MA 7007: Numerical Solution of Differential Equations I
Elliptic Partial Differential Equations

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

E-mail: syyang@math.ncu.edu.tw
Website: http://www.math.ncu.eud.tw/∼syyang/

Introduction

In two space dimensions a constant-coefficient elliptic equation has the form

$$a_1 u_{xx}(x,y) + a_2 u_{xy}(x,y) + a_3 u_{yy}(x,y) + a_4 u_x(x,y) + a_5 u_y(x,y) + a_6 u(x,y) = f(x,y),$$

for all $(x,y) \in \Omega$, where $\Omega \subseteq \mathbb{R}^2$ is typically an open bounded domain and the coefficients a_1, a_2, a_3 satisfy

$$a_2^2 - 4a_1 a_3 < 0.$$

This equation must be complemented with some boundary condition on the boundary $\partial\Omega$ such as the Dirichlet boundary condition

$$u(x,y) = g(x,y) \qquad \text{for all } (x,y) \in \partial\Omega.$$

Steady-state heat conduction

Heat conduction problem in two space dimensions:

$$\begin{cases} u_t = (\kappa u_x)_x + (\kappa u_y)_y + \psi, \quad t \in (0,T),\ (x,y) \in \Omega, \\ \text{“Initial and boundary conditions.”} \end{cases}$$

where $\kappa(x,y) > 0$ is the diffusivity and $\psi(t,x,y)$ is a source function. If the boundary conditions and the source term are independent of time t, then we expect a steady state to exist,

$$(\kappa u_x)_x + (\kappa u_y)_y = -\psi := f \quad \text{in } \Omega + \text{“boundary conditions.”}$$

Let $\kappa(x,y) \equiv 1$ for all $(x,y) \in \Omega$.

1. Poisson equation: $u_{xx} + u_{yy} = f$.
2. Laplace equation: $u_{xx} + u_{yy} = 0$.
   Solutions to the Laplace equation are called harmonic functions.

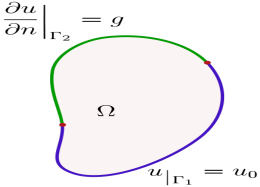
Notation and boundary conditions

Notation: $\nabla := [\partial_x, \partial_y]^\top$.

1. gradient operator: $\nabla u = [u_x, u_y]^\top$.
2. divergence operator: $\nabla \cdot [u, v]^\top = u_x + v_y$.
3. Laplacian operator: $\nabla^2 u := \nabla \cdot \nabla u = u_{xx} + u_{yy} := \Delta u$.

Boundary conditions:

1. Dirichlet BC: $u(x,y) = g(x,y), \quad \forall (x,y) \in \partial\Omega$
2. Neumann BC: $\dfrac{\partial u}{\partial \boldsymbol{n}}(x,y) \left(:= \nabla u(x,y) \cdot \boldsymbol{n}(x,y) \right) = g(x,y), \quad \forall (x,y) \in \partial\Omega$
3. Robin BC: $au(x,y) + b\dfrac{\partial u}{\partial \boldsymbol{n}}(x,y) = g(x,y), \quad \forall (x,y) \in \partial\Omega$
4. Mixed BC:



Centered difference scheme

For example, we consider the Poisson equation with the Dirichlet BC:

$$\begin{aligned} \nabla^2 u &= f \quad \text{in } \Omega := (0,1)\times(0,1), \\ u &= g \quad \text{on } \partial\Omega. \end{aligned}$$

We will use the uniform Cartesian grid: (x_i, y_j), where $x_i = i\Delta x$ and $y_j = j\Delta y$, Δx and Δy are the grid sizes in $x-$ and $y-$ directions.
Let u_{ij} represent an approximation to $u(x_i, y_j)$ and $f_{ij} := f(x_i, y_j)$.

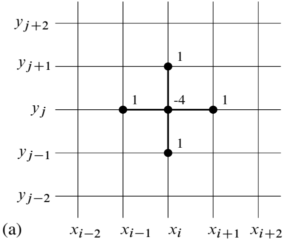
$$\frac{1}{(\Delta x)^2}(u_{i-1,j} - 2u_{ij} + u_{i+1,j}) + \frac{1}{(\Delta y)^2}(u_{i,j-1} - 2u_{ij} + u_{i,j+1}) = f_{ij}.$$

For simplicity, we set $\Delta x = \Delta y = h$. Then we have

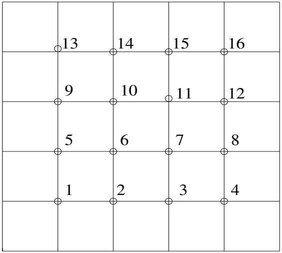
$$\nabla_5^2 u_{ij} := \frac{1}{h^2}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{ij}) = f_{ij}.$$

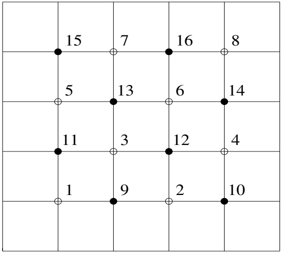
Let $0 = x_0 < x_1 < \cdots < x_m < x_{m+1} = 1$ and $0 = y_0 < y_1 < \cdots < y_m < y_{m+1} = 1$ be the partitions. Then $h = 1/(m+1)$. From the above equations, we have an $m^2 \times m^2$ linear system $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{F}$ of m^2 unknowns u_{ij} for $1 \le i \le m, 1 \le j \le m$, where $\boldsymbol{A}$ is sparse (Roughly speaking, at least $\frac{2}{3} \uparrow$ zeros).

Computational grid: 5-point stencil and 9-point stencil



Ordering the unknowns and equations

(a)


(b)


(a) The rowwise ordering. (b) The red-black ordering.

The rowwise ordering

Let

$$\boldsymbol{u} = \begin{bmatrix} u^{[1]} \\ u^{[2]} \\ \vdots \\ u^{[m]} \end{bmatrix}, \quad u^{[j]} = \begin{bmatrix} u_{1j} \\ u_{2j} \\ \vdots \\ u_{mj} \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} f^{[1]} \\ f^{[2]} \\ \vdots \\ f^{[m]} \end{bmatrix} + BV, \quad f^{[j]} = \begin{bmatrix} f_{1j} \\ f_{2j} \\ \vdots \\ f_{mj} \end{bmatrix}.$$

Then

$$\boldsymbol{A} = \frac{1}{h^2} \begin{bmatrix} T & I & & & \\ I & T & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & T & I \\ & & & I & T \end{bmatrix}, \quad T = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix}.$$

Accuracy and stability

The local truncation error τ_{ij} at the grid point (i, j) is defined by

$$\tau_{ij} := \frac{1}{h^2}\Big(u(x_{i-1}, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1}) - 4u(x_i, y_j)\Big) - f(x_i, y_j).$$

By the Taylor expansion, we have

$$\tau_{ij} = \frac{1}{12}h^2\Big(u_{xxxx}(x_i, y_j) + u_{yyyy}(x_i, y_j)\Big) + O(h^4)$$

and

$$\boldsymbol{A}\boldsymbol{u}^{exact} = \boldsymbol{F} + \boldsymbol{\tau},$$

where $\boldsymbol{A}$ is the discretization matrix corresponding to the rowwise ordering. Letting the global error $E_{ij} := u_{ij} - u(x_i, y_j)$ and noting that $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{F}$, we obtain

$$\boldsymbol{A}\boldsymbol{E} = -\boldsymbol{\tau} \Longrightarrow \boldsymbol{E} = \boldsymbol{A}^{-1}(-\boldsymbol{\tau}).$$

The method will be globally second order accurate in some grid function norm provided that $\|\boldsymbol{A}^{-1}\|$ is uniformly bounded as $h \to 0^+$.

Accuracy and stability (continued)

We consider the 2-norm for the discretization matrix $\boldsymbol{A}$. By further computations, one can show that for $p, q = 1, 2, \cdots, m$, the eigenvector $\boldsymbol{u}^{p,q}$ has the m^2 elements,

$$u_{ij}^{p,q} = \sin(p\pi ih)\sin(q\pi ih)$$

and the corresponding eigenvalue is

$$\lambda_{p,q} = \frac{2}{h^2}\Big((\cos(p\pi h) - 1) + (\cos(q\pi h) - 1)\Big) < 0. \qquad (\text{Note that } h = \frac{1}{m+1})$$

Thus, the one closest to origin is $\lambda_{1,1} = -2\pi^2 + O(h^2)$. (Hint: By Taylor expansion: $\cos(x) = 1 - x^2/2! + x^4/4! - \cdots$) The spectral radius of $\boldsymbol{A}^{-1}$ is

$$\rho(\boldsymbol{A}^{-1}) = \frac{1}{|\lambda_{1,1}|} \approx \frac{1}{2\pi^2} \quad \text{as } h \to 0^+,$$

and then as $h \to 0^+$,

$$\|\boldsymbol{A}^{-1}\|_2 = \sqrt{\rho(\boldsymbol{A}^{-\top}\boldsymbol{A}^{-1})} = \sqrt{\rho((\boldsymbol{A}^{-1})^2)} = \sqrt{(\rho(\boldsymbol{A}^{-1}))^2} = \rho(\boldsymbol{A}^{-1}) \approx \frac{1}{2\pi^2},$$

which is uniformly bounded.

Accuracy and stability (continued)

From the centered difference scheme with uniform mesh size, $\nabla_5^2 u_{ij} = f_{ij}$, we obtain an $m^2 \times m^2$ linear system of m^2 unknowns u_{ij} for $1 \le i \le m, 1 \le j \le m$,

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{F},$$

(or more precisely, $\boldsymbol{A}^h \boldsymbol{u}^h = \boldsymbol{F}^h$). Now suppose source term $\boldsymbol{F}$ is perturbed by a small vector $\boldsymbol{p}$ (say, $\|\boldsymbol{p}\|_2 < \delta$ for a small $\delta > 0$) and the corresponding solution is denoted by $\widetilde{\boldsymbol{u}}$. Then we have

$$\boldsymbol{A}\widetilde{\boldsymbol{u}} = \boldsymbol{F} + \boldsymbol{p},$$

and

$$\boldsymbol{A}(\widetilde{\boldsymbol{u}} - \boldsymbol{u}) = \boldsymbol{p},$$

which implies

$$\widetilde{\boldsymbol{u}} - \boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{p} \Longrightarrow \|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2 \le \|\boldsymbol{A}^{-1}\|_2 \|\boldsymbol{p}\|_2 < \|\boldsymbol{A}^{-1}\|_2 \delta,$$

where $\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2$ and $\|\boldsymbol{p}\|_2$ are grid function norms. Since $\|\boldsymbol{A}^{-1}\|_2$ is uniformly bounded, we have $\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2 \le C\delta$. Hence, the centered difference scheme for the Poisson problem is *stable*.

Condition number

The 2-norm condition number of the discretization matrix $\boldsymbol{A}$ is defined by

$$\kappa(\boldsymbol{A}) := \|\boldsymbol{A}\|_2 \|\boldsymbol{A}^{-1}\|_2.$$

Notice that as $h \to 0^+$,

$$\|\boldsymbol{A}\|_2 = \rho(\boldsymbol{A}) = \max_{1\le p,q\le m} |\lambda_{p,q}| = |\lambda_{m,m}| = \frac{4}{h^2}\left|\cos(\frac{m}{m+1}\pi) - 1\right| \approx \frac{4}{h^2}|-2| = \frac{8}{h^2}.$$

Therefore

$$\kappa(\boldsymbol{A}) \approx \frac{4}{\pi^2 h^2} = O(h^{-2}) \quad \text{as } h \to 0^+.$$

The discretization matrix $\boldsymbol{A}$ is very ill-conditioned as we refine the grid.

The 9-point Laplacian

By the Taylor expansion, we have

$$\begin{aligned}
&\nabla_5^2 u(x_i,y_j) = \nabla^2 u(x_i,y_j) + \frac{1}{12}h^2\frac{\partial^4 u}{\partial x^4}(x_i,y_j) + \frac{1}{12}h^2\frac{\partial^4 u}{\partial y^4}(x_i,y_j) + O(h^4)\\
&\Longrightarrow \nabla_5^2 u(x_i,y_j) + \frac{2}{12}h^2\frac{\partial^4 u}{\partial y^2\partial x^2}(x_i,y_j)\\
&\qquad = \nabla^2 u(x_i,y_j) + \frac{1}{12}h^2\left\{\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2\partial y^2} + \frac{\partial^4 u}{\partial y^4}\right\}(x_i,y_j) + O(h^4)\\
&\qquad = \nabla^2 u(x_i,y_j) + \frac{1}{12}h^2\nabla^2 f(x_i,y_j) + O(h^4)\\
&\Longrightarrow \nabla_5^2 u(x_i,y_j) + \frac{2}{12}h^2\frac{\partial^4 u}{\partial y^2\partial x^2}(x_i,y_j) - \frac{1}{12}h^2\nabla^2 f(x_i,y_j) = \nabla^2 u(x_i,y_j) + O(h^4).
\end{aligned}$$

The 9-point Laplacian (continued)

$$\nabla_5^2 u(x_i,y_j) + \frac{2}{12}h^2\frac{\partial^4 u}{\partial y^2 \partial x^2}(x_i,y_j) - \frac{1}{12}h^2\nabla^2 f(x_i,y_j) = \nabla^2 u(x_i,y_j) + O(h^4)$$

$$\Longrightarrow \nabla_5^2 u(x_i,y_j) + \frac{h^2}{6h^4}\Big\{u(x_{i-1},y_{j-1}) - 2u(x_{i-1},y_j) + u(x_{i-1},y_{j+1})$$

$$-2u(x_i,y_{j-1}) + 4u(x_i,y_j) - 2u(x_i,y_{j+1})$$

$$+u(x_{i+1},y_{j-1}) - 2u(x_{i+1},y_j) + u(x_{i+1},y_{j+1})\Big\} + O(h^4)$$

$$-\frac{1}{12}h^2\nabla^2 f(x_i,y_j) = \nabla^2 u(x_i,y_j) + O(h^4).$$

$$\therefore \qquad \nabla_9^2 u_{ij} := \frac{1}{6h^2}\Big\{4u_{i-1,j} + 4u_{i+1,j} + 4u_{i,j-1} + 4u_{i,j+1} + u_{i-1,j-1} + u_{i-1,j+1}$$

$$+u_{i+1,j-1} + u_{i+1,j+1} - 20u_{ij}\Big\} = f_{ij} + \frac{1}{12}h^2\nabla^2 f(x_i,y_j)$$

is a finite difference scheme for the Poisson problem with local truncation error $O(h^4)$. The term $\frac{1}{12}h^2\nabla^2 f(x_i,y_j)$ can be exactly computed or approximated by $\frac{1}{12}h^2\nabla_5^2 f(x_i,y_j)$.

Estimates from the true solution

Suppose we know the true solution. Let $E(h)$ denote the error function of grid size h, i.e., $E(h) = \|U(h) - \widehat{U}(h)\|$, where $U(h)$ is the numerical solution vector and $\widehat{U}(h)$ is the true solution evaluated on the same grid.

If the method is p-th order accurate, i.e., $E(h) = Ch^p + O(h^{p+1})$ as $h \to 0$, then for $0 < h_2 < h_1$ sufficiently small, we expect $E(h_1) \approx Ch_1^p$ and $E(h_2) \approx Ch_2^p$. The order of convergence can be estimated using

$$p \approx \frac{\log(E(h_1)/E(h_2))}{\log(h_1/h_2)},$$

this is because

$$\log \frac{E(h_1)}{E(h_2)} \approx \log \frac{Ch_1^p}{Ch_2^p} = \log\left(\frac{h_1}{h_2}\right)^p = p \log \frac{h_1}{h_2}.$$

Estimates from a fine-grid solution

Now suppose we don't know the exact solution but that we can afford to run the problem on a very fine grid, say $\bar{h}$, and use the numerical solution $U(\bar{h})$ as a reference solution.

Let $U(h)$ be the numerical solution on a coarser grid h, and $\overline{U}(h)$ be the restriction of $U(\bar{h})$ to the h-grid. Define the approximate error and the true error as

$$\overline{E}(h) = \|U(h) - \overline{U}(h)\| \quad \text{and} \quad E(h) = \|U(h) - \widehat{U}(h)\|,$$

respectively. Then consider

$$U(h) - \overline{U}(h) = (U(h) - \widehat{U}(h)) + (\widehat{U}(h) - \overline{U}(h)).$$

If the method is supposed to be p-th order accurate and $\bar{h}^p \ll h^p$, then we will have $U(h) - \overline{U}(h) \approx U(h) - \widehat{U}(h)$ since the second term $\widehat{U}(h) - \overline{U}(h)$ should be negligible compared to the first term $U(h) - \widehat{U}(h)$. In this case, the approximate error $\overline{E}(h)$ can be used as a good estimate of the true error $E(h)$.

L^p-norm and discrete L^p-norm for grid functions, $1 \le p \le \infty$

1. **L^p-norm:** Let $U(x)$ be an approximate solution of $u(x)$ on $\overline{\Omega} = [a, b]$ and let $e(x) := U(x) - u(x)$, where $U(x)$ and $u(x)$ are smooth enough. Then

$$\|e\|_{L^\infty(\Omega)} := \max_{a \le x \le b} |e(x)| \quad \text{and} \quad \|e\|_{L^p(\Omega)} := \left(\int_a^b |e(x)|^p dx \right)^{1/p}, \quad p \ge 1.$$

2. **Discrete L^p-norm of grid function e:** Let $U_i \approx u(x_i)$, $1 \le i \le N$. Let $e_i = U_i - u(x_i)$ and $e = (e_1, \cdots, e_N)^\top$. Then

$$\|e\|_\infty := \max_{1 \le i \le N} |e_i| \quad \text{and} \quad \|e\|_p := \left(h \sum_{i=1}^{N} |e_i|^p \right)^{1/p}, \quad p \ge 1.$$

3. **2-D discrete L^p-norm of grid function e:**

$$\|e\|_\infty := \max_{1 \le i,j \le N} |e_{ij}| \quad \text{and} \quad \|e\|_p := \left(h^2 \sum_i \sum_j |e_{ij}|^p \right)^{1/p}, \quad p \ge 1.$$

Review: Vector norm

Let $\boldsymbol{V}$ be a vector space over $\mathbb{R}$, e.g., $\boldsymbol{V} = \mathbb{R}^n$. A norm is a real-valued function $\|\cdot\| : \boldsymbol{V} \to \mathbb{R}$ that satisfies

1. $\|\boldsymbol{x}\| \geq 0, \forall\, \boldsymbol{x} \in \boldsymbol{V}$, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$;
2. $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|, \forall\, \boldsymbol{x} \in \boldsymbol{V}$ and $\lambda \in \mathbb{R}$;
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|, \forall\, \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{V}$ (triangle inequality).

Note: $\|\boldsymbol{x}\|$ is called the norm of $\boldsymbol{x}$, the length or magnitude of $\boldsymbol{x}$.

Some vector norms on $\mathbb{R}^n$

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$:

1. The 2-norm (Euclidean norm, or ℓ^2 norm):

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

2. The infinity norm (ℓ^∞-norm):

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le n} |x_i|.$$

3. The 1-norm (ℓ^1-norm):

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

The difference between the above norms

1. Take three vectors $\boldsymbol{x} = (4, 4, -4, 4)^\top$, $\boldsymbol{v} = (0, 5, 5, 5)^\top$, $\boldsymbol{w} = (6, 0, 0, 0)^\top$:

	$\|\cdot\|_1$	$\|\cdot\|_2$	$\|\cdot\|_\infty$
$\boldsymbol{x}$	16	8	4
$\boldsymbol{v}$	15	8.66	5
$\boldsymbol{w}$	6	6	6

2. What is the unit ball $\{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\| \leq 1\}$ for the three norms above?
    - 2-norm: a circle;
    - ∞-norm: a square;
    - 1-norm: a diamond.

Matrix norm

Let $\boldsymbol{A}$ be an $n \times n$ real matrix. If $\|\cdot\|$ is any norm on $\mathbb{R}^n$, then

$$\|\boldsymbol{A}\| := \sup\{\|\boldsymbol{A}\boldsymbol{x}\| : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\} \Big(\Longleftrightarrow \|\boldsymbol{A}\| := \sup\{\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} : \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}\} \Big)$$

defines a norm on the vector space of all $n \times n$ real matrices.
(This is called the matrix norm associated with the given vector norm)

Proof:

- $\because \|\boldsymbol{A}\boldsymbol{x}\| \geq 0 \ \forall \ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1. \therefore \|\boldsymbol{A}\| \geq 0.$
  Moreover, one can check that $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \boldsymbol{0}$.
- $\|\lambda \boldsymbol{A}\| = \sup\{\|\lambda \boldsymbol{A}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} = \sup\{|\lambda|\|\boldsymbol{A}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\}$
  $= |\lambda| \sup\{\|\boldsymbol{A}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} = |\lambda|\|\boldsymbol{A}\|.$
- $\|\boldsymbol{A} + \boldsymbol{B}\| = \sup\{\|(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} \leq \sup\{\|\boldsymbol{A}\boldsymbol{x}\| + \|\boldsymbol{B}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\}$
  $\leq \sup\{\|\boldsymbol{A}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} + \sup\{\|\boldsymbol{B}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|.$

Some additional properties

1. $\|\boldsymbol{Ax}\| \le \|\boldsymbol{A}\| \|\boldsymbol{x}\|, \forall\, \boldsymbol{x} \in \mathbb{R}^n$.

   Proof:

   Let $\boldsymbol{x} \neq \boldsymbol{0}$. Then $\boldsymbol{v} = \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ is of norm 1. $\quad \therefore \|\boldsymbol{A}\| \ge \|\boldsymbol{Av}\| = \dfrac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|}$.

2. $\|\boldsymbol{I}\| = 1$.
3. $\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\| \|\boldsymbol{B}\|$.

   Proof:

   $\|\boldsymbol{AB}\| := \sup\{\|(\boldsymbol{AB})\boldsymbol{x}\| : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}$
   $\le \sup\{\|\boldsymbol{A}\| \|\boldsymbol{Bx}\| : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}$
   $\le \sup\{\|\boldsymbol{A}\| \|\boldsymbol{B}\| \|\boldsymbol{x}\| : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\} = \|\boldsymbol{A}\| \|\boldsymbol{B}\|$.

Some matrix norms

Let $A_{n\times n} = (a_{ij})$ be an $n \times n$ real matrix. Then

1. The ∞-matrix norm:
$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

2. The 1-matrix norm:
$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|.$$

3. The 2-matrix norm:
$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2.$$

The 2-matrix norm

1. $\|A\|_2$ is not easy to compute.
2. Since $A^\top A$ is symmetric, $A^\top A$ has n real eigenvalues, $\lambda_1, \lambda_2, \cdots, \lambda_n \in \mathbb{R}$. Moreover, one can prove that they are all nonnegative. Then

$$\rho(A^\top A) := \max_{1 \le i \le n} \{\lambda_i\} \ge 0.$$

   is called the spectral radius of $A^\top A$.
3. Then the 2-matrix norm of A is given by

$$\|A\|_2 = \sqrt{\rho(A^\top A)}.$$

4. The 2-matrix norm is also called the *spectral norm*.

Some error analysis

1. Suppose that we want to solve the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, but $\boldsymbol{b}$ is somehow perturbed to $\widetilde{\boldsymbol{b}}$ (this may happen when we convert a real $\boldsymbol{b}$ to a floating-point $\boldsymbol{b}$).
2. Then actual solution would satisfy a slightly different linear system

$$\boldsymbol{A}\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{b}}.$$

3. *Question*: Is $\widetilde{\boldsymbol{x}}$ very different from the desired solution $\boldsymbol{x}$ of the original system?
4. Of course, the answer should depend on *how good the matrix $\boldsymbol{A}$ is*.
5. Let $\|\cdot\|$ be a vector norm, we consider two types of errors:
   - absolute error: $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$?
   - relative error: $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\| / \|\boldsymbol{x}\|$?

The absolute error

For the absolute error, we have

$$\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\| = \|\boldsymbol{A}^{-1}\boldsymbol{b} - \boldsymbol{A}^{-1}\widetilde{\boldsymbol{b}}\| = \|\boldsymbol{A}^{-1}(\boldsymbol{b} - \widetilde{\boldsymbol{b}})\| \leq \|\boldsymbol{A}^{-1}\|\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\|.$$

Therefore, the absolute error of $\boldsymbol{x}$ depends on two factors: the absolute error of $\boldsymbol{b}$ and the matrix norm of $\boldsymbol{A}^{-1}$.

The relative error

For the relative error, we have

$$\begin{aligned}
\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\| &= \|A^{-1}\boldsymbol{b} - A^{-1}\widetilde{\boldsymbol{b}}\| = \|A^{-1}(\boldsymbol{b} - \widetilde{\boldsymbol{b}})\| \\
&\leq \|A^{-1}\|\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\| = \|A^{-1}\|\|A\boldsymbol{x}\|\frac{\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|} \\
&\leq \|A^{-1}\|\|A\|\|\boldsymbol{x}\|\frac{\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.
\end{aligned}$$

That is

$$\frac{\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}\|\|A\| \frac{\|\boldsymbol{b} - \widetilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.$$

Therefore, the relative error of $\boldsymbol{x}$ depends on two factors: the relative error of $\boldsymbol{b}$ and $\|A\|\|A^{-1}\|$.

Condition number

1. Therefore, we define a condition number of the matrix $\boldsymbol{A}$ as

$$\kappa(\boldsymbol{A}) := \|\boldsymbol{A}\| \|\boldsymbol{A}^{-1}\|.$$

   $\kappa(\boldsymbol{A})$ measures how good the matrix $\boldsymbol{A}$ is.

2. Example: Let $\varepsilon > 0$ and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix} \Longrightarrow \boldsymbol{A}^{-1} = \varepsilon^{-2} \begin{bmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{bmatrix}.$$

   Then $\|\boldsymbol{A}\|_\infty = 2+\varepsilon$, $\|\boldsymbol{A}^{-1}\|_\infty = \varepsilon^{-2}(2+\varepsilon)$, and $\kappa(\boldsymbol{A}) = \left(\frac{2+\varepsilon}{\varepsilon}\right)^2 \geq \frac{4}{\varepsilon^2}$.

Condition number (continued)

1. For example, if $\varepsilon = 0.01$, then $\kappa(\boldsymbol{A}) \geq 40000$.
2. What does this mean?
   It means that the relative error in $\boldsymbol{x}$ can be 40000 times greater than the relative error in $\boldsymbol{b}$.
3. If $\kappa(\boldsymbol{A})$ is large, we say that $\boldsymbol{A}$ is *ill-conditioned*, otherwise $\boldsymbol{A}$ is *well-conditioned*.
4. In the ill-conditioned case, the solution is very sensitive to the small changes in the right-hand vector $\boldsymbol{b}$ (higher precision in $\boldsymbol{b}$ may be needed).

Another way to measure the error

Consider the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}(\neq \boldsymbol{0})$. Let $\widetilde{\boldsymbol{x}}$ be a computed solution (an approximation to $\boldsymbol{x}$).

1. Residual vector:
$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\widetilde{\boldsymbol{x}}.$$
2. Error vector:
$$\boldsymbol{e} = \boldsymbol{x} - \widetilde{\boldsymbol{x}}.$$
3. They satisfy
$$\boldsymbol{A}\boldsymbol{e} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\widetilde{\boldsymbol{x}} = \boldsymbol{b} - \boldsymbol{A}\widetilde{\boldsymbol{x}} = \boldsymbol{r}.$$
4. Moreover, we have
$$\frac{1}{\kappa(\boldsymbol{A})}\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} \leq \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \leq \kappa(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}.$$

   (Theorem on bounds involving condition number)

Proof of the Theorem

$\because \boldsymbol{A}\boldsymbol{e} = \boldsymbol{r}$.

$\therefore \boldsymbol{e} = \boldsymbol{A}^{-1}\boldsymbol{r}$.

$\therefore \|\boldsymbol{e}\|\|\boldsymbol{b}\| = \|\boldsymbol{A}^{-1}\boldsymbol{r}\|\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}^{-1}\|\|\boldsymbol{r}\|\|\boldsymbol{A}\|\|\boldsymbol{x}\|$.

$\therefore \dfrac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \leq \kappa(\boldsymbol{A}) \dfrac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}$.

On the other hand, we have $\|\boldsymbol{r}\|\|\boldsymbol{x}\| = \|\boldsymbol{A}\boldsymbol{e}\|\|\boldsymbol{A}^{-1}\boldsymbol{b}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{e}\|\|\boldsymbol{A}^{-1}\|\|\boldsymbol{b}\|$.

$\therefore \dfrac{1}{\kappa(\boldsymbol{A})} \dfrac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} \leq \dfrac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|}$.