CMSC666 Numerical Analysis I Final (40%) + Midterms (25%) + Homework (35%), Final: 8am-10am, Dec 17 (Fri)

Introduction to Numerical Anaylsis, J. Stoer, R. Bulirsch (3rd edition) Background (Ch.1,4); Interpolation (Ch.2); Integration (Ch.3); Eigen Problems (Ch.6); Linear System (Ch.8)

---

1 Interpolation

- Sample values at points $\{x_i, y_i\}_{i=1}^n$ (assumes continuity C).
- Sample derivative values, $y_i = f'(x_i), \{x_i, y_i, y'_i\}_{i=1}^n$
- Sample averages over subintervals, $y_{i+\frac{1}{2}} = \frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} f(x)dx$

Linear Interpolation: Polynomial, trigonometric, splines; $\Phi(x; P_1, \cdots, P_n) = \sum_{k=1}^n P_k\Phi(x)$

Non-linear: Rational; $\Phi(x; P_1, \cdots, P_n) = \frac{a_0+a_1x+\cdots+a_nx^n}{1+b_1x+\cdots+b_nx^n}$

1.1 Polynomial Interpolation

Theorem 1.1 (Lagrange formula) *Given $n+1$ distinct points $\{x_0, \cdots, x_n\}$ and $n+1$ associated values $(x_i, y_i), i = 0, \cdots, n$. $\exists$ a unique polynomial $P(x) = a_0 + a_1x + \cdots + a_nx^n$, s.t. (i) $\deg(P) \le n$ and (ii) $P(x_i) = y_i, i = 0, \cdots, n$.*

Proof: (Uniquness) Let P_1 and P_2 satisfy (i) and (ii). Then, $Q = P_1 - P_2$ is a polynomial with degree $\le n$. $Q(x_i) = P_1(x_i) - P_2(x_i) = 0$ for $i = 0, \cdots, n$. Q has $n+1$ roots $\Rightarrow Q = 0$.

(Existence) Let $W(x) = \prod_{i=1}^n (x - x_i)$.

Define **Lagrange polynomial** $L_i(x) = \frac{w(x)}{(x-x_i)w'(x_i)} = \frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}$. Then, $L_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{, if } i = j \\ 0 & \text{, otherwise.}\end{cases}$ Set $P(x) = \sum_{i=0}^n y_iL_i(x)$. $P(x_j) = \sum_{i=0}^n y_iL_i(x_j) = \sum_{i=0}^n y_i\delta_{ij} = y_j$. $\mathcal{QED}$

Error estimates: Given an arbitrary functin f and points $\{x_i\}_{i=0}^n$. Let $P(x) = \sum_{i=0}^n f(x_i)L_i(x)$. How well does P approximate f over $[x_L, x_R]$?

Theorem 1.2 *Assume $f \in C^{n+1}([x_L, x_R])$. Let $\{x_i\}_{i=0}^n$ be distinct with $x_L \le x_0 < x_1 < \cdots < x_n \le x_R$. Let $P(x) = \sum_{i=0}^n f(x_i)L_i(x)$. Then, $\forall x \in [x_L, x_R], \exists\xi \in [x_L, x_R]$, s.t. $f(x) - P(x) = \frac{1}{(n+1)!}W(x)f^{(n+1)}(\xi)$*

Proof: Assume $W(x) \ne 0$ (i.e. $x \ne x_i$). Let $K(x) = \frac{f(x)-P(x)}{W(x)}$. Define $g(t) = f(t) - P(t) - W(t)K(x)$. Then, $g(x) = 0$ and $g(x_i) = 0$ for $i = 0, \cdots, n$. g has $n+2$ 0's and $g(t) \in C^{n+1}([x_L, x_R])$. By Rolle theorem, $g^{(i)}$ has i 0's for $i = 0, \cdots, n+1$. Therefore, $\exists\xi, g^{(n+1)}(\xi) = 0$. But $g^{(n+1)}(t) = f^{(n+1)}(t) - 0 - (n+1)!K(x)$. Set $t = \xi \Rightarrow K(x) = \frac{1}{(n+1)!}f^{(n+1)}(\xi)$. $\mathcal{QED}$

Remark: If f is a polynomial of degree $\le n+1$, then $f^{(n+1)}$ is constant. So the error formula has no unknowns.

C **norm** $\|G\|_{C([x_L,x_R])} = \sup\{|G(x)| \mid x \in [x_L, x_R]\} = \|G\|_C$ (shorthand).

Corollary 1.3

$$\|f - P\|_{C([x_L,x_R])} \le \frac{1}{(n+1)!}\|W\|_C \left\|f^{(n+1)}\right\|_C.$$

Remarks: This bound is sharp when f is a polynomial with degree $\le n+1$. $\|W\|_C$ depends on only $\{x_i\}_{i=0}^n$ while $\left\|f^{(n+1)}\right\|_C$ depends only on f.

Chebyshev Interpolation Given $[x_L, x_R]$ and n, how to choose $\{x_i\}_{i=0}^n$, so that $\|W\|_C$ is minimum?

Remark: It's enough to consider $[x_L, x_R] = [-1, 1]$. If $\{z_i\}_{i=0}^n$ solve the problem for $[-1, 1]$, then $x_i = \frac{x_R+x_L}{2} + \frac{x_R-x_L}{2}z_i$ solve the problem for $[x_L, x_R]$.

Let $x = \cos\theta$, where $\theta \in [0, \pi]$. $\sin\theta = \sqrt{1-x^2} \ge 0$. $\cos(n\theta) + i\sin(n\theta) = (\cos\theta + i\sin\theta)^n = (x - i\sqrt{1-x^2})^n$. Define **Chebyshev polynomial** of degree n,

$$\begin{aligned}
T_n(x) &= \cos(n\theta) = \cos(n\cos^{-1}x) \\
&= x^n - \binom{n}{2}x^{n-2}(1-x^2) + \binom{n}{4}x^{n-4}(1-x^2)^2 - \cdots \\
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
&\vdots \\
T_n(x) &= 2^{n-1}x^n + \text{ lower terms} \\
T_n(-x) &= \begin{cases} T_n(x) & \text{, when } n \text{ even,} \\ -T_n(x) & \text{, when } n \text{ odd.}\end{cases} \\
T_n(\overline{x}_j) &= (-1)^j, \text{ when } \overline{x}_j = \cos(\frac{j\pi}{n}) \\
\|T_n\|_C &= 1
\end{aligned}$$

The solution for our problem is $x_j = \cos((\frac{2j+1}{n})\frac{\pi}{2})$. Then, $T_{n+1}(x_j) = 0$ and x_j are the roots of T_{n+1}. In this case $W(x) = \frac{T_{n+1}(x)}{2^n}$. $\|W\|_{C([-1,1])} = \frac{1}{2^n}$. In general $\|W\|_{C([x_L,x_R])} = \frac{|x_R-x_L|^{n+1}}{2^{2n+1}}$.

Suppose $\exists\{x_j\}_{j=0}^n$, s.t. $\|W\|_C \le \frac{1}{2^n}$. Set $Q(x) = \frac{1}{2^n}T_{n+1}(x) - W(x)$. $Q(\overline{x}_j) = \frac{(-1)^j}{2^n} - W(\overline{x}_j)$. Then, $Q(\overline{x}_j) \ge 0$ when j even and $Q(\overline{x}_j) \le 0$ when j odd. Q have at least $n+1$ roots but $\deg(Q) \le n$. Hence $Q = 0$.

Example: Let $f(x) = \frac{1}{1+25x^2}$ and $e_n = \left\|f - P^{(n)}\right\|_C$.

n	Uniformly pick points e_n	Chebyshev e_n
2	0.96	0.93
4	0.71	0.75
6	0.43	0.56
8	0.25	0.39
10	0.30	0.27
12	0.56	0.18
14	1.07	0.12
⋮	⋮	⋮
20	8.57 (diverges)	0.03

Neville's Algorithm Given $\{(x_i, y_i)\}_{i=0}^n$ and $\overline{x} \in [x_L, x_R]$, how should one compute $P(\overline{x})$ in a way that is stable and fast as possible? Neville is the best for few (one) evaluations.

Let $P_{i_0,\cdots,i_k}(x)$ be a polynomial with degree $\leq k$ and $\forall i = 0, \cdots, n, P_{i_0,\cdots,i_k}(x_i) = y_i$. These partial interpolants can be computed by

$$\begin{aligned} P_{i_0}(x) &= y_{i_0}, \\ P_{i_0,\cdots,i_k}(x) &= \frac{(x - x_{i_0})P_{i_1,\cdots,i_k}(x) + (x_{i_k} - x)P_{i_0,\cdots,i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \end{aligned}$$

The algorithm is pictured as a tableau:

$$\begin{array}{lllll} x_0 & y_0 & & & \\ & & > P_{01}(\overline{x}) & & \\ x_1 & y_1 & & > P_{012}(\overline{x}) & \\ & & > P_{12}(\overline{x}) & & \\ x_2 & y_2 & & \cdots & > P_{0,1,\cdots,n}(\overline{x}) \\ \vdots & \vdots & & & \\ \vdots & \vdots & > P_{n-1,n}(\overline{x}) & & \\ x_n & y_n & & & \end{array}$$

This computes $P(\overline{x})$ with $n(n+1)$ multiplications and $\frac{n(n-1)}{2}$ divisions.

Newton's Interpolation Formula It is better for many evaluation of P since it first computes P, then evaluates $P(\overline{x})$ for many $\overline{x}$'s. Write P as $P(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \cdots + b_n(x - x_0)\cdots(x - x_{n-1}) = b_0 + (x - x_0)(b_1 + (x - x_1)(b_2 + \cdots + (x - x_{n-2})(b_{n-1} + (x - x_{n-1})b_n)\cdots))$. One can evaluate $P(\overline{x})$ by the **Horner scheme** which involves n multiplications to find b_i's.

$$\begin{aligned} y_0 = P(x_0) &= b_0 \\ y_1 = P(x_1) &= b_0 + b_1(x_1 - x_0) \\ y_2 = P(x_2) &= b_0 + b_1(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1) \\ &\vdots \end{aligned}$$

More efficient is the method of divided differences.

$$\begin{aligned} P_{0,\cdots,k}(x) &= P_{0,\cdots,k-1}(x) + y_{0,\cdots,k}(x - x_0)\cdots(x - x_{k-1}) \\ &= P_{1,\cdots,k}(x) + y_{0,\cdots,k}(x - x_1)\cdots(x - x_k) \\ y_{0,\cdots,k} &= \frac{y_{1,\cdots,k} - y_{0,\cdots,k-1}}{x_k - x_0} \end{aligned}$$

Consider the tableau:

$$\begin{array}{lllll} x_0 & y_0 = b_0 & & & \\ & & > y_{01} = b_1 & & \\ x_1 & y_1 & & > y_{012} = b_2 & \\ & & > y_{12} & & \\ x_2 & y_2 & & \cdots & > y_{0,1,\cdots,n} = b_n \\ \vdots & \vdots & & & \\ \vdots & \vdots & > y_{n-1,n} & & \\ x_n & y_n & & & \end{array}$$

L^2 Approximation Let $I = (x_L, x_R)$ be an interval and $w(x) > 0$ be continues weight over I. Define

$$L^2(wdx) = \left\{ f \;\middle|\; \int_I f(x)^2 w(x)dx < \infty \right\}.$$

Define the $L^2(wdx)$ **inner product**

$$(f \mid g) = \int_I f(x)g(x)w(x)dx.$$

Clearly $(f, g) \mapsto (f \mid g)$ is (i) linear in f and g (bilinear), i.e. $(\alpha f_1 + f_2 \mid g) = \alpha(f_1 \mid g) + (f_2 \mid g)$, (ii) commutative, i.e. $(f \mid g) = (g \mid f)$, and (iii) $(f \mid f) \geq 0$ with $(f \mid f) = 0 \Leftrightarrow f = 0$. Define the $L^2(wdx)$ **norm** by $\|f\| = (f \mid f)^{\frac{1}{2}}$.

Theorem 1.4 (Cauchy-Schwarz)

$$\|f\| \, \|g\| \geq |(f \mid g)|.$$

The equality holds $\Leftrightarrow$ f is a scaler multiple of g.

Proof: Let $G = \begin{pmatrix} (f \mid f) & (f \mid g) \\ (f \mid g) & (g \mid g) \end{pmatrix}$. $\forall \alpha, \beta, 0 \leq (\alpha f + \beta g \mid \alpha f + \beta g) = \alpha^2(f \mid f) + 2\alpha\beta(f \mid g) + \beta^2(g \mid g) = \begin{pmatrix} \alpha & \beta \end{pmatrix} G \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. So G is **positive semidefinite** (or non-negative definite). Hence, $0 \leq \det(G) = (f \mid f)(g \mid g) - (f \mid g)^2$. The equality holds $\Leftrightarrow \exists \alpha, \beta$, s.t. $\alpha f + \beta g = 0$. $\mathcal{QED}$

Example: For $I = [-1, 1]$ and $w(x) = 1$, $\int_{-1}^1 f(x)^2 dx \int_{-1}^1 g(x)^2 dx \geq \left(\int_{-1}^1 f(x)g(x)dx \right)^2$.

I	$w(x)$	Orthogonal polynomials
$[-1, 1]$	$\frac{1}{\sqrt{1-x^2}}$	$T_n(x)$ Chebyshev
$[0, \infty]$	e^{-x}	$L_n(x)$ Laguerre
$[-\infty, \infty]$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	$H_n(x)$ Hermite

One can use Cauchy-Schwarz to show

1. $\|f + g\| \leq \|f\| + \|g\|$,
2. $\|\alpha f\| = \alpha \|f\|$,
3. $\|f\| \geq 0$, and
4. $\|f\| = 0 \Leftrightarrow f = 0$.

L^2 polynomial Approximation Suppose $x^n \in L^2(wdx), \forall n \in \mathbb{N} = \{0, 1, 2, \cdots\}$, i.e. $\int_I x^{2n} w(x)dx \leq \infty$. This holds for all examples given before. Let $P^n = \{\text{polynomials with degree} \leq n\} = \{p \mid p(x) = \alpha_0 + \cdots \alpha_n x^n\}$. P^n is a linear sub-space of $L^2(wdx)$ of dimension $n + 1$.

Given $f \in L^2(wdx)$, how to find the polynomial $p \in P^n$ that best approximates f? We want to find $p \in P^n$, s.t.

$$\|f - p\|^2 \leq \inf \left\{ \|f - q\|^2 \mid q \in P^n \right\}. \tag{1.1}$$

Theorem 1.5 *$\exists p$ solves (1.1)* $\Leftrightarrow \forall q \in P^n, (f - p \mid q) = 0$.

Proof: Define **Gram matrix**, $G = \left((x^i \mid x^j) \right)_{i,j=0}^n$

$$= \begin{pmatrix} \int_I wdx & \int_I xwdx & \cdots & \int_I x^n wdx \\ \int_I xwdx & \int_I x^2 wdx & \cdots & \int_I x^{n+1} wdx \\ \vdots & \vdots & \ddots & \vdots \\ \int_I x^n wdx & \int_I x^{n+1} wdx & \cdots & \int_I x^{2n} wdx \end{pmatrix}$$

$\in \mathbb{R}^{(n+1)\times(n+1)}$. $\begin{pmatrix} \alpha_0 & \cdots & \alpha_n \end{pmatrix} G \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix} = \int_I p^2 wdx \geq 0$, where $p = \alpha_0 + \alpha_1 x + \cdots \alpha_n x^n$. The equality holds $\Leftrightarrow p = 0 \Leftrightarrow \begin{pmatrix} \alpha_0 & \cdots & \alpha_n \end{pmatrix} = 0$. Therefore, G is positive semidefinite. $\exists$ an orthogonal matrix O, s.t. $O^T GO$ is a positive diagonal matrix. Using **Gram-Schmidt**, we can construct orthogonal polynomials

$$\begin{aligned} p_0(x) &= \alpha_{00}, \\ p_1(x) &= \alpha_{10} + \alpha_{11}x, \\ &\vdots \\ p_n(x) &= \alpha_{n0} + \alpha_{n1}x + \cdots + \alpha_{nn}x^n, \end{aligned}$$

where $\alpha_{ii} \neq 0$. $(p_i \mid p_j) = 0$ when $i \neq j$.

Clam: $p(x) = \sum_{i=0}^n a_i p_i(x)$, where $a_i = \frac{(f \mid p_i)}{(p_i \mid p_i)}$. Check $(f - p \mid p_j) = (f \mid p_j) - a_j(p_j \mid p_j)$. $\mathcal{QED}$

1.2 Trigonometric Interpolation

Let $T^n = \mathsf{span}\{1, \cos(kx), \sin(kx)\}_{k=1}^n$. Given $f(x) \in L^2$, 2π-periodic, find $S_n f(x) \in T^n$, s.t. $\forall t(x) \in T^n, \|f(x) - S_n f(x)\| \leq \|f(x) - t(x)\|$. The answer is **Fourier expansion**

$$\begin{aligned} S_n f(x) &= \frac{a_0}{2} + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)), \\ \text{where } a_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)dx \text{ and} \\ b_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(kx)dx. \end{aligned}$$

are **Fourier series**.

If $f(x)$ is p-periodic, then $g(x) = f(\frac{p}{2\pi}x)$ is 2π-periodic. If $f(x)$ is defined on $[a, b]$, then $g(x) = f(\frac{b-a}{2\pi}+1)$ is difined on $[0, 2\pi]$. By the transfromations

$$\begin{aligned} e^{ikx} &= \cos(kx) + i\sin(kx), \\ \cos(kx) &= \frac{e^{ikx} + e^{-ikx}}{2}, \text{ and} \\ \sin(kx) &= \frac{e^{ikx} - e^{-ikx}}{2i}, \end{aligned}$$

$T^n = \mathsf{span}\left\{e^{ikx}\right\}_{k=-n}^n$. The Fourier expansion becomes

$$S_\infty f(x) = \sum_{k=-\infty}^{\infty} \hat{f}_k e^{ikx}. \text{ with } \hat{f}_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}dx.$$

$f(x)$ is the phase polynomial with phase x and $\hat{f}_k$ is the Fourier series. The partial sum (truncated expansion) $S_n f(x) = \sum_{k=-n}^n \hat{f}_k e^{ikx}$ is the best L^2 approximation of $f(x)$ among all $t(x) \in T^n$, i.e. $\forall t \in T^n, \|f(x) - S_n f(x)\| \leq \|f(x) - t(x)\|$. Also, $\lim_{n\to\infty} \|f(x) - S_n f(x)\| = \lim_{n\to\infty} \left(\int_0^{2\pi} (f(x) - S_n f(x))^2 dx\right)^{\frac{1}{2}} = 0$.

Derivation of the formula: Equip the linear space T^n with L^2 inner product,

$$(f(x) \mid g(x)) = \int_0^{2\pi} f(x)\overline{g(x)}dx.$$

The complex conjugate is necessary for complex-valued function since L^2 norm $\|f(x)\| = (f(x) \mid f(x))^{\frac{1}{2}} = \left(\int_0^{2\pi} f(x)\overline{f(x)}dx\right)^{\frac{1}{2}} = \left(\int_0^{2\pi} |f(x)|^2 dx\right)^{\frac{1}{2}} \geq 0, \forall f$. Then, for integers j and k,

$$\begin{aligned} (e^{ijx} \mid e^{ikx}) &= \int_0^{2\pi} e^{ijx}\overline{e^{ikx}}dx \\ &= \int_0^{2\pi} e^{i(j-k)x}dx \\ &= \begin{cases} [x]_0^{2\pi} & \text{, if } j = k \\ \left[\frac{e^{i(j-k)x}}{i(j-k)}\right]_0^{2\pi} & \text{, if } j \neq k \end{cases} \\ &= \begin{cases} 2\pi & \text{, if } j = k \\ 0 & \text{, if } j \neq k \end{cases}. \end{aligned}$$

For norm is minimal, $\forall t(x) \in T^n$,

$$\begin{aligned} & \|f(x) - S_n f(x)\| \leq \|f(x) - t(x)\| \\ \Rightarrow\ & f(x) - S_n f(x) \text{ and } t(x) - S_n f(x) \text{ are orthogonal} \\ \Rightarrow\ & (f(x) - S_n f(x) \mid t(x) - S_n f(x)) = 0 \\ \Rightarrow\ & (f(x) - S_n f(x) \mid t(x)) = 0 \\ \Rightarrow\ & (f(x) - S_n f(x) \mid e^{ikx}) = 0 \\ \Rightarrow\ & (f(x) \mid e^{ikx}) = (S_n f(x) \mid e^{ikx}) \\ & = (\sum_{j=-n}^n \hat{f}_j e^{ijx} \mid e^{ikx}) = \sum_{j=-n}^n \hat{f}_j (e^{ijx} \mid e^{ikx}) = 2\pi \hat{f}_k \\ \Rightarrow\ & \hat{f}_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}dx. \end{aligned}$$

Theorem 1.6 (Parseval equality)

$$\|f(x)\|^2 = 2\pi \sum_{k=-\infty}^{\infty} |\hat{f}_k|^2$$

Theorem 1.7 (Bessel inequality)

$$\|S_n f(x)\|^2 = 2\pi \sum_{k=-n}^n |\hat{f}_k|^2 \leq \|f(x)\|^2$$

By Bessel inequality, $\hat{f}_k \to 0$ as $k \to -\infty, \infty$.

Formally, $f'(x) = \sum_{k=-\infty}^{\infty} \hat{f}_k k e^{ikx} i$. If $f'(x) \in L^2$, $\hat{f}_k k i$ is the Fourier coefficient of $f'(x)$. By Bessel inequality, $\lim_{k\to-\infty,\infty} |\hat{f}_k k i|^2 = 0 \Rightarrow |\hat{f}_k| \leq \frac{c}{|k|}$ for some constant c.

Assume $f'(x) \in L^2$, $\hat{f}'_k = \frac{1}{2\pi}\int_0^{2\pi} f'(x)e^{-ikx}dx = \frac{1}{2\pi}\left(\left[f(x)e^{-ikx}\right]_0^{2\pi} + ik\int_0^{2\pi} f(x)e^{-ikx}dx\right) = \hat{f}_k k i$.

Assume $f''(x) \in L^2$. Similarly, $|\hat{f}_k| \leq \frac{c'}{k^2}$. $|f(x) - S_n f(x)| = \left|\sum_{|k|>n} \hat{f}_k e^{ikx}\right| \leq \sum_{|k|>n} |\hat{f}_k| \leq \sum_{|k|>n} \frac{c'}{k^2} \to 0$ as $n \to \infty$.

Let $f(x) = \begin{cases} 1 & \text{, if } x \geq \pi \\ 0 & \text{, if } x < pi \end{cases}$. $\|f(x) - S_n f(x)\| \to 0$ as $n \to \infty$. However, it does not converge at $x = \pi$. This is called **Gibbs phenomenon**. The difference is $0.09(f(\pi^+) - f(\pi^-))$.

Discrete Fourier Series Let $x_v = \frac{2\pi v}{2n+1}$ be support points with $v = -n, \cdots, n$. For any function $f(x)$ defined on $[-\pi, \pi]$, suppose we only know $f(x_v)$. In $l^2(\mathbb{C}^{2n+1}) =$

span $\{\vec{w}_k\}_{k=-n}^n$, $f(x)$ can be discreatized in form of vector $\vec{f} = \begin{pmatrix} f(x_{-n}) \\ \vdots \\ f(x_n) \end{pmatrix}$, where $\vec{w}_k = \begin{pmatrix} e^{ikx_{-n}} \\ \vdots \\ e^{ikx_n} \end{pmatrix}$, The inner product $(\vec{f} \mid \vec{g}) = \sum_{v=-n}^n f(x_v)\overline{g(x_v)}$. Then, the orthogonality of basis is preserved, i.e. $(\vec{w}_j \mid \vec{w}_k) = \begin{cases} 0 & \text{, if } j \neq k \\ 2n+1 & \text{, if } j = k \end{cases}$.

Then, $\hat{f}_k$ can be approximated by

$$\tilde{f}_k = \frac{(\vec{f} \mid \vec{w}_k)}{(\vec{w}_k \mid \vec{w}_k)} = \frac{1}{2n+1} \sum_{v=-n}^n f(x_v) e^{-ikx_v}.$$

The Fourier expansion is $D_n f(x) = \sum_{k=-n}^n \tilde{f}_k e^{ikx}$. $\forall v, D_n f(x_v) = f(x_v)$. $\forall t \in T^n, \left\| \vec{f} - \overrightarrow{D_n f} \right\| \leq \left\| \vec{f} - \vec{t} \right\|$.

$f(x) - D_n f(x) = (f(x) - S_n f(x)) + (S_n f(x) - D_n f(x))$. The first part is truncation error due to cutoff of high frequency components, $|f(x) - S_n f(x)| \leq \frac{c_s}{n^{s-\sigma}}, c_s = \left\| f^{(s)}(x) \right\|$. The second part is discretization error (aliasing error), $S_n f(x) - D_n f(x) = \sum_{k=-n}^n (\hat{f}_k - \tilde{f}_k) e^{ikx}$. $\tilde{f}_k = \frac{1}{2n+1} \sum_{v=-n}^n \left(\sum_{j=-\infty}^\infty \hat{f}_j e^{ijx_v} \right) e^{-ikx_v} = \frac{1}{2n+1} \sum_{j=-\infty}^\infty \hat{f}_j \sum_{v=-n}^n e^{i(j-k)x_v}$. Since $x_v = \frac{2\pi v}{2n+1}$, $\sum_{v=-n}^n e^{i(j-k)x_v} = \begin{cases} 2n+1 & \text{, if } (2n+1)|(j-k) \\ 0 & \text{, otherwise.} \end{cases}$
Then, $\tilde{f}_k - \hat{f}_k = \sum_{j \neq 0} \hat{f}_{k+(2n+1)j}$.

Fast Fourier Transforms Let $n = 2^m$, $w = \frac{2\pi}{n}$, $f_j = f(x_j)$ and $\vec{f} = (f_0, \cdots, f_{n-1})$. Then,

$$\begin{aligned} \tilde{f}_k &= (f_0, f_1, \cdots, f_{n-1})_k \\ &= \frac{1}{n} \sum_{v=0}^{n-1} f_v e^{-ikwv} \\ &= \frac{1}{n} \left(\sum_{v=0}^{n/2} f_{2v} e^{-ikw(2v)} + \sum_{v=0}^{n/2} f_{2v+1} e^{-ikw(2v+1)} \right) \\ &= \frac{1}{2} \left((f_0, f_2, \cdots, f_{n-2})_k + e^{-ikw} (f_1, f_3, \cdots, f_{n-1})_k \right) \end{aligned}$$

The running of the algroithm is $N \log_2 N$.

1.3 Spline Interpolation

One of the simplest is continuous, piecewise linear interpolation (connect the dots). Given data $(x_j, y_j)_{j=0}^n$, $Y(x) = \sum_{j=0}^n y_j T_j(x)$, where $T_j(x)$ is a tent function,

$$T_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & \text{, if } x_{j-1} < x < x_j \\ \frac{x - x_{j+1}}{x_j - x_{j+1}} & \text{, if } x_j \leq x < x_{j+1} \\ 0 & \text{, otherwise.} \end{cases}$$

This is the best second order accurate $\sim (\Delta x)^2$ if $y_i = f(x_i)$ with $f \in C^2$.

The idea of splines is to patch together higher degree polynomials with greater regularity. What is the most regularity we can impose using cubics? Assume $Y(x)$ is piecewise cubic. There are $4n$ unknowns: $2n$ interpolation constraints, $n-1$ continuity of derivative constraints, $n-1$ continuity of second order derivative constraints and setting $Y'(x) = 0$ at x_0, x_n.

Consider the problem

$$\min \left\{ \frac{1}{2} \int_{x_0}^{x_n} (Y''(x))^2 dx \;\middle|\; \forall j, y_j = Y(x_j) \right\}.$$

Using **Lagrange multipliers**, let

$$Q(Y, \vec{\lambda}) = \frac{1}{2} \int_{x_0}^{x_n} (Y''(x))^2 dx - \sum_{j=0}^n \lambda_j (Y(x_j) - y_j).$$

Then, $\forall \tilde{Y} \in C^2$ with $\forall j, \tilde{Y}(x_j) = 0$,

$$\begin{aligned} 0 &= \tilde{Y} \nabla_Y Q \\ &= \frac{d}{ds} Q(Y + s\tilde{Y}, \lambda) \Big|_{s=0} \\ &= \frac{d}{ds} \left(\frac{1}{2} \int_{x_0}^{x_n} (Y''(x) + s\tilde{Y}''(x))^2 dx \right. \\ &\qquad \left. - \sum_{j=0}^n \lambda_j (Y(x_j) + s\tilde{Y}(x_j) - y_j) \right) \Bigg|_{s=0} \\ &= \int_{x_0}^{x_n} Y''(x) \tilde{Y}''(x) dx \\ &= \sum_{j=1}^n \left(\left[Y''(x) \tilde{Y}'(x) \right]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} Y'''(x) \tilde{Y}'(x) dx \right) \\ &= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} Y^{(4)}(x) \tilde{Y}(x) dx \\ &\quad + \sum_{j=1}^n \left[Y''(x) \tilde{Y}'(x) - Y'''(x) \tilde{Y}(x) \right]_{x_{j-1}}^{x_j} \\ &= \sum_{j=1}^n \left(Y''(x_j^-) \tilde{Y}'(x_j) - Y''(x_{j-1}^+) \tilde{Y}'(x_{j-1}) \right) \\ &= Y''(x_n^-) \tilde{Y}'(x_n) - Y''(x_0^+) \tilde{Y}'(x_0) \\ &\quad + \sum_{j=1}^{n-1} \left(Y''(x_j^-) - Y''(x_j^+) \right) \tilde{Y}'(x_j) \end{aligned}$$

$$\Leftrightarrow \begin{cases} Y''(x_n^-) = Y''(x_0^+) = 0 \text{ and} \\ Y''(x_j^-) = Y''(x_j^+) \text{ for } j = 1, \cdots, n \end{cases}$$

$$\Leftrightarrow (*) \begin{cases} Y''(x_n^-) = Y''(x_0^+) = 0 \text{ and} \\ Y'' \text{ is continuous.} \end{cases}$$

Assume condition $(*)$ and let $\tilde{Y}$ be arbitrary.

$$\begin{aligned} 0 &= \frac{d}{ds} Q(Y + s\tilde{Y}, \lambda) \Big|_{s=0} \\ &= \sum_{j=1}^n \left[Y''(x) \tilde{Y}'(x) - Y'''(x) \tilde{Y}(x) \right]_{x_{j-1}}^{x_j} - \sum_{j=0}^n \lambda_j \tilde{Y}(x_j) \\ &= \sum_{j=1}^{n-1} (Y'''(x_j^+) - Y'''(x_j^-) - \lambda_j) \tilde{Y}(x_j) \\ &\quad + (Y'''(x_0^+) - \lambda_0) \tilde{Y}(x_0) - (Y'''(x_n^-) + \lambda_n) \tilde{Y}(x_n) \end{aligned}$$

Then, $\lambda_j = \begin{cases} Y'''(x_0^+) & \text{, if } j=0 \\ -Y'''(x_n^-) & \text{, if } j=n \\ Y'''(x_j^+) - Y'''(x_j^-) & \text{, otherwise.} \end{cases}$

Let $M_0 = M_n = 0, M_j = Y''(x_j)$ and $\Delta_j = x_j - x_{j-1}$. Y'' is piecewise linear. Consider Y'' over $(x_{j-1}, x_j,$

$$\begin{aligned} Y''(x) &= M_j \frac{x - x_{j-1}}{\Delta_j} + M_{j-1}\frac{x_j - x}{\Delta_j} \\ Y(x) &= M_j \frac{(x-x_{j-1})^3}{6\Delta_j} + M_{j-1}\frac{(x_j - x)^3}{6\Delta_j} \\ &\quad + A_j(x - x_{j-1}) + B_j(x_j - x) \\ y_{j-1} = Y(x_{j-1}) &= M_{j-1}\frac{\Delta_j^2}{6} + B_j\Delta_j \\ y_j = Y(x_j) &= M_j \frac{\Delta_j^2}{6} + A_j\Delta_j \\ A_j &= \frac{y_j}{\Delta_j} - M_j\frac{\Delta_j}{6} \\ B_j &= \frac{y_{j-1}}{\Delta_j} - M_{j-1}\frac{\Delta_j}{6} \\ A_j - B_j &= \frac{y_j - y_{j-1}}{\Delta_j} - (M_j - M_{j-1})\frac{\Delta_j}{6} \\ Y'(x_{j-1}) &= -M_{j-1}\frac{\Delta_j}{2} + A_j - B_j \\ Y'(x_j) &= M_j\frac{\Delta_j}{2} + A_j - B_j \end{aligned}$$

Since Y' is continuous,

$$\begin{aligned} -M_j\frac{\Delta_{j+1}}{2} + A_{j+1} - B_{j+1} &= M_j\frac{\Delta_j}{2} + A_j - B_j \\ \frac{\Delta_j}{6}M_{j-1} + \frac{\Delta_j + \Delta_{j+1}}{3}M_j + \frac{\Delta_{j+1}}{6}M_{j+1} &= \frac{y_{j+1}-y_j}{\Delta_{j+1}} - \frac{y_j - y_{j-1}}{\Delta_j} \end{aligned}$$

Let $c_j = \dfrac{\frac{y_{j+1}-y_j}{\Delta_{j+1}} - \frac{y_j - y_{j-1}}{\Delta_j}}{\frac{\Delta_j + \Delta_{j+1}}{2}}$

$$= \frac{\Delta_j}{3(\Delta_j + \Delta_{j+1})}M_{j-1} + \frac{2}{3}M_j + \frac{\Delta_{j+1}}{3(\Delta_j + \Delta_{j+1})}M_{j+1}$$

Assume $\forall j, \Delta_j = \Delta$ (uniform intervals).

$$\frac{1}{6}M_{j-1} + \frac{2}{3}M_j + \frac{1}{6}M_{j+1} = \frac{y_{j+1} - 2y_j + y_{j-1}}{\Delta^2}$$

$$\begin{pmatrix} \frac{2}{3} & \frac{1}{6} & & \\ \frac{1}{6} & \ddots & \ddots & \\ & \ddots & & \frac{1}{6} \\ & & \frac{1}{6} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{pmatrix}$$

This tridiagonal systems can be solved by **Gaussian elimination**.

Theorem 1.8 *Let $f \in C^2([x_L, x_R])$ and $y_j = f(x_j)$ where $j = 0, \cdots, n$ and $x_L = x_0 < x_1 < \cdots < x_n = x_R$. Let Y be the continuous, piecewise linear spline. Then, $\|f - Y\|_C \le k\|f''\|_C \Delta^2$, where $\Delta = \max\{\Delta_j\}_j$ and k is a constant.*

Theorem 1.9 *Let $f \in C^4([x_L, x_R])$. Let Y be the cubic spline. Then, $\|f - Y\|_C \le k_3 \left\|f^{(4)}\right\|_C \Delta^4$.*

2 Numerical Quadrature

Given integrand f, evaluate $\int_{x_L}^{x_R} f(x)dx$. Let $x_L = x_0 < x_1 < \cdots < x_n = x_R$ be a partition of $[x_L, x_R]$. Evaluate $y_i = f(x_i)$ and build an interpolant $P(x)$, then evaluate $\int_{x_L}^{x_R} P(x)dx$.

2.1 Newton-Coles formula

If P is the polynomial interpolation of degree n of (x_i, y_i), then $P(x) = \sum_{i=0}^n y_i L_i(x) \Rightarrow \int_{x_L}^{x_R} f(x)dx \approx \sum_{i=0}^n y_i w_i$, where the weights $w_i = \int_{x_L}^{x_R} L_i(x)dx$. Since $\sum_{i=0}^n L_i(x) = 1$, $\sum_{i=0}^n w_i = x_R - x_L$. It is not so clear that each $w_i > 0$. Note w_i's depend only on $\{x_i\}_{i=0}^n$ and $[x_L, x_R]$, but not f.
Example: For $n = 2$ and $\Delta_i = \Delta$, $\int_{x_L}^{x_R} f(x)dx \approx (\frac{1}{3}y_0 + \frac{4}{3}y_1 + \frac{1}{3}y_2)\Delta$ (Simpson's rule).

Trapezoidal rule If P is the linear spline, $\int_{x_L}^{x_R} f(x)dx \approx \sum_{i=1}^n \frac{y_i + y_{i-1}}{2}\Delta_i = \frac{\Delta_1}{2}y_0 + \sum_{i=1}^{n-1}\frac{\Delta_i + \Delta_{i+1}}{2}y_i + \frac{\Delta_n}{2}y_n$, where $\Delta_i = x_i - x_{i-1}$. Then, $w_i = \begin{cases} \frac{\Delta_1}{2} & \text{, if } i = 0 \\ \frac{\Delta_i + \Delta_{i+1}}{2} & \text{, if } i = 1, \cdots, n-1 \\ \frac{\Delta_n}{2} & \text{, if } i = n \end{cases}$. Note $\sum_{i=0}^n w_i = x_R - x_L$ and $w_i > 0$.

Error estimate Given $\{x_i\}_{i=0}^n$ and $\{w_i\}_{i=0}^n$, how accurate $\sum_{i=0}^n y_i w_i$? Let $\Delta = \max\{\Delta_i\}$. We estimate the error $E(f)$ when approximating $I(f) = \int_{x_L}^{x_R} f(x)dx$ by the numerical quadrature $Q(f) = \sum_{i=0}^n y_i w_i$. Clearly, $E(f) = I(f) - Q(f)$. $|E(f)| \le M\left\|f^{(k)}\right\|_C \Delta^m$ for some constant M depended only on $[x_L, x_R]$, or in the asympotic form $\left|\int_{x_L}^{x_R} f(x)dx - \sum_{i=0}^n y_i w_i\right| \sim c\Delta^m$ for c depended on f and $[x_L, x_R]$.

Right-hand rule Let f be continous and non-decreasing. Then, $\sum_{i=1}^n y_i\Delta_i - \int_{x_L}^{x_R} f(x)dx \le \sum_{i=1}^n (y_i - y_{i-1})\Delta_i \le \Delta\sum_{i=1}^n (y_i - y_{i-1}) = \Delta(y_n - y_0) = \Delta(f(x_n) - f(x_0)) \le (x_R - x_L)\|f'\|_C \Delta$.

Consider $\frac{1}{2}\int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1})f''(x)dx = \frac{1}{2}\int_{x_{i-1}}^{x_i} (x - x_{i-1} - x_i + x)f'(x)dx = \frac{\Delta_i}{2}(f(x_{i-1}) + f(x_i)) - \int_{x_{i-1}}^{x_i} f(x)dx$. Therefore, $E(f) = \sum_{i=1}^n \frac{1}{2}\int_{x_{i-1}}^{x_i}(x_i - x)(x - x_{i-1})f''(x)dx = \int_{x_L}^{x_R} K(x)f''(x)dx$, where $K(x) = \frac{1}{2}(x_i - x)(x - x_{i-1}) \ge 0$ for $x_{i-1} \le x \le x_i$. One can estimate this error as

$$\begin{aligned} |E(f)| &\le \int_{x_L}^{x_R} K(x)dx\,\|f''\|_{C([x_L,x_R])} \\ \text{or} \quad |E(f)| &\le \|K(x)\|_{C([x_L,x_R])}\int_{x_L}^{x_R}|f''(x)|dx \\ \text{or} \quad |E(f)| &\le \left(\int_{x_L}^{x_R}(K(x))^2dx\right)^{\frac{1}{2}}\left(\int_{x_L}^{x_R}|f''(x)|^2dx\right)^{\frac{1}{2}} \\ &\cdots \end{aligned}$$

Pick the best given what you know about f''.

$\frac{1}{2}\int_{x_{i-1}}^{x_i}(x_i-x)(x-x_{i-1})dx = \frac{\Delta_i^3}{12}$. Hence, $\int_{x_L}^{x_R} K(x)dx = \sum_{i=1}^{n}\frac{\Delta_i^3}{12} \leq \frac{x_R-x_L}{12}\Delta^2$, where $\Delta = \max\{\Delta_i\}$. Then, $|E(f)| \leq \frac{x_R-x_L}{12}\Delta^2 \|f''\|_{C([x_L,x_R])}$ for the first inequality.

For the second inequality, $\|K(x)\|_{C([x_{i-1},x_i])} = \max\left\{\frac{1}{2}(x_i-x)(x-x_{i-1}) \mid x \in [x_{i-1},x_i]\right\} = \frac{1}{2}(x_i - \frac{x_i+x_{i-1}}{2})(\frac{x_i+x_{i-1}}{2} - x_{i-1}) = \frac{1}{8}\Delta_i^2$. Hence, $|E(f)| \leq \frac{1}{8}\Delta^2\int_{x_L}^{x_R}|f''(x)|dx$.

Theorem 2.1 (Peano's Kernel) *Suppose $Q(f)$ integrates ploynomials of degree m or less exactly for $f(x) \in C^m_{([x_L,x_R])}$, i.e. $\int_{x_L}^{x_R}|f^{(m+1)}(x)|dx < \infty$. Then, $\exists K(x)$, s.t. $E(f) = \int_{x_L}^{x_R}K(x)f^{(m+1)}(x)dx$.*

Proof: Let $\Phi^k(x) = \sum_{i=0}^{n}\frac{w_i}{k!}(x-x_i)^k H(x-x_i) - \frac{(x-x_L)^{k+1}}{(k+1)!}$, where $H(x) = \begin{cases} 1 & \text{, if } x \geq 0 \\ 0 & \text{, otherwise} \end{cases}$ is the Heavisick function. Note that $\frac{d\Phi^k(x)}{dx} = \Phi^{k-1}(x)$, $\Phi^k(x_L) = 0$ and $\Phi^k(x_R) = \sum_{i=0}^{n}\frac{w_i}{k!}(x_R-x_i)^k - \frac{(x_R-x_L)^{k+1}}{(k+1)!} = Q(\frac{(x_R-x)^k}{k!}) - I(\frac{(x_R-x)^k}{k!}) = 0$ for $k \leq m$.

Then, $E(f) = \int_{x_L}^{x_R}\left(\sum_{i=0}^{n}w_i\delta(x-x_i)-1\right)f(x)dx = \int_{x_L}^{x_R}\frac{d\Phi^0(x)}{dx}f(x)dx = \int_{x_L}^{x_R}(-1)^{m+1}\Phi^m(x)f^{(m+1)}(x)dx$. $\mathcal{QED}$

For the trapezoidal rule, $K(x) = \frac{1}{2}(x_i-x)(x-x_{i-1})$ for $x \in [x_{i-1},x_i]$. The key to showing this is

$$\begin{aligned} E_i(f) =& \frac{\Delta_i}{2}(f(x_{i-1})+f(x_i)) - \int_{x_{i-1}}^{x_i}f(x)dx \\ =& \int_{x_{i-1}}^{x_i}\frac{1}{2}(x_i-x)(x-x_{i-1})f''(x)dx \\ =& \frac{\Delta_i^2}{12}(f'(x_i)-f'(x_{i-1})) + \int_{x_{i-1}}^{x_i}(\psi_2(x)-\frac{\Delta_i^2}{12})f''(x)dx, \end{aligned}$$

where $\psi_2(x) = K(x)$ and $\|K(x)\| = \frac{\Delta_i^2}{12}$. Let $\psi_4(x) = \frac{1}{24}(x_i-x)^2(x-x_{i-1})^2 \geq 0$. $\psi_4''(x) = -\psi_2(x) + \frac{\Delta_i^2}{12}$ and $\psi_4(x_{i-1}) = \psi_4'(x_{i-1}) = \psi_4'(x_i) = \psi_4(x_i) = 0$. Then,

$$E_i(f) = \frac{\Delta_i^2}{12}(f'(x_i)-f'(x_{i-1})) - \int_{x_{i-1}}^{x_i}\psi_4(x)f^{(4)}(x)dx.$$

Similarly, we can find $\psi_6(x) \geq 0$, s.t. $\psi_6(x) = -\psi_4(x) + \frac{\Delta_i^4}{30(4!)}$, where $\frac{\Delta_i^4}{30(4!)}$ is the mean of $\psi_4(x)$. So,

$$\begin{aligned} E_i(f) =& \frac{\Delta_i^2}{12}(f'(x_i)-f'(x_{i-1})) \\ & -\frac{\Delta_i^4}{30(4!)}(f'''(x_i)-f'''(x_{i-1})) + \int_{x_{i-1}}^{x_i}\psi_6(x)f^{(6)}(x)dx \end{aligned}$$

The last term can be bounded by $\int_{x_{i-1}}^{x_i}\psi_6(x)dx\,\|f^{(6)}(x)\|_{C[x_{i-1},x_i]} = \frac{\Delta_i^7}{42(6!)}\|f^{(6)}(x)\|_C$.

Consider the trapezoidal rule with uniform subintervals. Then, $E(f) = \frac{\Delta_i^2}{12}(f'(x_R)-f'(x_L)) - \frac{\Delta^4}{30(4!)}(f'''(x_R)-f'''(x_L)) + e_6(f)$, where $|e_6(f)| \leq \frac{\Delta_i^6}{42(6!)}(x_R-x_L)\|f^{(6)}(x)\|_C = O(\Delta^6)$.

Euler-Maclaurin formula For $f \in C^{2m+2}$,

$$E(f) = \sum_{j=1}^{m}\frac{B_{2j}}{(2j)!}\Delta^{2j}(f^{(2j-1)}(x_R)-f^{(2j-1)}(x_L)) + e_{2m+2}(f)$$

with $e_{2m+2}(f) \leq \frac{B_{2m+2}}{(2m+2)!}\Delta^{2m+1}(x_R-x_L)\|f^{(2m+2)}(x)\|_C$, where B_{2j} are Bernoulli numbers. $B_2 = \frac{1}{6}, B_4 = \frac{-1}{30}, B_6 = \frac{1}{42}, B_2 = \frac{-1}{30}$. The formula is an asymptotic expension.

Suppose f is periodic and $[x_L,x_R]$ is a multiple of the period. The formula becomes $|E(f)| = |E_{2m+2}(f)| \sim O(\Delta^{2m+2})$. The trapezoidal rule has spectral accuracy for $f \in C^\infty$ (i.e. it converges faster than any Δ^{2m+2}).

Extrapolation & Rombery Intergration Let $Q(f)$ denote the quadrature by the trapezoidal rule with uniform subintervals of length $\Delta = \frac{x_R-x_L}{n}$. The Euler-Maclaurin formula gives $Q_\Delta(f) = I(f) + \alpha_2\Delta^2 + \alpha_4\Delta^4 + \cdots + \alpha_{2m}\Delta^{2m} + O(\Delta^{2m+2})$. Suppose n is even. $Q_{2\Delta}(f) = I(f) + 4\alpha_2\Delta^2 + 16\alpha_4\Delta^4 + \cdots + 2^{2m}\alpha_{2m}\Delta^{2m} + O(\Delta^{2m+2})$. Then, $\frac{4Q_\Delta(f)-Q_{2\Delta}(f)}{3} = I(f) + 4\alpha_4\Delta^4 + \cdots + \alpha_{2m}\Delta^{2m} + O(\Delta^{2m+2})$, which is 4^{th} order. What scheme is this? $\frac{4}{3}Q_\Delta(f) - \frac{1}{3}Q_{2\Delta}(f) = \frac{4}{3}(\frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n))\Delta - \frac{1}{3}(\frac{1}{2}f(x_0) + f(x_2) + f(x_4) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n))2\Delta$, which is **Simpson's rule** (or Newton-Coles for $n = 2$).

Suppose n is divisible by 4. We look for a combination of $Q_\Delta(f), Q_{2\Delta}(f), Q_{4\Delta}(f)$ that eliminate Δ^2 and Δ^4 terms.

$$\begin{aligned} Q_\Delta(f) &= I(f) + \alpha_2\Delta^2 + \alpha_4\Delta^4 + O(\Delta^6) \\ Q_{2\Delta}(f) &= I(f) + 4\alpha_2\Delta^2 + 16\alpha_4\Delta^4 + O(\Delta^6) \\ Q_{4\Delta}(f) &= I(f) + 16\alpha_2\Delta^2 + 256\alpha_4\Delta^4 + O(\Delta^6) \end{aligned}$$

Then, $\frac{64Q_\Delta(f)-20Q_{2\Delta}(f)+Q_{4\Delta}(f)}{45} = I(F) + O(\Delta^6)$, which is **Milie's rule**.

$Q_\Delta(f)$
  $> \frac{4Q_\Delta(f)-Q_{2\Delta}(f)}{3} = S_\Delta$
$Q_{2\Delta}(f)$ $> \frac{16S_\Delta - S_{2\Delta}}{15}$
  $> \frac{4Q_{2\Delta}(f)-Q_{4\Delta}(f)}{3} = S_{2\Delta}$
$Q_{4\Delta}(f)$

When n is divisible by 6, we can use $Q_\Delta(f), Q_{2\Delta}(f), Q_{3\Delta}(f), Q_{6\Delta}(f)$ to eliminate $\Delta^2, \Delta^4, \Delta^6$ terms. This leads to **Weddle's rule** (or Newton-Coles for $n = 6$), $\frac{41}{140}f(x_0) + \frac{216}{140}f(x_1) + \cdots$.

Romberg considers $n = 2^m$. Let T_k be the trapezoidal rule with k uniform subintervals and M_k be the mid-point rule with k uniform subintervals. Then, $T_1 = \frac{1}{2}(f(x_L)+f(x_R))(x_R-x_L)$, $M_1 = f(\frac{x_L+x_R}{2})(x_R-x_L)$, $T_2 = \frac{1}{2}(T_1+M_1)$ and $T_{2k} = \frac{1}{2}(T_k+M_k)$. Let $T_{2k}^{(l)} = \frac{4^l T_{2k}^{(l-1)} - T_k^{(l-1)}}{4^l - 1}$ be the l^{th} level extrapolent. We have **Neville's algorithm**

T_1
  $> T_2^{(1)} = \frac{4T_2-T_1}{3}$
T_2 $> T_4^{(2)} = \frac{16T_4^{(1)}-T_2^{(1)}}{15}$
  $> T_4^{(1)} = \frac{4T_4-T_2}{3}$ $> T_8^{(3)} = \frac{64T_8^{(2)}-T_4^{(2)}}{63}$
T_4 $> T_8^{(2)} = \frac{16T_8^{(1)}-T_4^{(1)}}{15}$
  $> T_8^{(1)} = \frac{4T_8-T_4}{3}$
T_8

2.2 Gaussian intergration

Let $-\infty \le a < b \le \infty$ and $f(x)$ be a function defined on (a,b). We consider integrals of the form $I(f) = \int_a^b f(x)w(x)dx$, where $w(x) > 0$ is a weight function. Then, f continuous, $f \ge 0$ and $\int_a^b f(x)w(x)dx = 0 \Rightarrow f(x) = 0$. Assume $\int_a^b (f(x))^2 w(x)dx \le \infty$.

Orthogonal polynomials $1, x, x^2, \cdots$ are linearly independent $\Rightarrow 1, x, \cdots, x^n$ are linearly independent for each n. If we apply **Gram-Sehnuld** to $1, x, x^2, \cdots$, we get a sequence of orthogonal polynomials $\phi_0, \phi_1, \cdots$, s.t. $(\phi_m, \phi_n)_2 = \int_a^b \phi_m(x)\phi_n(x)w(x)dx = \delta_{mn}$. ϕ_n is a polynomial of degree n. ϕ_n are uniquely determined by a, b, w if the coefficient of x^n is positive. $\phi_0, \phi_1, \cdots$ are the orthogonal polynomials with respect to w. These polynomials have many properties:

1. $\int_a^b \phi_n(x)p(x)w(x)dx = 0$ if p is a polynomial with degree $< n$.
2. $\phi_n(x)$ has n simple zeros in (a,b) for $n \ge 1$.

Proof: Suppose $\phi_n(x)$ does not change sign on (a,b). Then, $\int_a^b \phi_n(x)w(x)dx = 0$ since $\phi_n \perp \phi_0 \Rightarrow \phi_n(x) = 0$, which leads to a contradiction. Therefore, $\phi_n(x)$ has at least one zero on (a,b).

Let $x_1, \cdots, x_r$ be the zero of odd multiplicity of $\phi_n(x)$ on (a,b) and suppose $r < n$. Then, $\phi(x) = \phi_n(x)(x - x_1)\cdots(x - x_r)$ does not change sign on (a,b). But $\int_a^n \phi_n(x)(x-x_1)\cdots(x-x_r)w(x)dx = 0$ since $\phi_n \perp \phi_r \Rightarrow \phi_n(x)(x-x_1)\cdots(x-x_r) = 0 \Rightarrow \phi_n = 0$. It leads to a contradiction. Therefore $r = n$. $\mathcal{QED}$

Examples:

1. **Legendre polynomial**, $P_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}(x^2-1)^n$ for $(a,b) = (-1,1)$ and $w(x) = 1$. The leading coefficient is 1 and $\|P_n\| = \frac{2^n(n!)^2}{(2n)!}\sqrt{\frac{2}{2n+1}}$. P_n are orthogonal and are ϕ_n, to within a constant multiple. $\phi_0(x) = \sqrt{\frac{1}{2}}, \phi_1(x) = \sqrt{\frac{3}{2}}x, \phi_2(x) = \frac{1}{2}\sqrt{\frac{5}{2}}(3x^2-1)$ and $\phi_n(x) = \frac{P(x)}{\|P\|}$. The leading coefficient is $\frac{1}{\|P\|}$.
2. **Chehyshev polynomial**, $T_n(x) = \cos(n\cos^{-1}x)$ for $(a,b) = (-1,1)$ and $w(x) = \frac{1}{\sqrt{1-x^2}}$. These are mutually orthogonal and are ϕ_n, to within a constant multiple. We know that their zeros are in $(-1,1)$, $\|T_n\|^2 = \frac{\pi}{2}$ and the leading coefficient is 2^{n-1}. $\phi_n(x) = \frac{T_n(x)}{\|T_n\|} = \sqrt{\frac{2}{\pi}}\cos(n\cos^{-1}x)$ and the leading coefficient is $2^{n-1}\sqrt{\frac{2}{\pi}}$.
3. **Laguerre polynomial**, $L_n(x) = \frac{e^x}{n!}\frac{d^n}{dx^n}(x^n e^{-x})$, for $(a,b) = (0,\infty)$ and $w(x) = e^{-x}$. $L_0(x) = 1, L_1(x) = -x+1, L_2(x) = x^2-4x+2, L_3(x) = -x^3-19x^2-18x+6$. Then, $(L_m, L_n) = \int_0^\infty L_n(x)L_m(x)e^{-x}dx = \delta_{mn}(m!)(n!)$. $\phi_n(x) = \frac{(-1)^n L_n(x)}{n!}$, the leading coefficient is $\frac{(-1)^n}{n!}$.
4. **Hermite polynomial**, $H_n(x) = e^{x^2}\frac{d^n}{dx^n}e^{-x^2}$ for $(a,b) = (-\infty,\infty)$ and $w(x) = e^{-x^2}$. $H_0(x) = 1, H_1(x) = 2x, H_2(x) = 4x^2-2, H_3(x) = 8x^3-12x, H_4(x) = 16x^4-48x^2+12$. $\|H_n\|^2 = 2^n n!\sqrt{\pi}$, the leading coefficient is 2^n and $\phi_n(x) = \frac{H_n(x)}{(2^n n!\sqrt{\pi})^{\frac{1}{2}}}$.

Gaussian rule For ≥ 1, let $x_1, \cdots, x_n$ be the zeros of $\phi_n(x)$ and P_{n-1} interpolate $f(x)$ at $x_1, \cdots, x_n$. Then, we approximate $I(f) = \int_a^b f(x)w(x)dx \approx \int_a^b P_{n-1}(x)w(x)dx = \int_{x_{i-1}}^{x_i}\sum_{i=1}^n f(x_i)L_i(x)w(x)dx = \sum_{i=1}^n f(x_i)w_i$, where $L_i(x) = \prod_{j\ne i}\frac{x-x_j}{x_i-x_j}$ and $w_i = \int_{x_{i-1}}^{x_i} L_i(x)w(x)dx$.

Theorem 2.2 *A Gaussian rule of order n is exact on polynomials of degree $\le 2n-1$.*

Proof: Suppose p is a polynomial of degree $\le 2n-1$. By long division, $p(x) = q(x)\phi_n(x) + r(x)$ with $\deg q \le n-1$ and $\deg r < n$. Then, $\int_a^b p(x)w(x)dx = \int_a^b (q(x)\phi_n(x) + r(x))w(x)dx = \int_a^b r(x)w(x)dx = \sum_{i=1}^n r(x_i)w_i = \sum_{i=1}^n (p(x_i) - q(x_i)\phi_n(x_i))w_i = \sum_{i=1}^n p(x_i)w_i$ since $\phi_n(x_i) = 0$. $\mathcal{QED}$

Order of k or degree at precision k if an integration rule is exact on polynomials of degree $\le k$ but not higher degree polynomials.

Examples:

1. **Legendre:** For $f(x) = x^n$,

$$\int_{-1}^1 x^n dx = \left[\frac{x^{n+1}}{n}\right]_{-1}^1 = \begin{cases} \frac{2}{n} & \text{, if } n \text{ is even} \\ 0 & \text{, if } n \text{ is odd.} \end{cases}$$

For $n = 1$, $\phi_1(x)$ has a root $x_1 = 0$. $w_1 = \int_{-1}^1 1dx = 2$. Then, $\int_{-1}^1 f(x)dx \approx w_1 f(0) = 2f(0)$ (midpoint rule) is exact on polynomials with degree $\le 2\cdot 1 - 1 = 1$ (linears).

For $n = 2$, the roots of $\phi_2(x)$ are $x_1 = -\frac{1}{\sqrt{3}}, x_2 = -\frac{1}{\sqrt{3}}$. $\int_{-1}^1 f(x)dx \approx w_1 f(-\frac{1}{\sqrt{3}}) + w_2 f(\frac{1}{\sqrt{3}})$ is exact for degree $\le 2\cdot 2 - 1 = 3$. For $f(x) = 1$, $2 = w_1 + w_2$. For $f(x) = x$, $0 = -\frac{w_1}{\sqrt{3}} + \frac{w_2}{\sqrt{3}}$. Then, $w_1 = w_2 = 1$.

For $n = 3$, the roots of $\phi_3(x)$ are $x_1 = -\sqrt{\frac{3}{5}}, x_2 = 0, x_3 = \sqrt{\frac{3}{5}}$. $\int_{-1}^1 f(x)dx \approx w_1 f(-\sqrt{\frac{3}{5}}) + w_2 f(0) + w_3 f(\sqrt{\frac{3}{5}})$ is exact for degree $\le 2\cdot 3 - 1 = 5$. $\begin{cases} w_1 + w_2 + w_3 = 2 \text{ for } f(x) = 1, \\ -w_1 + w_3 = 0 \text{ for } f(x) = x, \\ \frac{3}{5}w_1 + \frac{3}{5}w_3 = \frac{2}{3} \text{ for } f(x) = x^2. \end{cases}$ Then, $w_1 = w_3 = \frac{5}{9}$ and $w_2 = \frac{8}{9}$.

2. **Chehyshev:** For $n = 1$, $T_1(x)$ has a root $x_1 = 0$. $\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}}dx \approx w_1 f(0)$ is exact for degree ≤ 1. For $f(x) = 1$, $w_1 = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}}dx = \left[\sin^{-1}x\right]_{-1}^1 = \pi$.
3. **Laguerre:** For $n = 2$, the roots of $L_2(x)$ are $x_1 = 2 - \sqrt{2}$ and $x_2 = 2 + \sqrt{2}$. $\int_0^\infty f(x)e^{-x}dx \approx w_1 f(2-\sqrt{2}) + w_2 f(2-\sqrt{2})$. For $f(x) = 1$, $w_1 + w_2 = \int_0^\infty e^{-x}dx = [-e^{-x}]_0^\infty = 1$. For $f(x) = x$, $(2-\sqrt{2})w_1 + (2+\sqrt{2})w_2 = \int_0^\infty xe^{-x}dx = [-xe^{-x}]_0^\infty + \int_0^\infty xe^{-x}dx = 1$. Then, $w_1 = \frac{2+\sqrt{2}}{4}$ and $w_2 = \frac{2-\sqrt{2}}{4}$.
4. **Hemite:** For $n = 1$, $H_1(x)$ has a root $x_1 = 0$. $\int_{-\infty}^\infty f(x)e^{-x^2}dx \approx w_1 f(0) = \sqrt{\pi}f(0)$.

For $n = 2$, the roots of $H_2(x)$ are $x_1 = -\frac{1}{\sqrt{2}}$ and $x_2 = \frac{1}{\sqrt{2}}$. $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx w_1 f(-\frac{1}{\sqrt{2}}) + w_2 f(\frac{1}{\sqrt{2}}) = \frac{\sqrt{\pi}}{2}\left(f(-\frac{1}{\sqrt{2}}) + f(\frac{1}{\sqrt{2}})\right)$ since $w_1 + w_2 = \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$ and $-\frac{w_1}{\sqrt{2}} + \frac{w_2}{\sqrt{2}} = \int_{-\infty}^{\infty} xe^{-x^2}dx = \left[-\frac{e^{-x^2}}{2}\right]_{-\infty}^{\infty} = 0$.

Theorem 2.3 (Mean value theorem for integral) *$\int_a^b f(x)dx = f(\xi)(b-a)$ for $\xi \in (a,b)$. More generally, $\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx$ if $g(x)$ is at one sign.*

Error estimate Suppose $f \in C^{2n}(a,b)$. Let h be a polynomial with degree $\leq 2n-1$ that interpolates f at $x_1, x_1, x_2, x_2, \cdots, x_n, x_n$, i.e. $h(x_i) = f(x_i)$ and $h'(x_i) = f'(x_i)$. Then, $\int_a^b h(x)w(x)dx = \sum_{i=1}^n w_i h(x_i) = \sum_{i=1}^n w_i f(x_i)$. $f(x) - h(x) = f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, x]\prod_{i=1}^n (x - x_i)^2 \Rightarrow E_n(f) = \int_a^b f(x)w(x)dx - \sum_{i=1}^n w_i f(x_i) = \int_a^b (f(x) - h(x))w(x)dx = \int_a^b f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, x]\prod_{i=1}^n (x - x_i)^2 w(x)dx = f[x_1, x_1, x_2, x_2, \cdots, x_n, x_n, \xi]\int_a^b \prod_{i=1}^n (x - x_i)^2 w(x)dx = \frac{f^{(2n)}(\eta)}{(2n)!}\int_a^b \prod_{i=1}^n (x - x_i)^2 w(x)dx = \frac{f^{(2n)}(\eta)}{(2n)!A_n^2}\int_a^b \phi_n^2(x)w(x)dx = \frac{f^{(2n)}(\eta)}{(2n)!A_n^2}$ for some $\xi, \eta \in (a,b)$, where A_n is the leading coefficient of $\phi_n(x)$.
Example: For **Gauss-Legendre** formula, $A_n = \frac{(2n)!}{2^n(n!)^2}\sqrt{\frac{2n+1}{2}}$. Then, $E_n(f) = \frac{2^{2n+1}(n!)^4}{(2n+1)((2n)!)^3}f^{(2n)}(\eta)$.

Integrands with singularities Quadrature rule generally work well (or best) if the integrand is smooth, i.e. it has several derivatives of modulate size (huge $\frac{1}{1+x^2}$). If it is not the case, sometimes the integral on (a,b) can be subdivided by $a = a_0 < a_1 < \cdots < a_m = b$ in such a way that the integral is smooth on each $[a_{i-1}, a_i]$ and continuous. $\int_a^b f(x)dx = \sum_{i=1}^m \int_{a_{i-1}}^{a_i} f(x)dx$.
Examples: Let $f(x) = \sqrt{x}\sin x$. $f'(x) = \frac{\sin x}{2\sqrt{x}} + \sqrt{x}\cos x$. $f''(x) = \frac{\cos x}{2\sqrt{x}} - \frac{\sin x}{4x^{\frac{3}{2}}} - \sqrt{x}\sin x$. Let $t = \sqrt{x}$, $dx = 2tdt$. $\int_0^1 \sqrt{x}\sin x dx = \int_0^1 2t^2 \sin t^2 dt$.

Or $\int_0^1 \sqrt{x}\sin x dx = \int_0^\xi \sqrt{x}\sin x dx + \int_\xi^1 \sqrt{x}\sin x dx = \int_0^\xi \sqrt{x}(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots)dx + \int_\xi^1 \sqrt{x}\sin x dx = \sum_{i=0}^\infty \frac{(-1)^i \xi^{2i+\frac{5}{2}}}{(2i+1)!(2i+\frac{5}{2})} + \int_\xi^1 \sqrt{x}\sin x dx$.

For $\int_1^\infty f(x)dx$, let $x = \frac{1}{t}, dx = \frac{-1}{t^2}dt$. Then, $\int_1^\infty f(x)dx = \int_0^1 f(\frac{1}{t})\frac{1}{t^2}dt$.

3 Linear Systems

One faced with solving $N \times N$ systems $Ax = b$, where N is very large ($\approx 10^6$ or 10^7. Such systems can be effectively solved by iterative methods. The idea is to construct a sequence of approximate solutions $x^{(0)}, x^{(1)}, \cdots, x^{(n)}$. At each step, the error is $e^{(n)} = x^{(n)} - x$. To specify an iterative method, one needs (i) a rule for constructing an approximation $\tilde{e}^{(n)}$ to $e^{(n)}$, so that one set $x^{(n+1)} = x^{(n)} - \tilde{e}^{(n)}$, and (ii) a stopping criterion, ideally based on bounds on $\left\|e^{(n)}\right\|$ or better on relative error $\frac{\left\|e^{(n)}\right\|}{\|x\|}$.

Residual is defined to be $r^{(n)} = b - Ax^{(n)} = Ax - Ax^{(n)} = -Ae^{(n)}$. $\left\|A^{-1}r^{(n)}\right\| \leq \left\|A^{-1}\right\|\left\|r^{(n)}\right\|$ and $\frac{1}{\|A^{-1}b\|} \leq \frac{\|A\|}{\|b\|} \Rightarrow \frac{\left\|e^{(n)}\right\|}{\|x\|} = \frac{\left\|A^{-1}r^{(n)}\right\|}{\|A^{-1}b\|} \leq \|A\|\left\|A^{-1}\right\|\frac{\left\|r^{(n)}\right\|}{\|b\|}$. $\|A\|\left\|A^{-1}\right\|$ is the **condition number** of A. If the condition number of A is bounded, then a stopping criterion might be that $\frac{\left\|r^{(n)}\right\|}{\|b\|}$ below a tolerance for a certain number of iterations.

3.1 Vector and matrix norm

Vector norm on a linear space is a mapping $\|\cdot\|$, s.t.

1. $\|x\| \geq 0$,
2. $\|x\| = 0 \Leftrightarrow x = 0$,
3. $\|x + y\| \leq \|x\| + \|y\|$, and
4. $\|\alpha x\| = |\alpha|\,\|x\|$.

The distance between x, y is $\|x - y\|$. Some common vector norms for R^N are

1. $\|x\|_1 = \sum_{i=1}^N w_i|x_i|$,
2. $\|x\|_2 = \left(\sum_{i=1}^N w_i x_i^2\right)^{\frac{1}{2}}$, and
3. $\|x\|_\infty = \max_i\{|x_i|\}$,

where $w = \begin{pmatrix} w_1 & \cdots & w_n \end{pmatrix}$ is vector of positive weights.

Matrix norm is assoicated with vector norm,

$$\|A\| = \sup\left\{\frac{\|Ax\|}{\|x\|} \;\middle|\; x \neq 0\right\}.$$

Adjoint of A with respect to the inner product $(x \mid y) = \sum_{i=1}^N x_i y_i w_i$ is A^*, s.t. $\forall x, y, (A^*x \mid y) = (x \mid Ay)$. Then

1. $\|A\|_1 = \max_j\left\{\sum_{i=1}^N |a_{ij}|w_i\right\}$,
2. $\|A\|_2 = \max\left\{\lambda^{\frac{1}{2}} \;\middle|\; \lambda \text{ is an eigenvalue of } A^*A\right\}$,
3. $\|A\|_\infty = \max_i\left\{\sum_{j=1}^N |a_{ij}|w_j\right\}$.

For all matrix norm, we have

1. $\|I\| = 1$,
2. $\|Ax\| \leq \|A\|\,\|x\|$, and
3. $\|AB\| \leq \|A\|\,\|B\|$.

3.2 Spectral theory

λ is an eigenvalue of $A \in \mathbb{C}^{N\times N}$ if $\exists 0 \neq x \in \mathbb{C}^N$, s.t. $Ax = \lambda x \Leftrightarrow \det(A - \lambda I) = 0 \Leftrightarrow A - \lambda I$ is not invertible.

Spectrum of A, $\mathsf{sp}(A) = \{\lambda \in \mathbb{C} \mid A - \lambda I \text{ is not invertible}\}$. Let $\lambda \in \mathsf{sp}(A)$, s.t. $\exists e, Ae = \lambda e$. For any matrix norm, $|\lambda| = \frac{\|\lambda e\|}{\|e\|} = \frac{\|Ae\|}{\|e\|} \leq \max\left\{\frac{\|Ax\|}{\|x\|} \;\middle|\; x \neq 0\right\} = \|A\|$.

Spectral radius is $\rho_{\mathsf{sp}}(A) = \max\{|\lambda| \mid \lambda \in \mathsf{sp}(A)\} \leq \|A\|$. The spectral radius formula $\rho_{\mathsf{sp}}(A) = \lim_{n\to\infty}\|A^n\|^{\frac{1}{n}}$.

3.3 Stationary iterative methods

In these motheds, the rule for finding $x^{(n+1)}$ from $x^{(n)}$ is the same for each n. Suppose we have an approximation B to A, s.t. B^{-1} is cheap to compute. Then, set $\tilde{e}^{(n)} = -B^{-1}r^{(n)}$, so that $x^{(n+1)} = x^{(n)} + B^{-1}r^{(n)}$.

When does this converge? Notice $e^{(n+1)} = x^{(n+1)} - x = x^{(n)} - x + B^{-1}r^{(n)} = e^{(n)} - B^{-1}Ae^{(n)} = (I - B^{-1}A)e^{(n)}$. Hence, $e^{(n)} = (I - B^{-1}A)^n e^{(0)} = G^n e^{(0)}$, where $G = I - B^{-1}A$ is the **growth matrix**.

Theorem 3.1 *This converges for all $x^{(0)} \Leftrightarrow \rho_{\mathsf{sp}}(G) < 1$.*

Proof: Clearly, if $\rho_{\mathsf{sp}}(G) \geq 1$, then set $e^{(0)}$ = eigenvector of λ with $|\lambda| \geq 1 \Rightarrow$ no convergence.

$\frac{\|e^{(n)}\|}{\|e^{(0)}\|} = \frac{\|G^n e^{(0)}\|}{\|e^{(0)}\|} \leq \|G^n\|$ but $\lim_{n\to\infty} \|G^n\|^{\frac{1}{n}} = \rho_{\mathsf{sp}}(G) < 1$. Pick $\delta > 0$ with $\rho_{\mathsf{sp}}(G) < 1 - \delta$ for some n_0, s.t. $\forall n \geq n_0, \|G^n\|^{\frac{1}{n}} < 1 - \delta \Rightarrow \|G^n\| \leq (1-\delta)^n$. $\mathcal{QED}$

Many classical choice for B are based on the decompositions $A = D - W = D - L - U$, where D is diagonal, W is off-diagonal, L and U are strictly lower and upper triangular respectively. Every entry of D is non-zero.

	Jocobi	Gauss-Seidel	Successive overrelaxation
B	D	$D-L$	$\frac{1}{\omega}D - L$
G	$D^{-1}W$	$(D-L)^{-1}U$	$(D-\omega L)^{-1}((1-\omega)D + \omega U)$

When $\omega = 1$, SOR is Gauss-Seidel.

Row diagonally dominant if $A = (a_{ij})$, $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$ for $i = 1, \cdots, n$. A is **column diagonally dominant** if $|a_{ii}| \geq \sum_{j\neq i} |a_{ji}|$ for $i = 1, \cdots, n$. A is strictly row/column diagonally dominant if all the corresponding inequalities are strict (>).

Theorem 3.2 *If A is strictly diagonally dominant, the the Jacobi method converges.*

Proof: For the row case, $\rho_{\mathsf{sp}}(G_{\mathsf{J}}) \leq \|G_{\mathsf{J}}\|_\infty = \|D^{-1}W\|_\infty = \max_i \left\{ \frac{1}{|a_{ii}|} \sum_{j\neq i} a_{ij} \right\} < 1$.

Similarly, for the column case, $\rho_{\mathsf{sp}}(G_{\mathsf{J}}) \leq \|G_{\mathsf{J}}\|_1 = \|D^{-1}W\|_1 = \max_i \left\{ \frac{1}{|a_{ii}|} \sum_{j\neq i} a_{ji} \right\} < 1$. $\mathcal{QED}$

Strict diagonally dominant is a strong condition. Consider the problem $-u'' = f$ of $[0,1]$ with $u[0] = u[1] = 0$. Approximate it as $\frac{-u_{i+1} + 2u_i - u_{i-1}}{\Delta^2} = f_i$ for $u_0 = u_n = 0$ and $\Delta = \frac{1}{n}$. That is

$$\frac{1}{\Delta^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \end{pmatrix}$$

The matrix is irreducible but not strictly diagonal dominant.

Irreducible $A = (a_{ij})_{N\times N}$ is irreducible if there is no permutation matrix P, s.t. $P^TAP = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$, where A_{11} is $N_1 \times N_1$, A_{12} is $N_1 \times N_2$ and A_{22} is $N_2 \times N_2$ with $N_1 + N_2 = N$ and $N_1, N_2 > 0$.

There is a simple graphical test for irreducibility. Create N nodes. Connect node i to node j by an oriented arc if $a_{ij} \neq 0$. The matrix is irreducible $\Leftrightarrow$ there is a oriented path connecting node every pair of nodes.

Irreducibly diagonally dominant if (i) irreducible, (ii) diagonally dominant, and (iii) at least one of the diagonally dominant inequalities is strict.

Theorem 3.3 *If $A = (a_{ij})$ is irreducibly diagonally dominant, then the Jacobi method converges.*

Proof: Similar to theorem 3.2, $\|D^{-1}W\|_\infty \leq 1$ for the row case or $\|D^{-1}W\|_1 \leq 1$ for the column case $\Rightarrow \rho_{\mathsf{sp}}(D^{-1}W) \leq 1$.

Suppose $\rho_{\mathsf{sp}}(D^{-1}W) = 1$. For the row case, $\exists e \in \mathbb{C}^N, \lambda \in \mathbb{C}, |\lambda| = 1$, s.t. $D^{-1}We = \lambda e$. $\|e\|_\infty = 1 \Rightarrow \exists i, e_i = 1$. $a_{ii}e_i = \frac{1}{\lambda}\sum_{j\neq i} a_{ij}e_j$. $|a_{ii}| \leq \sum_{j\neq i} |a_{ij}||e_j|$. By the irreducibility, $\forall i, |e_i| = 1$. $\mathcal{QED}$

Lemma 3.4 *If a matrix A is either strictly diagonal dominant or irreducibly diagonal dominant, A is invertible.*

Proof: Suppose $Ae = 0$ with $e \neq 0$. Then, $|e_i|\sum_{j\neq i} |a_{ij}| \leq |a_{ii}e_i| = \left|\sum_{j\neq i} a_{ij}e_j\right|$. WLOG, assume $\|e\|_\infty = 1$. $\exists i, e_i = 1$. $\Rightarrow \sum_{j\neq i} |a_{ij}| \leq |a_{ii}| = \left|\sum_{j\neq i} a_{ij}e_j\right| \leq \sum_{j\neq i} |a_{ij}e_j| \leq \sum_{j\neq i} |a_{ij}| \Rightarrow |e_j| = 1$ for all j when $a_{ij} \neq 0 \Rightarrow a_{ii} = \sum_{j\neq i} |a_{ij}|$ for all $i \Rightarrow A$ is only diagonal dominant but all inequalities are not strict which leads to a contradiction. $\mathcal{QED}$

Theorem 3.5 Jacobi, Gauss-Seidel, SOR convergence theorem:

1. *If A is either strictly diagonal dominant or irreducibly diagonal dominant, then both Jacobi and Gauss-Seidel methods converge.*
2. *If $A = D - W$ with $W \geq 0$ entrywise and the Jacobi method converges (i.e. $\rho_{\mathsf{sp}}(G_{\mathsf{J}}) < 1$), then Gauss-Seidel method converages with $\rho_{\mathsf{sp}}(G_{\mathsf{GS}}) < \rho_{\mathsf{sp}}(G_{\mathsf{J}})$.*
3. *If A is symmetric positive definite, then Gauss-Seidel method converges and the SOR method converges for $\omega \in (0,2)$.*
4. *The SOR method diverges for $\omega \notin (0,2)$.*

Proof: *1.* Let $\lambda \in \mathsf{sp}(G_{\mathsf{GS}})$ and $G_{\mathsf{GS}}e = \lambda e$ for $0 \neq e \in \mathbb{C}^N$. Then, $(D-L)^{-1}Ue = \lambda e \Rightarrow Ue = \lambda De - \lambda Le \Rightarrow (\lambda D - \lambda L - U)e = 0$. Define $A(\lambda) = \lambda D - \lambda L - U$.

For $\lambda \geq 1$, $A = A(1)$ is strictly diagonal dominant $\Rightarrow |\lambda||a_{ii}| > |\lambda|\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{N} |a_{ij}| \Rightarrow A(\lambda)$ is strictly diagonal dominant. On the other hand, if $A = A(1)$ is irreducibly diagonal dominant, $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$ for every i and the inequality is strict for some i. $|\lambda| \geq 1 \Rightarrow |\lambda||a_{ii}| - |\lambda|\sum_{j=1}^{i-1} |a_{ij}| - \sum_{j=i+1}^{N} |a_{ij}| \geq |\lambda|(|a_{ii}| - \sum_{j\neq i} |a_{ij}|) \begin{cases} \geq 0 & \text{, for every } i \\ > 0 & \text{, for some } i \end{cases} \Rightarrow A(\lambda)$ is diagonal dominant. $A(\lambda)$ is clearly irreducible for $\lambda \neq 0$ since the non-zero entries are the same as before. Therefore, $A(\lambda)$ is irreducibly diagonal dominant.

By lemma 3.4, $A(\lambda)$ is invertible for $\lambda \geq 1$. $A(\lambda)e = 0 \Rightarrow A(\lambda)$ is not invertible $\Rightarrow |\lambda| < 1 \Rightarrow \rho_{\mathsf{sp}}(G_{\mathsf{GS}}) < 1$.

The proof for Jacobi is similar.

4. Observe $D - \omega L$ is lower triangular and $(1-\omega)D + U$ is upper triangular. $\prod_{\lambda \in \mathsf{sp}(G_{\mathsf{SOR}})} |\lambda| = |\det(G_{\mathsf{SOR}}(\omega))| = |\det((D-\omega L)^{-1})\det((1-\omega)D+U)| = \left|\frac{1}{\det(D)}(1-\omega)^N \det(D)\right| = |1-\omega|^N$. $\rho_{\mathsf{sp}}(G_{\mathsf{SOR}}) \geq |1-\omega| \geq 1$ if $\omega \notin (0,2)$. $\mathcal{QED}$

Hermitian symmetric if $A \in \mathbb{C}^{N\times N}$, $A^* = A$, where A^* is the complex transpose of A. When A is real, $A^* = A^T$, so A is Hermitian symmetric if A is symmetric.

Euclidean inner product on $\mathbb{C}^N$, $(x \mid y) = x^* y$. Then, $(x \mid Ay) = x^*Ay = (A^*x)^*y = (A^*x \mid y)$.

Self-adjoint with respect to $(x \mid y)$ if $(x \mid Ay) = (Ax \mid y) \Leftrightarrow A^* = A$. Then, $\max\{\text{eigenvalues of } A\} = \max\left\{\frac{(x \mid Ax)}{(x \mid x)} \;\middle|\; x \neq 0\right\}$ and $\min\{\text{eigenvalues of } A\} = \min\left\{\frac{(x \mid Ax)}{(x \mid x)} \;\middle|\; x \neq 0\right\}$.

Non-negative definite (≥ 0) if $A \in \mathbb{C}^{N\times N}$, $A^* = A$ and $x^*Ax \geq 0$. A is **positive definite** (> 0) if, in addition, $x^*Ax = 0 \Rightarrow x = 0$.

If $A \geq 0$, then $D \geq 0$ for both entrywise and as form. If $A > 0$, then $D > 0$.

Theorem 3.6 (Spectral mapping theorem) *Let $p(x)$ be a rational function. If $\lambda \in \mathsf{sp}(M)$, $p(\lambda) \in \mathsf{sp}(p(M))$.*

Theorem 3.7 *Let $A^* = A$ and $D > 0$. Then,*

1. *Jacobi converges $\Leftrightarrow -D < W < D$.*
2. *SOR converges $\Leftrightarrow |\omega - 1| < 1$ and $A > 0$.*

Proof: *1.* Since $A^* = A$, we have $D^* = D$ and $W^* = W$. $D^{-1}W$ is self-adjoint with respect to $(x \mid y)_D = x^*Dy$, i.e. $(x \mid D^{-1}Wy)_D = x^*DD^{-1}Wy = x^*WD^{-1}Dy = (D^{-1}Wx)^*Dy = (D^{-1}Wx \mid y)_D$. Then, $\rho_{\mathsf{sp}}(G_{\mathsf{J}}) = \rho_{\mathsf{sp}}(D^{-1}W) = \max\left\{\frac{|(x \mid D^{-1}Wx)_D|}{(x \mid x)_D} \;\middle|\; x \neq 0\right\} = \max\left\{\frac{|x^*Wx|}{x^*Dx} \;\middle|\; x \neq 0\right\}$. $\rho_{\mathsf{sp}}(G_{\mathsf{J}}) < 1 \Leftrightarrow \forall x \neq 0, \frac{|x^*Wx|}{x^*Dx} < 1 \Leftrightarrow \forall x \neq 0, -x^*Dx < x^*Wx < x^*Dx \Leftrightarrow -D < W < D$.

2. $(\Rightarrow)$ By theorem 3.5 part 4, SOR converges $\Rightarrow |\omega - 1| < 1$. (show $A > 0$).

$(\Leftarrow)$ Let $M(\omega) = 2A^{-1}B_{\mathsf{SOR}}(\omega) - I$. $G_{\mathsf{SOR}}(\omega) = I - B_{\mathsf{SOR}}^{-1}(\omega)A$. If $\alpha \in \mathsf{sp}(B_{\mathsf{SOR}}^{-1}(\omega)A)$, then $\beta = \frac{2}{\alpha} - 1 \in \mathsf{sp}(M(\omega))$ and $\gamma = 1 - \alpha \in \mathsf{sp}(G_{\mathsf{SOR}}(\omega))$ by theorem 3.6. (see notes) $\mathcal{QED}$

Comparitive rates of convergence Let P be a permutation matrix. Observe that the entries of diagonal of a matrix do not change (up to permutation) when the indices of a matrix are permuted. Then, $PAP^{-1} = PDP^{-1} + PWP^{-1}$, where PDP^{-1} and PWP^{-1} are still diagonal and off diagonal respectively. However, it does mix up L and U. So, Gauss-Seidel and SOR depend on the ordering.

Consistently ordered when the spectrum of $J(\alpha) \stackrel{\text{def}}{=} \alpha D^{-1}L + \frac{1}{\alpha}D^{-1}L$ for $\alpha \neq 0$, is independent of α.

Theorem 3.8 *Let A be consistently ordered. Then*

1. *$\mu \in \mathsf{sp}(G_{\mathsf{J}}) \Leftrightarrow -\mu \in \mathsf{sp}(G_{\mathsf{J}})$.*
2. *$\mu \in \mathsf{sp}(G_{\mathsf{J}})$ and λ satisfies*
$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2 \tag{3.2}$$
$\Rightarrow \lambda \in \mathsf{sp}(G_{\mathsf{SOR}}(\omega))$.
3. *$\lambda \in \mathsf{sp}(G_{\mathsf{SOR}}(\omega)) \setminus \{0\}$ and μ satisfies equation 3.2 $\Rightarrow \mu \in \mathsf{sp}(G_{\mathsf{J}})$.*

Proof: *1.* Observe $G_{\mathsf{J}} = J(1)$ and $-G_{\mathsf{J}} = J(-1)$. Since $\mathsf{sp}(J(1)) = \mathsf{sp}(J(-1))$, $\mu \in \mathsf{sp}(G_{\mathsf{J}}) = \mathsf{sp}(J(1)) = \mathsf{sp}(J(-1)) = \mathsf{sp}(-G_{\mathsf{J}}) \Rightarrow -\mu \in \mathsf{sp}(G_{\mathsf{J}})$.

3. $0 \neq \lambda \in \mathsf{sp}(G_{\mathsf{SOR}}(\omega)) \Leftrightarrow G_{\mathsf{SOR}}(\omega)e = \lambda e$ for some $e \in \mathbb{C}^N \setminus \{0\} \Leftrightarrow ((1-\omega)D + \omega U)e = \lambda(D - \omega L)e \Leftrightarrow ((\lambda + \omega - 1)D - \lambda\omega L - \omega U)e = 0 \Leftrightarrow \left(\frac{\lambda+\omega-1}{\sqrt{\lambda}\omega}D - \sqrt{\lambda}L - \frac{1}{\sqrt{\lambda}}U\right)e = 0 \Leftrightarrow \mu = \frac{\lambda+\omega-1}{\sqrt{\lambda}\omega} \in \mathsf{sp}(J(\sqrt{\lambda})) = \mathsf{sp}(J(1))$.

2. If $\lambda \neq 0$, set $\sqrt{\lambda}$, s.t. $\mu = \frac{\lambda+\omega-1}{\sqrt{\lambda}\omega}$.

If $\lambda = 0$, then equation 3.2 implies $(\omega - 1)^2 = 0$ i.e. $\omega = 1$. But $\det(G_{\mathsf{SOR}}(1)) = \det((D-L)^{-1}U) = 0 \Rightarrow \lambda = 0 \in \mathsf{sp}(G_{\mathsf{SOR}}(\omega))$. $\mathcal{QED}$

Corollary 3.9 *Let A be consistently ordered. Then $\rho_{\mathsf{sp}}(G_{\mathsf{GS}}) = (\rho_{\mathsf{sp}}(G_{\mathsf{J}}))^2$. (i.e. Gauss-Seidel converges twice as fast as Jacobi.)*

Proof: In theorem 3.8, when $\omega = 1$, $\lambda = \mu^2$. $\mathcal{QED}$

Theorem 3.10 *Let A be consistently ordered, $\mathsf{sp}(G_{\mathsf{J}})$ be real and $\rho_{\mathsf{J}} \stackrel{\text{def}}{=} \rho_{\mathsf{sp}}(G_{\mathsf{J}}) < 1$. Then, $\rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega))$*

$$= \begin{cases} \omega - 1 & \text{, for } \omega_{\mathsf{opt}} \leq \omega < 2, \\ 1 - \omega + \frac{\omega^2\rho_{\mathsf{J}}^2}{2} + \omega\rho_{\mathsf{J}}\sqrt{1 - \omega + \frac{\omega^2\rho_{\mathsf{J}}^2}{4}} & \text{, for } 0 < \omega \leq \omega_{\mathsf{opt}}, \end{cases}$$

where $\omega_{\mathsf{opt}} = \frac{2}{1+\sqrt{1-\rho_{\mathsf{J}}^2}} > 1$. Moreover, $\left(\frac{\rho_{\mathsf{J}}}{1+\sqrt{1-\rho_{\mathsf{J}}^2}}\right)^2 = \rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega_{\mathsf{opt}})) \leq \rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega)) < 1$.

Proof: For $\omega = 1$, $\rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(1)) = \rho_{\mathsf{J}}^2$ by corollary 3.9.

Consider $\omega \neq 1$. By theorem 3.8 part 2, $\mathsf{sp}(G_{\mathsf{SOR}}(\omega)) = \left\{\lambda_\pm \;\middle|\; \lambda_\pm = 1 - \omega + \frac{\omega^2\mu^2}{2} \pm \omega\mu\sqrt{d(\omega,\mu)}, 0 \leq \mu \in \mathsf{sp}(G_{\mathsf{J}})\right\}$, where $d(\omega,\mu) = 1 - \omega + \frac{\omega^2\mu^2}{4}$. If $d(\omega,\mu) \leq 0$, then $|\lambda_\pm| = \left|1 - \omega + \frac{\omega^2\mu^2}{2} \pm i\omega\mu\sqrt{-d(\omega,\mu)}\right| = |1-\omega|$. If $d(\omega,\mu) > 0$, $\lambda_\pm$ are both real with $\lambda_+\lambda_- = (1-\omega)^2$ and $\lambda_+ > \lambda_-$. λ_+ is an increasing function of μ. The largest value is reached when $\mu = \rho_{\mathsf{J}}$.

If $d(\omega,\rho_{\mathsf{J}}) \leq 0$, then $2 > \omega \geq \omega_{\mathsf{opt}} > 1$. It is always in the first case $\Rightarrow \rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega)) = |1-\omega| = \omega - 1$. If $d(\omega,\rho_{\mathsf{J}}) > 0$, then $0 < \omega < \omega_{\mathsf{opt}}$. Sometimes it is in the second case. Since $1 - \omega + \frac{\omega^2\rho_{\mathsf{J}}^2}{2} + \omega\rho_{\mathsf{J}}\sqrt{d(\omega,\rho_{\mathsf{J}})} > |1-\omega|$, $\rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega)) = 1 - \omega + \frac{\omega^2\rho_{\mathsf{J}}^2}{2} + \omega\rho_{\mathsf{J}}\sqrt{d(\omega,\rho_{\mathsf{J}})}$. $\mathcal{QED}$

Observe that ω_{opt} is an increasing function of ρ_{J}. We find ρ_*, s.t. $\rho_{\mathsf{J}} \leq \rho_* < 1$. Set $\omega = \frac{2}{1+\sqrt{1-\rho_*^2}} \geq \omega_{\mathsf{opt}}$.

Example: Consider the block tridiagonal matrix

$$\begin{pmatrix} D_1 & A_{12} & 0 & \cdots & 0 \\ A_{21} & D_2 & A_{23} & \ddots & \vdots \\ 0 & A_{32} & D_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & A_{m-1,m} \\ 0 & \cdots & 0 & A_{m,m-1} & D_m \end{pmatrix},$$

where each D_j is diagonal. This is consistently ordered since $J(\alpha) = \alpha D^{-1}L + \frac{1}{\alpha}D^{-1}U$

$$= -\begin{pmatrix} 0 & \frac{1}{\alpha}D_1^{-1}A_{12} & 0 & \cdots & 0 \\ \alpha D_2^{-1}A_{21} & 0 & \frac{1}{\alpha}D_2^{-1}A_{23} & \ddots & \vdots \\ 0 & \alpha D_3^{-1}A_{32} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{\alpha}D_{m-1}^{-1}A_{m-1,m} \\ 0 & \cdots & 0 & \alpha D_m^{-1}A_{m,m-1} & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \alpha & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha^{m-1} \end{pmatrix} J(1) \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \alpha & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha^{m-1} \end{pmatrix}^{-1}.$$

Block property A if $\exists P$, a permutation matrix, s.t. $PAP^T = \begin{pmatrix} D_1 & A_{12} \\ A_{21} & D_2 \end{pmatrix}$, where D_1 and D_2 are diagonal.

Example: Consider the matrix using the form $-\Delta u \stackrel{\text{def}}{=} \nabla^2 u = g$ on $[0,L]^2$ with $u = 0$ on boundary. Let $\Delta = \frac{L}{N+1}$. Consider $\frac{1}{\Delta^2}(-u_{i,j+1} - u_{i+1,j} + 4u_{i,j} - u_{i,j-1} - u_{i-1,j})$.

Estimating rates of convergence Suppose $\rho_{\mathsf{J}} \stackrel{\text{def}}{=} \rho_{\mathsf{sp}}(G_{\mathsf{J}}) < 1$. If we can find ρ_*, s.t. $\rho_{\mathsf{J}} \le \rho_* < 1$, then if A is consistently ordered, $\rho_{\mathsf{GS}} \stackrel{\text{def}}{=} \rho_{\mathsf{sp}}(G_{\mathsf{GS}}) = \rho_{\mathsf{J}}^2 \le \rho_*^2 < 1$. Be setting $\omega_* = \frac{2}{1+\sqrt{1-\rho_*^2}}$, $\rho_{\mathsf{SOR}}(\omega_*) \stackrel{\text{def}}{=} \rho_{\mathsf{sp}}(G_{\mathsf{SOR}}(\omega_*)) = \left(\frac{\rho_*}{1+\sqrt{1-\rho_*^2}}\right)^2$. Note that $\omega_{\mathsf{opt}} \le \omega_* < 2$.

Example: Consider the BVP, $-\frac{d}{dx}\left(a(x)\frac{du}{dx}\right) + c(x)u = g(x)$ for $x \in [0,L]$, $u(0) = u(L) = 0$, $a(x) > 0$ and $c(x) > 0$. Consider the differencing $\frac{-1}{\delta}\left(a_{j+\frac{1}{2}}\frac{u_{j+1}-u_j}{\delta} - a_{j-\frac{1}{2}}\frac{u_j-u_{j-1}}{\delta}\right) + c_j u_j = g_j$, where $\delta = \frac{L}{N+1}$, the nodes $x_j = j\delta$ for $j = 1,\cdots,N$, the mid-points $x_{j+\frac{1}{2}} = \left(j+\frac{1}{2}\right)\delta$ for $j = 0,\cdots,N$, $a_{j+\frac{1}{2}} = a(x_{j+\frac{1}{2}})$, $c_j = c(x_j)$ and $g_j = g(x_j)$. This yields the linear system $A\begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} g_1 \\ \vdots \\ g_N \end{pmatrix}$, where $A =$

$$\begin{pmatrix} \frac{a_{\frac{1}{2}}+a_{\frac{3}{2}}}{\delta^2}+c_1 & \frac{-1}{\delta^2}a_{\frac{3}{2}} & & & \\ \ddots & \ddots & \ddots & & \\ & \frac{-1}{\delta^2}a_{j-\frac{1}{2}} & \frac{a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}}{\delta^2}+c_j & \frac{-1}{\delta^2}a_{j+\frac{1}{2}} & \\ & & \ddots & \ddots & \ddots \\ & & & \frac{a_{N-\frac{1}{2}}+a_{N+\frac{1}{2}}}{\delta^2}+c_N & \frac{-1}{\delta^2}a_{N+\frac{1}{2}} \end{pmatrix}$$

is symmetric, strictly diagonal dominant with the diagonal elements > 0, and, therefore, positive definite. Then,

$$G_{\mathsf{J}} = D^{-1}W = \begin{pmatrix} 0 & \frac{a_{\frac{3}{2}}}{a_{\frac{1}{2}}+a_{\frac{3}{2}}+c_1\delta^2} & & & \\ \ddots & \ddots & \ddots & & \\ & \frac{a_{j-\frac{1}{2}}}{a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}+c_j\delta^2} & 0 & \frac{a_{j+\frac{1}{2}}}{a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}+c_j\delta^2} & \\ & & \ddots & \ddots & \ddots \\ & & & \frac{a_{N-\frac{1}{2}}}{a_{N-\frac{1}{2}}+a_{N+\frac{1}{2}}+c_N\delta^2} & 0 \end{pmatrix}.$$

$\rho_{\mathsf{J}} \le \max\left\{\frac{a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}}{a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}+c_j\delta^2} \;\middle|\; j = 1,\cdots,N\right\}$. Let $\rho_* = \sqrt{1-\epsilon^2}$, where $\epsilon \ll 1$. Hence, $\rho_{\mathsf{J}} \le \sqrt{1-\epsilon^2}$, $\rho_{\mathsf{GS}} \le 1-\epsilon^2$ and $\rho_{\mathsf{SOR}}(\omega_*) = \left(\frac{\sqrt{1-\epsilon^2}}{1+\epsilon}\right)^2 = \frac{1-\epsilon}{1+\epsilon}$. For $\epsilon = 0.01$, $\rho_{\mathsf{J}} \le \sqrt{0.9999} \approx 0.999950$, $\rho_{\mathsf{GS}} \le 0.9999$ and $\rho_{\mathsf{SOR}}(\omega_*) = \frac{0.99}{1.01} \approx 0.98$.

Alternating methods So far we have studied stationary methods $x^{(n+1)} = x^{(n)} - \tilde{e}^{(n)}$, where the rule for this is $\tilde{e}^{(n)} = -B^{-1}r^{(n)}$, where $B \approx A$ in the sense $\rho_{\mathsf{sp}}(I - B^{-1}A) < 1$.

Now suppose you have two guesses, B_1 and B_2. Consider $x^{(2n+1)} = x^{(2n)} + B_1^{-1}r^{(2n)}$ and $x^{(2n+2)} = x^{(2n+1)} + B_2^{-1}r^{(2n+1)}$. Then, $\tilde{e}^{(2n+2)} = (I - B_2^{-1}A)(I - B_1^{-1}A)\tilde{e}^{(2n)}$. The method converges $\Leftrightarrow \rho_{\mathsf{sp}}((I-B_2^{-1}A)(I-B_1^{-1}A)) < 1$.

Example: ADI (Alternating Direction Implicit) method: $-\Delta u = g$ on $\Omega = [0,L]^2$ with $u = 0$ on $\partial\Omega$.

3.4 Conjugate gradient method

Proposition 3.11 *Consider the equation*

$$Ax = b, \tag{3.3}$$

where $A^T = A > 0$. The solution to equation 3.3 is also the solution at

$$f(x) = \min\left\{f(y) \;\middle|\; y \in \mathbb{R}^N\right\}, \tag{3.4}$$

where $f(y) = \frac{1}{2}(y \mid Ay) - (b \mid y)$.

Proof: Suppose x solves equation 3.3. $\forall y = x+z \in \mathbb{R}^N$. $f(y) = \frac{1}{2}(x+z \mid A(x+z)) - (b \mid x+z) = f(x) + (z \mid Ax - b) + \frac{1}{2}(z \mid Az) = f(x) + \frac{1}{2}(z \mid Az) \ge f(x)$ by $A > 0$.

Conversely, suppose x solves equation 3.4. $\forall z \in \mathbb{R}^N, t \in \mathbb{R}$, $f(x) \le f(x+tz) = f(x) + t(z \mid Ax - b) + \frac{t^2}{2}(z \mid Az)$. The parabola have a minimum at $t = 0$. Then, $0 = \left.\frac{df(x+tz)}{dx}\right|_{t=0} = (z \mid Ax - b) \Rightarrow Ax - b = 0$. $\mathcal{QED}$

An iterative method to solves equation 3.3: $x^{(n+1)} = x^{(n)} - \tilde{e}^{(n)}$, where $\tilde{e}^{(n)} \approx e^{(n)} = x^{(n)} - x$. Notice $f(x^{(n)}) = f(x + e^{(n)}) = f(x) + \frac{1}{2}(e^{(n)} \mid Ae^{(n)})$. So $(e^{(n)} \mid Ae^{(n)})$ is a measure of the size of the error. Hence, minimizing $f(x^{(n)})$ is the same as minimizing the ***A*-norm** of $e^{(n)}$,

$$\left\|e^{(n)}\right\|_A = \sqrt{(e^{(n)} \mid Ae^{(n)})}.$$

Suppose $\tilde{e}^{(n)} = -\alpha p^{(n)}$, where $p^{(n)} \in \mathbb{R}^N$ is given. What is the best choice of α? We pick α to minimize $f(x^{(n+1)}) = f(x^{(n)} + \alpha p^{(n)}) = f(x^{(n)}) + \alpha(p^{(n)} \mid Ax^{(n)} - b) + \frac{\alpha^2}{2}(p^{(n)} \mid Ap^{(n)}) = f(x^{(n)}) - \alpha(p^{(n)} \mid r^{(n)}) + \frac{\alpha^2}{2}(p^{(n)} \mid Ap^{(n)})$, where $r^{(n)} = b - Ax^{(n)}$ is the residual.

This happen at $\alpha = \alpha_n = \frac{(p^{(n)} \mid r^{(n)})}{(p^{(n)} \mid Ap^{(n)})}$. Then, $f(x^{(n+1)}) = f(x^{(n)}) - \frac{(p^{(n)} \mid r^{(n)})}{2(p^{(n)} \mid Ap^{(n)})} \Rightarrow \frac{1}{2}(e^{(n+1)} \mid Ae^{(n+1)}) = \frac{1}{2}(e^{(n)} \mid Ae^{(n)}) - \frac{(p^{(n)} \mid r^{(n)})}{2(p^{(n)} \mid Ap^{(n)})} \Rightarrow \left\|e^{(n+1)}\right\|_A^2 = \left\|e^{(n)}\right\|_A^2 - \frac{(p^{(n)} \mid r^{(n)})}{(p^{(n)} \mid Ap^{(n)})}$. This is the maximum norm can be reduced for a given $p^{(n)}$.

Remark: We saw in Prof. Osborn's lecture that for SOR, $\left\|e^{(n+1)}\right\|_A^2 = \left\|e^{(n)}\right\|_A^2 - \left(\frac{1}{\omega^*} + \frac{1}{\omega} - 1\right)(d^{(n)} \mid Dd^{(n)})$. So, one idea to improve SOR (or other fixed methods) is setting $x^{(n+1)} = x^{(n)} + \alpha_n p^{(n)}$, where $p^{(n)} = B^{-1}r^{(n)}$ and $\alpha_n = \frac{(p^{(n)} \mid r^{(n)})}{(p^{(n)} \mid Ap^{(n)})} = \frac{(r^{(n)} \mid Qr^{(n)})}{(p^{(n)} \mid Ap^{(n)})}$, where $Q = B^{-1}$ with $Q^* = Q > 0$. If $Q = I$, then this is the method at steepest descents $-\nabla_y f(x^{(n)}) = r^{(n)} - p^{(n)}$.

3.5 Conjugate gradient method II

Let $0 < A \in \mathbb{R}^{N\times N}$ and $Q > 0$, s.t. $Q \approx A^{-1}$. The conjugate gradient iteration goes as follows:

CG Choose $x^{(0)} \in \mathbb{R}^N$. Set $r^{(0)} = b - Ax^{(0)}$, $p^{(0)} = Qr^{(0)}$. Begin loop on n until "converges"

$\alpha_n = \frac{(r^{(n)} \mid p^{(n)})}{(p^{(n)} \mid Ap^{(n)})}$
$x^{(n+1)} = x^{(n)} + \alpha_n p^{(n)}$
$r^{(n+1)} = r^{(n)} - \alpha_n Ap^{(n)}$
test for "convergence" here
$\beta_n = \frac{(r^{(n+1)} \mid Qr^{(n+1)})}{(r^{(n)} \mid Qr^{(n)})}$
$p^{(n+1)} = Qr^{(n+1)} + \beta_n p^{(n)}$

Lemma 3.12 *$\forall n$, s.t. $x^{(n)} \neq x$, we have*

1. *$\forall m < n, (r^{(n)} \mid p^{(m)}) = (r^{(n)} \mid Qr^{(m)}) = 0.$*
2. *$\forall m < n, (p^{(n)} \mid Ap^{(m)}) = 0.$*
3. *$\mathsf{span}\left\{p^{(0)}, \cdots, p^{(n)}\right\} = \mathsf{span}\left\{Qr^{(0)}, \cdots, Qr^{(n)}\right\} = \mathsf{span}\left\{p^{(0)}, QAp^{(0)}, \cdots, (QA)^n p^{(0)}\right\} = \mathcal{K}_{n+1}(p^{(0)}, QA)$ ($n+1^{th}$* **Krylov subspace**).

Proof: (*1 & 2*) They are trivially true for $n = 0$. Suppose they are true for n. $(r^{(n+1)} \mid p^{(m)}) = (r^{(n)} \mid p^{(m)}) - \alpha_n(Ap^{(n)} \mid p^{(m)}) = \begin{cases} 0 & \text{for } m < n \text{ by induction} \\ 0 & \text{for } m = n \text{ by the definition of } \alpha_n \end{cases}$. For $m > 0$, $(r^{(n+1)} \mid p^{(m)}) = (r^{(n+1)} \mid Qr^{(m)}) + \beta_n(r^{(n+1)} \mid p^{(m-1)}) = (r^{(n+1)} \mid Qr^{(m)})$. (*1*) is true for $n+1$. $(p^{(n+1)} \mid Ap^{(m)}) = (Qr^{(n+1)} \mid Ap^{(m)}) + \beta_n(p^{(n)} \mid Ap^{(m)}) = \frac{1}{\alpha_m}(r^{(n+1)} \mid Q(r^{(m)} - r^{(m+1)})) + \beta_n(p^{(n)} \mid Ap^{(m)}) = 0.$ $\mathcal{QED}$

Theorem 3.13 *Let $A, Q \in \mathbb{R}^{N\times N}$ with $A > 0$ and $Q > 0$, $x^{(0)} \in \mathbb{R}^N$ and $p^{(0)} = Q(b - Ax^{(0)})$. The following are equivalent:*

1. *$x^{(n)}$ is the n^{th} iteration of CG,*
2. *$\forall y \in x^{(0)} + \mathcal{K}_n(p^{(0)}, QA), f(x^{(n)}) \leq f(y)$,*
3. *$\left\|x^{(n)} - x\right\|_A \leq \|y - x\|_A$,*
4. *$b - Ax^{(n)} \perp \mathcal{K}_n(p^{(0)}, QA)$,*

where $f(y) = \frac{1}{2}(y \mid Ay) - (b \mid y)$ and $\|z\|_A = \sqrt{(z \mid Az)}$.

Proof: (*1⇒2*) By CG, $x^{(n)} \in x^{(0)} + \mathcal{K}_n(p^{(0)}, QA)$. Let $y = x^{(n)} + z \in x^{(0)} + \mathcal{K}_n(p^{(0)}, QA)$ for some $z \in \mathcal{K}_n(p^{(0)}, QA)$. $f(y) = f(x^{(n)} + z) = f(x^{(n)}) + (z \mid Ax^{(n)} - b) + \frac{1}{2}(z \mid Az)$. By lemma 3.12 part 1, $(z \mid Ax^{(n)} - b) = 0$.

(*2⇒3*) $\frac{1}{2}\left\|x^{(n)} - x\right\|_A^2 = f(x^{(n)}) - f(x) \leq f(y) - f(x) = \frac{1}{2}\|y - x\|_A^2$, $\forall y \in x^{(0)} + \mathcal{K}_n(p^{(0)}, QA)$.

(*3⇒4*) Let $z \in \mathcal{K}_n(p^{(0)}, QA)$ and $y = x^{(n)} + tz$. Then, $\frac{1}{2}\left\|x^{(n)} - x\right\|_A^2 \leq \frac{1}{2}\|y - x\|_A^2 = \frac{1}{2}\left\|x^{(n)} - x\right\|_A^2 + t(z \mid x^{(n)} - x)_A + \frac{t^2}{2}\|z\|_A^2 \Rightarrow 0 = (z \mid x^{(n)} - x)_A = (z \mid Ax^{(n)} - b)$.

(*4⇒1*) Show it! $\mathcal{QED}$

Convergence rate Since $(y \mid QAz)_A = (y \mid AQAz) = (QAy \mid Az) = (QAy \mid z)_A$, QA is self-adjoint w.r.t. the A-inner product. Hence $\|QA\|_A = \rho_{\mathsf{sp}}(QA)$ and $\left\|(QA)^{-1}\right\|_A = \rho_{\mathsf{sp}}((QA)^{-1})$. $y \neq 0 \Rightarrow Ay \neq 0 \Rightarrow (y \mid QAy)_A = (Ay \mid QAy) > 0 \Rightarrow QA$ is positive definite w.r.t. the A-inner product. Let $\lambda_{\max} = \max\{\lambda \in \mathsf{sp}(QA)\} = \rho_{\mathsf{sp}}(QA)$ and $\lambda_{\min} = \min\{\lambda \in \mathsf{sp}(QA)\} = \frac{1}{\rho_{\mathsf{sp}}((QA)^{-1})}$ by theorem 3.6. Then, the condition number,

$$\kappa^2 = \mathsf{Cond}_A(QA) \stackrel{\mathsf{def}}{=} \|QA\|_A \left\|(QA)^{-1}\right\|_A = \frac{\lambda_{\max}}{\lambda_{\min}} > 1.$$

Theorem 3.14 *Let $e^{(n)} = x^{(n)} - x$ be the error of n^{th} iterate of CG. Then, $\left\|e^{(n)}\right\|_A \leq 2\left(\frac{\kappa-1}{\kappa+1}\right)^n \left\|e^{(0)}\right\|_A$.*

What has this brought us? For stationary iteration, $x^{(n+1)} = x^{(n)} + Qr^{(n)}$. Let $\rho_Q = \rho_{\mathsf{sp}}(I - QA)$. Then, $\left\|e^{(n)}\right\| \leq \rho_Q^n \left\|e^{(0)}\right\|_A$. This will converge *iff* $\mathsf{sp}(QA) \subset (0, 2)$. $\mathsf{sp}(QA) \subset [\lambda_{\min}, \lambda_{\max}] \Rightarrow \mathsf{sp}(I - QA) \subset [1 - \lambda_{\max}, 1 - \lambda_{\min}] \Rightarrow \rho_Q = \max\{|\lambda_{\max} - 1|, |1 - \lambda_{\min}|\} \Rightarrow \lambda_{\max} \leq 1 + \rho_Q$ and $\lambda_{\min} \geq 1 - \rho_Q \Rightarrow \kappa^2 = \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{1+\rho_Q}{1-\rho_Q}$. Hence $\frac{\kappa-1}{\kappa+1} \leq \frac{\sqrt{1+\rho_Q} - \sqrt{1-\rho_Q}}{\sqrt{1+\rho_Q} + \sqrt{1-\rho_Q}} = \frac{\rho_Q}{1+\sqrt{1-\rho_Q^2}} \leq \rho_Q$. So when $\rho_Q^2 = 1 - \delta^2$ with $\delta \ll 1$, $\frac{\kappa-1}{\kappa+1} \leq \frac{\sqrt{1-\delta^2}}{1+\delta}$. The denominator helps. Remarks: Conjugate gradient always converges. The game is to find a Q that makes $\mathsf{Cond}(QA)$ as small as possible. This is called precondition.

3.6 Krylov space methods

Consider solving

$$Ax = b, \tag{3.5}$$

where $b \in \mathbb{R}^N$ and $A \in \mathbb{R}^{N\times N}$ with $\det A \neq 0$. We know equation 3.5 has a unique solution $x \in \mathbb{R}^N$. Can we narrow it down more?

Let $q(\lambda) = a_n\lambda^n + \cdots + a_1\lambda + a_0$ be a polynomial. Then $q(A) \stackrel{\mathsf{def}}{=} a_nA^n + \cdots + a_1A + a_0I$. Let $\psi(\lambda) = \det(\lambda I - A)$ be the characteristic polynomial of A. The **Cayly-Hamilton theorem** states $\psi(A) = 0$.

Let $M = \min\{\deg(q) \mid q(A) = 0\}$. By Cayly-Hamilton, $M \leq N$. There is a unique monic polynomial $m(\lambda) = \lambda^M + \mu_{M-1}\lambda^{M-1} + \cdots + \mu_0$, s.t. $m(A) = 0$. Moveover, $\mu_0 \neq 0$ because $\lambda \in \mathsf{sp}(A) \Leftrightarrow m(\lambda) = 0$. Hence, $A^{-1} = \frac{-1}{\mu_0}(A^{M-1} + \mu_{M-1}A^{M-2} + \cdots + \mu_1 I)$ and $x = A^{-1}b = \frac{-1}{\mu_0}(A^{M-1}b + \mu_{M-1}A^{M-2}b + \cdots + \mu_1 b) \in \mathcal{K}_M(b, A)$. $A(x-$

$x^{(0)}) = b - Ax^{(0)} = r^{(0)}$. Hence $x \in x^{(0)} + \mathcal{K}_M(r^{(0)}, A)$. Clearly, for $n > 0$, $1 \leq \dim \mathcal{K}_n(r^{(0)}, A) \leq n$. Let $K \leq M$ be the largest, s.t. $\forall n \leq K, \dim \mathcal{K}_n(r^{(0)}, A) = n$. Then, $A^n \mathcal{K}_K(r^{(0)}, A)$ and $x \in x^{(0)} + \mathcal{K}_K(r^{(0)}, A)$.

Krylov space iterative methods have the general form:

1. Choose $x^{(0)} \in R^N$.
2. Pick $x^{(n)} \in x^{(0)} + \mathcal{K}_n(r^{(0)}, A)$, s.t. some norm of the error is minimized.

Lemma 3.15 (Orthogonality Lemma) *Let $A^* = A$ w.r.t. $(\ |\ )$. If $\forall n \leq K, \dim \mathcal{K}_n(p^{(0)}, A) = n$ for some $p^{(0)} \in \mathbb{R}^N \backslash \{0\}$, then $\mathcal{K}_n(p^{(0)}, A) = \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$, where $p^{(1)} = Ap^{(0)} - \beta_0 p^{(0)}$, $p^{(n+1)} = Ap^{(n)} - \beta_n p^{(n)} - \gamma_n p^{(n-1)}$ for $1 \leq n \leq K-2$, $\beta_n = \frac{(p^{(n)} \mid Ap^{(n)})}{(p^{(n)} \mid p^{(n)})}$ and $\gamma_n = \frac{(p^{(n-1)} \mid Ap^{(n)})}{(p^{(n-1)} \mid p^{(n-1)})}$. Moveover,*

(i) For $m < n < K$, $(p^{(m)} \mid p^{(n)}) = 0$.

(ii) For $m < n < K$, $(p^{(m)} \mid Ap^{(n)}) = (p^{(m+1)} \mid p^{(n)})$.

Proof: See notes. $\mathcal{QED}$

3.7 Minimum residual method

Recall that if $x^{(n+1)} = x^{(n)} + \alpha p^{(n)}$, the value of α that minimizes $\left\|e^{(n+1)}\right\|_E$ is $\alpha = -\frac{(e^{(n)} \mid p^{(n)})_E}{(p^{(n)} \mid p^{(n)})_E}$ since $\left\|e^{(n+1)}\right\|_E^2 = \left\|e^{(n)}\right\|_E^2 + 2\alpha(e^{(n)} \mid p^{(n)})_E + \alpha^2 (p^{(n)} \mid p^{(n)})_E$. Of course, we do not know $e^{(n)}$. The game is to find an inner product for which we can compute $(e^{(n)} \mid p^{(n)})_E$ without knowing $e^{(n)}$. Recall that we know $r^{(n)} = -Ae^{(n)}$. The idea is to use $(y \mid z)_E \stackrel{\mathsf{def}}{=} (y \mid z)_{A^2} = (Ay \mid Az)$. Then, $\alpha = \frac{(r^{(n)} \mid Ap^{(n)})}{(Ap^{(n)} \mid Ap^{(n)})}$. Recall that if $\det A \neq 0$ and $A^* = A$, then $A^2 > 0$.

Return to the general setting $x^{(n+1)} = x^{(0)} + \alpha_0 p^{(0)} + \cdots + \alpha_n p^{(n)}$. Suppose $\left\{p^{(k)}\right\}_{k=0}^n$ is orthogonal w.r.t. $(\ |\ )_E$. Then, $\left\|e^{(n+1)}\right\|_E^2 \leq \min\left\{ \|y - x\|_E^2 \;\middle|\; y \in x^{(0)} + \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n)}\right\}\right\}$ since α_k are coefficients of orthogonal projection onto $\mathsf{span}\left\{p^{(0)}, \cdots, p^{(n)}\right\}$. Putting this together with lemma 3.15, we get

MINRES Suppose $A^* = A$ w.r.t. $(\ |\ )$ and $\det A \neq 0$. Choose $x^{(0)} \in \mathbb{R}^N$. Set $r^{(0)} = b - Ax^{(0)}$, $p^{(0)} = r^{(0)}$ and $q^{(0)} = Ar^{(0)}$. Begin a loop on n until stopping

$$\alpha_n = \frac{(r^{(n)} \mid q^{(n)})}{(q^{(n)} \mid q^{(n)})} = -\frac{(e^{(n)} \mid p^{(n)})_{A^2}}{(p^{(n)} \mid p^{(n)})_{A^2}}.$$
$$x^{(n+1)} = x^{(n)} + \alpha_n p^{(n)}$$
$$r^{(n+1)} = r^{(n)} - \alpha_n q^{(n)}$$

Check for stopping.

$$\beta_n = \frac{(q^{(n)} \mid Aq^{(n)})}{(q^{(n)} \mid q^{(n)})} = \frac{(p^{(n)} \mid Ap^{(n)})_{A^2}}{(p^{(n)} \mid p^{(n)})_{A^2}}$$
$$\gamma_n = \frac{(q^{(n)} \mid q^{(n)})}{(q^{(n-1)} \mid q^{(n-1)})} = \frac{(p^{(n)} \mid p^{(n)})_{A^2}}{(p^{(n-1)} \mid p^{(n-1)})_{A^2}}$$
$$p^{(n+1)} = q^{(n)} - \beta_n p^{(n)} - \gamma_n p^{(n-1)}$$
$$q^{(n+1)} = Aq^{(n)} - \beta_n q^{(n)} - \gamma_n q^{(n-1)}$$

3.8 Optimal error methods

(like CG and MINRES) To solve $Ax = b$ with $\det A \neq 0$. The iterative scheme has the form $x^{(n+1)} = x^{(n)} + \alpha_n p^{(n)}$.

Let $X_n = \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$, $\overline{X} = \bigcup_n X_n \subset \mathbb{R}^N$ with $\dim \overline{X} = \max_n \{\dim X_n\}$ and $\overline{n} = \min\left\{n \mid \dim X_n = \dim \overline{X}\right\}$. One could show $X_n = \overline{X}$ for $n \geq \overline{n}$. Then, $x^{(n)} = x^{(0)} + \alpha_0 p^{(0)} + \cdots + \alpha_{n-1} p^{(n-1)} \in x^{(0)} + X_n \subset x^{(0)} + \overline{X}$. For the convergence of $x^{(n)} \to x$, we need $x \in x^{(0)} + \overline{X}$.

Given such X_n, pick $x^{(n)}$ to be optimal over $x^{(0)} + X_n$, where "optimal" means $\forall y \in x^{(0)} + X_n, \left\|x^{(n)} - x\right\|_G \leq \|y - x\|_G$ with $G^* = G > 0$. Let P_n be the projection onto X_n that is orthogonal w.r.t. $(\ |\ )_G$, i.e. $P_n^2 = P_n$, $\mathsf{Im}\, P_n = X_n$ and $GP_n = P_n^* G$. By lemma 3.15, $x^{(n)} - x = (I - P_n)(x^{(0)} - x)$ which is the same as $x^{(n)} = x^{(0)} - P_n e^{(0)}$.

Theorem 3.16 (Optimal error characterization theorem) *$\forall \tilde{x} \in x^{(0)} + X_n$, the following are equivalent:*

(i) $\tilde{x} = x^{(n)}$,

(ii) $(x - \tilde{x}) \perp X_n$ in G-inner product,

(iii) $\forall y \in x^{(0)} + X_n, f_G(\tilde{x}) \leq f_G(y)$, where $f_G(y) = (y \mid y)_G - 2(y \mid x)_G$.

We must find G, s.t. $(y \mid x)_G$ can be computed. There are two netural choices: $G = A$ when $A^* = A > 0$. Then, $f_A(y) = (y \mid Ay) - 2(y \mid b)$ (CG). $G = A^*A$ when $\det A \neq 0$. Then, $f_{A^*A}(y) = (Ay \mid Ay) - 2(Ay \mid b)$ (MINRES).

Observe that we can easily compute $P_n e^{(0)}$ if we can find a set of non-zero vectors $\left\{p^{(0)}, \cdots, p^{(\overline{n}-1)}\right\}$, s.t. $X_n = \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$ for $n \leq \overline{n}$. and $(p^{(m)} \mid p^{(n)})_G = 0$ for every $m < n < \overline{n}$. This means $\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$ is an orthogonal basis of X_n. Because $x \in x^{(0)} + \overline{X}$, $e^{(0)} = x^{(0)} - x \in \overline{X} \Rightarrow e^{(0)} = \sum_{k=0}^{\overline{n}-1} \alpha_k p^{(k)}$, where $\alpha_k = -\frac{(e^{(0)} \mid p^{(k)})_G}{(p^{(k)} \mid p^{(k)})_G}$. Hence, $-P_n e^{(0)} = \sum_{k=0}^{n-1} \alpha_k p^{(k)} - e^{(n)} = \sum_{k=0}^{\overline{n}-1} \alpha_k p^{(k)}$, so $\alpha_n = -\frac{(e^{(n)} \mid p^{(n)})_G}{(p^{(n)} \mid p^{(n)})_G}$. Hence $x^{(n+1)} = x^{(n)} + \alpha p^{(n)}$ for $0 \leq n < \overline{n}$ and $x^{(\overline{n})} = x$.

3.9 General minimum residual

Consider $Ax = b$, where $b \in \mathbb{R}^N$ and $A \in \mathbb{R}N \times N$ with $\det A \neq 0$. N is enormous ($\approx 10^7$). We have optimal error methods:

1. find a good norm;
2. identify subspaces (often Krylov);
3. find orthogonal vectors, s.t. $X_n = \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$.

For conjugate gradient, we have $A^* = A > 0$ and

1. $G = A$;
2. $X_n = \mathcal{K}_n(p^{(0)}, QA)$, where $p^{(0)} = Q(b - Ax^{(0)})$ and $Q^* = Q > 0$, s.t. $\mathsf{Cond}(QA) \ll \mathsf{Cond}(A)$;
3. CG algorithm.

QA is always positive definite w.r.t. A-norm.

For mininum residual, we have $A^* = A$, $\det A \neq 0$ and

1. $G = A^*A$;
2. $X_n = \mathcal{K}_n(r^{(0)}, A)$, where $r^{(0)} = b - Ax^{(0)}$;
3. MINRES algorithm.

It is hard to "pre-condition" it. It is hard to find Q, s.t. QA is self-adjoint w.r.t. A^*A-inner product.

For general minimum residual, $\det A \neq 0$ and

1. $G = A^*A$;
2. $X_n = \mathcal{K}_n(p^{(0)}, QA)$, where Q is invertible and $p^{(0)} = Q(b - Ax^{(0)}) = Qr^{(0)}$;
3. Arnoldi algorithm.

Lemma 3.17 (Arnoldi) *Let $G^* = G > 0$, $B \in \mathbb{R}^{N\times N}$ and $r \in \mathbb{R}^N$. Set $p^{(0)} = r$, $p^{(n+1)} = Bp^{(n)} - \sum_{m=0}^{n} \beta_{mn} p^{(m)}$, where $\beta_{mn} = \frac{(p^{(m)} \mid Bp^{(n)})_G}{(p^{(m)} \mid p^{(n)})_G}$. Then,*

(i) $p^{(n)} \in B^n p^{(0)} + \mathcal{K}_n(p^{(0)}, B)$;

(ii) $(p^{(m)} \mid p^{(n)})_G = 0$, *for* $m < n$;

(iii) $\mathcal{K}_n(p^{(0)}, B) = \mathsf{span}\left\{p^{(0)}, \cdots, p^{(n-1)}\right\}$.

Proof: Assume (i), (ii) and (iii) are trun for n. $p^{(n+1)} \in (B - \beta_{nn}I)p^{(n)} + \mathcal{K}_n(p^{(0)}, B) = Bp^{(n)} + \mathcal{K}_{n+1}(p^{(0)}, B) = B^{n+1}p^{(0)} + B\mathcal{K}_n(p^{(0)}, B) + \mathcal{K}_{n+1}(p^{(0)}, B) = B^{n+1}p^{(0)} + \mathcal{K}_{n+1}(p^{(0)}, B) \Rightarrow$ (i) holds.

$(p^{(m)} \mid p^{(n+1)})_G = (p^{(m)} \mid Bp^{(n)})_G - \sum_{k=0}^{n} \beta_{kn}(p^{(m)} \mid p^{(k)})_G = (p^{(m)} \mid Bp^{(n)})_G - \beta_{mn}(p^{(m)} \mid p^{(m)})_G = 0 \Rightarrow$ (ii) holds. $\mathcal{QED}$

3.10 Another algorithm

Consider $Ax = b$ with $\det A \neq 0$. Consider an iteration method that measure their error in $(\ \mid\ )_G$, where $G = A^*QA$ and $Q \approx (AA^*)^{-1}$ with $Q^* = Q > 0$, i.e. $\mathsf{Cond}(QAA^*)$ is as small as possible while multiplication by Q is quick. This means $(e \mid e) \approx (e \mid e)_G = (r \mid r)_Q$. So we need to construct a G-orthogonal set of vectors.

Lemma 3.18 *Let $G = A^*QA$ and $H = QAA^*Q$. Choose $v^{(0)} \in \mathbb{R}^N$ and set $u^{(0)} = A^*Qv^{(0)}$. Define $v^{(n+1)} = Au^{(n)} - \beta_n v^{(n)}$ and $u^{(n+1)} = A^*Qv^{(n+1)} - \gamma_n u^{(n)}$, where $\beta_n = \frac{(u^{(n)} \mid u^{(n)})_G}{(v^{(n)} \mid v^{(n)})_H}$ and $\gamma_n = \frac{(v^{(n+1)} \mid v^{(n+1)})_H}{(u^{(n)} \mid u^{(n)})_G}$. Then,*

(i) $(v^{(m)} \mid v^{(n)})_H = (u^{(m)} \mid u^{(n)})_G = 0$ *for* $m < n < \overline{n}$, *where $\overline{n}$ is the maximal Krylov subspace;*

(ii) $u^{(n)} \in (A^*QA)^n u^{(0)} + \mathcal{K}_n(u^{(0)}, A^*QA)$;

(iii) $v^{(n)} \in (AA^*Q)^n v^{(0)} + \mathcal{K}_n(v^{(0)}, AA^*Q)$;

(iv) $\mathcal{K}_n(u^{(0)}, A^*QA) = \mathsf{span}\left\{u^{(0)}, \cdots, u^{(n-1)}\right\}$;

(v) $\mathcal{K}_n(v^{(0)}, AA^*Q) = \mathsf{span}\left\{v^{(0)}, \cdots, v^{(n-1)}\right\}$.

Proof: By induction, $(v^{(m)} \mid v^{(n+1)})_H = (v^{(m)} \mid Au^{(n)})_H - \beta_n(v^{(m)} \mid v^{(n)})_H = (A^*Qv^{(m)} \mid u^{(n)})_G - \beta_n(v^{(m)} \mid v^{(n)})_H = (u^{(m)} \mid u^{(n)})_G + \gamma_{m-1}(u^{(m-1)} \mid u^{(n)})_G - \beta_n(v^{(m)} \mid v^{(n)})_H = 0$.

$(u^{(m)} \mid u^{(n+1)})_G = (u^{(m)} \mid A^*Qv^{(n+1)})_G - \gamma_n(u^{(m)} \mid u^{(n)})_G = (Au^{(m)} \mid v^{(n+1)})_H - \gamma_n(u^{(m)} \mid u^{(n)})_G = (v^{(m+1)} \mid v^{(n+1)})_H + \beta_n(v^{(m)} \mid v^{(n+1)})_H - \gamma_n(u^{(m)} \mid u^{(n)})_G = 0$. Therefore, (i) is true.

Because $u^{(0)} = A^*Qv^{(0)}$, $\mathcal{K}_n(u^{(0)}, A^*QA) = A^*Q\mathcal{K}_n(v^{(0)}, AA^*Q)$. $\mathcal{QED}$

Algorithm Choose $x^{(0)}$, initialize $r^{(0)} = b - Ax^{(0)}$, $v^{(0)} = r^{(0)}$ $q^{(0)} = A^*Qr^{(0)}$, $u^{(0)} = q^{(0)}$ and $p^{(0)} = Au^{(0)}$. Begin loop on n:

$\alpha_n = \frac{(p^{(n)} \mid r^{(n)})_Q}{(p^{(n)} \mid p^{(n)})_Q}$
$x^{(n+1)} = x^{(n)} + \alpha_n u^{(n)}$
$r^{(n+1)} = r^{(n)} - \alpha_n p^{(n)}$
Check for convergence

3.11 Preconditioning for Krylov optimal error methods

In general, these methods solve $Ax = b$ by picking $x^{(n)}$, s.t. $\left\|x^{(n)} - x\right\|_G = \min\left\{\|y - x\|_G \;\middle|\; y \in x^{(0)} + \mathcal{K}_n(p^{(0)}, K)\right\}$. One needs to find $G, p^{(n)}, K$, s.t.

(i) one can compute a G orthogonal basis of $\overline{\mathcal{K}}(p^{(0)}, K)$, s.t. $\mathcal{K}_n(p^{(0)}, K) = \mathsf{span}\left\{p^{(n)}, \cdots, p^{(n-1)}\right\}$;

(ii) $(e^{(n)} \mid p^{(n)})_G$ can be computed.

The three basic cases we have concerned:

1. K is G-positive definite, i.e. $GK = K^*G > 0$. e.g. CG, $A = A^* > 0$ for $G = A$ and $K = A$.
2. K is G-self-adjoint, i.e. $GK = K^*G$. e.g. MINRES, $G = A^2$ and $K = A$ (Lanczos).
3. A is invertible. e.g. GMRES, $G = A^*A$ and $K = A$ (Arnoldi).

Given $Ax = b$, find Q quick "inverse":

1. precondition CG, $Q^* = Q > 0$, $G = A$ and $K = QA$;
2. precondition MINRES, $Q^* = Q > 0$, $G = AQA$ and $K = QA$;
3. precondition GMRES, $G = A^*Q^*QA$ and $K = QA$ for Q invertible. Recall $\mathsf{Cond}_2(K) = \|K\|_2 \left\|K^{-1}\right\|_2$. Since $\|K\|_2 = \sqrt{\rho_{\mathsf{sp}}(K^*K)}$, $\mathsf{Cond}_2(K) = \sqrt{\frac{\max\{x \in \mathsf{sp}(K^*K)\}}{\min\{x \in \mathsf{sp}(K^*K)\}}}$, where $K^*K = A^*Q^*QA$ and $AK^*KA^{-1} = AA^*Q^*Q$.

Example: Let $A = D - L - L^*$ with $A^* = A > 0$ and $Q = (D - L^*)^{-1}D(D - L)^{-1}$ (symmetic Gauss-Seidel). SSOR: $Q = (D - \omega L^*)^{-1}\frac{1}{\omega}D(D - \omega L)^{-1}$.
Example: $A^* = A$, same as above because Q's are positive definite. For $A > 0$, $A = LL^*$, $A \sim L_I L_I^*$ and $Q = (L_I^*)^{-1}L_I^{-1}$.

4 Eigenvalue problems

Let $A \in \mathbb{R}^{N\times N}$ that is diagonalizable. How might you compute the eigenvalues and eigenvectors? Basic method to do this is the **power method**.

Suppose A is diagonalizable within complexes. i.e. $\exists \{v_i\}_{i=1}^N \subset \mathbb{C}^N$ that are linearly independent. $Av_i = \lambda_i v_i, \lambda_i \in \mathbb{C}$. Let $V = \begin{pmatrix} v_1 & \cdots & v_N \end{pmatrix}$ with $\det V \neq 0$. Then, $AV = V\Lambda$, where $\Lambda = \mathsf{Diag}(\lambda_1, \cdots, \lambda_N) \stackrel{\mathsf{def}}{=} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$. Then, $A = V\Lambda V^{-1}$ and $\Lambda = V^{-1}AV$.

$\forall r \in \mathbb{C}^N, r = \alpha_1 v_1 + \cdots + \alpha_N v_N$. Then, $A^k r = \alpha_1 \lambda_1^k v_1 + \cdots + \alpha_N \lambda_N^k v_N$. Suppose $|\lambda_1| > |\lambda_i|$ for $i > 1$. $\frac{1}{\lambda_1^k}A^k r = \alpha_1 v_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \alpha_N\left(\frac{\lambda_N}{\lambda_1}\right)^k v_N$. Clearly, as $k \to \infty$, $\frac{1}{\lambda_1^k}A^k r \to \alpha_1 v_1$, $\frac{1}{|\lambda_1|^k}\left\|A^k r\right\| \to |\alpha_1|\,\|v_1\|$, $\left(\frac{|\lambda_1|}{\lambda_1}\right)^k \frac{A^k r}{\|A^k r\|} \to \frac{\alpha_1}{|\alpha_1|}\frac{v_1}{\|v_1\|}$. $\frac{A^k r}{\|A^k r\|} \to \frac{v_1}{\|v_1\|}$, $\frac{A^{k+1}r}{\|A^k r\|} \to \lambda_1 \frac{v_1}{\|v_1\|}$, $\frac{(A^k r \mid A^{k+1}r)}{\|A^k r\|^2} \to \lambda_1$.

The game is to repeat this with $(A - \mu I)^{-1}$ in place of A, where μ is a guess at an eigenvalue. Recall $\mathsf{sp}((A - \mu I)^{-1}) = \left\{\frac{1}{\lambda - \mu} \;\middle|\; \lambda \in \mathsf{sp}(A)\right\}$.

Lemma 4.1 (Similarity transformation) *Let $A, B, U \in \mathbb{R}^{N\times N}$ with U invertible and $A = U^{-1}BU$. Then, $\mathsf{sp}(A) = \mathsf{sp}(B)$.*

Let $B = V^{-1}AV$ and $\delta \ll 1$. $\delta\tilde{B}_\delta = (V+\delta\tilde{V})^{-1}(A+\delta\tilde{A})(V+\delta\tilde{V}) - V^{-1}AV = \delta(V^{-1}\tilde{A}V + V^{-1}A\tilde{V} - V^{-1}\tilde{V}V^{-1}AV) + O(\delta^2)$. Hence, $\left\|\tilde{B}_\delta\right\| \approx \left\|V^{-1}\tilde{A}V + V^{-1}AV(V^{-1}\tilde{V}) - (V^{-1}\tilde{V})V^{-1}AV\right\| \leq \mathsf{Cond}(V)\left\|\tilde{A}\right\| + 2\|B\|\left\|V^{-1}\tilde{V}\right\|$. The best control is given when $\mathsf{Cond}(V) = 1$. This means $\left\|V^TV\right\| = \|V\|^2 = 1$. and $\left\|(V^TV)^{-1}\right\| = 1$. $\mathsf{sp}(V^TV) = \{1\} \Rightarrow V^TV = I$. Therefore, V is an orthogonal matrix.

Orthogonal matrix if $VV^T = V^TV = I$. Let Q be an orthogonal matrix. Then, $\det Q = \pm 1$ and $\mathsf{sp}(Q) \subset \{\lambda \in \mathbb{C} \mid |\lambda| = 1\}$. Permutations, rotations and reflection are examples of orthogonal matrice.

Householder matrix Simple reflections have the form $Q_H = I - 2\frac{uu^T}{|u|^2} = I - 2\hat{u}\hat{u}^T$, where $u \in \mathbb{R}^N \setminus \{0\}$ and $\hat{u} = \frac{u}{|u|}$. Let $e_j = \begin{pmatrix}0 & \cdots & 0 & \underbrace{1}_{j^{th}} & 0 & \cdots & 0\end{pmatrix}^T$. Then,

$$Q_H(e_j) = \begin{pmatrix}1 &&&&&& \\ & \ddots &&&&& \\ && 1 &&&& \\ &&& -1 &&& \\ &&&& 1 && \\ &&&&& \ddots & \\ &&&&&& 1\end{pmatrix}.$$

Givens matrix Simple rotations have the form $Q_G = e^{\theta\frac{u\wedge v}{|u\wedge v|}}$, where $u, v \in \mathbb{R}^N$ are linearly independent and $u\wedge v = uv^T - vu^T$ and $|u\wedge v|^2 = |u|^2|v|^2 - (u^Tv)^2$. $\mathsf{sp}(Q_G) = \{1, e^{i\theta}, e^{-i\theta}\}$ $1 \in \mathsf{sp}(Q_G)$ with multiplicity $N-2$. $Q_G = e^{\theta\hat{u}\wedge\hat{v}}$, where $\hat{u}\cdot\hat{v} = 0$. $Q_G(e_je_k) =$

$$\begin{pmatrix}1 &&&&&&&&&& \\ & \ddots &&&&&&&&& \\ && 1 &&&&&&&& \\ &&& \cos\theta &&&& \sin\theta &&& \\ &&&& 1 &&&&&& \\ &&&&& \ddots &&&&& \\ &&&&&& 1 &&&& \\ &&& -\sin\theta &&&& \cos\theta &&& \\ &&&&&&&& 1 && \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& 1\end{pmatrix}.$$

Hessenbreg form $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ is called upper (lower) Hessenbreg if $a_{ij} = 0$ for $i > j+1$ ($j > i+1$).

Lemma 4.2 *$\forall A \in \mathbb{R}^{N\times N}$, there exists an orthogonal matrix Q, s.t. Q^TAQ is upper Hessenberg.*

Proof: We will prove for $k = 1, \cdots, N-1$, $\exists Q_k \in \mathbb{R}^{N\times N}$ is orthogonal, s.t. $Q_k^TAQ_k = \begin{pmatrix}H_k & B_k^T \\ C_k & A_k\end{pmatrix}$, where $H_k \in \mathbb{R}^{k\times k}$ is upper Hessenberg, $A_k \in \mathbb{R}^{(N-k)\times(N-k)}$, $B_k, C_k \in \mathbb{R}^{(N-k)\times k}$ with $C_k = \begin{pmatrix}0 & c_k\end{pmatrix}$, $c_k \in \mathbb{R}^{N-k}$.

This is clearly true for $k = 1$. We have $H_1 = (a_{11})$, $C_1 = \begin{pmatrix}a_{21} \\ \vdots \\ a_{N1}\end{pmatrix}$, $B_1 = \begin{pmatrix}a_{12} \\ \vdots \\ a_{1N}\end{pmatrix}$, $A_1 = \begin{pmatrix}a_{22} & \cdots & a_{2N} \\ \vdots && \\ a_{N2} & \cdots & a_{NN}\end{pmatrix}$.

Suppose it is true for some $k \geq 1$. For $k+1$, let $\hat{c_k} = \frac{c_k}{|c_k|}$, $\hat{e} = \begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^T \in \mathbb{R}^{N-k}$ and $\tilde{Q} = \begin{pmatrix}I & 0 \\ 0 & I - 2\hat{u}\hat{u}^T\end{pmatrix}$, where $\hat{u} = \begin{cases}\frac{\hat{c_k}-\hat{e}}{|\hat{c_k}-\hat{e}|} & \text{, if } \hat{c_k} \neq \hat{e} \\ 0 & \text{, otherwise}\end{cases}$. Set $Q_{k+1} = Q_k\tilde{Q}$. $Q_{k+1}^TAQ_{k+1} = \tilde{Q}Q_k^TAQ_k\tilde{Q} = \tilde{Q}\begin{pmatrix}H_k & B_k^T \\ C_k & A_k\end{pmatrix}\tilde{Q} = \begin{pmatrix}H_k & B_k^T(I-2\hat{u}\hat{u}^T) \\ (I-2\hat{u}\hat{u}^T)C_k & (I-2\hat{u}\hat{u}^T)A_k(I-2\hat{u}\hat{u}^T)\end{pmatrix}$. $\mathcal{QED}$

Corollary 4.3 *If $A^T = A$, Q^TAQ is symmetric tri-diagonal.*

4.1 QR method

Theorem 4.4 (QR factorization) *$\forall A \in \mathbb{R}^{N\times N}$, there exists an orthogonal matrix Q and an upper triangular matrix R, s.t. $A = QR$.*

Proof: We show that for $k = 1, \cdots, N-1$, $\exists Q_k$ orthogonal, s.t. $A = Q_kR_k$, where $R_k = \begin{pmatrix}\tilde{R}_k & B_k \\ 0 & \tilde{A}_k\end{pmatrix}$, $\tilde{R}_k$ is upper triangular.

For $k = 1$, let $A = \begin{pmatrix}a & \tilde{D}\end{pmatrix}$, where $a = \begin{pmatrix}a_{11} & \cdots & a_{N1}\end{pmatrix}^T$. Let $\hat{e} = \begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^T \in \mathbb{R}^N$ and $\hat{u} = \begin{cases}\frac{\hat{a}-\hat{e}}{|\hat{a}-\hat{e}|} & \text{, if } \hat{a} \neq \hat{e} \\ 0 & \text{, otherwise}\end{cases}$. Let $Q_1 = I - 2\hat{u}\hat{u}^T$ and $R_1 = \begin{pmatrix}|a|\hat{e} & R_1\end{pmatrix} = Q_1A$. Therefore, $A = Q_1R_1$.

For $k \geq 1$, set $Q_{k+1} = \begin{pmatrix}I & 0 \\ 0 & I - 2\hat{u_k}\hat{u_k}^T\end{pmatrix}Q_k$. $\mathcal{QED}$

One of the best method for computing eigenvalues and eigenvectors is the so-called QR-hmethod, which is based on QR decomposition.

QR method Given $A \in \mathbb{R}^{N\times N}$, construct a sequence $\{A_i\}_{i=0}^\infty$ as follow:

1. $A_0 = A$.
2. Decomposite $A_i = Q_iR_i$, where Q_i is orthogonal and R_i is upper triangular.
3. Set $A_{i+1} = R_iQ_i$.

Observe $A_{i+1} = Q_i^TA_iQ_i = R_iA_iR_i^{-1}$ (if A^{-1} exists, then R_i^{-1} exists). $A_{i+1} \sim A_i \sim A$. Hessenberg form and symmetric triangular form are preserved. $A^i = Q_0\cdots Q_{i-1}R_{i-1}\cdots R_0$ (relation to power method).

$A_i \to A_\infty$, where A_∞ is diagonal if $A^T = A$. In the case $A \in \mathbb{C}^{N\times N}$, A'_∞ is upper triangular, where $A = U^*A'_\infty U$ with U unitary by **Schur's lemma of linear algebra**.

In the real case,

$$A_\infty = \begin{pmatrix} \lambda_1 & * & \cdots & & & & & * \\ & \ddots & \ddots & & & & & \vdots \\ & & \lambda_{m_1} & * & \cdots & & & * \\ & & & u_1 & v_1 & * & \cdots & * \\ & & & -v_1 & u_1 & * & \ddots & \vdots \\ & & & & & \ddots & * & * \\ & & & & & & u_{m_2} & v_{m_2} \\ & & & & & & -v_{m_2} & u_{m_2} \end{pmatrix},$$

where $\lambda_1, \cdots, \lambda_{m_1}$ are real eigenvalues and $u_1 \pm iv_1, \cdots, u_{m_2} \pm iv_{m_2}$ are complex eigenvalues. A_∞ is unique up to permutations.

Shifted-QR method The QR method can be improved by shifting. $A_i - \sigma_i I = Q_i R_i$ and $A_{i+1} - \sigma_i I = R_i Q_i$.

4.2 Iso-spectral flows

Let $J(t) \in \mathbb{R}^{N\times N}$ be continuous and $Q(t) \in \mathbb{R}^{N\times N}$ with $Q(0) = I$, s.t.

$$\frac{dQ(t)}{dt} = J(t)Q(t). \tag{4.6}$$

$\det Q(t) = \det(Q(0))e^{\int_0^t \mathsf{Tr}(J(t'))dt'}$. $\det Q(t) = 1$ if $\forall t, \mathsf{Tr}(J(t)) = 0$.

Let $H_0 \in \mathbb{R}^{N\times N}$ and $H(t) = Q(t)H_0Q(t)^{-1}$. Note that $\mathsf{sp}(H(t)) = \mathsf{sp}(H_0)$, i.e. this is "iso-spectral". $\frac{dH(t)}{dt} = J(t)Q(t)H_0Q(t)^{-1} - Q(t)H_0Q(t)^{-1}J(t)$. Hence

$$\frac{dH(t)}{dt} = J(t)H(t) - H(t)J(t). \tag{4.7}$$

If H satisfies equation 4.7, then so does H^k and H^{-1}. $\frac{dH^T}{dt} = -J^TH^T + H^TJ^T$. If $J^T = -J$, then H^T satisfies equation 4.7. $JH - HJ$ is symmetric if $H^T = H$.

Proof: Since Q satisfies equation 4.6, $\frac{dQ^T}{dt} = Q^TJ^T = -Q^TJ$ and $\frac{dQ^{-1}}{dt} = -Q^{-1}\frac{dQ}{dt}Q^{-1} = -Q^{-1}J$. $Q^{-1}(0) = Q^T(0) = I$. Therefore, $Q^{-1}(t) = Q^T(t)$. When $J^T = -J$, then $(JH - HJ)^T = JH^T - H^TJ$. $\mathcal{QED}$

We will consider symmetric cases: find $J(H)$, s.t. $H(t) \to$ diagonal as $t \to \infty$.

We have to identify a mapping $\mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$, $H \mapsto J(H)$ with $J(H)^T = -J(H)$, s.t. $H(0) = H_0$, H and J satisfy equation 4.7 and $H(t) \to$ "simple" as $t \to \infty$. In particular, when $H_0^T = H_0$, then $H(t)^T = H(t)$ for every t and $H(t) \to$ diagonal as $t \to \infty$.

Example: Let $H = D + L + L^T$, where D is diagonal and L is strictly lower triangular. Set $J = L - L^T$. Then,

$$\frac{dH}{dt} = 2(LL^T - L^TL) + LD - DL + DL^T - L^TD. \tag{4.8}$$

where $LD - DL$ is strictly lower triangular and $DL^T - L^TD$ is strictly upper triangular. Assume H_0 has been reduced to tridiagonal. For $H = \begin{pmatrix} b_0 & a_1 & & \\ a_1 & b_1 & \ddots & \\ & \ddots & \ddots & a_{N-1} \\ & & a_{N-1} & b_{N-1} \end{pmatrix}$, $L^TL = \begin{pmatrix} 0 & & & \\ & a_1^2 & & \\ & & \ddots & \\ & & & a_{N-1}^2 \end{pmatrix}$, $LL^T = \begin{pmatrix} a_1^2 & & & \\ & \ddots & & \\ & & a_{N-1}^2 & \\ & & & 0 \end{pmatrix}$. $LL^T - L^TL$ is diagonal. Hence, the tridiagonal form is preserved by equation 4.8 with $\frac{dD}{dt} = 2(LL^T - L^TL)$, $\frac{dL}{dt} = LD - DL$,

$$\begin{cases} \frac{db_i}{dt} = 2(a_j^2 - a_{j+1}^2) \text{ for } j = 0, \cdots, N-1 \\ \frac{da_i}{dt} = a_j(b_{j-1} - b_j) \text{ for } j = 1, \cdots, N-1 \end{cases} \tag{4.9}$$

with $a_0 = a_N = 0$. The only stationary (fixed) points of equation 4.9 are when it is diagonal, i.e. $a_1 = \cdots = a_{N-1} = 0$. This will be asymptotically stable provided $b_0 < \cdots < b_{N-1}$. If $\forall j, a_j \neq 0$ for H_0, then the eigenvalues of H_0 are simple.

Example: For $H = \begin{pmatrix} b_0 & a \\ a & b_1 \end{pmatrix}$, $\frac{db_0}{dt} = -2a^2$, $\frac{db_1}{dt} = 2a^2$ and $\frac{da}{dt} = a(b_0 - b_1)$. Let $s = \frac{b_1 + b_0}{2}$ and $c = \frac{b_1 - b_0}{2}$. Then, $\frac{ds}{dt} = 0$, $\frac{dc}{dt} = 2a^2$ and $\frac{da}{dt} = -2ac$. The solution is $s(t) = s_0$, $c(t) = r\frac{c_0 + r\tanh(2rt)}{r + c_0\tanh(2rt)}$ and $a(t) = r\frac{a_0 \mathsf{sech}(2rt)}{r + c_0\tanh(2rt)}$, where $r = \sqrt{a_0^2 + c_0^2}$. (Remarks: This is "better" than QR method. For $A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ (reflection), $A = QR \Rightarrow Q = A$ and $R = I$.)

Theorem 4.5 *Let H_0 be symmetric tridiagonal and $H(t) = \begin{pmatrix} b_0 & a_1 & & \\ a_1 & b_1 & \ddots & \\ & \ddots & \ddots & a_{N-1} \\ & & a_{N-1} & b_{N-1} \end{pmatrix}$ satisfy equation 4.9. Then, $H(t) \to$ diagonal as $t \to \infty$.*

Proof: Let $s_m = \sum_{j=0}^{m-1} b_j$ be the m^{th} partial trace. $\frac{ds_m}{dt} = -2a_m^2 < 0$ for $m = 1, \cdots, N-1$. One can show that $\frac{d}{dt}\left(\sum_{j=0}^{N-1} b_j^2 + 2\sum_{j=1}^{N-1} a_j\right) = 0$. Hence, $a_j(t)$ and $b_j(t)$ are bounded $\Rightarrow s_m(t)$ are bounded. Therefore, $\lim_{t\to\infty} s_m(t) = s_m^\infty$ exists. Moreover, $\lim_{t\to\infty} s_m(t+h) = s_m^\infty$ uniformly in $h \geq 0$. $\lim_{t\to\infty} b_m(t+h) = \lim_{t\to\infty}(s_{m+1}(t+h) - s_m(t+h)) = s_{m+1}^\infty - s_m^\infty \stackrel{\mathsf{def}}{=} b_m^\infty$ uniformly in $h \geq 0$.

By equation 4.9, $a_j(t') = a_j(t)e^{\int_t^{t'} b_{j-1}(t'') - b_j(t'')dt''}$. Combine this with $\int_t^{t+h} a_m(t')^2 dt' = -\frac{1}{2}(s_m(t+h) - s_m(t))$. Therefore,

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{1}{h}\int_t^{t+h} a_m(t')^2 dt' \\ &= \lim_{t\to\infty} \frac{a(t)^2}{h}\int_t^{t+h} e^{2\int_t^{t'} b_{m-1}(t'') - b_m(t'')dt''} dt' \\ &\geq \lim_{t\to\infty} a_m(t)^2 \\ &= 0 \\ & \quad \frac{1}{h}\int_t^{t+h} e^{2\int_t^{t'} b_{m-1}(t'') - b_m(t'')dt''} dt' \\ &= \frac{1}{h}\int_0^h e^{2\int_0^{h'} b_{m-1}(t+h'') - b_m(t+h'')dh''} dh' \end{aligned}$$

$$
\begin{aligned}
&\rightarrow \quad \frac{1}{h}\int_0^h e^{2(b_{m-1}^\infty - b_m^\infty)h'}\,dh' \\
&= \quad \frac{e^{2(b_{m-1}^\infty - b_m^\infty)h}}{2(b_{m-1}^\infty - b_m^\infty)h} \\
&\geq \quad 1
\end{aligned}
$$

$\mathcal{QED}$

More generally, unitary iso-spectral flows have the form $\frac{dH}{dt} = JH - HJ$, where $J = L - L^T$, $f(H) = D + L + L^T$ and f is any function analytic in a neighbourhood of $\mathsf{sp}(H_0)$.

There is an amazing fact: $H(t) = Q(t)^T H_0 Q(t)$, where $e^{tf(H_0)} = Q(t)R(t)$, $Q(t)$ is orthogonal and $R(t)$ is strictly upper triangular with positive diagonal. Consider $f(z) = \log(z)$, $H_0^T = Q(t)R(t)$. Set $t = 1$, $H_0 = Q(1)R(1)$. $H(1) = Q(1)^T H_0 Q(1) = R(1)Q(1)$. (For tridiagonal, Toda Lattice.)

Theorem 4.6 (Gershgorin circle theorem) *Let* $A = D - W$ *and* $D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_N \end{pmatrix}$. *Then,* $|\lambda - d_j| \leq \|W\|$.

Let $A = D - tW$ for $t \in [0, 1]$. Then, $|\lambda - d_j| \leq t\,\|W\|$.