Numerical Methods for Differential Equations

Mathematical and Computational Tools

Tony Stillfjord, Gustaf Söderlind

Numerical Analysis, Lund University

Contents

Part 1. Vector norms, matrix norms and logarithmic norms

1. Vector norms

Definition *A vector norm* $\|\cdot\| : \mathbb{X} \to \mathbb{R}$ *satisfies*

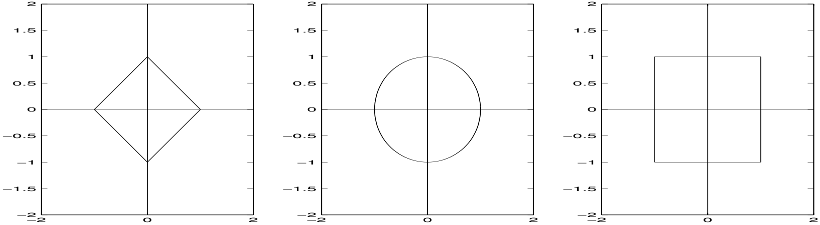
1. $\|u\| \geq 0\,;\qquad \|u\| = 0 \Leftrightarrow u = 0$
2. $\|\alpha u\| = |\alpha| \cdot \|u\|$
3. $\|u\| - \|v\| \leq \|u \pm v\| \leq \|u\| + \|v\|$

A norm generalizes the notion of *distance* between points

Vector norms

Definition *The* l^p *norms are defined* $\|x\|_p = \left(\sum_{k=1}^{N} |x_k|^p \right)^{1/p}$

Graph of the unit circle in $\mathbb{R}^2$ for l^p norms



Unit circles for $p = 1$, $p = 2$ *(Euclidean norm), and* $p = \infty$

2. Matrix norms

Definition *The operator norm associated with the vector norm* $\|\cdot\|$ *is defined by*

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

For every vector norm there is a corresponding matrix norm

Note

1. $\|Ax\| \leq \|A\| \cdot \|x\|$
2. $\|AB\| \leq \|A\| \cdot \|B\|$

Vector norm	Matrix norm
$\|x\|_1 = \sum_i \|x_i\|$	$\max_j \sum_i \|a_{ij}\|$
$\|x\|_2 = \sqrt{\sum_i \|x_i\|^2}$	$\sqrt{\rho[A^{\mathrm{H}} A]}$
$\|x\|_\infty = \max_i \|x_i\|$	$\max_i \sum_j \|a_{ij}\|$

Definition The *spectral radius* of a matrix is defined by

$$\rho[A] = \max |\lambda[A]|$$

3. Inner products

Definition A bilinear form $\langle \cdot, \cdot \rangle : \mathbb{X} \times \mathbb{X} \to \mathbb{C}$ satisfying

1. $\langle u, u \rangle \geq 0\,;\qquad \langle u, u \rangle = 0 \;\Leftrightarrow\; u = 0$
2. $\langle u, v \rangle = \overline{\langle v, u \rangle}$
3. $\langle u, \alpha v \rangle = \alpha \langle u, v \rangle$
4. $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$

An inner product generates the Euclidean norm $\langle u, u \rangle = \|u\|^2$

An inner product generalizes the notion of scalar product

1. Scalar product in $\mathbb{R}^m$: $\langle u, v\rangle = u^{\mathrm{T}} v$

with corresponding Euclidean vector norm $\|u\|_2^2 = \sum_{k=1}^{N} |u_k|^2$

2. General inner product in $\mathbb{C}^m$: $\langle u, v\rangle_G = u^{\mathrm{H}} G v$ for any symmetric positive definite matrix G

3. Inner product in $L^2[0,1]$: $\langle u, v\rangle = \int_0^1 uv \, \mathrm{d}x$

with corresponding L^2 norm of functions $\|u\|_{L^2}^2 = \int_0^1 |u|^2 \, \mathrm{d}x$

Inner products and operator norms

Theorem *Cauchy–Schwarz inequality*

$$|\langle u, v\rangle| \leq \|u\| \cdot \|v\|$$

Note the absolute value!

$\Rightarrow \langle u, v\rangle \geq -\|u\| \cdot \|v\|$ on real vector spaces

Definition *The operator norm associated with $\langle \cdot, \cdot \rangle$ is*

$$\|A\|^2 = \sup_{x \neq 0} \frac{\langle Au, Au\rangle}{\|u\|^2}$$

Hence $\langle Au, Au\rangle \leq \|A\|^2 \|u\|^2$

Classical stability for ODEs

$$\dot{x} = Ax; \quad x(0) = x_0$$

Characterization *Elementary stability conditions*

- $\mathrm{Re}\,\lambda_k < 0 \quad (\Leftrightarrow \mathrm{e}^{tA} \to 0 \text{ as } t \to \infty)$
- $\|\mathrm{e}^{tA}\| \le C \quad \text{for all } t \ge 0$
- $\mathrm{d}\|\mathrm{e}^{tA}\|/\mathrm{d}t \le 0 \quad (\Leftrightarrow \|\mathrm{e}^{tA}\| \le 1 \text{ for all } t \ge 0)$

Time-dependent linear systems

Consider non-autonomus linear system

$$\dot{x} = A(t)x; \quad x(0) = x_0$$

Stability is no longer characterized by eigenvalues

Even with constant eigenvalues in the left half plane the system can be unstable (Petrowski & Hoppenstedt)

Problem *Under what conditions does* $\|x(t)\|$ *remain bounded as* $t \to \infty$*?*

Note that for an inner product norm,

$$\frac{\mathrm{d}\|x\|^2}{\mathrm{d}t} = \frac{\mathrm{d}\langle x, x\rangle}{\mathrm{d}t} = 2\mathrm{Re}\langle x, \dot{x}\rangle$$

and that if $\dot{x} = Ax$, then

$$\mathrm{Re}\,\langle x, \dot{x}\rangle = \mathrm{Re}\,\langle x, Ax\rangle \leq \mu[A] \cdot \langle x, x\rangle$$

Definition *The logarithmic norm is defined by*

$$\mu[A] = \sup_{x \neq 0} \frac{\mathrm{Re}\langle x, Ax\rangle}{\|x\|^2}$$

4. The logarithmic norm $\mu[A]$

Definition *For general matrix norms the logarithmic norm is defined by*

$$\mu[A] = \lim_{h \to 0+} \frac{\|I + hA\| - 1}{h}$$

If $\dot{x} = Ax$, the following differential inequality holds

$$\mathrm{d}\|x\|/\mathrm{d}t \ \leq \ \mu[A] \cdot \|x\|$$

Note The logarithmic norm may be *negative*. Solution bound

$$\|x(t)\| \leq \mathrm{e}^{t\mu[A]} \cdot \|x(0)\| \; ; \quad t \geq 0$$

Why the logarithmic norm?

Crude estimate

$$\frac{\mathrm{d}\|x\|}{\mathrm{d}t} \leq \|\dot{x}\| = \|Ax\| \leq \|A\| \cdot \|x\|$$

Exponentially growing bound

$$\|x(t)\| \leq \mathrm{e}^{t\|A\|} \cdot \|x(0)\|$$

Note *Because* $\mu[A] \leq \|A\|$ *we always have*

$$\mathrm{e}^{t\mu[A]} \leq \mathrm{e}^{t\|A\|}$$

Vector norm	Matrix norm	Log norm $\mu[A]$
$\|x\|_1 = \sum_i \|x_i\|$	$\max_j \sum_i \|a_{ij}\|$	$\max_j [\mathrm{Re}\, a_{jj} + \sum_i' \|a_{ij}\|]$
$\|x\|_2 = \sqrt{\sum_i \|x_i\|^2}$	$\sqrt{\rho[A^{\mathrm{H}}A]}$	$\alpha[(A + A^{\mathrm{H}})/2]$
$\|x\|_\infty = \max_i \|x_i\|$	$\max_i \sum_j \|a_{ij}\|$	$\max_i [\mathrm{Re}\, a_{ii} + \sum_j' \|a_{ij}\|]$

Definition *Spectral abscissa* $\alpha[A] = \max \mathrm{Re}\, \lambda[A]$

5. Logarithmic norm properties

Theorem *The logarithmic norm has the following basic properties, which hold for all matrices* A and B

1. $\mu[A] \leq \|A\|$
2. $\mu[A + zI] = \mu[A] + \mathrm{Re}\, z$
3. $\mu[\alpha A] = \alpha\, \mu[A], \qquad \alpha \geq 0$
4. $\mu[A + B] \leq \mu[A] + \mu[B]$
5. $\|\mathrm{e}^{tA}\| \leq \mathrm{e}^{t\mu[A]}, \qquad t \geq 0$

Uniform Monotonicity Theorem

The condition $\mu[A] < 0$ is akin to A being *negative definite*. Then A has a bounded inverse. More precisely,

Theorem (Uniform Monotonicity Theorem) *If $\mu[A] < 0$ then A is nonsingular and*

$$\|A^{-1}\| \leq -1/\mu[A]$$

Proof (Here only for the Euclidean norm) Note that $\forall x$

$$x^{\mathrm{T}} A x \leq \mu_2[A] \cdot x^{\mathrm{T}} x$$

Suppose $\mu_2[A] < 0$. By the Cauchy-Schwarz inequality

$$-\|x\|_2 \cdot \|Ax\|_2 \leq x^{\mathrm{T}} Ax \leq \mu_2[A] \cdot \|x\|_2^2 < 0$$

for all $x \neq 0$. Hence $Ax \neq 0$, so A^{-1} exists! Put $x = A^{-1}y$ and rearrange to get

$$-\|y\|_2 \leq \mu_2[A] \cdot \|A^{-1}y\|_2 \quad \Rightarrow \quad \frac{\|A^{-1}y\|_2}{\|y\|_2} \leq -\frac{1}{\mu_2[A]}$$

Take maximum over y to see that $\|A^{-1}\|_2 \leq -1/\mu_2[A]$ $\square$

Consider Explicit Euler for $\dot{x} = Ax + f(t)$

$$x_{n+1} = x_n + hAx_n + hf(t_n)$$
$$x(t_{n+1}) = x(t_n) + hAx(t_n) + hf(t_n) - h^2 r_n$$

Global error $e_n = x_n - x(t_n)$

$$e_{n+1} = e_n + hAe_n + h^2 r_n \quad \Rightarrow$$

$$\|e_{n+1}\| \leq \|e_n\| + h\|A\| \cdot \|e_n\| + \|h^2 r_n\|$$

Classical convergence analysis

Recall (Chapter 1, p.15)

Lemma *If* $u_{n+1} \leq (1 + h\mu)u_n + ch^2$ *with* $u_0 = 0$, *then*

$$u_n \leq \frac{ch}{\mu}[(1 + h\mu)^n - 1] \leq ch\,\frac{\mathrm{e}^{\mu t_n} - 1}{\mu}$$

if $h\mu \geq 0$. *In case* $-1 < h\mu < 0$, *we have*

$$\max_n u_n \leq -\frac{ch}{\mu}$$

Classical convergence analysis . . .

$$\|e_{n+1}\| \leq (1 + h\|A\|) \cdot \|e_n\| + \|h^2 r_n\|$$

The lemma applies with $\mu = \|A\|$ and $c = \max_n \|r_n\|$

$$\|e_n\| \leq h \max_n \|r_n\| \, \frac{\mathrm{e}^{\|A\| t_n} - 1}{\|A\|}$$

"Convergence," but *exponentially growing bound*

(Hopeless!)

Modern convergence analysis

Explicit Euler for $\dot{x} = Ax + f(t)$

$$x_{n+1} = x_n + hAx_n + hf(t_n)$$
$$x(t_{n+1}) = x(t_n) + hAx(t_n) + hf(t_n) - h^2 r_n$$

Global error $e_n = x_n - x(t_n)$

$$e_{n+1} = e_n + hAe_n + h^2 r_n = (I + hA)e_n + h^2 r_n \quad \Rightarrow$$

$$\|e_{n+1}\| \leq \|I + hA\| \cdot \|e_n\| + \|h^2 r_n\|$$

Modern convergence analysis ...

Note that, using the log norm,

$$\mu[A] = \lim_{h \to 0+} \frac{\|I + hA\| - 1}{h}$$

we have

$$\|I + hA\| = 1 + h\mu[A] + \mathrm{O}(h^2)$$

as $h\|A\| \to 0$

Modern convergence analysis . . .

Now the lemma applies with $\mu \approx \mu[A]$ and $c = \max_n \|r_n\|$

$$\|e_n\| \lessapprox h \max_n \|r_n\| \frac{\mathrm{e}^{\mu[A]t_n} - 1}{\mu[A]} ; \quad \mu[A] \geq 0$$

and in case $-1 < h\mu[A] < 0$ we have

$$\max_n \|e_n\| \lessapprox - \frac{h \max_n \|r_n\|}{\mu[A]}$$

Convergence, with *"realistic" error bound*, as $\mu[A] \leq \|A\|$

Consider Implicit Euler for $\dot{x} = Ax + f(t)$

$$x_{n+1} = x_n + hAx_{n+1} + hf(t_{n+1})$$
$$x(t_{n+1}) = x(t_n) + hAx(t_{n+1}) + hf(t_{n+1}) - h^2 r_n$$

When is it possible to solve $(I - hA)x_{n+1} = x_n + hf(t_{n+1})$?

Uniform monotonicity theorem guarantees unique solution if

$$\mu[hA - I] < 0 \Leftrightarrow \mu[hA] < 1$$

Easily satisfied, *even without bound on* $\|hA\|$

Part 2. Interpolation

1. What is interpolation?

Interpolation is "the opposite" of discretization

Problem Given a discrete grid function (a vector) $F = \{f_j\}_0^N$ defined on a grid $\{x_j\}_0^N$, find a continuous function $f(x)$ with the *interpolating property* $f(x_j) = f_j$

Compare digital–to–analog conversion

Typically the function *f is sought among polynomials* or among trigonometric functions (Fourier analysis)

Naïve polynomial interpolation

$$P_n(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

$n + 1$ coefficients, $n + 1$ interpolation conditions $P_n(x_j) = f_j$

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}$$

Vandermonde matrix, nonsingular if $x_i \neq x_j$, unique solution

Tedious and often ill-conditioned approach

On a grid $\{x_0, x_1, \ldots, x_k\}$ construct a degree k polynomial basis

$$\{\varphi_i(x)\}_{i=0}^{k}$$

such that $\varphi_i(x_j) = \delta_{ij}$ (the Kronecker delta)

Theorem *If the values* $f_j = f(x_j)$ *are known for the function* $f(x)$*, then the degree* k *polynomial*

$$P(x) = \sum_{j=0}^{k} \varphi_j(x)\, f_j$$

interpolates $f(x)$ *on the grid:*

$$P(x_j) = f_j \quad \text{with} \quad P(x) \approx f(x) \quad \text{for all } x$$

$$\varphi_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$

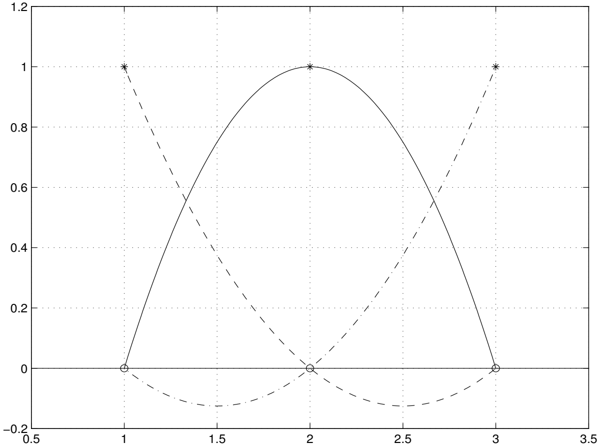


Table of Lagrange basis polynomials

x	φ_0	φ_1	φ_2	f	P_2
x_0	1	0	0	f_0	f_0
x_1	0	1	0	f_1	f_1
x_2	0	0	1	f_2	f_2

$$P_2(x) = f_0\,\varphi_0(x) + f_1\,\varphi_1(x) + f_2\,\varphi_2(x)$$

$$P_2(x) = f_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

4. Numerical integration

Numerical integration is the approximation of definite integrals

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

Problem For many functions *no primitive function is known.* The integral cannot be calculated analytically

Note For *polynomials* an integral $\int_a^b P(x)\mathrm{d}x$ can always be computed analytically

Idea Approximate $f(x) \approx P(x)$ and compute $\int_a^b P(x)\,\mathrm{d}x$

Numerical integration...

Approximate $f \approx P$ and substitute "infinite sum" by a finite sum

The integrand is sampled at a finite number of points

$$I(f) = \sum_{i=1}^{n} w_i f(x_i) + R_n$$

Here $R_n = I(f) - \sum_{i=0}^{n} w_i f(x_i)$ is the *integration error*

Numerical integration method

$$I(f) \approx \sum_{i=0}^{n} w_i f(x_i)$$

Approximate using Lagrange 2nd degree interpolant

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b \sum_{i=0}^{2} f(x_i)\varphi_i(x)\,\mathrm{d}x = \sum_{i=0}^{2} f(x_i) \int_a^b \varphi_i(x)\,\mathrm{d}x$$

Weights $w_i = \int_a^b \varphi_i(x)\,\mathrm{d}x$ can be computed once and for all

Numerical integration method

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{2} w_i f(x_i)$$

Contents

1. Solving nonlinear equations $f(x) = 0$

We can have a single equation

$$x - \cos x = 0$$

but in general we have systems

$$4x^2 - y^2 = 0$$
$$4xy^2 - x = 1$$

Nonlinear equations may have

- no solution
- one solution
- any finite number of solutions
- infinitely many solutions

Nonlinear equations are solved by *iteration*, computing a *sequence* $\{x^{[k]}\}$ of approximations to the root x^*

Definition *The error is defined by* $e^{[k]} = x^{[k]} - x^*$

Definition *The method converges if* $\lim_{k\to\infty} \|e^{[k]}\| = 0$

Definition *The convergence is*

- *linear* if $\|e^{[k+1]}\| \le c \cdot \|e^{[k]}\|$ with $0 < c < 1$
- *quadratic* if $\|e^{[k+1]}\| \le c \cdot \|e^{[k]}\|^p$ with $p = 2$
- *superlinear* if $p > 1$;
- *cubic* if $p = 3$, etc.

2. Fixed points

Definition *x is called a fixed point of the function g if*

$$x = g(x)$$

Definition *A function g is called contractive if*

$$\|g(x) - g(y)\| \leq L[g] \cdot \|x - y\|$$

with $L[g] < 1$ *for all x, y in the domain of g*

A contraction map reduces the distance between points

Fixed Point Theorem

Theorem *Assume that g is Lipschitz continuous on the compact interval I. Further,*

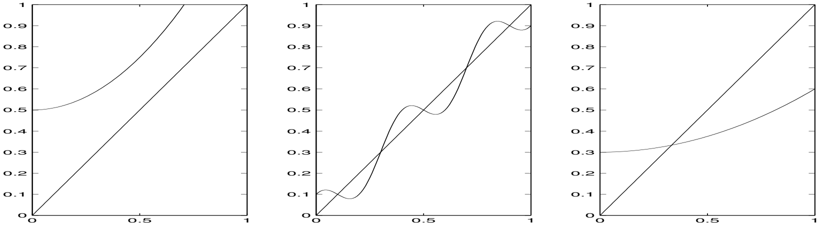
- *If $g : I \to I$ there exists an $x^* \in I$ such that $x^* = g(x^*)$*
- *If in addition $L[g] < 1$ on I, then x^* is unique, and*

$$x_{n+1} = g(x_n)$$

  converges to the fixed point x^* for all $x_0 \in I$

Note Both conditions are absolutely essential!

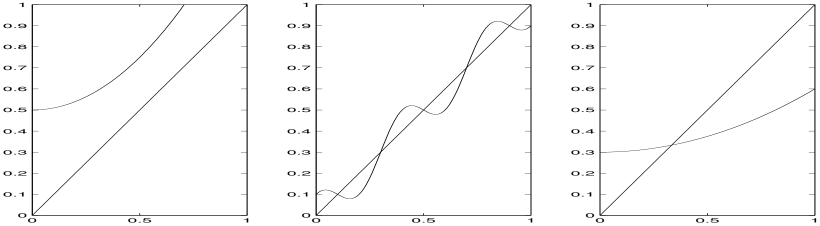
Fixed Point Theorem — *Existence and uniqueness*



Left	No condition satisfied – no x^*
Center	First condition satisfied – maybe multiple x^*
Right	Both conditions satisfied – unique x^*

Fixed Point Theorem

Existence and uniqueness



Left	No condition satisfied – no x^*
Center	First condition satisfied – maybe multiple x^*
Right	Both conditions satisfied – unique x^*
Exercise	Only second condition satisfied – ?

Error bound in fixed point iteration

By the Lipschitz condition

$$\begin{aligned} x^{[k+1]} - x^* &= g(x^{[k]}) - g(x^*) \\ &= g(x^{[k]}) - g(x^{[k+1]}) + g(x^{[k+1]}) - g(x^*) \end{aligned}$$

we have

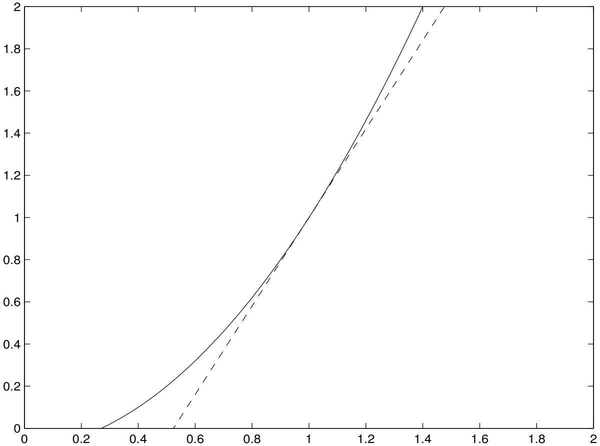
$$\|x^{[k+1]} - x^*\| \leq L[g] \cdot \|x^{[k]} - x^{[k+1]}\| + L[g] \cdot \|x^{[k+1]} - x^*\|$$

Theorem *If $L[g] < 1$, then the error in fixed point iteration is bounded by*

$$\|x^{[k+1]} - x^*\| \leq \frac{L[g]}{1 - L[g]} \|x^{[k]} - x^{[k+1]}\|$$

3. Newton's method

Newton's method solves $f(x) = 0$ using *repeated linearizations*.
Linearize at the point $(x^{[k]}, f(x^{[k]}))$



Newton's method ...

Straight line equation

$$y - f(x^{[k]}) = f'(x^{[k]}) \cdot (x - x^{[k]})$$

Define $x = x^{[k+1]} \;\Rightarrow\; y = 0$, *so that*

$$-f(x^{[k]}) = f'(x^{[k]}) \cdot (x^{[k+1]} - x^{[k]})$$

Solve for $x = x^{[k+1]}$, to get *Newton's method*

$$x^{[k+1]} = x^{[k]} - \frac{f(x^{[k]})}{f'(x^{[k]})}$$

Expand $f(x^{[k+1]})$ in a Taylor series around $x^{[k]}$

$$
\begin{aligned}
f(x^{[k+1]}) &= f(x^{[k]} + (x^{[k+1]} - x^{[k]})) \\
&\approx f(x^{[k]}) + f'(x^{[k]}) \cdot (x^{[k+1]} - x^{[k]}) := 0 \\
&\Rightarrow \\
x^{[k+1]} &= x^{[k]} - \left(f'(x^{[k]})\right)^{-1} f(x^{[k]})
\end{aligned}
$$

Definition $f'(x^{[k]})$ is the *Jacobian matrix* of f, defined by

$$f'(x) = \left\{ \frac{\partial f_i}{\partial x_j} \right\}$$

Write Newton's method as a fixed point iteration $x^{[k+1]} = g(x^{[k]})$ with iteration function

$$g(x) := x - f(x)/f'(x)$$

Note Newton's method *converges fast if* $f'(x^*) \neq 0$, because $g'(x^*) = f(x^*)f''(x^*)/f'(x^*)^2 = 0$

Expand $g(x)$ in a Taylor series around x^*

$$g(x^{[k]}) - g(x^*) \approx g'(x^*)(x^{[k]} - x^*) + \frac{g''(x^*)}{2}(x^{[k]} - x^*)^2$$

$$x^{[k+1]} - x^* \approx \frac{g''(x^*)}{2}(x^{[k]} - x^*)^2$$

Define the error by $\varepsilon^{[k]} = x^{[k]} - x^*$, then

$$\varepsilon^{[k+1]} \sim \left(\varepsilon^{[k]}\right)^2$$

Newton's method is *quadratically convergent*

Fixed point iterations are typically only linearly convergent

$$\varepsilon^{[k+1]} \approx g'(x^*) \cdot \varepsilon^{[k]}$$

A problem with Newton's method is that *starting values need to be close enough* to the root

Convergence order and rate

Definition *The convergence order is p with (asymptotic) error constant C_p, if*

$$0 < \lim_{k\to\infty} \frac{\|\varepsilon^{[k+1]}\|}{\|\varepsilon^{[k]}\|^p} = C_p < \infty$$

Special cases

$p = 1$ Linear convergence
Fixed point iteration $C_p = |f'(x^*)|$

$p = 2$ Quadratic convergence
Newton iteration $C_p = \left|\frac{f''(x^*)}{2f'(x^*)}\right|$

As $y_{n+1} = y_n + hf(y_{n+1})$ we need to solve an equation

$$y = hf(y) + \psi$$

Note All implicit methods lead to an equation of this form

Theorem *Fixed point iterations converge if* $L[hf] < 1$, *restricting the step size to* $h < 1/L[f]$

Note For stiff equations $L[hf] \gg 1$ so fixed point iterations will not converge; it is *necessary to use Newton's method*