Summary Sheet: MATH 3NA3 - Numerical Linear Algebra

Floating Point (FP) Number Systems

FP system definition: $\begin{cases} \beta & : \text{Base} \\ P & : \text{Precision} \\ L & : \text{Exponent min} \\ U & : \text{Exponent max} \end{cases}$

$x = \pm(d_0.d_1d_2\dots d_{P-1}) \times \beta^E, \quad E \in [L, U]$

$x = \pm\left(\sum_{i=0}^{P-1} \frac{d_i}{\beta^i}\right)\beta^E, \quad d_i \in [0, \beta-1]$

$\epsilon_{mach} = \beta^{1-P}$

Absolute rounding error $= |fl(x) - x| \leq |x|\epsilon_{mach}$

RRE (Relative representation error) $= \frac{|fl(x)-x|}{|x|}$

maxRRE(x) $= \begin{cases} \epsilon_{mach} & \text{if round up/down} \\ \frac{\epsilon_{mach}}{2} & \text{if round to nearest} \end{cases} \leq \epsilon_{mach}$

$fl(x) = x(1+\delta), \delta = \frac{fl(x)-x}{x}, |\delta| \leq \epsilon_{mach}$

Min value representable $> 0 = \beta^L$

Max value representable $= \beta^{U+1}(1-\beta^{-P})$

$|\mathbb{F}| = 2(\beta-1)(\beta^{P-1})(U-L+1)+1$

Properties of FP systems:

1. Finite: $\exists$ overflow and underflow
2. Discrete: $\exists$ gaps btwn nums $\in \mathbb{F}$
3. Non-Uniform: Nums $\in \mathbb{F}$ $\neg$(evenly distributed)

Floating Point Operations

$x \circledast y := fl(x \star y) = (x \star y)(1+\delta) \quad |\delta| < \epsilon$

Fundamental Axiom: $\frac{|x \circledast y - (x \star y)|}{|x \star y|} \leq \epsilon = \frac{1}{2}\epsilon_{mach}$

Cancellation Error: subtract similar sized nums

General Algebra

$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

eigvals of $A^TA \in \mathbb{R}^{n\times n} = [\sigma_1 \dots \sigma_n]^2$ from A's SVD

- $a^2 - b^2 = (a-b)(a+b)$
- Singular matrix:= not invertible

SPD:

- A SPD iff $A^T = A$ & (strict diag. dom $\Leftrightarrow \lambda_{min} > 0$)
- A SPD iff B^TAB is SPD for nonsingular B
- if A SPD, principle submatrices SPD

Gershgorin's thm:

- any eigenvalue of A is in at least one of the closed disks $D(a_{ii}, R_{ii}, R_{ii} = \sum_{j\neq i}|A_{ij}$

Diagonal dominance: properties:

1. If A strict diag dom, A invertible
2. $A^T = A$, if A strict diag dom, and $A_{ii} > 0$, then A SPD.

pf of 1 by contradiction: sps non-invertible, then $\exists$ row of 0's, that row's diag not greater than sum of others, so contradiction.

Matrix Norms & SVD

Matrix Norm Properties:

1. $||A|| \geq 0, ||A|| = 0$ iff $A = 0$
2. $||cA|| = |c| \times ||A||$
3. $||A+B|| \leq ||A|| + ||B||$

$||A||_p = \max_{||\vec{x}||_p=1} ||A\vec{x}||_p = \max_{\vec{x}\neq 0} \frac{||A\vec{x}||_p}{||\vec{x}||_p}$

$||A||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$ (max abs col sum)

$||A||_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$ (max abs row sum)

$||A||_2 = \sqrt{\lambda_{max}(A^TA))} = \sigma_{max}(A)$

$||AA^T||_2 = ||A^TA||_2 = \lambda_{max}(A^TA)) = \sigma_{max}(A)^2$

Induced Matrix Norm Properties:

1. $||A\vec{x}|| \leq ||A|| \times ||\vec{x}||$
2. $||AB|| \leq ||A|| \times ||B||$
3. $||Q_1AQ_2||_2 = ||A||_2$
4. $||Q||_2 = 1$
5. $||A^T||_2 = ||A||_2$
6. $||I|| = 1$

Vector Norms: $||\vec{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

1. $||\vec{x}|| \geq 0, ||\vec{x}|| = 0$ iff $\vec{x} = 0$
2. $||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$
3. $||\alpha\vec{x}|| = |\alpha| \times ||\vec{x}||$

- $||\cdot||_a$, $||\cdot||_b$ equiv. iff $\exists c$'s s.t. $c_1||\vec{x}||_b \leq ||\vec{x}||_b \leq c_2||\vec{x}||_b$, meaning can exchange in p-norm applications
- $||\vec{x}||_\infty = \max_{1\leq j\leq n} |x_j|$
- $||\vec{x}||_2 \leq ||\vec{x}||_1 \leq \sqrt{n}||\vec{x}||_2$
- $||\vec{x}||_\infty \leq ||\vec{x}||_1 \leq n||\vec{x}||_\infty$

SVD:

- $A = U\Sigma V^T \Leftrightarrow A^{-1} = V\Sigma^{-1}U^T$

  A symm. pos. definite $\Rightarrow$ diagonalization = SVD
- Singular values of $AA^T = A^TA = \sigma_1^2, \dots, \sigma_n^2$
- $\det(A) = \prod_{i=1}^{n} \sigma_i$

MATLAB

Command	Purpose
realmax/realmin	return max/min float
eps	return ϵ_{mach}
norm$(\vec{x}, p)$, norm(A, p)	$\|\|\vec{x}\|\|_p$, $\|\|A\|\|_p$
cond(A, p)	$\kappa_p(A)$
pinv(A)	pseudo-inv A

Error, Sensitivity & Big O

input/output perturbation: $x + \delta x$, $f + \delta f$

Absolute Condition Number: $\hat{\kappa} = ||f'(x)||$

Relative Condition Number: $\kappa = \frac{||f'(x)|| ||x||}{||f(x)||}$

Absolute Error: $||\tilde{f}(x) - f(x)||$, $\quad \tilde{f} :=$ num mthd outpt

Relative Error: $\frac{||\tilde{f}(x) - f(x)||}{||f(x)||}$

Algo accurrate iff $\frac{||\tilde{f}(x) - f(x)||}{||f(x)||} = O(\epsilon_{mach})$

Backward Error: $|\tilde{x} - x|$

- Attribute output err to Δ inp
- $\tilde{f}(x) = f(\tilde{x})$, solve for $\tilde{x}$
- rel fwd err $\leq \kappa$ rel back err

Solutions of Linear Equations 1

Condition Number: $\kappa_p(A) = ||A||_p||A^{-1}||_P$

1. $\kappa_p(A) \geq 1$
2. $\kappa_p(I) = 1$
3. $\kappa_p(\alpha A) = \kappa_p(A) \forall$ scalars α

- $\kappa_2(A) = \frac{\sigma_{max}}{\sigma_{min}}$

Condition Number of solving $A\vec{x} = \vec{b}$:

Math: $f(A, \vec{b}) = \vec{x} \Leftrightarrow A\vec{x} = \vec{b}$

Compute: $\tilde{f}(A, \vec{b}) = \tilde{\vec{x}} = \vec{x} + \delta\vec{x} \Leftrightarrow (A + \delta A) = \vec{b} + \delta\vec{b}$

- Relative Error: $\frac{||\delta\vec{x}||}{||\vec{x}||}$
- Relative Backward Error: $\frac{||\delta\vec{b}||}{||\vec{b}||}$, $\frac{||\delta A||}{||A||}$
- Algo Backward Stable iff $\frac{||\delta\vec{b}||}{||\vec{b}||}$, $\frac{||\delta A||}{||A||} = O(\epsilon_{mach})$

Residual Properties: $\vec{r} = \vec{b} - \tilde{\vec{x}}$

1. $\frac{||\delta\vec{x}||}{\vec{x}} \leq \kappa(A)\frac{\vec{r}}{\vec{b}}$
2. $\frac{||\delta A||}{A} \geq \frac{\vec{r}}{A\vec{x}}$

- Problem: $A\vec{x} = \vec{b}$
- Computation: $(A + \delta A)\tilde{\vec{x}} = \vec{b}(1+\delta)$
- $\tilde{\vec{x}} = \vec{x} + \delta\vec{x}$

Solutions of Linear Equations 2: LU

LU factorization: $A_{n\times n} = LU$

Steps:

1. Initialize L as identity matrix.
2. Initialize U as zero matrix.
3. For each column j:
   a. Set elements of U in row i up to j.
   b. Set elements of L from row $j+1$ to n.

Pivoting:

With partial pivoting: $PA = LU$.
P - Permutation matrix.

$$M_k = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & \vdots & \dots & 0 \\ 0 & \dots & -m_{k+1} & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -m_n & 0 & \dots & 1 \end{bmatrix}$$

- $L_k = M_k^{-1} = I + \vec{m_k}\vec{e_k}^T$
- $U = M_{n-1}M_{m-2}\dots M_2M_1A$
- $L = M_1^{-1}M_2^{-1}\dots M_{m-2}^{-1}M_{n-1}^{-1}$
- LU factorization not backw stable, PLU is.
- LU factorization $O(\frac{2}{3}n^3)$ flops

Cholesky: $A = LL^T$ (unique, for SPD A)

$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$ diag elements

$l_{ij} = \frac{aij - \sum_{k=1}^{j-1} l_{ik}l_j}{l_{jj}}$ other elements

- Cholesky factorization $O(\frac{1}{3}n^3)$ flops

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \; L = \begin{bmatrix} \sqrt{a} & 0 \\ \frac{b}{\sqrt{a}} & \sqrt{c - \left(\frac{b}{\sqrt{a}}\right)^2} \end{bmatrix}$$

Iterative Methods

$A = M - N$, M nonsingular

$\vec{x}^{(k+1)} = M^{-1}N\vec{x}^{(k)} + M^{-1}\vec{b} = G\vec{x}^{(k)} + \vec{c}$

Terminate when $||\vec{r}^k|| = ||\vec{b} - A\vec{x}^{(k)}|| \le$ tol

Properties of $\rho(A)$:

1. $\rho(A) \le ||A||_k \forall k$
2. Spectral radius $\rho(A) = \max|\lambda(A)|$
3. $\lim_{n\to\infty} A_{n\times n}^n = 0$ iff $\rho(A) < 1$
4. Iterative mthd converges iff $\rho(G) < 1$

- Iter mthd converges iff $||G||_a < 1$ for some a

Jacobi Mthd: $A = D+L+U (D = M, L+U = -N)$

- each iter $O(2kn^2)$, good for large, sparse A
- $D :=$ Diag entries of A
- $D :=$ Strictly lower diagonal entries of A
- $D :=$ Strictly upper diagonal entries of A
- $\vec{x} = D^{-1}(-(L+U)\vec{x} + \vec{b})$
- $\vec{x}^{(k+1)} = D^{-1}(-(L+U)\vec{x}^{(k)} + \vec{b})$
- $G_j = D^{-1}(L+U)$

Gauss-Seidel Mthd: $L + D = M, -U = N$

- $\vec{x}^{(k+1)} = D^{-1}(\vec{b} - L\vec{x}^{(k+1)} - U\vec{x}^{(k)}$
- $\vec{x}^{(k+1)} = (L+D)^{-1}(\vec{b} - U\vec{x}^{(k)})$
- $G_{gs} = -(L+D)^{-1}U$
- $\vec{c}_{gs} = (L+D)^{-1}\vec{b}$

SOR Mthd: (equivalent to GS mthd for $\omega = 1$

$G_{sor} = (D+\omega L)^{-1}((1-\omega)D - \omega U)\vec{x}^{(k)} + \omega(D+\omega L)^{-1}\vec{b}$

- $G_{sor} = (D+\omega L)^{-1}((1-\omega)D - \omega U)$
- if SOR converges, then $0 < \omega < 2$

Convergence:

- Convergence rate $\rho(G) = \gamma = \lim_{k\to\infty} \frac{||\vec{x}^{(k+1)} - \vec{x}^{(*)}||}{||\vec{x}^{(k)} - \vec{x}^{(*)}||^q}$
- $\lim_{k\to\infty} \vec{x}^{(k)} = \vec{x}^{(*)}$
- $q = 1$, $0 < \gamma < 1$ linear convergence
- each iter gain $-\log_{10}(\gamma)$ correct digits
- smaller $\gamma \Rightarrow$ faster convergence
- A strict diag. dom. $\Rightarrow$ Jacobi & G-S convrg (1)
- A SPD $\Rightarrow$ SOR converges iff $0 < \omega < 2$

pf of (1)J (G-S) same idea - end of soln's lin eqns:
by contr. sps. G_J has $|\lambda| \ge 1 \Rightarrow \det(\lambda I - G_J)$=0

- Tridiagonal A: $\omega_{opt} = \frac{2}{1+\sqrt{1-\rho(G_j)^2}}$
- $\rho(G_{sor\omega opt}) = \frac{1-\sqrt{1-\rho(G_J)^2}}{1+\sqrt{1+\rho(G_J)^2}}$

Least Squares

Goal: find $\underset{\vec{x}\in\mathbb{R}^n}{\operatorname{argmin}} ||\vec{b} - A\vec{x}||_2$

Solution set: $\chi_{ls} = \{\vec{x} \in \mathbb{R}^n : \vec{x} = \underset{\vec{x}\in\mathbb{R}^n}{\operatorname{argmin}} ||\vec{b} - A\vec{x}||_2\}$

$\chi_{ls} = \vec{x}_{ls} + \text{null}(A^TA) = \vec{x}_{ls} + \text{null}(A)$

Theorems:

- $\vec{x} \in \chi_{ls} \Leftrightarrow A^TA\vec{x} = A^T\vec{b}$ (normal equations)
- $\exists$ unique solution if rank$(A) = n$

Pseudo-Inverse:

- $A^\dagger = V\Sigma^\dagger U^T$, $\sigma_i^\dagger = \frac{1}{\sigma_i}$ if $\sigma_i \ne 0$, else 0
- $\vec{x}_{ls} = A^\dagger\vec{b}$

QR factorization: $A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1R$

- $A \in \mathbb{R}^{m\times n}, m \ge n \Rightarrow A$ has QR factorization
- $Q_{m\times m} = \begin{bmatrix} Q_{1\mathbb{R}m\times n} & Q_{2\mathbb{R}m\times m-n} \end{bmatrix}$ orthogonal
- $R_{m\times n} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}$, $\hat{R}_{n\times n}$ upper triangular
- $\vec{x}_{ls} = R^{-1}Q_1^T\vec{b}$
- $||\vec{b} - \vec{x}_{ls}|| = ||Q_2^T\vec{b}||_2$

Householder Transformation:

idea: $H_n \dots H_2H_1A = R$ (upper triangular), $H_{m\times m}$

- $H\vec{x}$ is reflection of $\vec{x}$ in plane orthog to $\vec{v}$
- H is orthogonal
- $H = I - 2\vec{v}\vec{v}^T \frac{1}{\vec{v}^T\vec{v}}$, $||\vec{v}||_2 = 1$
- $H = H_1^T H_2^T \dots H_n^T$
- $Q = H_1H_2 \dots H_n$