\operatorname{rank}(A) = n$. Then there exists a unique reduced QR-factorization $A = \tilde{Q}\tilde{R}$ of A, or the diagonal entries of $\tilde{R}$ are not all strictly positive. Then also there exists a unique full QR-factorization.*

Proof. We have already kind of proven this, because we have constructed a QR-factorization. We shall now prove uniqueness. Suppose $A = Q_1 R_1$ and $A = Q_2 R_2$. Suppose that the diagonal entries of R_1 and R_2 are strictly positive. Then they are also both invertible. Hence A must be invertible. Therefore, $Q_2^{-1} Q_1 = R_2^{-1} R_1$. These compositions are then both Hermitian and upper triangular. But then $R_2^{-1} R_1 = cI$ for some $c \in \mathbb{R}$. Hence $R_1 = cR_2$ and $Q_1 = cQ_2$. Thus $c^2 Q_2 R_2 = A$ and finally $c = 1$, finishing the proof. □

6 Polynomial interpolation

Theorem. *Let $n \in \mathbb{N}^*$ and $x_1, \ldots, x_n \in \mathbb{R}$ be distinct and $y_1, \ldots, y_n \in \mathbb{R}$. Then there exists a unique polynomial $f : \mathbb{R} \to \mathbb{R}$ of degree at most $n-1$ such that for all $i = 1, \ldots, n$, $f(x_i) = y_i$.*

Remark. Before we prove this, we shall try to construct it. Consider that

$$x \mapsto \prod_{j=1}^{n} (x - x_j)$$

is a polynomial with roots $x_1, \ldots, x_n$, let us give it the name L. If we remove the ith root, i.e. the continuous extension of the map

$$x \mapsto \frac{L(x)}{x - x_i}$$

we get a polynomial that is 0 at $x_1, \ldots, x_n$ except at x_i, there it cannot be zero. Consider that its value at x_i is then

$$\prod_{j=1 \wedge j \neq i}^{n} (x_i - x_j)$$

We conclude that the map

$$L_i(x) := \prod_{j=1 \wedge j \neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

satisfies $L_i(x_j) = \delta_i^j$. We have now found a degree $n-1$ polynomial that is an 'indicator' of a data point.

Proof. We consider the polynomial $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) := \sum_{i=1}^{n} y_i L_i(x)$$

Consider that then f is of degree at most $n-1$ and also for all $j = 1, \ldots, n$ we have

$$f(x_j) = \sum_{i=1}^{n} y_i \delta_j^i = y_j$$

To prove uniqueness, consider another $g : \mathbb{R} \to \mathbb{R}$ satisfying the requirements. Then $h := f - g$ is a polynomial of degree at most $n-1$ with n roots. By induction, h is the zero map. Therefore $f = g$. □

Remark. Observe that trivially the Lagrange polynomials form a basis over the space of polynomials, considering that they are linearly independent. We can also figure out a dual basis: consider the interpolation points $x_1, \ldots x_n$ and for a certain l_i (basis element), consider that $l_i(x_j) = \delta_j^i$. Now if you then have a polynomial

$$p(x) = \sum_{i=1}^{n} p^i l_i(x)$$

It is clear that

$$p(x_j) = p^i \delta_i^j = p^j$$

So

$$\varepsilon^j(p) := p(x_j)$$

forms a dual basis, and indeed $\varepsilon^j(l_i) = l_i(x_j) = \delta_i^j$.

Theorem. *Let $I \subseteq \mathbb{R}$ be an interval, let $x_1, \ldots, x_n \in \mathbb{R}$ be distinct, let $y_1, \ldots, y_n \in \mathbb{R}$, let $f : I \to \mathbb{R}$ be n-times continuously differentiable, let p be the unique degree $n-1$ polynomial that satisfies $p(x_i) = y_i$ for all $i = 1, \ldots, n$ and let $x \in I$. Define the **interpolation error** $E_f(x) := f(x) - p(x)$. Then there exists a $\xi \in I$ such that*

$$E_f(x) = \frac{f^{(n)}(\xi)}{n!} L(x)$$

Proof. Observe that the statement is clearly true when $x \in \{x_1, \ldots, x_n\}$. Consider now the map $h : \mathbb{R} \to \mathbb{R}$

$$x \mapsto CL(x)$$

and define $m(x) := f(x) - p(x) - h(x)$ (we will show that this is the zero map shortly). We choose $C \in \mathbb{R}$ such that $m(x) = 0$. Therefore, m is n times continuously differentiable and has $n+1$ zero's. Then $m^{(n)}$ has only one zero due to Rolle, call it $\xi \in I$. Then

$$0 = m^{(n)}(\xi) = f^{(n)}(\xi) - Cn!$$

This directly implies the desired result. □

Part II
Lectures by Schlottbom

7 Numerical integration

Goal: find sufficiently good approximations for 'nice' integrable functions $f : [a, b] \to \mathbb{R}$, $a < b$. By convention, we define:

$$I(f) := \int_a^b f(x)\,dx$$

The way to go about this generally is:

- Find sequences of easily integrable functions whose integral is arbitrarily close to that of f.
- Analytically find a closed form for the approximate integral, and derive a formula.

One approximation we already know of: take a sufficiently fine partition of $[a, b]$, and maybe the average of the lower and upper sums. Another method is the Riemann sum. Suppose f is Riemann / Darboux integrable. Let P be a partition of $[a, b]$, e.g. $a = x_0 < x_1 < \ldots < x_{n-1} = b$ of n points. Pick for every $i = 1, \ldots, n-1$ a point $\xi_i \in (x_{i-1}, x_i)$. We define

$$I_n(f) := \sum_{i=1}^{n-1} (x_i - x_{i-1}) f(\xi_i)$$

Now for every $\varepsilon > 0$ there exists $n \in \mathbb{N}^*$, a partition P such that $|P| = n$ and $\xi_1, \ldots, \xi_{n-1} \in \mathbb{R}$ restricted like above such that

$$|I_n(f) - I(f)| < \varepsilon$$

E.g. in a sense, as $n \to \infty$ we have $I_n(f) \to I(f)$. This is a property we would like to have for any method of numerical integration.

Now we know that a partition might exist, but how do we actually find one? And furthermore, given a concrete partition and sample points, can we quantify the error $|I_n(f) - I(f)|$ in terms of properties of f and the 'fineness' of the partition, that is

$$h := \max_{i=1,\ldots,n-1} (x_i - x_{i-1})$$

The general procedure for studying these approximations will be

- Pick a concrete partition P
- Consider for every $n := |P|$ easily integrable functions f_n such that $f_n \to f$ as $n \to \infty$
- Approximate $I(f)$ with a **quadrature rule** (fancy word for numerical integration) $I_n(f) := I(f_n)$ (note that this is not unique as it depends on choices of partition)

The following definitions help us to quantify properties of the quadrature rules that we will derive.

Definition. Let $I_n : I([a, b]) \to \mathbb{R}$ be a quadrature rule on integrable functions on $[a, b]$, $a < b$. We assume (from now on for all quadrature rules) that constant functions can always be exactly integrated, that is, for all $c \in \mathbb{R}$, $I_n(x \mapsto c) = c(b - a)$.
The largest $m \in \mathbb{N}$ such that for all $j \in \mathbb{N}$ with $0 \leq j \leq m$, we have

$$I_n(x^j) = I(x^j)$$

is called the **degree of exactness** of I_n.

Definition. Let I_n be a quadrature rule on a compact interval $I \subseteq \mathbb{R}$. Let $J \subseteq \mathbb{R}$ also be compact and $J \subseteq I$. The **restriction** of I_n to J, $I_n|_J : I(J) \to \mathbb{R}$, is given by

$$f \mapsto I_n\left(x \mapsto \begin{cases} f(x) & \text{if } x \in J \\ 0 & \text{otherwise} \end{cases}\right)$$

Definition. Let I_n be a quadrature rule on $[a,b]$, $a < b$, let P be a partition of size n on $[a,b]$, and let $f_n : [a,b] \to \mathbb{R}$ be simple functions, to be specified by the quadrature rule, e.g. such that $I(f_n) = I_n(f)$, for some integrable f on $[a,b]$. For every $i = 1, \ldots, n-1$, we define the **local error** e_i, by

$$e_i := I|_{[x_{i-1},x_i]}(f) - I_n|_{[x_{i-1},x_i]}(f) = \int_{x_{i-1}}^{x_i} f(x)\,dx - \int_{x_{i-1}}^{x_i} f_n(x)\,dx$$

Remark. Consider that independently of choice of fineness h of partition P or for that matter the size n of P, if a quadrature rule I_n has degree of exactness $m \in \mathbb{N}$, then also I_2 restricted to $[0,1]$ should have degree of exactness n, the converse is also true. This is much easier to check typically.

7.1 Composite midpoint rule

Let a partition already be given and define for all $i = 1, \ldots, n-1$, $\xi_i := (x_{i-1} + x_i)/2$, e.g. the midpoint of the i-th interval. The simple function we now consider is a step function f_n with partition P such that

$$f_n(x) := f(\xi_i) \quad , \quad x \in (x_{i-1}, x_i)$$

We had already seen this in analysis for proving a result for regulated functions. Now $I_n(f) := I(f_n)$ which is just a Riemann sum, e.g.

$$I_n^{\text{mid}}(f) := \sum_{i=1}^{n} (x_i - x_{i-1}) f(\xi_i)$$

Lemma. *The degree of exactness of I_n is 1. Moreover, if f is an integrable and twice continuously differentiable function on $[a, b]$, we have*

$$|I_n(f) - I(f)| \leq \frac{h^2(b-a)}{24} \|f''\|_\infty$$

Proof. Constant functions are clearly integrated correctly. Consider now the identity function $x \mapsto x$ on $[0, 1]$, we know that its integral should just be $1/2$. Choose the trivial partition $P := \{0, 1\}$ and we get $I_2(f_n) = I(x \mapsto x)$. It is easy to verify that this fails for $n = 2$. The other claim can be verified using Taylor's theorem. Use a specific form of a remainder theorem and do a first-order expansion centered at each midpoint. Estimate the local error and sum them up. □

7.2 Composite trapezoidal rule

For all $i = 1, \ldots, n-1$ we define for $x \in (x_{i-1}, x_i)$, $f(x)$ by picking first the Lagrange polynomials

$$l_{i,0}(x) := (x - x_i)/(x_{i-1} - x_i) \quad l_{i,1}(x) := (x - x_{i-1})/(x_i - x_{i-1})$$

and we define

$$f_n(x) := f(x_{i-1}) l_{i,0}(x) + f(x_i) l_{i,1}(x), \qquad x \in (x_{i-1}, x_i)$$

It is easy to verify that

$$I_n(f) = I(f_n) = \sum_{i=1}^{n} (x_i - x_{i-1})(f(x_i) + f(x_{i-1}))$$

Lemma. *The degree of exactness of I_n is 1. Moreover, if f is an integrable and twice continuously differentiable function on $[a, b]$, we have*

$$|I_n(f) - I(f)| \leq \frac{h^2(b-a)}{12} \|f''\|_\infty$$

7.3 Even more rules

Skipped due to it being very repetitive.

Lemma. *The degree of exactness of I_n^{CS} is 3. Moreover, if f is an integrable and 4 times continuously differentiable function on $[a, b]$, we have*

$$|I_n(f) - I(f)| \leq \frac{h^4(b-a)}{180} \|f''''\|_\infty$$

7.4 Newton-Cotes rule

On every interval (x_{i-1}, x_i), we consider $N \in \mathbb{N}^*$ evenly spaced points, and construct a Lagrange interpolation on that interval through the evenly spaced points evaluated on f. This is a generalization of all previous rules. The rule, which underlies a partition of n points, and has N interpolating points, is denoted by $I_{n,N}^{NC}$.

Theorem. *$I_{n,N}^{NC}$ has degree of exactness*

$$D := \begin{cases} N & \text{if } N \text{ is odd} \\ N+1 & \text{if } N \text{ is even} \end{cases}$$

and furthermore, there exists $C > 0$ such that for f in $C^{D+1}([a,b];\mathbb{R})$ we have that

$$\left|I(f) - I_{n,N}^{NC}(f)\right| \leq Ch^{D+1}(b-a)\left\|f^{(D+1)}\right\|_\infty$$

Remark. If, for the points of interpolation of Lagrange polynomials within an interval $[x_i, x_{i+1}]$, $x_{i,0}, \ldots, x_{i,N}$ for some $N \in \mathbb{N}$ to be decided, it holds that $x_{i,0} = x_i$ and $x_{i,N} = x_{i+1}$, then the quadrature rule is said to be **closed** (i.e. including boundary points). Otherwise, it is said to be **open**. An example of an open rule is the midpoint rule.

Remark. Due to the theorem, $I_{n,N}^{NC} \to I(f)$ as $h \to 0$ which is true if and only if $n \to \infty$, assuming evenly spaced partitions and sufficiently nice f. Looking at the formula, if $f \in C^\infty([a,b];\mathbb{R})$, it could make sense to also claim this as $N \to \infty$, and requiring $h < 1$, however, this might not be true as the sequence $\left\|f^{(n)}\right\|_\infty$ might not converge. But take for example $f = \exp$. Then $f^{(n)} = f$ and hence the sequence does converge, namely to the maximum on the interval of integration.

7.5 Stability of quadratures

Consider any quadrature rule with sample points $x_0, \ldots, x_n$ and coefficients (depending on $f : [a, b] \to \mathbb{R}$) $w_0, \ldots, w_n$, such that

$$I_n(f) = (b-a)\sum_{i=0}^{n} w_i f(x_i)$$

Many quadrature rules can be written in this form.

Lemma. *If $f, g \in C([a, b]; \mathbb{R})$, and the coefficients and sample points of I_n are independent of choice of f, g, then*

$$|I_n(f) - I_n(g)| \leq (b-a)\|f - g\|_\infty \sum_{i=0}^{n} |w_i|$$

Proof.

$$|I_n(f) - I_n(g)| = (b-a)\left|\sum_{i=0}^{n} w_i (f-g)(x_i)\right|$$

which by the triangle inequality directly proves the result. □

Theorem. *Let I_n like above and with degree of exactness m and $f \in C([a, b]; \mathbb{R})$. Let P_m be the set of polynomials with degree at most m. Then*

$$|I(f) - I_n(f)| \leq (b-a) \inf_{p \in P_m} \|f - p\|_\infty \left(1 + \sum_{i=0}^{n} |w_i|\right)$$

Proof. Using the triangle inequality, the previous lemma, the fact that I_n is a linear operator, and that

$$|I(f) - I(p)| \leq (b-a)\|f - p\|_\infty$$

by the properties of the integral, for all $p \in P_m$, we get the desired result. □

Remark. Note that $I_n : C([a, b]; \mathbb{R}) \to \mathbb{R}$ is an operator between normed vector spaces, where we equip the domain with the ∞-norm, and $\mathbb{R}$ with the absolute value. Then we can find the norm of I_n, which is clearly bounded above by

$$(b-a)\sum_{i=0}^{n} |w_i|$$

We say that I_n is **stable** with the above stability constant, i.e. small changes in f give relatively small changes in the approximated integral.

Additionally, if I_n has degree of exactness greater than or equal to 0 (e.g. constant functions are exactly integrated) and $w_k \geq 0$ then also $\sum_{i=0}^{n} w_k = 1$, which tells you that the stability constant becomes just $b - a$. Consider that for the regular integral, we have

$$|I(f) - I(g)| \leq \|f - g\|(b-a)$$

where equality can be reached as well, e.g. if f and g are constants, thereby making $(b-a)$ the absolute condition number / stability constant of I.

We conclude that if $\sum_{i=0}^{n} |w_k| \leq 1$, I_n is well-conditioned, and otherwise not, e.g. if w_k becomes negative, which can happen for the NC-rule if N is large.

7.6 Gaussian quadrature

Another way of going about things is this: what if we want to maximize the degree of exactness m of some rule

$$I_n(f) := \sum_{i=0}^{n} w_i f(x_i)$$

Well, we already require that, for all $p \in P_m$, we have $I(p) = I_n(p)$. Taking a basis for P_m we can already find $m+1$ conditions, while we have $n+1$ w_i's and the same amount of x_i's, so we need n first of all such that $2(n+1) = m+1$ in order to have a chance at uniquely determining optimal choices for weights and sample points.

7.6.1 Orthogonal polynomials

Let $w : [a,b] \to \mathbb{R}$ be integrable and strictly positive, let $f, g : [a,b] \to \mathbb{R}$ be polynomials, then define

$$\langle f, g \rangle := \int_a^b w(x) f(x) g(x)\, dx$$

Definition. In this inner product, a sequence $(p_j)_{j\in\mathbb{N}}$ of polynomials is said to be orthogonal if they are already orthogonal (e.g. the inner product is zero for differing indices) and for all $j \in \mathbb{N}$ we have that the degree of p_j is j.

Remark. Picking $p_0 := 1$, the constant 1 polynomial, you can recursively define more orthogonal polynomials with the Gramm-Schmidt procedure.

Definition. Let $f : I \to \mathbb{R}$ be a function, $I \subseteq \mathbb{R}$. Let $x_0 \in I$ such that $f(x_0) = 0$, e.g. x_0 is a root of f. x_0 is then said to be a **simple root** of f if f changes sign there, e.g. f is differentiable and $f'(x_0) \neq 0$.

Theorem. *Let $(p_j)_{j\in\mathbb{N}}$ be a sequence of orthogonal polynomials. Then for all $j \in \mathbb{N}$, p_j has j roots.*

7.6.2 The rule

Definition. Consider $p_0, \ldots, p_{n+1}$ orthogonal polynomials, let $x_0, \ldots, x_n \in \mathbb{R}$ be roots of p_{n+1}, let w be the weight of the inner product. Let for all $i = 0, \ldots, n$

$$w_i := \int_a^b w(x) l_i(x)\, dx$$

where $l_i : \mathbb{R} \to \mathbb{R}$ is given by

$$l_i(x) := \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$$

and finally define

$$I_{n,w}(f) := \sum_{i=0}^{n} w_i f(x_i)$$

Lemma. *$I_{n,w}$ has degree of exactness at least n, where we measure exactness w.r.t.*

$$I_w(f) := \int_a^b w(x) f(x)\, dx$$

Proof. Let $f \in P_n$, then consider that for all x

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

due to the basis results of the Lagrange polynomials from a while ago. Hence

$$I_w(f) = \sum_{i=0}^{n} f(x_i) \int_a^b w(x) l_i(x)\, dx = \sum_{i=0}^{n} f(x_i) w_i = I_{w,n}(f)$$

□

Theorem. *$I_{n,w}$ has degree of exactness $2n+1$ (as desired in an earlier upshot) and there exists a $C > 0$ such that if f is C^{2n+2}, then*

$$|I_{n,w}(f) - I_w(f)| \leq \frac{C}{(2n+2)!} \left\| f^{(2n+2)} \right\|_\infty$$

Lemma. *The weights w_i are positive.*

Proof. Consider a specific $j \in \{0, \ldots, n\}$ and observe that

$$0 < \int_a^b w(x) l_j(x)^2 \, dx = \sum_{i=0}^{n} w(x) l_j(x_i)^2 = \sum_{i=0}^{n} w(x) (\delta_j^i)^2 = w_i$$

□

Remark. Consider that for the constant 1 polynomial, its unique Lagrange interpolating polynomial of order n can be generated by picking any distinct sampling points $x_1, \ldots, x_n$, defining $l_1, \ldots, l_n$ and setting $x \mapsto 1 = l_1 + \ldots + l_n$.

Remark. Consider therefore that

$$\sum_{i=0}^{n} |w_i| = \int_a^b w(x) \, dx$$

We could have picked $w(x)$ at the start such that that equals 1, which makes this rule well-conditioned.

TODO: write notes on the Gauss-Legendre rules.

8 Solving nonlinear equations

Theorem. *Let X, Y be normed spaces, let $f : X \to Y$ be continuously partially differentiable and $x' \in X$, $y' \in Y$ such that $f(x') = y'$ and $\det df(x') \neq 0$. Then, within the metric topology induced on X and Y, there exist open sets $x' \in X' \in O_X$ and $y \in Y' \in O_Y$, and $C > 0$, such that for all $y \in Y'$ there exists a unique $x \in X'$ such that $f(x) = y$ and*

$$\|x - x'\| \leq C\|y - y'\|$$

Remark. I realized that I made a mistake here by generalizing to normed spaced, because $\det df$ is not defined if $X \neq Y$, so require that here and move on.

This shows that there essentially exists on a restricted neighborhood an inverse f^{-1} of f given 'sign change' occurs (this reduces to $f'(x') \neq 0$ in the single-variable case), which can also be applied to root finding by setting $y' = 0$. Therefore, root finding under these conditions is well-posed.

Remark. Consider that finding a root is relatively ill-conditioned in the case that the multiplicity m of a root $x_0 \in I$ is greater than 1 for some function $f : I \to \mathbb{R}$ with I an interval. We can see this by remarking that for all $\mathbb{N} \ni j < m$ we have that $f^{(j)}(x_0) = 0$ and observing that then in an open set $X \ni x_0$ with $\varepsilon > 0$ such that $x \in X \iff |f(x)| < \varepsilon \wedge |f'(m)| \geq c$ for some $c > 0$, that then for all $x \in X$ there exists a $\xi \in X$ such that

$$f(x) = \frac{1}{m!} f^{(m)}(\xi)(x - x_0)^m$$

due to Taylor's theorem. This immediately implies that there exists some $C > 0$ such that

$$|x - x_0| < C\varepsilon^{1/m}$$

which is to say that, for pertubations $\varepsilon < 1$ in the output, the neighborhood is generally larger as m grows, and hence it better conditioned when $m = 1$.

8.1 Newton-Raphson method

Say we want to find roots of a sufficiently nice $f : I \to \mathbb{R}$, I an interval, or alternatively $f : \mathbb{R} \to \mathbb{R}$ such that it has support I, and we want to find roots $x' \in I$ such that $f(x') = 0$, which is given to exist. Then, the idea is to start with an initial guess $x_0 \in I$ and try to iteratively improve it.

The main idea is to find the linearization of f at x_i, find its root, and define that to be x_{i+1}. If $f \in C^2(\mathbb{R}; \mathbb{R})$, then its linearization L at x_i is given by

$$L(x) := f(x_i) + f'(x_i)(x - x_i)$$

Its unique root is given by

$$L(x) = 0 \iff x = x_i - \frac{f(x_i)}{f'(x_i)} =: x_{i+1}$$

This gives you an iterative scheme called the **Newton-Raphson method**. If you instead want to find more generally when $f(x') = y'$ for some $y' \in \mathbb{R}$, we can define $g : \mathbb{R} \to \mathbb{R}$ by $x \mapsto f(x) - y'$ and find its root, e.g. $x' \in \mathbb{R}$ such that $g(x') = 0$, which is equivalent to $f(x') = y'$. So in more generality:

$$x_{i+1} := x_i - \frac{g(x_i)}{g'(x_i)} = x_i + \frac{y' - f(x_i)}{f'(x_i)}$$

If we require that $f'(x') \neq 0$ we have that on some neighborhood of f, the problem is well-posed, so let us analyze convergence in that case.

Theorem. *Let $f \in C^2(\mathbb{R}; \mathbb{R})$. Let $y' \in \mathbb{R}$ and suppose there exists $x' \in \mathbb{R}$ such that $f(x') = y'$ and require that there exists $c > 0$ such that $|f'(x')| = c$ (e.g. finding x' is well-posed). Let $x_0 \in \mathbb{R}$ such that $|x_0 - x'| \leq \frac{c}{2\|f''\|_\infty} < 1$ and define*

$$x_{i+1} := x_i - \frac{g(x_i)}{g'(x_i)} = x_i + \frac{y' - f(x_i)}{f'(x_i)}$$

for all $i \in \mathbb{N}$. *Then*

$$\lim_{n\to\infty} x_n = x' \wedge |x_{k+1} - x'| \leq \frac{c}{2\|f''\|_\infty}|x_k - x'|^2$$

Remark. In other words, if the problem is well posed and $|f'(x')| < 2\|f''\|_\infty$, we get convergence and the rate of convergence is quadratic.
If f' is Lipschitz continuous, it is quite easy to then show that there exists $c > 0$ such that

$$|f'(x') - f'(x_k)| \leq c|x' - x_k|$$

and

$$|x_{k+1} - x'| \leq c|x_k - x'|^2$$

The main idea of the omitted proof is to use standard analysis to show required inequalities like the ones above, and then show convergence by already knowing that $|x_k - x'|^2 < 1$ by induction which is ultimately true by assumption, and then using induction to prove that $|x_{k+1} - x'| \leq a_k|x_0 - x'|$ for some $a_k \to 0$ as $k \to \infty$, which proves convergence.

The final conclusion is that if f is sufficiently nice and the initial guess x_0 is at most 1 away from the root you are looking for, you are guaranteed to find a root by taking limits.

Remark. We can generalize this method to general $f : X \to X$ when X is a normed space. Let $x_0 \in X$ be an initial guess, require to find $x' \in X$ for some $y' \in X$ such that $f(x') = y'$ and define

$$x_{k+1} = x_k + (df(x_k))^{-1}(y' - f(x_k))$$

Convergence then works the exact same, except you work with norms instead of absolute values, and instead you require that $\det df(x_k) \neq 0$.

8.2 Fixed point iteration

We can generalize these iterative methods. Let $f : X \to X$ with X a normed space, let $x \in X$ be the solution such that for some $y \in X$ we have $f(x) = y$. Let $M : X \to \mathrm{Aut}(X)$ be arbitrary, e.g. for every $\tilde{x} \in X$ we have that $M(x)$ is an invertible linear mapping from X to X, e.g. an automorphism on X. Then

$$f(x) = y \iff x = x + M(x)(y - f(x))$$

(this almost seems like a tautology). We have seen schemes like this before, e.g. we define the map $\phi : X \to X$ with

$$\phi(x) := x + M(x)(y - f(x))$$

We shall, for completeness, remind ourselves about contractions and the Banach fixed-point theorem.

Definition. Let T be an endomorphism in any concrete category $\mathcal{C}$. An object $c \in \mathcal{C}$ is said to be a **fixed point** of T if $T(c) = c$.

Definition. Let T be an endomorphism on a Banach space X. We say that T is a **contraction** if there exists a $q \in [0, 1)$ such that for all $x, y \in X$ we have

$$\|T(x) - T(y)\| \leq q\|x - y\|$$

Lemma. *Contractions are continuous.*

Proof. In normed spaces, endomorphisms are continuous if and only if they are bounded. Take $x \in X$ arbitrarily and $y = 0_X$, then $\|T(x)\| \leq q\|x\|$ e.g. contractions are bounded. □

Lemma. *Fixed points of contractions are unique.*

Proof. Suppose $x, y \in X$ are fixed points. Then

$$\|x - y\| = \|T(x) - T(y)\| \leq q\|x - y\|$$

Then either $\|x - y\| = 0$ or $q \geq 1$, the latter being a contradiction, hence $x - y = 0$. □

Theorem. *Let T be a contraction on a Banach space X. Then T has a unique fixed point.*

Proof. Let $x_0 \in X$. We define the following sequence $(x_n)_{n\in\mathbb{N}}$ recursively by $x_{n+1} := T(x_n)$. Let $\varepsilon > 0$. Let $m \in \mathbb{N}$ be such that

$$q^m \frac{\|x_1 - x_0\|}{1 - q} < \varepsilon$$

which exists due to the fact that $q < 1$, so $q^m \to 0$ as $m \to \infty$. Let $m < n \in \mathbb{N}$. Observe that

$$\begin{aligned}
\|x_n - x_m\| &= \left\| \sum_{i=m}^{n-1} (x_i - x_{i-1}) \right\| \\
&\leq \sum_{i=m}^{n-1} \|x_i - x_{i-1}\| \\
&\leq \|x_1 - x_0\| \sum_{i=m}^{n-1} q^i < \varepsilon
\end{aligned}$$

E.g. x_n is Cauchy, therefore is converges. Consider that then

$$L := \lim_{n\to\infty} x_n = \lim_{n\to\infty} T(x_{n-1}) = T(L)$$

Uniqueness follows from the lemma. □

Corollary. *There exists $q < 1$ such that for the limit $L \in X$ and all $k \in \mathbb{N}^*$*

$$\|x_k - L\| \leq \frac{q^k}{1-q}\|x_1 - x_0\| \qquad \text{(a priori estimate)}$$

and all $k \geq 2$:

$$\|x_k - L\| \leq \frac{q}{1-q}\|x_k - x_{k-1}\| \qquad \text{(a posteriori estimate)}$$

and also

$$\|x_{k+1} - L\| \leq q\|x_k - L\| \qquad \text{(monotonically decreasing error)}$$

Returning to our original setup, we have that if the normed space X is also a complete (e.g. Banach), and ϕ is a contraction, then there exists a unique solution for any initial guess x_0 to $f(x) = y$.

Remark. If $M(x) := (df(x))^{-1}$ for all $x \in X$, and f is Frechet differentiable, then we obtain the Newton-Raphson method again.

Remark. In the special case that $X \subseteq R^n$ for some n and X is nonempty and closed, X is no longer a Banach space, because it is not a vector space, but the proof is furthermore exactly the same, considering all quantities are guaranteed to be in X, and by closedness of X the limit of a sequence in X, if it exists, is in X.

If X is then also convex, and ϕ is continuously partially differentiable, then

$$q \leq \|d\phi\|_\infty := \sup_{x \in X} \|d\phi(x)\|$$

equipping X with a norm on $\mathbb{R}^n$. Remember that this $q > 0$ is the smallest q such that for all $x, y \in X$,

$$\|\phi(x) - \phi(y)\| \leq q\|x - y\|$$

E.g. it suffices to show that $\|d\phi(x)\|$ is a lower bound for q, but this is clear from the definition.

9 One-step methods

Recall from the differential equations course the IVP

$$y'(t) = f(t, y(t))$$

which is supposed to hold for all $t \in [0, T]$ for some $T > 0$ and also, $y(0) = y_0 \in \mathbb{R}$. The goal is to approximate the unique $y : [0, T] \to \mathbb{R}$ that solves this, assuming sufficiently nice f, that is, $f : [0, T] \times \mathbb{R} \to \mathbb{R}$ is continuous, and additionally, Lipschitz in the second argument, that is, there exists $L > 0$ such that for all $x, y \in \mathbb{R}$ and $t \in [0, T]$ we have

$$|f(t, x) - f(t, y)| \leq L|x - y|$$

Therefore, for any $t \in [0, T]$ and $\tilde{t} \in [0, t]$ it must hold that

$$y(t) = y(\tilde{t}) + \int_{\tilde{t}}^{t} f(s, y(s))\, ds$$

9.1 Definition

The goal is to, instead of giving $y(t)$ for all $t \in [0, T]$, to instead sample it at certain points, that is, we are given a partition $P = \{t_0, \ldots, t_n\}$ such that

$$0 = t_0 < t_1 < \ldots < t_n = T$$

Then, we consider that $y(t_0)$ is given, and for all other $i = 0, \ldots, n - 1$ we have

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds$$

Notice that we can then apply 'quadrature rules' to approximate the integral. If $h_i := t_{i+1} - t_i$ is sufficiently small, you can reasonably approximate the integral by the area of a rectangle, for which the height can be chosen to fit the DE. Notice that in the above equation, $y(t_i)$ is already known, so if we approximate the integral, we can find approximations for $y(t_{i+1})$, and recursively continue from there.

We first assume that for all i, $h_i \leq 1$. The idea is to pick a function $\phi : [0, T] \times \mathbb{R} \times \mathbb{R} \times [0, 1] \to \mathbb{R}$ such that we define

$$y(t_{i+1}) \approx y_{i+1} := y_i + h_i \phi(t_i, y_i, y_{i+1}, h_i)$$

It is not immediately clear that this is well-defined. This gives you can equation that might not have a unique solution for y_{i+1}.

Definition. The above scheme is called a **one-step method**, where ϕ is called the **increment function**. If there exists $\psi : [0, T] \times \mathbb{R} \times [0, 1] \to \mathbb{R}$ such that for all t, y, z, h it holds that $\phi(t, y, z, h) = \psi(t, y, h)$, then the one-step method is said to be **explicit**, otherwise it is said to be **implicit**.

Definition. Picking $\phi(t, y, z, h) := f(t, y)$ gives the **explicit Euler method**, e.g. we approximate

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h_i f(t_i, y_i)$$

Definition. Picking $\phi(t, y, z, h) := f(t + h, z)$ gives the **implicit Euler method**, e.g. we approximate

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h_i f(t_{i+1}, y_{i+1})$$

Definition. Picking $\phi(t, y, z, h) := \frac{1}{2}(f(t, y) + f(t + h, z))$ gives the **(implicit) Crank-Nicholson method**, e.g. we take the average of the explicit and implicit Euler increments.

Theorem. *Suppose ϕ, the increment function of some scheme, is Lipschitz in the third argument, that is, there exists some $L > 0$ such that for all $t \in [0,T]$, $y, z, z' \in \mathbb{R}$ and $h \in [0,1]$, it holds that*

$$|\phi(t,y,z,h) - \phi(t,y,z',h)| \leq L|z - z'|$$

and additionally $hL < 1$. Then there exists a unique $y_{i+1} \in [0,1]$ that solves the implicit condition.

Proof. Let t_i, h_i, y_i be given, e.g. pick a certain i / let it be arbitrary and suppose indeed that $h_i L < 1$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be given by $\psi(z) := h_i \phi(t_i, y_i, z, h_i) + y_i$. Consider that

$$|\psi(z) - \psi(z')| = h_i |\phi(\ldots, z, \ldots) - \phi(\ldots, z', \ldots)) \leq h_i L < 1$$

which makes ψ a contraction with unique fixed point y_{i+1}. Hence, implicit schemes can be evaluated with fixed-point iteration. The initial guess might either be y_i or something else that you can compute using an explicit method, which is likely closer. □

Remark. Remember that convergence happens on the order of h^k (assuming fixed grid size), when you take k steps, in the worst case. So you can reasonably determine stopping criteria, consider some tolerance $\varepsilon > 0$, you then require $k \geq \lceil \log(\varepsilon)/\log(h) \rceil$.

9.2 Analysis

We simplify the conditions by assuming that $n+1$ points are evenly spaced, e.g. $T = (n+1)h$ where $h > 0$ is the size of the step. We can then collect the grid points instead in a special partition $\mathcal{T}_h$:

$$\mathcal{T}_h := \{ih \in \mathbb{R} \mid 0 \leq i \leq T/h = n+1\}$$

9.2.1 Consistency

Definition. Let $y : [0,T] \to \mathbb{R}$ be an (exact) solution to

$$y'(t) = f(t, y(t)), \qquad y(0) = y_0$$

Let ϕ be an increment function. We define the *local truncation error* at $t \in [0, T-h]$ by

$$\tau_h(t,y) := \frac{y(t+h) - y(t)}{h} - \phi(t, y(t), y(t+h), h)$$

e.g. you calculate one increment, an approximation of y', and subtract it from the local relative change. We define the *local error* by

$$e_h(t,y) := h\tau_h(t,y) = y(t+h) - y(t) - h\phi(t, y(t), y(t+h), h)$$

One can already see that if ϕ provides any good error, we have

$$e_h(t,y) \approx y(t+h) - y(t) - \int_t^{t+h} f(s, y(s))\, ds = 0$$

The scheme with step ϕ is called **consistent** if for all solutions y

$$\|\tau_h\|_\infty := \max_{t \in \mathcal{T}_h} |\tau_h(t,y)| \overset{h \to 0}{\longrightarrow} 0$$

The scheme is said to be of consistency order $p \in \mathbb{N}$ if there exists $C > 0$ such that for all h,

$$\|\tau_h\| \leq Ch^p$$

Lemma. *Consistency of ϕ with a problem $y'(t) = f(t, y(t))$ is equivalent to, for all $t \in [0,T]$ and $y \in \mathbb{R}$,*

$$\lim_{h \to 0} \phi(t, y, y, h) = f(t, y)$$

Proof. Clear due to

$$f(t,y) = y'(t) = \lim_{h\to 0} \frac{y(t+h) - y(t)}{h}$$

□

Example. Consider the consistency of explicit Euler. By differentiability of y, we know that

$$y(t+h) = y(t) + hy'(t) + \psi(h)$$

where ψ is a function on an appropriate neighborhood of 0 to $\mathbb{R}$ such that $\psi(h)/h \to 0$ as $h \to 0$. Then

$$\tau_h(t,y) = y'(t) + \psi(h)/h - \phi(t, y(t), y(t+h), h) = f(t, y(t)) + \psi(h)/h - f(t, y(t))$$

e.g. $\tau_h(t,y) = \psi(h)/h$ which goes to 0. Thereby, the explicit Euler method is consistent. By Taylor's theorem, for fixed h, we can find $\xi \in [t, t+h]$ such that $\psi(h)/h = y''(\xi)h/2$, which gives us a consistency order of 1. But this is guaranteed if y'' is bounded, which depends on smoothness of f.

9.3 Zero-stability

Definition. Let $\text{lerp} : \mathbb{R} \times \mathbb{R} \times [0,1]$ be defined by

$$(a, b, t) \mapsto (1-t)a + bt$$

called the **linear interpolation function** (short: lerp).

Definition. A time-discrete approximation $y_h : I \to \mathbb{R}$ for some closed interval $I \subseteq \mathbb{R}$ with partition $\mathcal{T}_h = \{t_0, \ldots, t_n\}$ of I is called **piecewise linear** if for all $x \in I$ and corresponding $t_i, t_{i+1} \in \mathcal{T}_h$ such that $x \in [t_i, t_{i+1}]$, we have that $y_h(x) = \text{lerp}\Big(y_i, y_{i+1}, \frac{t_{i+1}-x}{t_{i+1}-t_i}\Big)$, and additionally define $y_i := y_h(t_i)$ for all $0 \leq i \leq n$.

Definition. Let $y_h, \tilde{y}_h$ be piecewise linear time discrete approximations, defined by (for all $i \geq 0$) for some increment function ϕ

$$y_{i+1} = y_i + h\phi(t_i, y_i, y_{i+1}, h)$$
$$\tilde{y}_{i+1} = \tilde{y}_i + h(\phi(t_i, \tilde{y}_i, \tilde{y}_{i+1}, h) + \theta_i)$$

with piecewise linear θ_h. A one step method defined like for y_h is called **zero-stable** if there exists $C > 0$ such that for all $h > 0$ and θ we have

$$\|y_h - \tilde{y}_h\|_\infty \leq C(|y_0 - \tilde{y}_0| + \|\theta_h\|_\infty)$$

Theorem. *Let $\phi : [0,T] \times \mathbb{R} \times \mathbb{R} \times [0,1]$ be an increment function that is Lipschitz in the second and third argument, i.e. there exists $C > 0$ such that for all $t \in [0,T]$ and $h \in [0,1]$, $y, \tilde{y}, z, \tilde{z} \in \mathbb{R}$ we have*

$$|\phi(t, y, z, h) - \phi(t, \tilde{y}, \tilde{z}, h)| \leq C(|y - \tilde{y}| + |z - \tilde{z}|)$$

Then for all $h \in (0, 1/C)$ there exists a $\psi : [0,T] \times \mathbb{R} \times [0,1]$ such that for appropriately defined t_i, y_i, y_{i+1} we have

$$\psi(t_i, y_i, h) = \phi(t_i, y_i, y_{i+1}, h)$$

Proof. Basically a corollary of a previous theorem. We additionally prove that there is a bound on the Lipschitz constant of ψ in the second argument. Consider arbitrary $y_i, \tilde{y}_i$ from two one-step methods. Then

$$\begin{aligned}
|\psi(t, y_i, h) - \psi(t, \tilde{y}_i, h)| &= |\phi(t_i, y_i, y_{i+1}, h) - \phi(t_i, \tilde{y}_i, \tilde{y}_{i+1}, h)| \\
&\leq C(|y_i - \tilde{y}_i| + |y_{i+1} - \tilde{y}_{i+1}|) \\
&= C(|y_i - \tilde{y}_i| + |y_i + h\psi(t_i, y_i, h) - \tilde{y}_i - h\psi(t_i, \tilde{y}_i, h)|) \\
&\leq C(2|y_i - \tilde{y}_i| + h|\psi(t_i, y_i, h) - \psi(t_i, \tilde{y}_i, h)|) \\
&= 2C|y_i - \tilde{y}_i| + Ch|\psi(t_i, y_i, h) - \psi(t_i, \tilde{y}_i, h)| \\
\implies |\psi(t, y_i, h) - \psi(t, \tilde{y}_i, h)| &\leq \frac{2C}{1 - hC}|y_i - \tilde{y}_i|
\end{aligned}$$

□