

Contents

1	Euler’s Method and Beyond	3
1.1	Ordinary differential equations and Lipschitz condition	3
1.2	Euler’s method	4
1.3	The trapezoidal rule	5
1.4	The theta method	6
2	Multistep Method	6
3	8 Finite Differences Schemes	6
3.1	8.1 Finite differences	6
3.2	8.2 The five-point formula for $\nabla^2 u = f$	9

Basics

1. **(Taylor Expansion)** Given $f \in C^\infty(\mathbb{R})$, the Taylor expansion of f at a is given by

$$\begin{aligned} T(a) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n \\ &= f(a) + \frac{f'(a)}{1!} (x-a) + \frac{f''(a)}{2!} (x-a)^2 + \dots \end{aligned}$$

2. **(Taylor's Theorem)** Let $k \geq 1$ and function $f : \mathbb{R} \rightarrow \mathbb{R}$ be k times differentiable at a point $a \in \mathbb{R}$, then exists $h_k : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$f(x) = \sum_{n=0}^k \frac{f^{(n)}(a)}{n!} (x-a)^n + h_k(x)(x-a)^k$$

and $\lim_{x \rightarrow a} h_k(x) = 0$, i.e. the reminder term $R_k(x) = f(x) - P_k(x)$ is asymptotically trivial. If f is $k+1$ times differentiable on the open interval and $f^{(k)}$ continuous on the closed interval $[a, x]$, then the Lagrange remind is given by

$$R_k(x) = \frac{f^{(k+1)}(\zeta)}{(k+1)!} (x-a)^{k+1}$$

for soem $\zeta \in [a, x]$ by the mean value theorem

3. **(Taylor's Theorem for multivariable function)** If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ are k times differentiable function at point $\mathbf{a} \in \mathbb{R}^n$ then exists $h_\alpha : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$f(\mathbf{x}) = \sum_{|\alpha| \leq k} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha + \sum_{|\alpha|=k} h_\alpha(\mathbf{x})(\mathbf{x} - \mathbf{a})^\alpha \quad \lim_{\mathbf{x} \rightarrow \mathbf{a}} h_\alpha(\mathbf{x}) = 0$$

where

$$|\alpha| = \alpha_1 + \dots + \alpha_n \quad \alpha! = \alpha_1! \dots \alpha_n! \quad \mathbf{x}^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n} \quad D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}$$

4. **(\mathcal{O} notation)** $f(x) = \mathcal{O}(g(x))$ describes asymptotic behavior of function f

- (a) (as $x \rightarrow \infty$) if there exists $M \geq 0$ and $x_0 \in \mathbb{R}$ such that $|f(x)| \leq M g(x)$ for all $x > x_0$.
 - (b) (as $x \rightarrow a$) if there exists $M \geq 0$ and $\delta \in \mathbb{R}$ such that $|f(x)| \leq M g(x)$ when $0 < |x - a| < \delta$.
- Alternatively we can say

$$\limsup_{x \rightarrow a} \left| \frac{f(x)}{g(x)} \right| < \infty$$

5. **(binomial theorem)**

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k$$

6. (power series expansions)

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \dots$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots$$

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$$

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} \quad (\text{convergent if } |x| < 1)$$

$$\ln(1-x) = \sum_{n=1}^{\infty} \frac{x^n}{n} = x + \frac{x^2}{2} + \frac{x^3}{3} \quad (\text{convergent if } |x| < 1)$$

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{1}{3!} x^3 + \frac{1}{5!} x^5 + \dots$$

$$\sin^2 x = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} 2^{2n-1} x^{2n}}{(2n)!} = \frac{2x^2}{2!} - \frac{8x^4}{4!} + \frac{32x^6}{6!} + \dots$$

7. (trigonometric identities)

$$\sin(\theta + \phi) = \sin(\theta) \cos(\phi) + \sin(\phi) \cos(\theta)$$

$$\sin(\theta + \phi) + \sin(\theta - \phi) = 2 \sin(\theta) \cos(\phi)$$

$$1 - \cos(\theta) = 2 \sin^2 \left(\frac{\theta}{2} \right)$$

1 Euler's Method and Beyond

1.1 Ordinary differential equations and Lipschitz condition

1. **(Goal)** Approximate solution to

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \quad \text{with initial condition} \quad \mathbf{y}(t_0) = \mathbf{y}_0$$

where $t > t_0$ and $\mathbf{f} : [t_0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a sufficiently well behaved function

2. **(Lipschitz condition)** Given \mathbf{f} and norm $\|\cdot\|$, the Lipschitz condition is defined by

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \leq \lambda \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, t > t_0$$

where $\lambda \in \mathbb{R}$ is called Lipschitz constant.

3. **(Picard Lindelof theorem)** Consider initial value problem

$$y'(t) = f(t, y(t)) \quad y(t_0) = y_0$$

If f is uniformly Lipschitz continuous in y and continuous in t , then for some $\epsilon > 0$, there exists unique solution $y(t)$ to the initial value problem on the interval $[t_0 - \epsilon, t_0 + \epsilon]$

4. **(Analytic function)** A function \mathbf{f} is an analytic function if it is a function that is locally given by a convergent power series, i.e. an infinitely differentiable function such that at any point $(t, \mathbf{y}_0) \in [0, \infty) \times \mathbb{R}^d$ in its domain, the Taylor series converges to $\mathbf{f}(\mathbf{x})$ for \mathbf{x} in a neighborhood of (t, \mathbf{y}_0) .

- (a) (example) polynomial, exponential, trigonometric, logarithm, power function
- (b) (note) if \mathbf{f} is analytic, solution \mathbf{y} to the initial value problem is also analytic

1.2 Euler's method

Definition. (Euler's Method) Given initial value problem $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ for $t \geq t_0$ and initial value $\mathbf{y}(t_0) = \mathbf{y}_0$. If we assume $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \approx \mathbf{f}(t_0, \mathbf{y}(t_0))$ for $t \in [t_0, t_0 + h]$ (i.e. derivative in $[t_n, t_{n+1}]$ is approximated by value of derivative at t_n) for some sufficiently small time step $h > 0$, we can approximate the value of $\mathbf{y}(t)$ by

$$\begin{aligned}\mathbf{y}(t) &= \mathbf{y}(t_0) + \int_{t_0}^t \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau \\ &\approx \mathbf{y}_0 + (t - t_0) \mathbf{f}(t_0, \mathbf{y}_0)\end{aligned}$$

Given a sequence of times $(t_n)_{n \in \mathbb{N}} = (t_0, t_0 + h, \dots)$ we have numerical approximation $(\mathbf{y}_n)_{n \in \mathbb{N}}$ by

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \mathbf{f}(t_n, \mathbf{y}_n)$$

1. (intuition) euler's method is a time-stepping numerical method that covers interval by an equidistant grid and produce numerical solution at the grid points. we can show that euler's method is convergent, i.e. as $h \rightarrow 0$, grid is refined, the numerical solution tends to exact solution

Definition. (convergent method) Given a time-stepping numerical method on a compact interval $[t_0, t_0 + t^*]$, we can compute numerical solutions dependent upon h

$$\mathbf{y}_n = \mathbf{y}_{n,h} \quad \text{for } n = 0, 1, \dots, \lfloor t^*/h \rfloor$$

A method is said to be convergent if for every ODE with Lipschitz function \mathbf{f} , the numerical solution tends to the true solution as the grid becomes increasingly fine. More rigorously, if every ODE with Lipschitz function \mathbf{y} and for every $t^* > 0$, then following holds

$$\lim_{h \rightarrow 0^+} \max_{n=0,1,\dots,\lfloor t^*/h \rfloor} \|\mathbf{y}_{n,h} - \mathbf{y}(t_n)\| = 0$$

Theorem. (Euler's method is convergent)

Proof. Assume \mathbf{f} and therefore also \mathbf{y} is analytic, i.e. convergent Taylor expansion. Let $\mathbf{e}_{n,h} = \mathbf{y}_{n,h} - \mathbf{y}(t_n)$ be the numerical error. Show $\lim_{h \rightarrow 0^+} \max_n \|\mathbf{e}_{n,h}\| = 0$. By Taylor's theorem

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + h \mathbf{y}'(t_n) + \mathcal{O}(h^2) = \mathbf{y}(t_n) + h \mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathcal{O}(h^2)$$

given \mathbf{y} continuously differentiable, $\mathcal{O}(h^2)$ can be bounded uniformly for all $h > 0$ by a term ch^2 for some $c > 0$. Subtract previous from iterative formula of euler's method

$$\mathbf{e}_{n+1,h} = \mathbf{e}_{n,h} + h (\mathbf{f}(t_n, \mathbf{y}(t_n) + \mathbf{e}_{n,h}) - \mathbf{f}(t_n, \mathbf{y}(t_n))) + \mathcal{O}(h^2)$$

By triangle inequality and Lipschitz condition

$$\begin{aligned}\|\mathbf{e}_{n+1,h}\| &\leq \|\mathbf{e}_{n,h}\| + h \|\mathbf{f}(t_n, \mathbf{y}(t_n) + \mathbf{e}_{n,h}) - \mathbf{f}(t_n, \mathbf{y}(t_n))\| + ch^2 \\ &\leq (1 + h\lambda) \|\mathbf{e}_{n,h}\| + ch^2\end{aligned}$$

for $n = 0, 1, \dots, \lfloor t^*/h \rfloor - 1$. By induction on n , we can show $\|\mathbf{e}_{n,h}\| \leq \frac{c}{\lambda} h ((1 + h\lambda)^n - 1)$. Since $1 + h\lambda < e^{h\lambda}$ we have $(1 + h\lambda)^n < e^{nh\lambda} < e^{\lfloor t^*/h \rfloor h\lambda} \leq e^{t^*\lambda}$. Therefore

$$\|\mathbf{e}_{n,h}\| \leq \frac{c}{\lambda} (e^{t^*\lambda} - 1) h$$

for $n = 0, 1, \dots, \lfloor t^*/h \rfloor$. This is an upper bound on the error that is independent of h , hence $\lim_{h \rightarrow 0} \|\mathbf{e}_{n,h}\| = 0$. from which we can infer that error decays globally as $\mathcal{O}(h)$ \square

Definition. (order p method) Given arbitrary time-stepping method

$$\mathbf{y}_{n+1} = \mathcal{Y}_n(\mathbf{f}, h, \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_n) \quad n = 0, 1, \dots$$

for initial value problem, it is of order p if

$$\mathbf{y}(t_{n+1}) - \mathcal{Y}_n(\mathbf{f}, h, \mathbf{y}(t_0), \mathbf{y}(t_1), \dots, \mathbf{y}(t_n)) = \mathcal{O}(h^{p+1})$$

for every analytic \mathbf{f} and $n = 0, 1, \dots$. Intuitively, a method is of order p if it recovers exactly every polynomial solution of degrees p or less.

1. (intuition) order of a method gives information about local behavior, i.e. advancing from t_n to t_{n+1} where $h > 0$ is sufficiently small, we are incurring an error of $\mathcal{O}(h^{p+1})$. Generally want the global (convergence) behavior of the method instead.
2. (**fact**) euler's method is order 1

Proof. Euler's method can be written as $\mathbf{y}_{n+1} - (\mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n)) = 0$. Replace \mathbf{y}_k by $\mathbf{y}(t_k)$ and expand terms of Taylor series about t_n we have

$$\mathbf{y}(t_{n+1}) - (\mathbf{y}(t_n) + h\mathbf{f}(t_n, \mathbf{y}(t_n))) = (\mathbf{y}(t_n) + h\mathbf{y}'(t_n) + \mathcal{O}(h^2)) - (\mathbf{y}(t_n) + h\mathbf{y}'(t_n)) = \mathcal{O}(h^2)$$

□

1.3 The trapezoidal rule

Definition. (Trapezoidal Rule) Instead of approximating derivative by a constant in $[t_n, t_{n+1}]$, namely by its value at t_n , the trapezoidal rule approximates the value of the derivate by average of values at the endpoints. We can approximate solution $\mathbf{y}(t)$ by

$$\begin{aligned}\mathbf{y}(t) &= \mathbf{y}(t_n) + \int_{t_n}^t \mathbf{f}(\tau, \mathbf{f}(\tau))d\tau \\ &\approx \mathbf{y}(t_n) + \frac{1}{2}(t - t_n)(\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathbf{f}(t, \mathbf{y}(t)))\end{aligned}$$

Given a sequence of times $(t_n)_{n \in \mathbb{N}} = (t_0, t_0 + h, \dots)$ we have numerical approximation $(\mathbf{y}_n)_{n \in \mathbb{N}}$ by

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}h(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}))$$

1. (**theorem**) order of trapezoidal rule is 2

Proof. Compute by performing Taylor expansion on $\mathbf{y}(t_{n+1})$ and $\mathbf{y}'(t_{n+1})$ about t_n

$$\mathbf{y}(t_{n+1}) - \left\{ \mathbf{y}(t_n) + \frac{1}{2}h \{ \mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1})) \} \right\} = \mathcal{O}(h^3)$$

□

2. (**theorem**) trapezoidal rule is convergent

Proof. Detail of proof [here](#) . We can show error is bounded by

$$\|e_{n,h}\| \leq \frac{ch^2}{\lambda} \exp\left(\frac{t^*\lambda}{1 - \frac{1}{2}h\lambda}\right)$$

from which we can infer that error decays globally as $\mathcal{O}(h^2)$

□

3. (note) euler's method is explicit, since we can compute \mathbf{y}_{n+1} with a few arithmetic operations by computing \mathbf{f} , a function of a known \mathbf{y}_n . Trapezoidal rule is implicit, i.e. finding \mathbf{y}_{n+1} is not trivial and \mathbf{f} is a function of both \mathbf{y}_n and \mathbf{y}_{n+1} . We might need to solve a nonlinear equation of \mathbf{y}_{n+1}

$$\mathbf{y}_{n+1} - \frac{1}{2}h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) = \mathbf{v}$$

where $\mathbf{v} = \mathbf{y}_n + \frac{1}{2}h\mathbf{f}(t_n, \mathbf{y}_n)$ can be evaluated easily from assumptions.

1.4 The theta method

Definition. (*theta method*) is a generalization of Euler's method ($\theta = 1$) and the trapezoidal rule ($\theta = 1/2$), whereby the derivatives are assumed to be piecewise constant and provided by a linear combination of derivatives at the endpoints of each interval. The numerical approximates are,

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(\theta \mathbf{f}(t_n, \mathbf{y}_n) + (1 - \theta) \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})) \quad n = 0, 1, \dots$$

for some fixed $\theta \in [0, 1]$

1. (*fact*) theta method is explicit for $\theta = 1$ and implicit otherwise
2. (*theorem*) theta method is of order 2 for $\theta = 1/2$ and order 1 otherwise.
3. (*theorem*) theta method is convergent for every $\theta \in [0, 1]$

2 Multistep Method

3 8 Finite Differences Schemes

3.1 8.1 Finite differences

1. (**Finite difference operators**) Given real sequences $\mathbf{z} = \{z_k\}_{k \in \mathbb{Z}} = z(kh)$ for $k \in \mathbb{Z}$ as discrete sampling of a function z for some $h > 0$. Let $x_k = kh$. We can define finite difference operators mapping the space $\mathbb{R}^{\mathbb{Z}}$ of all such sequences to itself.

$$\begin{aligned} (\mathcal{E}\mathbf{z})_k &= z_{k+1} && \text{(shift)} \\ (\Delta_+\mathbf{z})_k &= z_{k+1} - z_k && \text{(forward difference)} \\ (\Delta_-\mathbf{z})_k &= z_k - z_{k-1} && \text{(backward difference)} \\ (\Delta_0\mathbf{z})_k &= z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}} && \text{(central difference)} \\ (\Upsilon_0\mathbf{z})_k &= \frac{1}{2}(z_{k-\frac{1}{2}} + z_{k+\frac{1}{2}}) && \text{(averaging)} \end{aligned}$$

Finite difference operators are composed under function composition.

- (a) (**fact**) $\mathcal{T} \in \{\mathcal{E}, \Delta_+, \Delta_-, \Delta_0, \Upsilon_0, \mathcal{D}\}$ are linear operators

$$\mathcal{T}(a\mathbf{w} + b\mathbf{z}) = a\mathcal{T}\mathbf{w} + b\mathcal{T}\mathbf{z} \quad \text{for } \mathbf{w}, \mathbf{z} \in \mathbb{R}^{\mathbb{Z}}, \quad a, b \in \mathbb{R}$$

- (b) (**convention**) $\mathcal{T}z_k$ stands for $(\mathcal{T}\mathbf{z})_k$

2. (**Differential operator**) The goal is to approximate derivatives \mathcal{D} by expressing it with a linear combination of values along the grid.

$$(\mathcal{D}\mathbf{z})_k = z'(kh) \quad \text{(differential)}$$

3. (**Functions of operators**) Finite difference operators are functions of h . Given an analytic function as Taylor series, $g(x) = \sum_{j=0}^{\infty} a_j x^j$, we can expand g about $\{\mathcal{E} - \mathcal{I}, \Upsilon_0 - \mathcal{I}, \Delta_+, \Delta_-, \Delta_0, h\mathcal{D}\}$,

$$g(\Delta_+)\mathbf{z} = \left(\sum_{j=0}^{\infty} a_j \Delta_+^j \right) \mathbf{z} = \sum_{j=0}^{\infty} a_j (\Delta_+^j \mathbf{z})$$

4. (**Asymptotics of operators**)

$$\{\mathcal{E} - \mathcal{I}, \Upsilon_0 - \mathcal{I}, \Delta_+, \Delta_-, \Delta_0, h\mathcal{D}\} \xrightarrow{h \rightarrow 0^+} \mathcal{O}$$

- (a) (**example**)

$$\Delta_+ z_k = z_{k+1} - z_k = z(x_k + h) - z(x_k) = h z'(\eta_k) = \mathcal{O}(h)$$

by some $\eta_k \in [x_k, x_{k+1}]$ by mean value theorem

5. (Operator $\mathcal{E}^{1/2}$)

$$(\mathcal{E}^{1/2}z)_k = z((k + \frac{1}{2})h)$$

by defining it with a power series expansion of $g(x) = \sqrt{1+x}$

$$\mathcal{E}^{1/2} = (\mathcal{I} + (\mathcal{E} - \mathcal{I}))^{1/2} = \mathcal{I} + \sum_{j=0}^{\infty} \frac{(-1)^{j-1}}{2^{2j-1}} \frac{(2j-2)!}{(j-1)!j!} (\mathcal{E} - \mathcal{I})^j$$

6. (Operator commutativity) Idea is all operator can be expressed w.r.t. \mathcal{E}

$$\begin{aligned}\Delta_+ &= \mathcal{E} - \mathcal{I} \\ \Delta_- &= \mathcal{I} - \mathcal{E}^{-1} \\ \Delta_0 &= \mathcal{E}^{1/2} - \mathcal{E}^{-1/2} \\ \Upsilon_0 &= \frac{1}{2}(\mathcal{E}^{1/2} + \mathcal{E}^{-1/2}) \\ \mathcal{I} &= \mathcal{E}^0 \\ h\mathcal{D} &= \ln \mathcal{E}\end{aligned}$$

Proof. Rest are trivial. To show $h\mathcal{D} = \ln \mathcal{E}$, note

$$\mathcal{E}z(x) = z(x+h) = \sum_{i=0}^{\infty} \frac{h^i}{i!} \frac{d^i z(x)}{dx^i} = \left[\sum_{i=0}^{\infty} \frac{1}{i!} (h\mathcal{D})^i \right] z(x) = e^{h\mathcal{D}} z(x)$$

□

7. (rewrite \mathcal{D} in terms of Δ_+ , Δ_- , Δ_0)

$$\begin{aligned}h\mathcal{D} &= \ln(\mathcal{I} + \Delta_+) \\ h\mathcal{D} &= -\ln(\mathcal{I} - \Delta_-) \\ h\mathcal{D} &= 2\ln\left(\frac{1}{2}\Delta_0 + \sqrt{\mathcal{I} + \frac{1}{4}\Delta_0^2}\right)\end{aligned}$$

Proof. From previous, $\mathcal{E} = \mathcal{I} + \Delta_+ = (\mathcal{I} - \Delta_-)^{-1}$. For the last expression, consider

$$\begin{aligned}\Delta_0 &= \mathcal{E}^{1/2} - \mathcal{E}^{-1/2} \\ \mathcal{E}^{1/2}\Delta_0 &= \mathcal{E} - \mathcal{I} \\ (\mathcal{E}^{1/2})^2 - \mathcal{E}^{1/2}\Delta_0 - \mathcal{I} &= 0 \\ \mathcal{E}^{1/2} &= \frac{1}{2}\Delta_0 \pm \sqrt{\mathcal{I} + \frac{1}{4}\Delta_0^2} \quad (+ \text{ is correct}) \\ \mathcal{E} &= \left(\frac{1}{2}\Delta_0 + \sqrt{\mathcal{I} + \frac{1}{4}\Delta_0^2}\right)^2\end{aligned}$$

□

8. (approximate \mathcal{D} and its powers) To approximate \mathcal{D} with Δ_+ , we can expand $\ln(\mathcal{I} + \Delta_+)$ by power series expansion of $\ln(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i!}$

$$\begin{aligned}\mathcal{D} &= \frac{1}{h} \ln(\mathcal{I} + \Delta_+) = \frac{1}{h} \left[\Delta_+ - \frac{1}{2}\Delta_+^2 + \frac{1}{3}\Delta_+^3 + \mathcal{O}(\Delta_+^4) \right] \\ &= \frac{1}{h} \left(\Delta_+ - \frac{1}{2}\Delta_+^2 + \frac{1}{3}\Delta_+^3 \right) + \mathcal{O}(h^3) \quad h \rightarrow 0\end{aligned}$$

where $\Delta_+ = \mathcal{O}(h)$ as $h \rightarrow 0$ shown perviously. Use binomial theorem on \mathcal{D} repeatedly and collect terms to $\mathcal{O}(h^3)$,

$$\mathcal{D}^s = \frac{1}{h^s} \left[\Delta_+^s - \frac{1}{2}s\Delta_+^{s+1} + \frac{1}{24}s(3s+5)\Delta_+^{s+2} \right] + \mathcal{O}(h^3) \quad h \rightarrow 0$$

Inuitively, we can approximate $\mathcal{D}^s z_k = d^s z(kh)/dx^s$ up to $\mathcal{O}(h^3)$ with $s+3$ grid points in the positive direction, i.e. $z_k, z_{k+1}, \dots, z_{k+s+2}$. Similarly we can express \mathcal{D} in terms of grid points to the left with Δ_- .

$$\mathcal{D}^s = \frac{(-1)^s}{h^s} (\ln(\mathcal{I} - \Delta_-))^s = \frac{1}{h^s} \left[\Delta_-^s + \frac{1}{2}s\Delta_-^{s+1} + \frac{1}{24}s(3s+5)\Delta_-^{s+2} \right] + \mathcal{O}(h^3) \quad h \rightarrow 0$$

Similarly we can express \mathcal{D} in terms of grid points on the left and right with Δ_0 operator. Note, only even powers of Δ_0 maps $\mathbb{R}^{\mathbb{Z}} \rightarrow \mathbb{R}^{\mathbb{Z}}$, i.e. onto grid points. ($\Delta_0^2 z_k = z_{k+1} - 2z_k + z_{k-1}$ and for any power to $2s$, $\Delta_0^{2s} = (\Delta_0^2)^s$). We consider Maclaurin expansion of function $g(\xi) = \ln(\xi + \sqrt{1 + \xi^2})$

$$g(\xi) = 2 \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \binom{2j}{j} \left(\frac{1}{2}\xi \right)^{2j+1}$$

Let $\xi = \frac{1}{2}\Delta_0$, we have power series expansion of \mathcal{D} in terms of Δ_0

$$\mathcal{D} = \frac{2}{h} g\left(\frac{1}{2}\Delta_0\right) = \frac{4}{h} \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \binom{2j}{j} \left(\frac{1}{4}\Delta_0 \right)^{2j+1}$$

However powers of Δ_0 are all odd, we raise power to $2s$ to keep output of operator on the grid

$$\mathcal{D}^{2s} = \frac{1}{h^{2s}} \left[(\Delta_0^2)^s - \frac{s}{12} (\Delta_0^2)^{s+1} + \frac{s(11+5s)}{1440} (\Delta_0^2)^{s+2} \right] + \mathcal{O}(h^6) \quad h \rightarrow 0$$

approximates \mathcal{D} up to $\mathcal{O}(h^6)$

9. **(comparing Δ_+ and Δ_0 for approximating \mathcal{D})** To attain $\mathcal{O}(h^{2p})$ error, Δ_+ requires $2s+2p$ grid points and Δ_0 requires $2s+2p-1$ grid points. However Δ_+ would have a smaller error constant. (exercise 8.3)

10. **(express \mathcal{R}_0 in terms of Δ_0)**

$$\mathcal{R}_0 = \left(\mathcal{I} + \frac{1}{4}\Delta_0^2 \right)^{1/2}$$

Proof.

$$\begin{aligned} \mathcal{R}_0 &= \frac{1}{2} \left(\mathcal{E}^{1/2} + \mathcal{E}^{-1/2} \right) & \longrightarrow & 4\mathcal{R}_0^2 = \mathcal{E} + 2\mathcal{I} + \mathcal{E}^{-1} \\ \Delta_0 &= \mathcal{E}^{1/2} - \mathcal{E}^{-1/2} & \longrightarrow & \Delta_0^2 = \mathcal{E} - 2\mathcal{I} + \mathcal{E}^{-1} \end{aligned}$$

Therefore $4\mathcal{R}_0 - \Delta_0^2 = 4\mathcal{I}$ and result follows □

11. **(approximate odd derivatives with \mathcal{R}_0 with central difference)**

$$\mathcal{D} = \frac{1}{h} (\mathcal{R}_0 \Delta_0) \left[\sum_{j=0}^{\infty} (-1)^j \binom{2j}{j} \left(\frac{1}{16}\Delta_0^2 \right)^j \right] \left[\sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1} \binom{2i}{i} \left(\frac{1}{16}\Delta_0^2 \right)^i \right]$$

which are constructed from even powers of Δ_0 and $\mathcal{R}_0 \Delta_0$ which will make image of \mathcal{D} operator reside on the grid.

$$\mathcal{R}_0 \Delta_0 z_k = \mathcal{R}_0 \left(z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}} \right) = \frac{1}{2} (z_{k+1} - z_{k-1})$$

Raising powers of \mathcal{D} yield

$$\mathcal{D}^2 = \frac{1}{h^2} (\mathcal{R}_0 \Delta_0)^2 (\mathcal{I} - \frac{1}{3}\Delta_0^2) + \mathcal{O}(h^4)$$

12. **(practical use)** instead of mixing difference operators, opt for finite difference grids. For finite grids, one-sided finite differences can be employed to evaluate \mathcal{D} near boundaries.

3.2 8.2 The five-point formula for $\nabla^2 u = f$

1. **(consistent)** A method is consistent if the truncation error goes to 0 as step size goes to zero
2. **(order of consistency)** of $\mathcal{O}(\Delta x^p) + \mathcal{O}(\Delta y^q)$ is p in x and q in y .
3. **(theorem)** If a method is consistent and stable, then it is convergent and order of convergence will be same as the order of consistency

1. (Poisson Equation)

$$\nabla^2 u = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) u = f \quad (x, y) \in \Omega$$

and $f = f(x, y)$ is continuous, domain $\Omega \subset \mathbb{R}^2$ is bounded, open, and connected and has a piecewise-smooth boundary. Assume *Dirichlet condition*, i.e.

$$u(x, y) = \phi(x, y) \quad (x, y) \in \partial\Omega$$

2. **(Setup)** Inscribe a grid that is axis-aligned with equal spacing of Δx in both direction, i.e. pick $\Delta x > 0$, $(x_0, y_0) \in \Omega$ and let $\Omega_{\Delta x}$ be

$$\Omega_{\Delta x} = \{x_0 + k\Delta x, y_0 + l\Delta x\} \subset \Omega$$

Denote

$$\begin{aligned} \mathbf{I}_{\Delta x} &= \{(k, l) \in \mathbb{Z}^2 \mid (x_0 + k\Delta x, y_0 + l\Delta x) \in \Omega\} \\ \mathbf{I}_{\Delta x}^\circ &= \{(k, l) \in \mathbb{Z}^2 \mid (x_0 + k\Delta x, y_0 + l\Delta x) \in \Omega^\circ\} \end{aligned}$$

and for every $(k, l) \in \mathbf{I}_{\Delta x}^\circ$, let $u_{k,l}$ be *approximation* to the solution $u(x_0 + k\Delta x, y_0 + l\Delta x)$ of the Poisson equation at the relevant grid point. Note there is no need to approximate points in $\mathbf{I}_{\Delta x} \setminus \mathbf{I}_{\Delta x}^\circ$ since they lie on $\partial\Omega$ and their exact values given by ϕ .

3. **(internal, near-boundary, boundary points)** A point on the grid $(k, l) \in \mathbf{I}_{\Delta x}$ whereby $(k \pm 1, l)$ and $(k, l \pm 1)$ are in $\mathbf{I}_{\Delta x}$ is called *internal point*. A point $(k, l) \in \mathbf{I}_{\Delta x}$ where we can no longer employ a finite difference scheme (and so requires a special approach) is called *near-boundary points*. $(k, l) \in \partial\Omega$ are called *boundary points*
4. **(Central difference approximation)** given u sufficiently smooth, we can approximate ∇^2

$$\nabla^2 = \frac{1}{(\Delta x)^2} (\Delta_{0,x}^2 + \Delta_{0,y}^2) + \mathcal{O}((\Delta x)^2) \quad \text{where} \quad \frac{\partial^2 u}{\partial x^2} = \frac{1}{\Delta x^2} \Delta_{0,x}^2 u_{k,l} + \mathcal{O}((\Delta x)^2)$$

with central difference operators, i.e. $\Delta_{0,x}, \Delta_{0,y}$ along the x,y-axis. We can rewrite Poisson equation by the *five point* finite difference scheme. For every internal grid point (k, l) , we have

$$\frac{1}{(\Delta x)^2} (\Delta_{0,x}^2 + \Delta_{0,y}^2) u_{k,l} = f_{k,l}$$

where $f_{k,l} = f(x_0 + k\Delta x, y_0 + l\Delta x)$. Expanding expression, we have

$$u_{k-1,l} + u_{k+1,l} + u_{k,l-1} + u_{k,l+1} - 4u_{k,l} = (\Delta x)^2 f_{k,l}$$

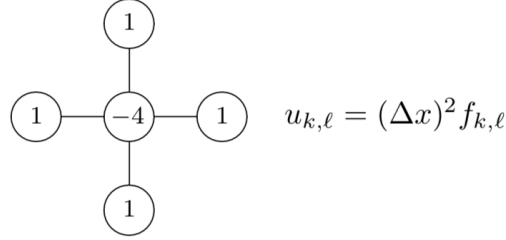
Intuitively, we have a linear combination of values of u at grid point and at immediate horizontal and vertical neighbors of this point.

5. (properties)

- (a) **(truncation error)** $\mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta y^2)$ (computed by substituting exact solution $\tilde{u}_{i,j} = u(x_0 + \Delta x, y_0 + \Delta y)$ to finite difference formula in place of approximate values at grid points $u_{i,j}$ to compute the truncation error)

$$\begin{aligned} & \frac{\tilde{u}_{i+1,j} - 2\tilde{u}_{i,j} + \tilde{u}_{i-1,j}}{\Delta x^2} + \frac{\tilde{u}_{i,j+1} - 2\tilde{u}_{i,j} + \tilde{u}_{i,j-1}}{\Delta y^2} - f(x_i, y_i) \\ &= \frac{\partial u(x_i, y_j)}{\partial x^2} + \mathcal{O}(\Delta x^2) + \frac{\partial u(x_i, y_j)}{\partial y^2} + \mathcal{O}(\Delta y^2) - f(x_i, y_i) \\ &= \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta y^2) \end{aligned}$$

6. **(computational stencil / molecule)**

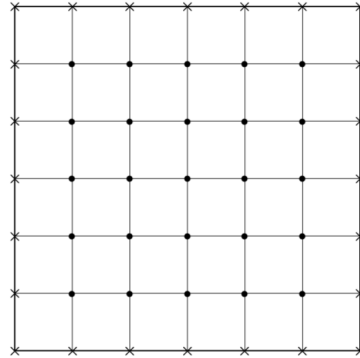


7. **(solving linear equation)** Main idea of finite difference method is to associate every grid point having an index in $\mathbf{I}_{\Delta x}^\circ$ a single linear equation. Solving a system of linear equations whose solution is our approximation $\mathbf{u} = (u_{k,l})_{(k,l) \in \mathbf{I}_{\Delta x}^\circ}$. Performance of finite difference is evaluated with

- (a) nonsingular linear system (such that \mathbf{u} exists and is unique)
- (b) for $\Delta x \rightarrow 0$, the convergence of \mathbf{u} to exact solution of Poisson equation and error
- (c) efficient and robust ways to solve sparse linear systems

8. **(A simplified grid)** over a square Ω . Let

$$\Omega = \{(x, y) \mid 0 < x, y < 1\} \quad \Delta x = 1/(m+1) \quad (x_0, y_0) = 0$$



9. **(numerical example on Laplace equation)** indicates that error decreases by a factor of 4 when number of steps m increase by a factor of 2.
10. **(discretization to a system of linear equations)** Rearrange $u_{k,l}$ to $\mathbf{u} \in \mathbb{R}^s$ where $s = m^2$ according to some permutation $\{(k_i, l_i)\}_{i=1,2,\dots,m}$ and write

$$A\mathbf{u} = \mathbf{b}$$

where A is $s \times s$ matrix and $\mathbf{b} \in \mathbb{R}^s$ includes both $(\Delta x)^2 f_{k,l}$ and boundary values.

11. **(lemma (unique solution to linear system))** A from previous is symmetric and the set of its eigenvalues is

$$\sigma(A) = \{\lambda_{\alpha,\beta} \mid \alpha, \beta = 1, 2, \dots, m\}$$

where

$$\lambda_{\alpha,\beta} = -4 \left\{ \sin^2 \left[\frac{\alpha\pi}{2(m+1)} \right] + \sin^2 \left[\frac{\beta\pi}{2(m+1)} \right] \right\}$$

Proof. Symmetry follows by examining matrix A . To find eigenvalues of A , find nonzero functions $(v_{k,l})_{k,l=0,1,\dots,m+1}$ such that $v_{k,0} = v_{k,m+1} = v_{0,l} = v_{m+1,l} = 0$ where $k, l = 1, 2, \dots, m$ such that

$$v_{k-1,l} + v_{k+1,l} + v_{k,l-1} + v_{k,l+1} - 4v_{k,l} = \lambda v_{k,l} \quad k, l = 1, 2, \dots, m$$

is satisfied for some λ . It follows that $(v_{k,l})$ is an eigenvector and λ is the corresponding eigenvalue of A . Given $\lambda_{\alpha,\beta}$ for some α, β , we show that

$$v_{k,l} = \sin \left(\frac{k\alpha\pi}{m+1} \right) \sin \left(\frac{l\beta\pi}{m+1} \right) \quad k, l = 0, 1, \dots, m+1$$

is the corresponding eigenvector by verifying above formula. \square

12. **(corollary)** The matrix A is negative definite and, therefore, nonsingular

Proof. A is symmetric and from previous lemma eigenvalues are negative, therefore it is negative definite and nonsingular \square

13. **(eigenvalues of the Laplace operator)** The function v , not identically zero, is said to be an *eigenfunction* of ∇^2 in a domain Ω and λ is the corresponding eigenvalue if v vanishes along $\partial\Omega$ and satisfies within Ω the equation $\nabla^2 v = \lambda v$. Note eigenvalues and eigenfunctions of the Laplace operator ∇^2 over $(0,1)^2$ are *related to* eigenvalues and eigenvectors of the matrix A . Given α, β , eigenvalue of ∇^2 and the corresponding eigenfunction is given by

$$\begin{aligned} \lambda_{\alpha,\beta} &= -(\alpha^2 + \beta^2)\pi^2 \\ v(x, y) &= \sin(\alpha\pi x) \sin(\beta\pi y) \quad x, y \in [0, 1] \end{aligned}$$

We can easily verify that

$$\begin{aligned} \nabla^2 v &= -\alpha^2 \pi \sin(\alpha\pi x) \sin(\beta\pi y) - \beta^2 \pi \sin(\alpha\pi x) \sin(\beta\pi y) \\ &= -(\alpha^2 + \beta^2)\pi^2 v \end{aligned}$$

v obeys boundary conditions. Note eigenvectors $v_{k,l}$ for A can be obtained by sampling of the eigenfunction v at grid points

$$\left\{ \left(\frac{k}{m+1}, \frac{l}{m+1} \right) \right\}_{k,l=0,1,\dots,m+1}$$

Note $(\Delta x)^{-2}\lambda_{\alpha,\beta}$ is a good approximation to $-(\alpha^2 + \beta^2)\pi^2$ provided α, β are small in comparison with m . Note we can expand \sin^2 in a power series

$$\begin{aligned} \frac{\lambda_{\alpha,\beta}}{(\Delta x)^2} &= -4 \left(\left\{ \left(\frac{\alpha\pi}{2(m+1)} \right)^2 - \frac{1}{3} \left(\frac{\alpha\pi}{2(m+1)} \right)^4 + \dots \right\} + \left\{ \left(\frac{\beta\pi}{2(m+1)} \right)^2 - \frac{1}{3} \left(\frac{\beta\pi}{2(m+1)} \right)^4 + \dots \right\} \right) \\ &= -(\alpha^2 + \beta^2)\pi^2 + \frac{1}{12}(\alpha^4 + \beta^4)\pi^4(\Delta x)^2 + \mathcal{O}((\Delta x)^4) \end{aligned}$$

14. **(theorem (convergence))** Subject to sufficient smoothness of the function f and the boundary conditions, there exists a number $c > 0$, independent of Δx , such that

$$\|e\| \leq c(\Delta x)^2 \quad \Delta x \rightarrow 0$$

or equivalently

$$\lim_{\Delta x \rightarrow 0} \|e\|_{\infty} = 0$$

where $e \in \mathbb{R}^s$, $s = m^2$ in same order as that of u . Denote $e_{k,l} = u_{k,l} - \tilde{u}_{k,l}$ as error of the five point formula at the (k,l) th grid point.

15. **(handle near boundary grid points)** approximate z'' at P in x direction as a linear combination of value of z at P, Q, T . The coefficient of terms can be determined via Taylor expansion of $z_{x_0-\Delta x}, z_{x_0}, z_{x_0+\tau\Delta x}$ at $a = x_0$ and solve a 3×3 linear system. The error of approximation is $\mathcal{O}(\Delta x)$

$$\begin{aligned} & \frac{1}{(\Delta x)^2} \left[\frac{2}{\tau+1} z(x_0 - \Delta x) - \frac{2}{\tau} z(x_0) + \frac{2}{\tau(\tau+1)} z(x_0 + \tau\Delta x) \right] \\ &= z''(x_0) + \frac{1}{3}(\tau-1)z'''(x_0)\Delta x + \mathcal{O}((\Delta x)^2) \end{aligned}$$

To achieve $\mathcal{O}((\Delta x)^2)$ order for the error, we use value of z at 4 grid points V, Q, P, T to approximate z'' . Coefficient to linear term can be determined with Taylor expansion and solve a 4×4 linear system.

$$\begin{aligned} & \frac{1}{(\Delta x)^2} \left[\frac{\tau-1}{\tau+2} z(x_0 - 2\Delta x) + \frac{2(2-\tau)}{\tau+1} z(x_0 - \Delta x) - \frac{3-\tau}{\tau} z(x_0) + \frac{6}{\tau(\tau+1)(\tau+2)} z(x_0 + \tau\Delta x) \right] \\ &= z''(x_0) + \mathcal{O}((\Delta x)^2) \end{aligned}$$

In total, a good approximation to $\nabla^2 u$ at P requires 6 points. For P corresponding to grid (k,l) , we obtain the following linear equation for constructing A and b for both first order and second order approximations

$$\begin{aligned} & \frac{2}{\tau+1} u_{k-1,l} + \frac{2}{\tau(\tau+1)} u_{k+\tau,l} + u_{k,l-1} + u_{k,l+1} - \frac{2+2\tau}{\tau} u_{k,l} = (\Delta x)^2 f_{k,l} \\ & \frac{\tau-1}{\tau+2} u_{k-2,l} + \frac{2(2-\tau)}{\tau+1} u_{k-1,l} + \frac{6}{\tau(\tau+1)(\tau+2)} u_{k+\tau,l} + u_{k,l-1} + u_{k,l+1} - \frac{3+\tau}{\tau} u_{k,l} = (\Delta x)^2 f_{k,l} \end{aligned}$$

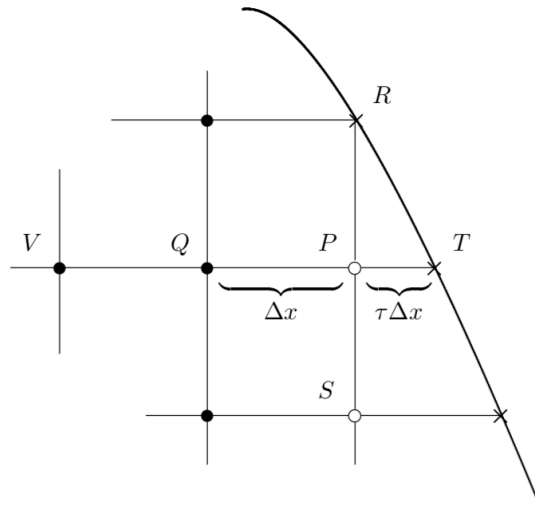


Figure 8.6 Computational grid near a curved boundary.