Méthodes numériques matricielles avancées: analyse et expérimentation

Marc Bonnet, Luiz Faria

Propagation des Ondes: Etudes Mathématiques et Simulation (POEMS)
UMR 7231 CNRS-INRIA-ENSTA
Unité de Mathématiques Appliquées
ENSTA Paris

mbonnet@ensta.fr
https://perso.ensta-paris.fr/∼mbonnet/enseignement.html

ANN 203, ENSTA PARIS, 2023-24

Plan général, organisation

Partie 1: Généralités
- Séance 1: Motivations et exemples. Généralités sur le calcul matriciel numérique
- Séance 1b: TP numérique 0 (prise en main de Julia) **(LF)**

Partie 2: Méthodes directes
- Séance 1: Généralités, factorisation LU
- Séance 2a: Factorisations LU et LDL^T
- Séance 2b: Factorisation QR et problèmes de moindres carrés
- Séance 3b: TP numérique 1 **(LF)**

Partie 3: Méthodes itératives
- Séance 3a: Méthodes de type point fixe; gradient conjugué
- Séances 4a, 4b: Gradient conjugué, GMRES
- Séance 5b: TP numérique 2 **(LF)**

Partie 4: Problèmes aux valeurs et vecteurs propres
- Séance 5a: Généralités, Puissances itérées, puissances inverses
- Séance 6a: Itérations orthogonales et algorithme QR

Partie 5: Systèmes linéaires mal conditionnés
- Séance 6b: Compression et approximation de systèmes mal conditionnés
- Séance 7a: Recherche de solutions parcimonieuses.

Evaluation: Rendus TP1 et TP2 (20% chacun), examen écrit (60%, poly autorisé, séance 7b).

Ressources: `https://perso.ensta-paris.fr/~mbonnet/ens001.html` (poly, examens, supports)
`https://github.com/maltezfaria/ANN203` (TPs)

Plan général

Partie 1: Généralités

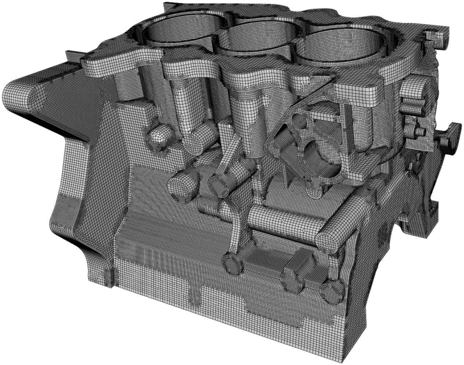
Partie 2: Méthodes directes

Partie 3: Méthodes itératives

Partie 4: Problèmes aux valeurs et vecteurs propres

Partie 5: Systèmes linéaires mal conditionnés

PDE discretization, finite elements



- Linear statics: $KU = F$
- Free vibrations: $(K - \omega^2 M)U = 0$
- Forced vibrations: $(K - \omega^2 M)U = F$

Solving non-linear equations

E.g. mechanical structure involving nonlinear material properties (or large strains, or contact, or. . .)

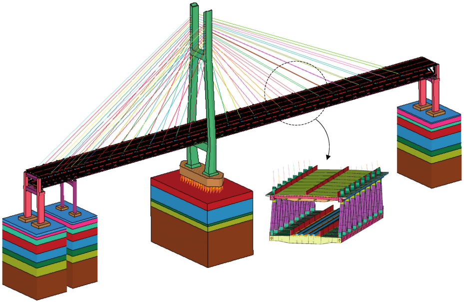
$$\mathcal{F}(U) = 0$$

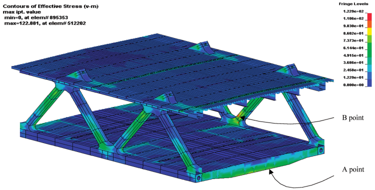
Newton-Raphson:

$$\mathcal{F}\big(U + \Delta U\big) = 0 \quad \Longrightarrow \quad \mathcal{F}(U) + \mathcal{F}'(U)\Delta U = 0$$

Iterations:

$$\boxed{K(U_k)\Delta U_k + \mathcal{F}(U_k) = 0, \quad U_{k+1} := U_k + \Delta U_k} \quad k = 1, 2, 3, \ldots$$





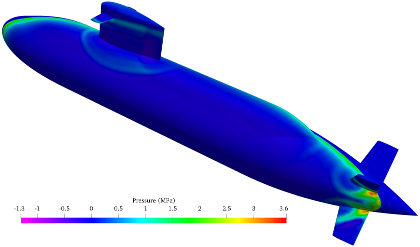
PDE discretization, boundary elements

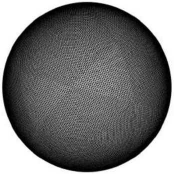
Boundary integral equation

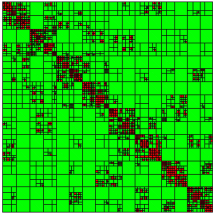
$$\int_S G(\boldsymbol{x},\boldsymbol{y})u(\boldsymbol{y})\,\mathrm{d}S(\boldsymbol{y}) = f(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x}\in S \qquad S\text{: surface}$$

Often used for wave propagation in large/unbounded media.

After boundary element discretization: $\boxed{GU = F}$







Inverse problems

Example: backward heat conduction problem (BHCP): quantify initial temperature field $\Theta(\boldsymbol{x}, 0)$ using later measurement

$$\underbrace{\Theta(\cdot, T)}_{\text{measurement at time } T} = \underbrace{\mathcal{A}([0, T])}_{\text{heat diffusion eq.}} \underbrace{\Theta(\cdot, 0)}_{\text{unknown}}$$

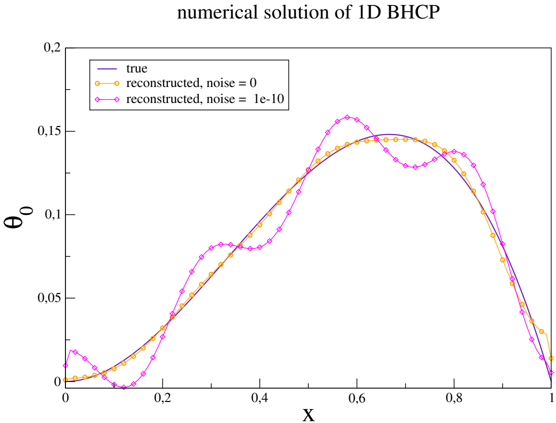
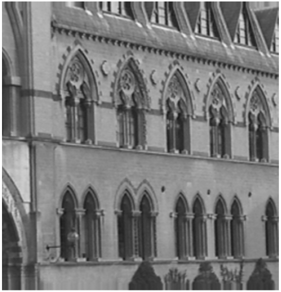


Image restoration

$$\boxed{\underbrace{\hat{f}(\boldsymbol{x})}_{\text{blurred}} = \int_Y \underbrace{k(\boldsymbol{x}-\boldsymbol{y})}_{\text{blurring}} \underbrace{f(\boldsymbol{y})}_{\text{image}} \,\mathrm{d}\boldsymbol{y}, \quad \boldsymbol{x} \in Y}$$

e.g. $k(\boldsymbol{z}) = Ce^{-|\boldsymbol{z}|^2/2\sigma^2}$ (atmospheric blur)

Pixel discretization: $\boxed{KF = \hat{F}}$ (K dense, ill-conditioned, numerically rank-deficient)



reference

blurred

restored

Correlation analysis

Basic problem:

- Population ($j = 1, \dots, m$ individuals)

  x_{ij} (explanatory variables), y_j

- Seek best affine model $y = a^\mathsf{T} x + b$ for an individual ($a \in \mathbb{R}^n,\ b \in \mathbb{R}$):

$$\boxed{\sum_{j=1}^{m} \left| y_j - a^\mathsf{T} x_j - b \right|^2 \ \to\ \min}$$

- Regression, least squares. . .

And much more: optimization, data analysis, machine learning. . .

Finite-precision computation

Floating-point representation of numbers:

$x = s\, m\, b^e$

s: sign (1 bit), x: mantissa (52 bits), b: basis (normally 2), e: exponent (11 bits)

Observations:

- Numbers subject to roundoff error (relative to orders of magnitude)
- Numbers not spaced evenly
- Estimating relative errors usually makes better sense.

IEEE 754 norm: floating point number representation ensuring

(a) for all $x \in \mathbb{R}$, exists ε, $|\varepsilon| < \varepsilon_{\text{mach}}$ $\quad \text{fl}(x) = x(1+\varepsilon)$

(b) for all $x, y \in \mathbb{F}$, $\quad x \circledast y = \text{fl}(x \star y)$ ($\star$ one of $+, -, \times, /, \surd$)

(c) for all $x, y \in \mathbb{F}$, exists ε, $|\varepsilon| < \varepsilon_{\text{mach}}$ $\quad \text{fl}(x \star y) = (x \star y)(1+\varepsilon)$

Vector spaces, matrices: a few reminders

Vector spaces A set E is a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ provided:

1. Addition of vectors is commutative: $xy = y + x$
2. Addition of vectors is associative: $x + (y + z) = (x + y) + z$
3. There exists a zero vector: $x + 0 = x$
4. Each vector has an opposite vector: $x + (-x) = 0$
5. Multiplication of scalars and with a vector are compatible: $(\alpha\beta)x = \alpha(\beta x)$
6. Scalar multiplication has a unit element: $1x = x$
7. Scalar multiplication is distributive w.r.t. scalar addition: $(\alpha + \beta)x = \alpha x + \beta x$
8. Scalar multiplication is distributive w.r.t. vector addition: $\alpha(x + y) = \alpha x + \alpha y$

Matrices: represent action of linear mappings $\mathcal{A} : \mathbb{K}^n \to \mathbb{K}^m$ (relative to bases chosen *a priori*):

$$y_i = \sum_{i=1}^{n} a_{ij} x_j, \qquad \text{or} \qquad \begin{Bmatrix} y_1 \\ \vdots \\ y_m \end{Bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{Bmatrix} x_1 \\ \vdots \\ x_n \end{Bmatrix} \quad \text{i.e.} \quad \boxed{y = Ax}$$

Terminology, notation

Terminology:

- $A \in \mathbb{K}^{n\times n}$ is called *square* (otherwise: *rectangular*)
- $A \in \mathbb{K}^{m\times n}$ with zeros in most entries is called *sparse* (otherwise: *dense* or *full*).

- *Transpose* $A^{\mathsf{T}} \in \mathbb{K}^{n\times m}$ of $A \in \mathbb{K}^{m\times n}$: $(A^{\mathsf{T}})_{ij} = A_{ji}$.
- *Conjugate transpose* $A^{\mathsf{H}} \in \mathbb{C}^{n\times m}$ of $A \in \mathbb{C}^{m\times n}$: $(A^{\mathsf{H}})_{ij} = \overline{A_{ji}}$, that is, $A^{\mathsf{H}} = \overline{A^{\mathsf{T}}}$.

- $A \in \mathbb{R}^{n\times n}$ verifying $A^{\mathsf{T}} = A$, $a_{ji} = a_{ij}$ is called *symmetric*.
- $A \in \mathbb{C}^{n\times n}$ verifying $A^{\mathsf{H}} = A$, $a_{ji} = \overline{a_{ij}}$ is called *Hermitian*.

- $A \in \mathbb{K}^{n\times n}$ Hermitian with $x^{\mathsf{H}}Ax > 0$ for all $x \neq 0$ is called *symmetric positive definite* (SPD).

Notation conventions (used throughout):

- Column vectors (e.g. $x \in \mathbb{K}^{n,1}$), (conjugate) transpose are row vectors. Consistent with

  $y = Ax$ (matrix-vector product), $(x, y) = x^{\mathsf{H}}y$ (scalar product).
- Vectors (matrices): lowercase (uppercase) letters, e.g. x (generic entry x_i),
  A (generic entry a_{ij}).

- MATLAB-like colon ":" to define submatrices by index ranges, e.g.

  $A_{k:\ell,p:q} := [a_{ij}]_{k\leq i\leq \ell,\, p\leq j\leq q}$ (rectangular submatrix of A),

  $A_{k:\ell,p} := [a_{ip}]_{k\leq i\leq \ell}$ (part of p-th column of A)

Vector norms

- Measuring "smallness/largeness" of vectors/matrices is essential (e.g. convergence of an algorithm: solution errors becoming "increasingly small").
- Magnitudes measured using (vector, matrix) norms. Defining requirements:

zero norm:	$\|x\| = 0$ if and only if $x = 0$,	
positive homogeneity:	$\|\lambda x\| = \|\lambda\|\|x\|$ for any $\lambda \in \mathbb{K}$	for all $x \in \mathbb{K}^n$
triangle inequality:	$\|x+y\| \le \|x\| + \|y\|$	

- Common vector norms ($p = 2$ is Euclidean 2-norm):

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \Big(\sum_{i=1}^{n} |x_i|^2 \Big)^{1/2}, \quad \|x\|_p := \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p}, \quad \|x\|_\infty := \max_{1 \le i \le n} |x$$

- All vector norms for finite-dimensional spaces are equivalent (fails as $n \to \infty$):

$$C_1 \|x\|_\alpha \le \|x\|_\beta \le C_2 \|x\|_\alpha, \quad \text{e.g.} \quad \begin{cases} \|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2 \\ \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty \\ \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty \end{cases} \quad \text{for all } x \in \mathbb{K}^n.$$

- Classical inequalities:

$$|x^{\mathsf{H}} y| \le \|x\|_2 \|y\|_2 \quad \text{(Cauchy-Schwarz)}, \qquad |x^{\mathsf{H}} y| \le \|x\|_p \|y\|_q \quad \Big(\text{Hölder}, \ \frac{1}{p} + \frac{1}{q} = 1\Big)$$

Matrix norms

- Matrix norms induced by vector norms:

$$\|A\|_p := \max_{x\in\mathbb{K}^n, x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p,$$

  Provides best upper bound on matrix-vector products: for any $A \in \mathbb{K}^{m\times n}$ and $x \in \mathbb{K}^m$,

$$\|Ax\|_p \leq \|A\|_p\|x\|_p \qquad \text{(with equality for at least one } x\text{)},$$

  (infinite-dimensional extension: operator norm, see e.g. MA 102)
- Another norm: Frobenius (not an induced norm)

$$\|A\|_{\mathrm{F}} := \Big(\sum_{i,j} |a_{ij}|^2 \Big)^{1/2} = \sqrt{\mathrm{Tr}(AA^{\mathrm{H}})}$$

Sub-multiplicativity: All induced matrix norms (and Frobenius norm) verify: $\|AB\| \leq \|A\|\,\|B\|$
Very important property for deriving (e.g. error) estimates.

Matrix norms are all equivalent. In particular:

$$\begin{aligned} \frac{1}{\sqrt{m}}\|A\|_1 &\leq \|A\|_2 \leq \sqrt{n}\|A\|_1 \\ \frac{1}{\sqrt{n}}\|A\|_\infty &\leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty \\ \frac{1}{\sqrt{\min(m,n)}}\|A\|_{\mathrm{F}} &\leq \|A\|_2 \leq \|A\|_{\mathrm{F}} \end{aligned} \qquad \text{for all } A\in\mathbb{K}^{m\times n}$$

Convention: Generic symbol $\|A\|$ always denotes an induced norm ($\|A\|_{\mathrm{F}}$ for Frobenius).

Accuracy and stability of computational solution methods

- Many scientific computing tasks boil down to:

  apply "function" $\mathcal{F}$ to data $x \in \mathcal{X}$, obtain $y = \mathcal{F}(x) \in \mathcal{Y}$

  Example: solve $y - f(y) = 0$ by fixed-point iterations from initial guess x:

  $\mathcal{F}(x) := \lim_{n\to\infty} f^n(x)$ (assuming f to be contracting!)

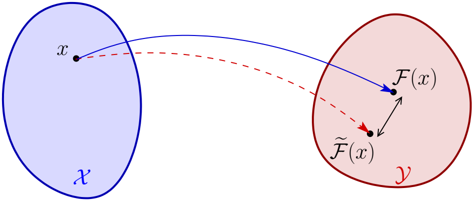
- In practice, data round-off, imperfect implementation of $\mathcal{F}$: $\boxed{\widetilde{y} := \widetilde{\mathcal{F}}(x)}$

  Example (no data round-off): $\widetilde{\mathcal{F}}(x) := \tilde{f}^{N+1}(x)$ with N such that $|\tilde{f}^{N+1}(x) - \tilde{f}^{N}(x)| < \varepsilon$

  How close to $y = \mathcal{F}(x)$ is the approximation $\widetilde{y} := \widetilde{\mathcal{F}}(x)$?

- **Relative solution accuracy:**

$$\boxed{e_{\text{rel}} := \frac{\|\widetilde{\mathcal{F}}(x) - \mathcal{F}(x)\|}{\|\mathcal{F}(x)\|}}$$



Best conceivable accuracy: $e_{\text{rel}} = O(\varepsilon_{\text{mach}})$ (achieved by individual floating-point operations).
Requirement $e_{\text{rel}} \approx \varepsilon_{\text{mach}}$ overly demanding (large-scale and/or ill-conditioned problems).

Accuracy and stability of computational solution methods

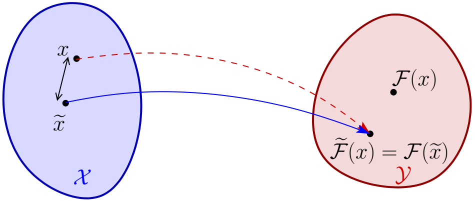
- **Stability:** a more-appropriate aim:

for each $x \in \mathcal{X}$: $\dfrac{\|\widetilde{\mathcal{F}}(x) - \mathcal{F}(\widetilde{x})\|}{\|\mathcal{F}(\widetilde{x})\|} = O(\varepsilon_{\text{mach}})$ for some $\widetilde{x}$ with $\dfrac{\|x - \widetilde{x}\|}{\|x\|} = O(\varepsilon_{\text{mach}})$

"A stable algorithm yields nearly the right answer if given a nearly correct data."

- Stronger requirement (replacing $O(\varepsilon_{\text{mach}})$ with zero): backward stability:

for each $x \in \mathcal{X}$: $\widetilde{\mathcal{F}}(x) = \mathcal{F}(\widetilde{x})$ for some $\widetilde{x}$ with $\dfrac{\|x - \widetilde{x}\|}{\|x\|} = O(\varepsilon_{\text{mach}})$

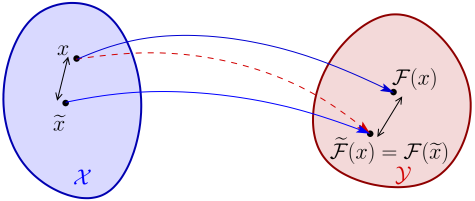


"A backward-stable algorithm yields the exact answer for some nearly correct data."

Conditioning, condition number

- Connect forward and backward errors by relative sensitivity:

$$\varrho = \varrho(\mathcal{F}; x, \widetilde{x}) := \frac{\|\widetilde{\mathcal{F}}(x) - \mathcal{F}(x)\|}{\|\mathcal{F}(x)\|} \left(\frac{\|\widetilde{x} - x\|}{\|x\|} \right)^{-1} = \frac{\|\mathcal{F}(\widetilde{x}) - \mathcal{F}(x)\|}{\|\mathcal{F}(x)\|} \, \frac{\|x\|}{\|\widetilde{x} - x\|}$$



- Condition number (of $\mathcal{F}$ at x): limiting value of ϱ for $\|\widetilde{x} - x\|$ small:

$$\boxed{\kappa(\mathcal{F}; x) := \lim_{\delta \to 0} \sup_{\|\widetilde{x} - x\| \le \delta} \varrho(\mathcal{F}; x, \widetilde{x})}$$

Explicit formula if $\mathcal{F}$ regular enough:

$$\kappa(\mathcal{F}, x) = \frac{\|\mathcal{F}'(x)\| \, \|x\|}{\|\mathcal{F}(x)\|},$$

- $\kappa(\mathcal{F}; x)$: dimensionless number;
- A solution process $\mathcal{F}$ is *well-conditioned* (*ill-conditioned*) if $\kappa = O(1)$ ($\kappa \gg 1$)

Condition number of linear systems

Solution of linear system $\boxed{Ay=b}$: sensitivity to data A, b ($A \in \mathbb{K}^{n\times n}$ invertible)

- Perturbation z of solution $y = A^{-1}b$ satisfies $(A+E)(y+z) = b+f$, i.e.

  $\boxed{(A+E)z = f - Ey}$.

- If $\|A^{-1}\|\|E\| < 1$ (perturbation of A small enough), $(A+E)^{-1} = A^{-1}(I + EA^{-1})^{-1}$ exists.

  $$\boxed{\|(A+E)^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|EA^{-1}\|} \leq \frac{\|A^{-1}\|}{1-\|E\|\|A^{-1}\|}}, \qquad \text{(using submultiplicativity)}$$

- Solution error estimate:

  $$\|z\| = \|(A+E)^{-1}(f-Ey)\| \implies \boxed{\|z\| \leq \frac{\|A^{-1}\|}{1-\|E\|\|A^{-1}\|}\big(\|f\| + \|E\|\|y\|\big)}.$$

  Formulate using relative errors:

  $$\frac{\|z\|}{\|y\|} \leq \frac{\|A^{-1}\|\|A\|}{1-\|E\|\|A^{-1}\|}\Big(\frac{\|f\|}{\|b\|}\frac{\|b\|}{\|A\|\|y\|} + \frac{\|E\|}{\|A\|}\Big) \leq \frac{\|A^{-1}\|\|A\|}{1-\|E\|\|A^{-1}\|}\Big(\frac{\|f\|}{\|b\|} + \frac{\|E\|}{\|A\|}\Big),$$

Relative sensitivity of solution w.r.t. data:

$$\frac{\|z\|/\|y\|}{\|f\|/\|b\| + \|E\|/\|A\|} \leq \frac{\kappa(A)}{1-\kappa(A)\|E\|/\|A\|} = \kappa(A) + O\big(\|E\|/\|A\|\big) \quad \text{with} \quad \kappa(A) := \|A^{-1}\|\|A\|.$$

Condition number of linear systems

Relative sensitivity of solution w.r.t. data:

$$\frac{\|z\|/\|y\|}{\|f\|/\|b\| + \|E\|/\|A\|} \leq \kappa(A) + O\big(\|E\|/\|A\|\big) \quad \text{with} \quad \kappa(A) := \|A^{-1}\|\|A\|.$$

Condition number $\kappa(A) = \|A^{-1}\|\|A\|$ of A: upper bound of condition number for solving $Ay = b$.

Properties of $\kappa(A)$:

- Always $\boxed{\kappa(A) \geq 1}$ ($\|A^{-1}\|\|A\| \geq \|A^{-1}A\| = \|I\| = 1$ for any induced norm).
- $\kappa(A)$ depends on choice of (matrix) norm.
- If A *normal* ($AA^{\text{H}} = A^{\text{H}}A$), we have $A = Q\Lambda Q^{\text{H}}$ for some Q unitary. Then:

  $$\|A\|_2 = |\lambda_{\max}|, \quad \|A^{-1}\|_2 = 1/|\lambda_{\min}|, \qquad \text{and hence} \quad \boxed{\kappa_2(A) = |\lambda_{\max}|/|\lambda_{\min}|}.$$

- $\|Q\|_2 = 1$ and $\|Q^{-1}\|_2 = 1$ if Q orthogonal or unitary. Consequently, $\boxed{\kappa_2(Q) = 1}$.
- For arbitrary $A \in \mathbb{K}^{m \times n}$, $\kappa_2(A)$ given in terms of either singular values or pseudo-inverse of A (see Part 3).

A simple numerical example

- Example (exact matrix inverse):

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \implies A^{-1} = \begin{bmatrix} 25 & 41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}$$

  Note: $AA^{\mathsf{T}} = A^{\mathsf{T}}A$ (i.e. A is normal)

- Effect of perturbations of A or b on solution of x of $Ax = b$:

$$b = [32 \;\; 23 \;\; 33 \;\; 31]^{\mathsf{T}} \implies x = [1 \;\; 1 \;\; 1 \;\; 1]^{\mathsf{T}}$$

$$\delta b = [0.1 \;\; -0.1 \;\; 0.1 \;\; -0.1]^{\mathsf{T}} \implies x = [9.2 \;\; -12.6 \;\; 4.5 \;\; -1.1]^{\mathsf{T}}$$

$$\delta A_{23} = 0.1 \implies x \approx [-4.86 \;\; -10.7 \;\; -1.43 \;\; -2.43]^{\mathsf{T}}$$

- Eigenvalues of A:

$$\Lambda \approx \text{Diag}[\, 30.29 \;\; 3.858 \;\; 0.8431 \;\; 0.01015 \,], \qquad \kappa_2(A) \approx 3\,10^3$$

  A is a rather ill-conditioned 4×4 matrix.