

Introduction to Scientific Computing

Raz Kupferman

September 30, 2008

Contents

1	Preliminaries	1
1.1	Review of calculus	1
1.2	Order of convergence	5
1.3	Floating point arithmetic	8
1.4	Stability and condition numbers	9
2	Nonlinear systems of equations	15
2.1	The bisection method	16
2.2	Iterative methods	18
2.3	Newton's method in \mathbb{R}	26
2.4	The secant method in \mathbb{R}	29
2.5	Newton's method in \mathbb{R}^n	31
2.6	A modified Newton's method in \mathbb{R}^n	35
3	Numerical linear algebra	41
3.1	Motivation	41
3.2	Vector and matrix norms	42
3.3	Perturbation theory and condition number	59
3.4	Direct methods for linear systems	63
3.4.1	Matrix factorization	63
3.4.2	Error analysis	70
3.5	Iterative methods	72

3.5.1	Iterative refinement	72
3.5.2	Analysis of iterative methods	74
3.6	Acceleration methods	79
3.6.1	The extrapolation method	79
3.6.2	Chebyshev acceleration	82
3.7	The singular value decomposition (SVD)	90
4	Interpolation	99
4.1	Newton's representation of the interpolating polynomial	99
4.2	Lagrange's representation	101
4.3	Divided differences	101
4.4	Error estimates	102
4.5	Hermite interpolation	102
5	Approximation theory	109
5.1	Weierstrass' approximation theorem	109
5.2	Existence of best approximation	112
5.3	Approximation in inner-product spaces	113
6	Numerical integration	119
7	More questions	121
7.1	Preliminaries	121
7.2	Nonlinear equations	121
7.3	Linear algebra	122
7.4	Interpolation	124
7.5	Approximation theory	124

Chapter 1

Preliminaries

1.1 Review of calculus

Theorem 1.1 (Mean value theorem) If $f \in C[a, b]$ is differentiable in (a, b) , then there exists a point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Notation: We denote by $C^k(\Omega)$ the set of functions that are k times continuously differentiable on the domain Ω .

Theorem 1.2 (Mean value theorem for integrals) Let $f \in C[a, b]$ and let g be integrable on $[a, b]$ and having constant sign. Then, there exists a point $c \in (a, b)$ such that

$$\int_a^b f(x)g(x) dx = f(c) \int_a^b g(x) dx.$$

If, in particular, $g(x) = 1$, then there exists a point where f equals to its average on the interval.

Theorem 1.3 (Taylor's theorem) let $f \in C^n[a, b]$ with $f^{(n+1)}$ existing on $[a, b]$ (but not necessarily differentiable). Let $x_0 \in [a, b]$. Then, for every $x \in [a, b]$ there exists a point $\xi(x)$ between x_0 and x such that

$$f(x) = P_n(x) + R_n(x),$$

where

$$P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

is the ***n*-th Taylor polynomial of f about x_0** , and


$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x - x_0)^{n+1}$$

is the **remainder term**.

Comment: It is often useful to think of x as $x_0 + h$; we know the function and some of its derivatives at a point x_0 and we want to estimate it at another point at a distance h . Then,

$$f(x_0 + h) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} h^k + \frac{f^{(n+1)}(x_0 + \theta(h)h)}{(n+1)!} h^{n+1},$$

where $0 < \theta(h) < 1$. Often, we approximate the function f by its n -th Taylor polynomial, in which case we refer to the remainder as the **truncation error**.

 *Exercise 1.1* (a) Approximate the function $f(x) = \cos x$ at the point $x = 0.01$ by its second and third Taylor polynomials about the point $x_0 = 0$. Estimate the error. (b) Use the third Taylor polynomial to estimate $\int_0^{0.1} \cos x \, dx$. Estimate the error.

Solution 1.1: (a) Since $f \in C^\infty(\mathbb{R})$ then Taylor's theorem applies everywhere on the line. Then,

$$\cos x = \cos x_0 + \frac{\sin x_0}{1!} x - \frac{\cos x_0}{2!} x^2 - \frac{\sin \xi(x)}{3!} x^3,$$

where the last term is the remainder. Substituting $x_0 = 0$ and $x = 0.01$ we find

$$\cos(0.01) = 1 - \frac{(0.01)^2}{2} - \sin(\xi(0.01)) \frac{(0.01)^3}{6}.$$

Since $|\sin x| \leq 1$, we immediately obtain that

$$|\cos(0.01) - 0.99995| \leq \frac{1}{6} \times 10^{-6}.$$

Since the third derivative of $\cos x$ vanishes at $x_0 = 0$, we can in fact derive a sharper error bound as

$$\cos(0.01) = 1 - \frac{(0.01)^2}{2} + \cos(\xi(0.01)) \frac{(0.01)^4}{24},$$

so that

$$|\cos(0.01) - 0.99995| \leq \frac{1}{24} \times 10^{-8}.$$

(b) Since

$$\cos(x) = 1 - \frac{x^2}{2} + \cos(\xi(x)) \frac{x^4}{24},$$

we may integrate both side from 0 to 0.1, and obtain

$$\int_0^{0.1} \cos x \, dx = \int_0^{0.1} \left(1 - \frac{x^2}{2}\right) dx + \frac{1}{24} \int_0^{0.1} x^4 \cos(\xi(x)) \, dx.$$

The polynomial is readily integrated giving $0.01 - \frac{1}{6}(0.1)^3$. The error is easily bounded as follows:

$$\left| \int_0^{0.1} \cos x \, dx - \left[0.01 - \frac{1}{6}(0.1)^3\right] \right| \leq \frac{1}{24} \left| \int_0^{0.1} x^4 \, dx \right| = \frac{10^{-5}}{120}.$$

Theorem 1.4 (Multi-dimensional Taylor theorem) Let f be n times continuously differentiable on a convex domain $\Omega \subseteq \mathbb{R}^k$, and all its $(n+1)$ st partial derivatives exist. Let $\mathbf{x}^0 = (x_1^0, \dots, x_k^0) \in \Omega$. Then for every $\mathbf{x} \in \Omega$

$$f(\mathbf{x}) = P_n(\mathbf{x}) + R_n(\mathbf{x}),$$


where

$$P_n(\mathbf{x}) = \sum_{i=1}^n \frac{1}{i!} \left[(x_1 - x_1^0) \frac{\partial}{\partial x_1} + \dots + (x_k - x_k^0) \frac{\partial}{\partial x_k} \right]^i f(\mathbf{x}^0),$$


is the n -th Taylor polynomial and

$$R_n(\mathbf{x}) = \frac{1}{(n+1)!} \left[(x_1 - x_1^0) \frac{\partial}{\partial x_1} + \dots + (x_k - x_k^0) \frac{\partial}{\partial x_k} \right]^{n+1} f(\mathbf{x}_0 + \theta(\mathbf{x} - \mathbf{x}_0)),$$

where $0 < \theta < 1$.

 **Exercise 1.2** Let k be a positive integer and let $0 < \alpha < 1$. To what class of functions $C^n(\mathbb{R})$ does the function $x^{k+\alpha}$ belong?

Solution 1.2: All its first k derivatives are continuous in \mathbb{R} and its $(k+1)$ -st derivative is singular at $x = 0$. Therefore, $x^{k+\alpha} \in C^k(\mathbb{R})$,

 **Exercise 1.3** For small values of x it is standard practice to approximate the function $\sin x$ by x itself. Estimate the error by using Taylor's theorem. For what range of x will this approximation give results accurate to six decimal places?


Solution 1.3: By Taylor's theorem:

$$\sin x = x - \frac{x^3}{3!} \sin(\theta x),$$

for some $0 < \theta < 1$. Thus,

$$\frac{|\sin x - x|}{|x|} \leq \frac{|x|^2}{6}.$$

The error is guaranteed to have a relative error of less than 10^{-6} if $|x|^2 \leq 6 \times 10^{-6}$.

 **Exercise 1.4** Find the first two terms in the Taylor expansion of $x^{1/5}$ about the point $x = 32$. Approximate the fifth root of 31.999999 using these two terms in the series. How accurate is your answer?

Solution 1.4: The Taylor expansion of $x^{1/5}$ about $x = 32$ is

$$(32 + h)^{1/5} = 32^{1/5} + \frac{32^{-4/5}}{5}h - \frac{4}{2 \cdot 25}(32 + \theta h)^{-9/5}h^2 = 2 + \frac{h}{80} - \frac{2}{25}(32 + \theta h)^{-9/5}h^2,$$

for some $0 < \theta < 1$. In the present case $h = 10^{-6}$, and the resulting error can be bounded by

$$|\text{Err}| \leq \frac{2}{25} \frac{10^{-12}}{512}.$$

 **Exercise 1.5** The **error function** defined by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

gives the probability that a trial value will lie within x units of the mean, assuming that the trials have a standard normal distribution. This integral cannot be evaluated in terms of elementary functions.

- ① Integrate Taylor's series for e^{-t^2} about $t = 0$ to show that

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)k!}$$

(more precisely, use the Taylor expansion for e^{-x}).

- ② Use this series to approximate $\operatorname{erf}(1)$ to within 10^{-7} .

Solution 1.5: The first part is trivial. For the second part, note that if we truncate the Taylor series at n , then the remainder can be bounded by

$$|R_n(x)| \leq \frac{2}{\sqrt{\pi}} \left| \int_0^1 \frac{[-\xi(t)]^{n+1}}{(n+1)!} dt \right| \leq \frac{2}{\sqrt{\pi} (n+1)!},$$

where $\xi(t) \in (0^2, 1^2)$. To ensure an error less than 10^{-7} it is sufficient to truncate the Taylor series at $n = 10$, so that within the required error

$$\operatorname{erf}(1) \approx \frac{2}{\sqrt{\pi}} \sum_{k=0}^{10} \frac{(-1)^k}{(2k+1)k!}.$$

1.2 Order of convergence

Convergence of sequences is a subject you all know from the first calculus course. Many approximation methods are based on the generation of sequences that eventually converge to the desired result. A question of major practical importance is to know how fast does a sequence approach its limit. This section introduces concepts pertinent to the notion of speed of convergence.

Definition 1.1 (Rate of convergence) Let (x_n) be a converging sequence with limit L . Its rate of convergence is said to be (at least) **linear** if there exist a constant $C < 1$ and an integer N , such that for all $n \geq N$,

$$|x_{n+1} - L| \leq C |x_n - L|.$$

The rate of convergence is said to be (at least) **superlinear** if there exists a sequence $\epsilon_n \rightarrow 0$, such that for all $n \geq N$,

$$|x_{n+1} - L| \leq \epsilon_n |x_n - L|.$$

The rate of convergence is said to be of order (at least) α if there exists a constant C (not necessarily smaller than 1) such that

$$|x_{n+1} - L| \leq C |x_n - L|^\alpha.$$

Comment: Can be generalized for sequences in a normed vector space.

Example 1.1 ① The convergence of $(1 + 1/n)^n$ to e satisfies

$$\frac{|x_{n+1} - e|}{|x_n - e|} \rightarrow 1,$$

i.e., the rate of convergence is worse than linear.

- ② The canonical sequence that converges linearly is $x_n = 1/2^n$. Note that linear rate of convergence really means exponentially fast convergence...
- ③ The sequence $2^{-n}/n$ is another example of a linear rate of convergence.
- ④ Consider the sequence (x_n) defined recursively by

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n},$$

with $x_1 = 1$. Then

$$\begin{aligned} 2x_n x_{n+1} &= x_n^2 + 2 \\ 2x_n x_{n+1} - 2\sqrt{2}x_n &= (x_n - \sqrt{2})^2 \\ 2x_n(x_{n+1} - \sqrt{2}) &= (x_n - \sqrt{2})^2, \end{aligned}$$

i.e.,

$$x_{n+1} - \sqrt{2} = \frac{(x_n - \sqrt{2})^2}{2x_n}.$$

Clearly, if the distance of the initial value from $\sqrt{2}$ is less than $1/2$, then the sequence converges. The rate is by definition quadratic. The following table gives the distance of x_n from $\sqrt{2}$ for various n

n	$x_n - \sqrt{2}$
1	-0.41×10^{-1}
2	8.58×10^{-2}
3	2.5×10^{-3}
4	2.12×10^{-6}
5	1.59×10^{-12}

Definition 1.2 Let (x_n) and (y_n) be sequences. We say that $x_n = O(y_n)$ if there exist C, N such that

$$|x_n| \leq C|y_n|$$

for all $n \geq N$. We say that $x_n = o(y_n)$ if


$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = 0.$$

Comments:

- ① Again generalizable for normed linear spaces.
- ② If $x_n = O(y_n)$ then there exists a $C > 0$ such that $\limsup x_n/y_n \leq C$.
- ③ $f(x) = O(g(x))$ as $x \rightarrow x_0$ means that there exists a neighborhood of x_0 in which $|f(x)| \leq C|g(x)|$. Also, $f(x) = o(g(x))$ if for every $\epsilon > 0$ there exists a neighborhood of x_0 where $|f(x)| \leq \epsilon|g(x)|$.

Example 1.2 ① Show that $x_n = O(z_n)$ and $y_n = O(z_n)$ implies that $x_n + y_n = O(z_n)$.


- ② Show that if $\alpha_n \rightarrow 0$, $x_n = O(\alpha_n)$ and $y_n = O(\alpha_n)$, then $x_n y_n = o(\alpha_n)$.

 **Exercise 1.6** Prove that if $x_n = O(\alpha_n)$ then $\alpha_n^{-1} = O(x_n^{-1})$. Prove that the same holds for the o -relation.

Solution 1.6: Let $x_n = O(\alpha_n)$. By definition there exist a $C > 0$ and an $N \in \mathbb{N}$ such that $|x_n| \leq C|\alpha_n|$ for all $n > N$. In particular, for all $n > N$ $\alpha_n = 0$ only if $x_n = 0$ as well. Taking the inverse of this inequality we get

$$\frac{1}{|\alpha_n|} \leq C \frac{1}{|x_n|},$$

where we accept the cases of $1/0 \leq 1/0$ and $1 \leq 1/0$.

 **Exercise 1.7** Let n be fixed. Show that

$$\sum_{k=0}^n x^k = \frac{1}{1-x} + o(x^n)$$

as $x \rightarrow 0$.

Solution 1.7: We have

$$\frac{1}{1-x} - \sum_{k=0}^n x^k = \sum_{k=n+1}^{\infty} x^k = \frac{x^{n+1}}{1-x},$$

and as $x \rightarrow 0$,

$$\lim_{x \rightarrow 0} \frac{x^{n+1}}{(1-x)x^n} = 0.$$

1.3 Floating point arithmetic

A real number in scientific notation has the following representation,

$$\pm(\text{fraction}) \times (\text{base})^{(\text{exponent})}.$$

Any real number can be represented in this way. On a computer, the base is always 2. Due to the finiteness of the number of bits used to represent numbers, the range of fractions and exponents is limited. A **floating point numbers** is a number in scientific notation that fits the format of a computer word, e.g.,

$$-0.1101 \times 2^{-8}.$$

A floating point is called **normalized** if the leading digit of the fraction is 1.

Different computers have different ways of storing floating point numbers. In addition, they may differ in the way they perform arithmetic operations on floating point numbers. They may differ in

- ① The way results are rounded.
- ② The way they deal with numbers very close to zero (underflow).
- ③ The way they deal with numbers that are too big (overflow).
- ④ The way they deal with operations such as $0/0$, $\sqrt{-1}$.

The most common choice of floating point arithmetic is the IEEE standard.

Floating point numbers in the IEEE standard have the following representation,

$$(-1)^s (1 + f) \times 2^{e-1023},$$

where the **sign**, s , takes one bit, the **fraction**, f , takes 52 bits, and the **exponent**, e , takes 11 bits. Because the number is assumed normalized, there is no need to store its leading one. We note the following:

- ① The exponent range is between $2^{-1023} \approx 10^{-308}$ (the underflow threshold), and $2^{1024} \approx 10^{308}$ (the overflow threshold).
- ② Let x be a number within the exponential range and $\text{fl}(x)$ be its approximation by a floating point number. The difference between x and $\text{fl}(x)$ scales with the exponent. The **relative representation error**, however, is bounded by

$$\frac{|x - \text{fl}(x)|}{|x|} \leq 2^{-53} \approx 10^{-16},$$

which is the relative distance between two consecutive floating point numbers. The bound in the relative representation error is known as the **machine- ϵ** .

IEEE arithmetic also handles $\pm\infty$ and NaN with the rules

$$\frac{1}{0} = \infty, \quad \infty + \infty = \infty, \quad \frac{x}{\pm\infty} = 0,$$

and

$$\infty - \infty = \text{NaN}, \quad \frac{\infty}{\infty} = \text{NaN}, \quad \sqrt{-1} = \text{NaN}, \quad x + \text{NaN} = \text{NaN}.$$

Let \odot be any of the four arithmetic operations, and let a, b be two floating point numbers. After the computer performs the operation $a \odot b$, the result has to be stored in a computer word, introducing a **roundoff error**. Then,

$$\frac{a \odot b - \text{fl}(a \odot b)}{a \odot b} = \delta,$$

where $|\delta| \leq \epsilon$. That is

$$\text{fl}(a \odot b) = a \odot b(1 + \delta).$$

1.4 Stability and condition numbers

Condition numbers Let X, Y be normed linear spaces and $f : X \rightarrow Y$. Suppose we want to compute $f(x)$ for some $x \in X$, but we may introduce errors in x and compute instead $f(x + \delta x)$, where $\|\delta x\|$ is “small”. A function is called **well-conditioned** if small errors in its input result in small errors in its output, and it is called **ill-conditioned** otherwise.

Suppose that f is differentiable. Then, under certain assumptions,

$$f(x + \delta x) \approx f(x) + Df(x) \delta x,$$

or,

$$\|f(x + \delta x) - f(x)\| \approx \|Df(x)\| \|\delta x\|.$$

The absolute output error scales like the absolute input error times a multiplier, $\|Df(x)\|$, which we call the **absolute condition number of f at x** . In addition,

$$\underbrace{\frac{\|f(x + \delta x) - f(x)\|}{\|f(x)\|}}_{\text{rel. output err.}} \approx \underbrace{\frac{\|Df(x)\| \|x\|}{\|f(x)\|}}_{\text{rel. cond. number}} \cdot \underbrace{\frac{\|\delta x\|}{\|x\|}}_{\text{rel. input err.}}.$$

Here we call the multiplier of the relative input and output errors the **relative condition number of f at x** . When the condition number is infinite the problem (i.e., the function) is called **ill-posed**. *The condition number is a characteristic of the problem, not of an algorithm.*

Backward stability Suppose next that we want to compute a function $f(x)$, but we use an approximating algorithm which yields instead a result $\text{alg}(x)$. We call $\text{alg}(x)$ a **backward stable algorithm for f** , if there exists a “small” δx such that

$$\text{alg}(x) = f(x + \delta x).$$

I.e., $\text{alg}(x)$ gives the exact solution for a slightly different problem. If the algorithm is backward stable, then

$$\text{alg}(x) \approx f(x) + Df(x) \delta x,$$

i.e.,

$$\|\text{alg}(x) - f(x)\| \approx \|Df(x)\| \|\delta x\|,$$

so that the output error is small provided that the problem is well-conditioned. To conclude, *for an algorithm to give accurate results, it has to be backward stable and the problem has to be well-conditioned.*

Example 1.3 Consider polynomial functions,

$$p(x) = \sum_{i=0}^d a_i x^i, \tag{1.1}$$

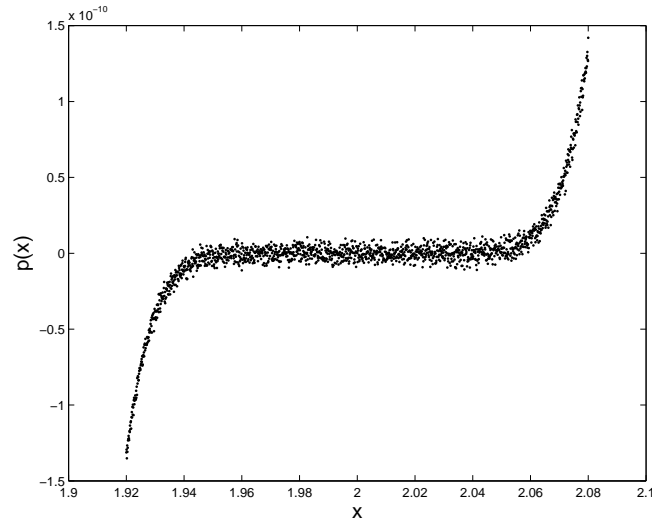


Figure 1.1: Results of calculation of the polynomial (1.1) using Horner's rule.

which are evaluated on the computer with **Horner's rule**:

Algorithm 1.4.1: POLYNOMIAL EVALUATION(x)

```

 $p = a_d$ 
for  $i = d - 1$  downto 0
  do  $p = x * p + a_i$ 
return ( $p$ )

```

The graph in Figure 1.1 shows the result of such a polynomial evaluation for $x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512 = (x - 2)^9$, on the interval $[1.92, 2.08]$.

We see that the behavior of the function is quite unpredictable in the interval $[1.05, 2.05]$, and merits the name of **noise**. In particular, try to imagine finding the root of $p(x)$ using the bisection algorithm.

Let's try to understand the situation in terms of condition numbers and backward stability. First, we rewrite Horner's rule as follows:

Algorithm 1.4.2: POLYNOMIAL EVALUATION(x)

```

 $p_d = a_d$ 
for  $i = d - 1$  downto 0
  do  $p_i = x * p_{i+1} + a_i$ 
return ( $p_0$ )

```

Then, insert a multiplicative term of $(1 + \delta_i)$ each time a floating point operations is done:

Algorithm 1.4.3: POLYNOMIAL EVALUATION(x)

```

 $p_d = a_d$ 
for  $i = d - 1$  downto 0
  do  $p_i = [x * p_{i+1}(1 + \delta_i) + a_i](1 + \delta'_i)$ 
return ( $p_0$ )

```

What do we actually compute? The coefficients a_i are in fact $a_i(1 + \delta'_i)$, and x is really $x(1 + \delta_i)(1 + \delta'_i)$, so that

$$p_0 = \sum_{i=0}^d \left[(1 + \delta'_i) \prod_{j=0}^{i-1} (1 + \delta_j)(1 + \delta'_j) \right] a_i x^i.$$

This expression can be simplified,

$$p_0 = \sum_{i=0}^d (1 + \bar{\delta}_i) a_i x^i,$$

where

$$(1 + \bar{\delta}_i) = (1 + \delta'_i) \prod_{j=0}^{i-1} (1 + \delta_j)(1 + \delta'_j).$$

Now,

$$\begin{aligned} (1 + \bar{\delta}_i) &\leq (1 + \epsilon)^{1+2i} \leq 1 + 2d\epsilon + O(\epsilon^2) \\ (1 - \bar{\delta}_i) &\geq (1 - \epsilon)^{1+2i} \geq 1 - 2d\epsilon + O(\epsilon^2), \end{aligned}$$

from which we deduce that $|\bar{\delta}_i| \leq 2d\epsilon$.

Thus, our algorithm computes exactly a polynomial with slightly different coefficients $\bar{a}_i = (1 + \bar{\delta}_i)a_i$, i.e., it is **backward stable** (the exact solution of a slightly different problem).

With that, we can compute the error in the computed polynomial:

$$\begin{aligned}
 |p(x) - p_0(x)| &= \left| \sum_{i=0}^d (1 + \bar{\delta}_i) a_i x^i - \sum_{i=0}^d a_i x^i \right| \\
 &= \left| \sum_{i=0}^d \bar{\delta}_i a_i x^i \right| \\
 &\leq 2d\epsilon \sum_{i=0}^d |a_i x^i|.
 \end{aligned}$$

This error bound is in fact attainable if the $\bar{\delta}_i$ have signs opposite to that of $a_i x^i$. The relative error (bound) in polynomial evaluation is

$$\frac{|p(x) - p_0(x)|}{|p(x)|} \leq 2d\epsilon \frac{\sum_{i=0}^d |a_i x^i|}{|\sum_{i=0}^d a_i x^i|}.$$

Since $2d\epsilon$ is a measure of the input error, the multiplier $\sum_{i=0}^d |a_i x^i| / |\sum_{i=0}^d a_i x^i|$ is the relative condition number for polynomial evaluation. The relative error bound can be computed directly:

Algorithm 1.4.4: POLYNOMIAL EVALUATION ERROR(x)

```


 $p = a_d$ 
 $\hat{p} = |a_d|$ 
for  $i = d - 1$  downto 0
  do  $\begin{cases} p = x * p + a_i \\ \hat{p} = |x| * \hat{p} + |a_i| \end{cases}$ 
return  $(2d\epsilon \hat{p} / |p|)$ 

```

From the relative error we may infer, for example, a lower bound number of correct digits,

$$n = -\log_{10} \frac{|p(x) - p_0(x)|}{|p(x)|}.$$

In Figure 1.2 we show this lower bound along with the actual number of correct digits. As expected, the relative error grows infinite at the root.

 *Computer exercise 1.1* Generate the two graphs shown in this example.

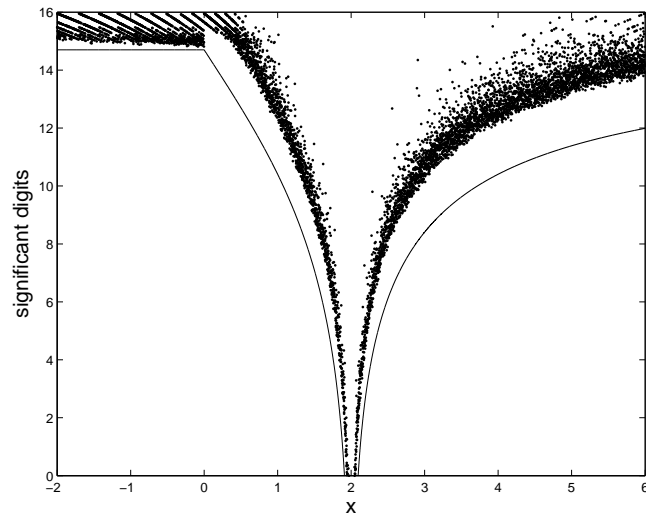


Figure 1.2: Number of significant digits in the calculation of the polynomial (1.1) using Horner's rule. The dots are the actual results and the solid line is the lower bound.

Chapter 2

Nonlinear systems of equations

A general problem in mathematics: X, Y are normed vector spaces, and $f : X \rightarrow Y$. Find $x \in X$ such that $f(x) = 0$.

Example 2.1 ① Find a non-zero $x \in \mathbb{R}$ such that $x = \tan x$ (in wave diffraction); here $f : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $f(x) = x - \tan x$.

② Find $(x, y, z) \in \mathbb{R}^3$ for which

$$\begin{aligned}z^2 - zy + 1 &= 0 \\x^2 - 2 - y^2 - xyz &= 0 \\e^y + 3 - e^x - 2 &= 0.\end{aligned}$$

③ Find a non-zero, twice differentiable function $y(t)$ for which

$$t y''(t) + (1 - t)y'(t) - y = 0.$$

Here $f : C^2(\mathbb{R}) \rightarrow C(\mathbb{R})$ is defined by $y \mapsto t y'' + (1 - t)y' - y$.

Comment:

- ① There are no general theorems of existence/uniqueness for nonlinear systems.
- ② Direct versus iterative methods.
- ③ Iterative algorithms: accuracy, efficiency, robustness, ease of implementation, tolerance, stopping criteria.

2.1 The bisection method

The bisection method applies for root finding in \mathbb{R} , and is based on the following elementary theorem:

Theorem 2.1 (Intermediate value theorem) Let $f \in C[a, b]$ such that (with no loss of generality) $f(a) < f(b)$. For every y such that $f(a) < y < f(b)$ there exists an $x \in (a, b)$ such that $f(x) = y$. In particular, if $f(a)f(b) < 0$, then there exists an $x \in (a, b)$ such that $f(x) = 0$.

The method of proof coincides with the root finding algorithm. Given a, b such that $f(a)f(b) < 0$, we set $c = \frac{1}{2}(a + b)$ to be the mid-point. If $f(a)f(c) < 0$ then we set $b := c$, otherwise we set $a := c$.

Stopping criteria:

- ① Number of iterations M .
- ② $|f(c)| < \epsilon$.
- ③ $|b - a| < \delta$.

Algorithm

Algorithm 2.1.1: BISECTION($a, b, M, \delta, \epsilon$)

```

 $f_a \leftarrow f(a)$ 
 $f_b \leftarrow f(b)$ 
 $\Delta \leftarrow b - a$ 
if  $f_a f_b > 0$  return (error)
for  $k \leftarrow 1$  to  $M$ 
     $\Delta \leftarrow \frac{1}{2}\Delta$ 
     $c \leftarrow a + \Delta$ 
     $f_c \leftarrow f(c)$ 
    do if  $|\Delta| < \delta$  or  $|f_c| < \epsilon$  return ( $c$ )
        if  $f_c f_a < 0$ 
            then  $b \leftarrow c, f_b \leftarrow f_c$ 
        else  $a \leftarrow c, f_a \leftarrow f_c$ 
return (error)

```

Comments:

- ① There is one evaluation of f per iteration (“cost” is usually measured by the number of function evaluations).
- ② There may be more than one root.

Error analysis Given (a, b) the initial guess is $x_0 = \frac{1}{2}(a + b)$. Let $e_n = x_n - r$ be the **error**, where r is the/a root. Clearly,

$$|e_0| \leq \frac{1}{2}|b - a| \equiv E_0.$$

After n steps we have

$$|e_n| \leq \frac{1}{2^{n+1}}|b - a| \equiv E_n.$$

Note that we don’t know what e_n is (if we knew the error, we would know the solution); we only have an **error bound**, E_n . The sequence of error bounds satisfies,

$$E_{n+1} = \frac{1}{2}E_n,$$

so that the bisection method converges linearly.

Discussion: The difference between **error** and **mistake**.

Complexity Consider an application of the bisection method, where the stopping criterion is determined by δ (proximity to the root). The number of steps needed is determined by the condition:

$$\frac{1}{2^{n+1}}|b - a| \leq \delta,$$

i.e.,

$$n + 1 \geq \log_2 \frac{|b - a|}{\delta}.$$

(If for example the initial interval is of length 1 and a tolerance of 10^{-16} is needed, then the number of steps exceeds $n = 50$.)

Advantages and disadvantages

Advantages	Disadvantages
always works	systems in \mathbb{R}^n
easy to implement	slow convergence
requires only continuity	requires initial data a, b

 *Exercise 2.1* Find a positive root of

$$x^2 - 4x \sin x + (2 \sin x)^2 = 0$$

accurate to two significant digits. *Use a hand calculator!*

2.2 Iterative methods

We are looking for roots r of a function $f : X \rightarrow Y$. Iterative methods generate an **approximating sequence** (x_n) by starting with an initial value x_0 , and generating the sequence with an **iteration function** $\Phi : X \rightarrow X$,

$$x_{n+1} = \Phi(x_n).$$

Suppose that each **fixed point** ζ of Φ corresponds to a root of f , and that Φ is continuous in a neighborhood of ζ , then **if** the sequence (x_n) converges, then by the continuity of Φ , it converges to a fixed point of Φ , i.e., to a root of f .

General questions (1) How to choose Φ ? (2) Will the sequence (x_n) converge? How fast will it converge?

Example 2.2 Set $\Phi(x) = x - f(x)$ so that

$$x_{n+1} = x_n - f(x_n).$$

If the sequence converges and f is continuous, then it converges to a root of f .

Example 2.3 (Newton's method in \mathbb{R}) If f is differentiable, Newton's method for root finding consists of the following iterations:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

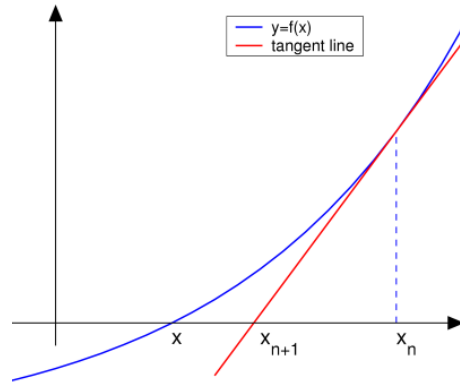


Figure 2.1: Illustration of Newton's iterative method for root finding in \mathbb{R} .


Figure 2.1 illustrates the idea behind this method.

Another way to get to the same iteration function is,

$$0 = f(r) = f(x_n) + (r - x_n)f'(x_n) + \frac{1}{2}(r - x_n)^2 f''(x_n + \theta(r - x_n)),$$

for some $\theta \in (0, 1)$. If we neglect the remainder we obtain

$$r \approx x_n - \frac{f(x_n)}{f'(x_n)}.$$

 **Computer exercise 2.1** Write a Matlab function which gets for input the name of a real-valued function f , an initial value x_0 , a maximum number of iterations M , and a tolerance ϵ . Let your function then perform iterations based on Newton's method for finding roots of f , until either the maximum of number iterations has been exceeded, or the convergence criterion $|f(x)| \leq \epsilon$ has been reached. Experiment your program on the function $f(x) = \tan^{-1} x$, whose only root is $x = 0$. Try to characterize those initial values x_0 for which the iteration method converges.

Example 2.4 (Newton's method in \mathbb{R}^n) Now we're looking for the root $r = (r_1, \dots, r_n)$ of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, which means

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0 \\ f_2(x_1, \dots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \dots, x_n) &= 0 \end{aligned}$$

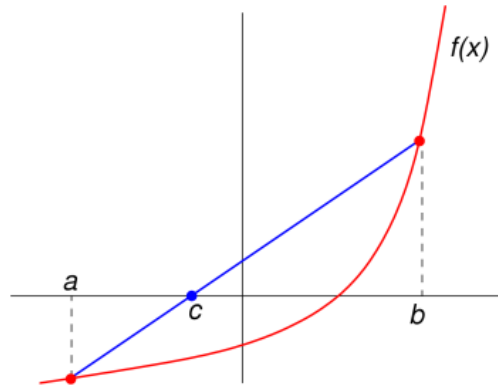


Figure 2.2: Illustration of the secant method for root finding in \mathbb{R} .

Using the same **linear approximation**:

$$0 = f(r) \approx f(x_n) + df(x_n) \cdot (r - x_n),$$

where df is the differential of f , from which we obtain

$$r \approx x_n - [df(x_n)]^{-1} \cdot f(x_n) \equiv x_{n+1}.$$

Example 2.5 (Secant method in \mathbb{R}) Slightly different format. The secant line is

$$y = f(x_n) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_n).$$

We define x_{n+1} to be the intersection with the x -axis:

$$x_{n+1} = x_n - \frac{f(x_n)}{[f(x_n) - f(x_{n-1})]/(x_n - x_{n-1})}$$

(see Figure 2.2). Think of it as an iteration

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = \Phi \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix},$$

which requires at startup the input of both x_0 and x_1 .

Definition 2.1 (Local and global convergence) Let Φ be an iteration function on a complete normed vector space $(X, \|\cdot\|)$, and let ζ be a fixed point of Φ . The

iterative method defined by Φ is said to be **locally convergent** if there exists a neighbourhood $\mathcal{N}(\zeta)$ of ζ , such that for all $x_0 \in \mathcal{N}(\zeta)$, the sequence (x_n) generated by Φ converges to ζ . The method is called **globally convergent** if $\mathcal{N}(\zeta)$ can be extended to the whole space X .

Definition 2.2 (Order of an iteration method) Let Φ be an iteration function on a complete normed vector space $(X, \|\cdot\|)$, and let ζ be a fixed point of Φ . If there exists a neighbourhood $\mathcal{N}(\zeta)$ of ζ , such that

$$\|\Phi(x) - \zeta\| \leq C\|x - \zeta\|^p, \quad \forall x \in \mathcal{N}(\zeta),$$

for some $C > 0$ and $p > 1$, or for $0 < C < 1$ and $p = 1$, then the iteration method is said to be of order (at least) p at the point ζ .

Theorem 2.2 Every iterative method Φ of order at least p at ζ is locally convergent at that point.

Proof: Let $\mathcal{N}(\zeta)$ be the neighbourhood of ζ where the iteration has order at least p . Consider first the case $C < 1$, $p = 1$, and take any open ball

$$B_r(\zeta) = \{x \in X : \|x - \zeta\| < r\} \subseteq \mathcal{N}(\zeta).$$

If $x \in B_r(\zeta)$ then

$$\|\Phi(x) - \zeta\| \leq C\|x - \zeta\| < \|x - \zeta\| < r,$$

hence $\Phi(x) \in B_r(\zeta)$ and the entire sequence lies in $B_r(\zeta)$. By induction,

$$\|x_n - \zeta\| \leq C^n \|x_0 - \zeta\| \rightarrow 0,$$

hence the sequence converges to ζ .

If $p > 1$, take $B_r(\zeta) \subseteq \mathcal{N}(\zeta)$, with r sufficiently small so that $Cr^{p-1} < 1$. If $x \in B_r(\zeta)$ then

$$\|\Phi(x) - \zeta\| \leq C\|x - \zeta\|^{p-1}\|x - \zeta\| < Cr^{p-1}\|x - \zeta\| < \|x - \zeta\|,$$

hence $\Phi(x) \in B_r(\zeta)$ and the entire sequence lies in $B_r(\zeta)$. By induction,

$$\|x_n - \zeta\| \leq (Cr^{p-1})^n \|x_0 - \zeta\| \rightarrow 0,$$

hence the sequence converges to ζ . ■

One dimensional cases Consider the simplest case where $(X, \|\cdot\|) = (\mathbb{R}, |\cdot|)$. If Φ is differentiable in a neighbourhood $\mathcal{N}(\zeta)$ of a fixed point ζ , with $|\Phi'(x)| \leq C < 1$ for all $x \in \mathcal{N}(\zeta)$, then

$$\Phi(x) = \Phi(\zeta) + \Phi'(\zeta + \theta(x - \zeta))(x - \zeta),$$

from which we obtain

$$|\Phi(x) - \zeta| \leq C|x - \zeta|,$$

i.e., the iteration method is at least first order and therefore converges locally. [Show geometrically the cases $\Phi'(x) \in (-1, 0)$ and $\Phi'(x) \in (0, 1)$.]

Example 2.6 Suppose we want to find a root ζ of the function $f \in C^1(\mathbb{R})$ with the iteration

$$x_{n+1} = x_n + \alpha f(x_n),$$

i.e., $\Phi(x) = x + \alpha f(x)$. Suppose furthermore that $f'(\zeta) = M$. Then, for every $\epsilon > 0$ there exists a neighbourhood $\mathcal{N}(\zeta) = (\zeta - \delta, \zeta + \delta)$ such that

$$|f'(x) - M| \leq \epsilon, \quad \forall x \in \mathcal{N}(\zeta).$$

In this neighbourhood,

$$|\Phi'(x)| = |1 + \alpha f'(x)|,$$

which is less than one provided that

$$-2 + |\alpha|\epsilon < \alpha M < -|\alpha|\epsilon.$$

Thus, the iteration method has order at least linear provided that α has sign opposite to that of $f'(\zeta)$, and is sufficiently small in absolute value.

If Φ is sufficiently often differentiable in a neighbourhood $\mathcal{N}(\zeta)$ of a fixed point ζ , with

$$\Phi'(\zeta) = \Phi''(\zeta) = \dots = \Phi^{(p-1)}(\zeta) = 0,$$

then for all $x \in \mathcal{N}(\zeta)$,

$$\Phi(x) = \Phi(\zeta) + \Phi'(\zeta)(x - \zeta) + \dots + \frac{\Phi^{(p)}(\zeta + \theta(x - \zeta))}{p!}(x - \zeta)^p,$$

i.e.,

$$|\Phi(x) - \zeta| = \frac{|\Phi^{(p)}(\zeta + \theta(x - \zeta))|}{p!}|x - \zeta|^p.$$

If $\Phi^{(p)}$ is bounded in some neighbourhood of ζ , say $|\Phi^{(p)}(x)| \leq M$, then

$$|\Phi(x) - \zeta| \leq \frac{M}{p!} |x - \zeta|^p,$$

so that the iteration method is at least of order p , and therefore locally convergent. Moreover,

$$\lim_{n \rightarrow \infty} \frac{|\Phi(x) - \zeta|}{|x - \zeta|^p} = \frac{|\Phi^{(p)}(\zeta)|}{p!},$$

i.e., the method is precisely of order p .

Example 2.7 Consider Newton's method in \mathbb{R} ,

$$\Phi(x) = x - \frac{f(x)}{f'(x)},$$


and assume that f has a simple zero at ζ , i.e., $f'(\zeta) \neq 0$. Then,

$$\Phi'(\zeta) = \left. \frac{f(x)f''(x)}{[f'(x)]^2} \right|_{x=\zeta} = 0,$$

and

$$\Phi''(\zeta) = \frac{f''(\zeta)}{f'(\zeta)},$$

the latter being in general different than zero. Thus, Newton's method is of second order and therefore locally convergent.


 *Exercise 2.2* The two following sequences constitute iterative procedures to approximate the number $\sqrt{2}$:

$$x_{n+1} = x_n - \frac{1}{2}(x_n^2 - 2), \quad x_0 = 2,$$

and


$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}, \quad x_0 = 2.$$

- ① Calculate the first six elements of both sequences.
- ② Calculate (numerically) the error, $e_n = x_n - \sqrt{2}$, and try to estimate the order of convergence.
- ③ Estimate the order of convergence by Taylor expansion.

 **Exercise 2.3** Let a sequence x_n be defined inductively by


$$x_{n+1} = F(x_n).$$

Suppose that $x_n \rightarrow x$ as $n \rightarrow \infty$ and that $F'(x) = 0$. Show that $x_{n+2} - x_{n+1} = o(x_{n+1} - x_n)$. (Hint: assume that F is continuously differentiable and use the mean value theorem.)

 **Exercise 2.4** Analyze the following iterative method,

$$x_{n+1} = x_n - \frac{f^2(x_n)}{f(x_n + f(x_n)) - f(x_n)},$$

designed for the calculation of the roots of $f(x)$ (this method is known as Steffensen's method). Prove that this method converges quadratically (order 2) under certain assumptions.

 **Exercise 2.5** Kepler's equation in astronomy is $x = y - \epsilon \sin y$, with $0 < \epsilon < 1$. Show that for every $x \in [0, \pi]$, there is a y satisfying this equation. (Hint: Interpret this as a fixed-point problem.)

Contractive mapping theorems General theorems on the convergence of iterative methods are based on a fundamental property of mapping: contraction.

Theorem 2.3 (Contractive mapping theorem) Let K be a closed set in a complete normed space $(X, \|\cdot\|)$, and let Φ be a continuous mapping on X such that (i) $\Phi(K) \subseteq K$, and there exists a $C < 1$ such that for every $x, y \in K$,

$$\|\Phi(x) - \Phi(y)\| \leq C\|x - y\|.$$

Then,

- ① The mapping Φ has a unique fixed point ζ in K .
- ② For every $x_0 \in K$, the sequence (x_n) generated by Φ converges to ζ .

Proof: Since $\Phi(K) \subseteq K$, $x_0 \in K$ implies that $x_n \in K$ for all n . From the contractive property of Φ we have

$$\|x_n - x_{n-1}\| \leq C\|x_{n-1} - x_{n-2}\| \leq C^{n-1}\|x_1 - x_0\|.$$

Now, write x_n as

$$x_n = x_0 + \sum_{j=1}^n (x_j - x_{j-1}).$$

For any $m < n$,

$$\begin{aligned} \|x_n - x_m\| &\leq \sum_{j=m+1}^n \|x_j - x_{j-1}\| \leq \sum_{j=m+1}^n C^{j-1} \|x_1 - x_0\| \\ &\leq \sum_{j=m+1}^{\infty} C^{j-1} \|x_1 - x_0\| \leq \frac{C^m}{1-C} \|x_1 - x_0\|, \end{aligned}$$

which converges to zero as $m, n \rightarrow \infty$. Thus (x_n) is a Cauchy sequence, and since X is complete it converges to a limit ζ , which must reside in K since K is closed. The limit point must on the other hand be a fixed point of Φ .

Uniqueness is immediate for if ζ, ξ are distinct fixed point in K , then

$$\|\zeta - \xi\| = \|\Phi(\zeta) - \Phi(\xi)\| \leq C\|\zeta - \xi\| < \|\zeta - \xi\|,$$

which is a contradiction. ■


Example 2.8 Consider for example the mapping

$$x_{n+1} = 3 - \frac{1}{2}|x_n|$$

on \mathbb{R} . Then,

$$|x_{n+1} - x_n| = \frac{1}{2}||x_n| - |x_{n-1}|| \leq \frac{1}{2}|x_n - x_{n-1}|.$$

Hence, for every x_0 the sequence (x_n) converges to the unique fixed point $\zeta = 2$.

 *Exercise 2.6* Let p be a positive number. What is the value of the following expression:


$$x = \sqrt{p + \sqrt{p + \sqrt{p + \cdots}}}.$$

By that, I mean the sequence $x_0 = p$, $x_{k+1} = \sqrt{p + x_k}$. (Interpret this as a fixed-point problem.)

 **Exercise 2.7** Show that the function


$$F(x) = 2 + x - \tan^{-1} x$$

satisfies $|F'(x)| < 1$. Show then that $F(x)$ doesn't have fixed points. Why doesn't this contradict the contractive mapping theorem?

 **Exercise 2.8** Bailey's iteration for calculating \sqrt{a} is obtained by the iterative scheme:

$$x_{n+1} = g(x_n) \quad g(x) = \frac{x(x^2 + 3a)}{3x^2 + a}.$$

Show that this iteration is of order at least three.

 **Exercise 2.9** (Here is an exercise which tests whether you *really* understand what root finding is about.) One wants to solve the equation $x + \ln x = 0$, whose root is $x \sim 0.5$, using one or more of the following iterative methods:

$$(i) \quad x_{k+1} = -\ln x_k \quad (ii) \quad x_{k+1} = e^{-x_k} \quad (iii) \quad x_{k+1} = \frac{x_k + e^{-x_k}}{2}.$$

- ① Which of the three methods *can* be used?
- ② Which method *should* be used?
- ③ Give an even better iterative formula; explain.

2.3 Newton's method in \mathbb{R}

We have already seen that Newton's method is of order two, provided that $f'(\zeta) \neq 0$, therefore locally convergent. Let's first formulate the algorithm

Algorithm 2.3.1: NEWTON(x_0, M, ϵ)

```

y ← f(x0)
if |y| < ε return (x0)
for k ← 1 to M
  do {
    x ← x0 - f(x0)/f'(x0)
    y ← f(x0)
    if |y| < ε return (x)
    x0 ← x
  }
return (error)

```

Note that in every iteration we need to evaluate both f and f' .

Newton's method does not, in general, converge globally [show graphically the example of $f(x) = x - \tan^{-1}x$.] The following theorem characterizes a class of functions f for which Newton's method converges globally:

Theorem 2.4 Let $f \in C^2(\mathbb{R})$ be monotonic, convex and assume it has a root. Then the root is unique and Newton's method converges globally.

Proof: The uniqueness of the root is obvious. It is given that $f''(x) > 0$, and assume, without loss of generality, that $f'(x) > 0$. If $e_n = x_n - \zeta$, then

$$0 = f(\zeta) = f(x_n) - e_n f'(x_n) + \frac{1}{2} e_n^2 f''(x_n - \theta e_n),$$

hence

$$e_{n+1} = e_n - \frac{f(x_n)}{f'(x_n)} = \frac{1}{2} \frac{f''(x_n - \theta e_n)}{f'(x_n)} e_n^2 > 0.$$

Thus, the iterates starting from e_1 are always to the right of the root. On the other hand, since

$$x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} < 0,$$

it follows that (x_n) is a monotonically decreasing sequence bounded below by ζ hence it converges. The limit must coincide with ζ by continuity. ■

Newton's method when f has a double root We now examine the local convergence of Newton's method when ζ is a double root, i.e., $f(\zeta) = f'(\zeta) = 0$. We assume that $f''(\zeta) \neq 0$, so that there exists a neighbourhood of ζ where $f'(x) \neq 0$. As above, we start with the relation

$$e_{n+1} = e_n - \frac{f(x_n)}{f'(x_n)}.$$

Using Taylor's expansion we have

$$0 = f(\zeta) = f(x_n) - e_n f'(x_n) + \frac{1}{2} e_n^2 f''(x_n - \theta e_n),$$

from which we extract $f(x_n)$ and substitute above to get

$$e_{n+1} = \frac{1}{2} e_n^2 \frac{f''(x_n - \theta e_n)}{f'(x_n)}.$$

The problem is that the denominator is not bounded away from zero. We use Taylor's expansion for f' :

$$0 = f'(\zeta) = f'(x_n) - e_n f''(x_n - \theta_1 e_n),$$

from which we extract $f'(x_n)$ and finally obtain

$$e_{n+1} = \frac{1}{2} e_n \frac{f''(x_n - \theta e_n)}{f''(x_n - \theta_1 e_n)}.$$

Thus, Newton's method is locally convergent, but the order of convergence reduces to first order. In particular, if the sequence (x_n) converges then

$$\lim_{n \rightarrow \infty} \frac{e_{n+1}}{e_n} = \frac{1}{2}.$$

The same result can be derived from an examination of the iteration function Φ . The method is at least second order if $\Phi'(\zeta) = 0$ and at least first order if $|\Phi'(\zeta)| < 1$. Now,

$$\Phi'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

In the limit $x \rightarrow \zeta$ we have, by our assumptions, $f(x) \sim a(x - \zeta)^2$, to that


$$\lim_{x \rightarrow \zeta} \Phi'(x) = \frac{1}{2}.$$


How can second order convergence be restored? The iteration method has to be modified into

$$x_{n+1} = x_n - 2 \frac{f(x_n)}{f'(x_n)}.$$


It is easily verified then that


$$\lim_{x \rightarrow \zeta} \Phi'(x) = 0.$$

 **Exercise 2.10** Your dog chewed your calculator and damaged the division key! To compute reciprocals (i.e., one-over a given number R) without division, we can solve $x = 1/R$ by finding a root of a certain function f with Newton's method. Design such an algorithm (that, of course, does not rely on division).

 **Exercise 2.11** Prove that if r is a root of multiplicity k (i.e., $f(r) = f'(r) = \dots = f^{(k-1)}(r) = 0$ but $f^{(k)}(r) \neq 0$), then the quadratic convergence of Newton's method will be restored by making the following modification to the method:

$$x_{n+1} = x_n - k \frac{f(x_n)}{f'(x_n)}.$$

 **Exercise 2.12** Similarly to Newton's method (in one variable), derive a method for solving $f(x)$ given the functions $f(x)$, $f'(x)$ and $f''(x)$. What is the rate of convergence?

 **Exercise 2.13** What special properties must a function f have if Newton's method applied to f converges cubically?

2.4 The secant method in \mathbb{R}

Error analysis The secant method is

$$x_{n+1} = x_n - (x_n - x_{n-1}) \frac{f(x_n)}{f(x_n) - f(x_{n-1})}.$$

If we want to analyze this method within our formalism of iterative methods we have to consider an iteration of a couple of numbers. To obtain the local convergence properties of the secant method we can resort to an explicit calculation.

Subtracting ζ from both side we get

$$\begin{aligned} e_{n+1} &= e_n - (e_n - e_{n-1}) \frac{f(x_n)}{f(x_n) - f(x_{n-1})} \\ &= -\frac{f(x_{n-1})}{f(x_n) - f(x_{n-1})} e_n + \frac{f(x_n)}{f(x_n) - f(x_{n-1})} e_{n-1} \\ &= \frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{f(x_n) - f(x_{n-1})} e_{n-1} e_n \\ &= \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{x_n - x_{n-1}} e_{n-1} e_n \end{aligned}$$

The first term can be written as

$$\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{1}{f'(x_{n-1} + \theta(x_n - x_{n-1}))}.$$

The second term can be written as

$$\frac{g(x_n) - g(x_{n-1})}{x_n - x_{n-1}} = g'(x_{n-1} + \theta_1(x_n - x_{n-1})),$$

where

$$g(x) = \frac{f(x)}{x - \zeta} = \frac{f(x) - f(\zeta)}{x - \zeta}.$$

Here comes a useful trick. We can write

$$f(x) - f(\zeta) = \int_{\zeta}^x f'(s) ds = (x - \zeta) \int_0^1 f'(s\zeta + (1-s)x) ds,$$

so that

$$g(x) = \int_0^1 f'(s\zeta + (1-s)x) ds.$$

We can then differentiate under the integral sign so get

$$g'(x) = \int_0^1 (1-s) f''(s\zeta + (1-s)x) ds,$$

and by the integral mean value theorem, there exists a point ξ between x and ζ such that

$$g'(x) = f''(\xi) \int_0^1 (1-s) ds = \frac{1}{2} f''(\xi).$$

Combining together, there are two intermediate points so that

$$e_{n+1} = \frac{f''(\xi)}{2 f'(\xi_1)} e_n e_{n-1},$$


and sufficiently close to the root,

$$e_{n+1} \approx C e_{n-1} e_n.$$

What is then the order of convergence? Guess the ansatz $e_n = a e_{n-1}^\alpha$, then

$$a e_n^\alpha = C (a^{-1} e_n)^{1/\alpha} e_n,$$

which implies that $\alpha^2 = \alpha + 1$, or $\alpha = \frac{1}{2}(1 + \sqrt{5}) \approx 1.62$ (the golden ratio). Thus, the order of convergence is super-linear but less than second order. On the other hand, each iteration requires only one function evaluation (compared to two for Newton)!

 **Exercise 2.14** The method of “false position” for solving $f(x) = 0$ starts with two initial values, x_0 and x_1 , chosen such that $f(x_0)$ and $f(x_1)$ have opposite signs. The next guess is then calculated by

$$x_2 = \frac{x_1 f(x_0) - x_0 f(x_1)}{f(x_0) - f(x_1)}.$$

Interpret this method geometrically in terms of the graph of $f(x)$.

2.5 Newton's method in \mathbb{R}^n

In the first part of this section we establish the local convergence property of the multi-dimensional Newton method.

Definition 2.3 (Differentiability) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. f is said to be differentiable at the point $x \in \mathbb{R}^n$, if there exists a linear operator on \mathbb{R}^n (i.e., an $n \times n$ matrix) A , such that

$$\lim_{y \rightarrow x} \frac{\|f(y) - f(x) - A(y - x)\|}{\|y - x\|} = 0.$$

We call the matrix A the differential of f at the point x and denote it by $df(x)$.

Comment: While the choice of norm of \mathbb{R}^n is not unique, convergence in one norm implies convergence in all norm for finite dimensional spaces. We will typically use here the Euclidean norm.

Definition 2.4 (Norm of an operator) Let $(X, \|\cdot\|)$ be a normed linear space and $\mathcal{B}(X)$ be the space of continuous linear transformations on X . Then, $\mathcal{B}(X)$ is a linear space which can be endowed with a norm,

$$\|A\| = \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}, \quad A \in \mathcal{B}(X).$$

In particular, every vector norm induces a subordinate matrix norm.

Comments:

- ① By definition, for all $x \in X$ and $A \in \mathcal{B}(X)$,

$$\|Ax\| \leq \|A\| \|x\|.$$

② We will return to subordinate matrix norms in depth in the next chapter.

Lemma 2.1 Suppose that $df(x)$ exists in a convex set K , and there exists a constant $C > 0$, such that

$$\|df(x) - df(y)\| \leq C\|x - y\| \quad \forall x, y \in K,$$

then

$$\|f(x) - f(y) - df(y)(x - y)\| \leq \frac{C}{2}\|x - y\|^2 \quad \forall x, y \in K.$$

Proof: Consider the function

$$\varphi(t) = f(y + t(x - y))$$

defined on $t \in [0, 1]$. Since K is convex then $\varphi(t)$ is differentiable on the unit segment, with

$$\varphi'(t) = df(y + t(x - y)) \cdot (x - y),$$

and

$$\|\varphi'(t) - \varphi'(0)\| \leq \|df(y + t(x - y)) - df(y)\|\|x - y\| \leq Ct\|x - y\|^2. \quad (2.1)$$

On the other hand,

$$\begin{aligned} \Delta &\equiv f(x) - f(y) - df(y)(x - y) = \varphi(1) - \varphi(0) - \varphi'(0) \\ &= \int_0^1 [\varphi'(t) - \varphi'(0)] dt, \end{aligned}$$

from which follows, upon substitution of (2.1),

$$\|\Delta\| \leq \int_0^1 \|\varphi'(t) - \varphi'(0)\| dt \leq \frac{C}{2}\|x - y\|^2.$$

■

With this lemma, we are in measure to prove the local quadratic convergence of Newton's method.

Theorem 2.5 Let $K \subseteq \mathbb{R}^n$ be an open set, and K_0 be a convex set, $\overline{K_0} \subset K$. Suppose that $f : K \rightarrow \mathbb{R}^n$ is differentiable in K_0 and continuous in K . Let $x_0 \in K_0$, and assume the existence of positive constants α, β, γ so that

- ① $\|df(x) - df(y)\| \leq \gamma\|x - y\|$ in K_0 .
- ② $[df(x)]^{-1}$ exists and $\|[df(x)]^{-1}\| \leq \beta$ in K_0 .
- ③ $\|[df(x_0)]^{-1}f(x_0)\| \leq \alpha$,

with

$$h \equiv \frac{\alpha\beta\gamma}{2} < 1,$$

and

$$B_r(x_0) \subseteq K_0,$$

where

$$r = \frac{\alpha}{1-h}.$$

Then,

- ① The Newton sequence (x_n) defined by

$$x_{n+1} = x_n - [df(x_n)]^{-1}f(x_n)$$

is well defined and contained in $B_r(x_0)$.

- ② The sequence (x_n) converges in the closure of $B_r(x_0)$ to a root ζ of f .
- ③ For all n ,

$$\|x_n - \zeta\| \leq \alpha \frac{h^{2^n-1}}{1-h^{2^n}},$$

i.e., the convergence is at least quadratic.

Proof: We first show that the sequence remains in $B_r(x_0)$. The third assumption implies

$$\|x_1 - x_0\| = \|[df(x_0)]^{-1}f(x_0)\| \leq \alpha < r,$$

i.e., $x_1 \in B_r(x_0)$. Suppose that the sequence remains in $B_r(x_0)$ up to the k -th element. Then x_{k+1} is well defined (by the second assumption), and

$$\begin{aligned}\|x_{k+1} - x_k\| &= \|[df(x_k)]^{-1}f(x_k)\| \leq \beta\|f(x_k)\| \\ &= \beta\|f(x_k) - f(x_{k-1}) - df(x_{k-1})(x_k - x_{k-1})\|,\end{aligned}$$

where we have used the fact that $f(x_{k-1}) + df(x_{k-1})(x_k - x_{k-1}) = 0$. Now, by the first assumption and the previous lemma,

$$\|x_{k+1} - x_k\| \leq \frac{\beta\gamma}{2}\|x_k - x_{k-1}\|^2.$$

From this, we can show inductively that

$$\|x_{k+1} - x_k\| \leq \alpha h^{2^{k-1}}, \quad (2.2)$$

since it is true for $k = 0$ and if it is true up to k , then

$$\|x_{k+1} - x_k\| \leq \frac{\beta\gamma}{2}\alpha^2(h^{2^{k-1}-1})^2 = \alpha\frac{\alpha\beta\gamma}{2}h^{2^k-2} < \alpha h^{2^k-1}.$$

From this we have

$$\begin{aligned}\|x_{k+1} - x_0\| &\leq \|x_{k+1} - x_k\| + \cdots + \|x_1 - x_0\| \\ &\leq \alpha(1 + h + h^3 + \cdots + h^{2^k-1}) < \frac{\alpha}{1-h} = r,\end{aligned}$$

i.e., $x_{k+1} \in B_r(x_0)$, hence the entire sequence.

Inequality (2.2) implies also that (x_n) is a Cauchy sequence, for

$$\begin{aligned}\|x_{n+1} - x_m\| &\leq \|x_{n+1} - x_n\| + \cdots + \|x_{m+1} - x_m\| \\ &\leq \alpha(h^{2^m-1} + \cdots + h^{2^n-1}) \\ &< \alpha h^{2^m-1}(1 + h^{2^m} + (h^{2^m})^3 + \cdots) < \alpha \frac{h^{2^m-1}}{1-h^{2^m}}.\end{aligned}$$


which tends to zero as $m, n \rightarrow \infty$. Thus the sequence (x_n) converges to a limit $\zeta \in \overline{B_r(x_0)}$. As a side results we obtain that

$$\|\zeta - x_m\| \leq \alpha \frac{h^{2^m-1}}{1-h^{2^m}}.$$

It remains to show that ζ is indeed a root of f . The first condition implies the continuity of the differential of f , so that taking limits:

$$\zeta = \zeta - [df(\zeta)]^{-1}f(\zeta),$$


and since by assumption, df is invertible, it follows that $f(\zeta) = 0$. ■

 *Computer exercise 2.2* Use Newton's method to solve the system of equations

$$\begin{aligned} xy^2 + x^2y + x^4 &= 3 \\ x^3y^5 - 2x^5y - x^2 &= -2. \end{aligned}$$

Start with various initial values and try to characterize the “basin of convergence” (the set of initial conditions for which the iterations converge).

Now, Matlab has a built-in root finder `fsolve()`. Try to solve the same problem using this functions, and evaluate whether it performs better or worse than your own program in terms of both speed and robustness.

 *Exercise 2.15* Go to the following site and enjoy the nice pictures:

<http://aleph0.clarku.edu/~djoyce/newton/newton.html>

(Read the explanations, of course....)

2.6 A modified Newton's method in \mathbb{R}^n

Newton's method is of the form

$$x_{k+1} = x_k - d_k,$$

where

$$d_k = [df(x_k)]^{-1} f(x_k).$$

When this method converges, it does so quadratically, however, the convergence is only guaranteed locally. A modification to Newton's method, which converges under much wider conditions is of the following form:

$$x_{k+1} = x_k - \lambda_k d_k,$$

where the coefficients λ_k are chosen such that the sequence $(h(x_k))$, where

$$h(x) = f^T(x)f(x) = \|f(x)\|^2,$$

is strictly monotonically decreasing (here $\|\cdot\|$ stands for the Euclidean norm in \mathbb{R}^n). Clearly, $h(x_k) \geq 0$, and if the sequence (x_k) converges to a point ζ , where $h(\zeta) = 0$ (i.e., a global minimum of $h(x)$), then $f(\zeta) = 0$. *The modified Newton method aims to minimize $h(x)$ rather than finding a root of $f(x)$.*

Definition 2.5 Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\|\cdot\|$ be the Euclidean norm in \mathbb{R}^n . For $0 < \gamma \leq 1$ we define

$$D(\gamma, x) = \left\{ s \in \mathbb{R}^n : \|s\| = 1, \frac{Dh(x)}{\|Dh(x)\|} \cdot s \geq \gamma \right\},$$

which is the set of all directions s which form with the gradient of h a not-too-acute angle.

Lemma 2.2 Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ be in C^1 in a neighbourhood $V(\zeta)$ of a point ζ . Suppose that $Dh(\zeta) \neq 0$ and let $0 < \gamma \leq 1$. Then there exist a neighbourhood $U(\zeta) \subseteq V(\zeta)$ and a number $\lambda > 0$, such that

$$h(x - \mu s) \leq h(x) - \frac{\mu\gamma}{4} \|Dh(\zeta)\|$$

for all $x \in U(\zeta)$, $s \in D(\gamma, x)$, and $0 \leq \mu \leq \lambda$.

Proof: Consider first the set

$$U_1(\zeta) = \left\{ x \in V(\zeta) : \|Dh(x) - Dh(\zeta)\| \leq \frac{\gamma}{4} \|Dh(\zeta)\| \right\},$$

which by the continuity of Dh and the non-vanishing of $Dh(\zeta)$ is a non-empty set and a neighbourhood of ζ . Let also

$$U_2(\zeta) = \left\{ x \in V(\zeta) : D(\gamma, x) \subseteq D(\frac{\gamma}{2}, \zeta) \right\},$$

which again is a non-empty neighbourhood of ζ . Indeed, it consists of all $x \in V(\zeta)$ for which

$$\left\{ s : \frac{Dh(x)}{\|Dh(x)\|} \cdot s \geq \gamma \right\} \subseteq \left\{ s : \frac{Dh(\zeta)}{\|Dh(\zeta)\|} \cdot s \geq \frac{\gamma}{2} \right\}.$$

Choose now a λ such that

$$\overline{B_{2\lambda}(\zeta)} \subseteq U_1(\zeta) \cap U_2(\zeta),$$

and finally set

$$U(\zeta) = \overline{B_\lambda(\zeta)}.$$

Now, for all $x \in U(\zeta)$, $s \in D(\gamma, x)$ and $0 \leq \mu \leq \lambda$, there exists a $\theta \in (0, 1)$ such that

$$\begin{aligned} h(x) - h(x - \mu s) &= \mu Dh(x - \theta\mu s) \cdot s \\ &= \mu \{(Dh(x - \theta\mu s) - Dh(\zeta)) \cdot s + Dh(\zeta) \cdot s\}. \end{aligned}$$

Now $x \in \overline{B_\lambda(\zeta)}$ and $\mu \leq \lambda$ implies that

$$x - \mu s, x - \theta \mu s \in \overline{B_{2\lambda}(\zeta)} \subseteq U_1(\zeta) \cap U_2(\zeta),$$

and by the membership in $U_1(\zeta)$,

$$(Dh(x - \theta \mu s) - Dh(\zeta)) \cdot s \geq -\|Dh(x - \theta \mu s) - Dh(\zeta)\| \geq -\frac{\gamma}{4}\|Dh(\zeta)\|,$$

whereas by the membership in $U_2(\zeta)$, $s \in D(\frac{\gamma}{2}, \zeta)$, hence

$$Dh(\zeta) \cdot s \geq \frac{\gamma}{2}\|Dh(\zeta)\|,$$

and combining the two,

$$h(x) - h(x - \mu s) \geq -\mu \frac{\gamma}{4}\|Dh(\zeta)\| + \mu \frac{\gamma}{2}\|Dh(\zeta)\| = \frac{\mu \gamma}{4}\|Dh(\zeta)\|.$$

This completes the proof. ■

Minimization algorithm Next, we describe an algorithm for the minimization of a function $h(x)$ via the construction of a sequence (x_k) .

- ① Choose sequences (γ_k) , (σ_k) , satisfying the constraints

$$\sup_k \gamma_k \leq 1, \quad \gamma \equiv \inf_k \gamma_k > 0, \quad \sigma \equiv \inf_k \sigma_k > 0,$$

as well as a starting point x_0 .

- ② For every k , choose a **search direction** $s_k \in D(\gamma_k, x_k)$ and set

$$x_{k+1} = x_k - \lambda_k s_k,$$

where $\lambda_k \in [0, \sigma_k \|Dh(x_k)\|]$ is chosen such to minimize $h(x_k - \lambda_k s_k)$.

Theorem 2.6 Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ be such that

- ① The set $K = \{x : h(x) \leq h(x_0)\}$ is compact.
 ② $h \in C^1$ in an open set containing K .

Then,

- ① The sequence (x_k) is in K and has at least one accumulation point ζ .
- ② Each accumulation point ζ is a critical point of h , $Dh(\zeta) = 0$.

Proof: Since, by construction, the sequence $(h(x_k))$ is monotonically decreasing then the $\{h(x_k)\}$ are all in K . Since K is compact, then the set $\{x_k\}$ has at least one accumulation point ζ .

Without loss of generality we can assume that $x_k \rightarrow \zeta$, otherwise we consider a converging sub-sequence. Assume that ζ is not a critical point, $Dh(\zeta) \neq 0$. From the previous lemma, we know that there exist a neighbourhood $U(\zeta)$ and a number $\lambda > 0$, such that

$$h(x - \mu s) \leq h(x) - \frac{\mu\gamma}{4} \|Dh(\zeta)\| \quad (2.3)$$

for all $x \in U(\zeta)$, $s \in D(\gamma, x)$, and $0 \leq \mu \leq \lambda$. Since $x_k \rightarrow \zeta$ and because Dh is continuous, it follows that for sufficiently large k ,

- ① $x_k \in U(\zeta)$.
- ② $\|Dh(x_k)\| \geq \frac{1}{2} \|Dh(\zeta)\|$.

Set now

$$\Lambda = \min\left(\lambda, \frac{1}{2}\sigma\|Dh(\zeta)\|\right), \quad \epsilon = \Lambda \frac{\gamma}{4} \|Dh(\zeta)\| > 0.$$

Since $\sigma_k \geq \sigma$ it follows that for sufficiently large k ,

$$[0, \Lambda] \subseteq [0, \sigma_k \frac{1}{2} \|Dh(\zeta)\|] \subseteq [0, \sigma_k \|Dh(x_k)\|],$$

the latter being the set containing λ_k in the minimization algorithm. Thus, by the definition of x_{k+1} ,

$$h(x_{k+1}) \leq h(x_k - \mu s_k),$$

for every $0 \leq \mu \leq \Lambda$. Since $\Lambda \leq \lambda$, $x_k \in U(\zeta)$, and $s_k \in D(\gamma_k, x_k) \subseteq D(\gamma, x_k)$, it follows from (2.3) that

$$h(x_{k+1}) \leq h(x_k) - \frac{\Lambda\gamma}{4} \|Dh(\zeta)\| = h(x_k) - \epsilon.$$

This means that $h(x_k) \rightarrow -\infty$ which contradicts its lower-boundedness by $h(\zeta)$. ■

The modified Newton algorithm The modified Newton algorithm works as follows: at each step

$$x_{k+1} = x_k - \lambda_k d_k, \quad d_k = [df(x_k)]^{-1} f(x_k),$$

where $\lambda_k \in (0, 1]$ is chosen such to minimize $h(x_k - \lambda_k d_k)$, where $h(x) = f^T(x)f(x)$.

Theorem 2.7 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$ satisfy the following properties:

- ① The set $K = \{x : h(x) \leq h(x_0)\}$ with $h(x) = f^T(x)f(x)$ is compact.
- ② $f \in C^1$ in some open set containing K .
- ③ $[df(x)]^{-1}$ exists in K .

Then, the sequence x_k defined by the modified Newton method is well-defined, and

- ① The sequence (x_k) is in K and has at least one accumulation point.
- ② Every such accumulation point is a zero of f .

Chapter 3

Numerical linear algebra

3.1 Motivation

In this chapter we will consider the two following problems:

- ① Solve linear systems $Ax = b$, where $x, b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.
- ② Find $x \in \mathbb{R}^n$ that minimizes

$$\sum_{i=1}^m (Ax - b)_i^2,$$

where $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$. When $m > n$ there are more equations than unknowns, so that in general, $Ax = b$ cannot be solved.

Example 3.1 (Stokes flow in a cavity) Three equations,

$$\begin{aligned}\frac{\partial p}{\partial x} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \\ \frac{\partial p}{\partial y} &= \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0,\end{aligned}$$

for the functions $u(x, y)$, $v(x, y)$, and $p(x, y)$; $(x, y) \in [0, 1]^2$. The boundary conditions are

$$\begin{aligned}u(0, y) = u(1, y) = u(x, 0) = 0, \quad u(x, 1) &= 1 \\ v(0, y) = v(1, y) = v(x, 0) = v(x, 1) &= 0.\end{aligned}$$

Solve with a staggered grid. A linear system in $n^2 + 2n(n - 1)$ unknowns. (And by the way, it is singular).

Example 3.2 (Curve fitting) We are given a set of m points (a_i, b_i) in the plane, and we want to find the best cubic polynomial through these points. I.e, we are looking for the coefficients x_1, x_2, x_3, x_4 , such that the polynomial

$$p(y) = \sum_{j=1}^4 x_j y^{j-1}$$

minimizes

$$\sum_{i=1}^m [p(y_i) - b_i]^2,$$

where the vector $p(y_i)$ is of the form Ax , and

$$A = \begin{pmatrix} 1 & y_1 & y_1^2 & y_1^3 \\ 1 & y_2 & y_2^2 & y_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_m & y_m^2 & y_m^3 \end{pmatrix}$$

3.2 Vector and matrix norms

Definition 3.1 (Norm) Let X be a (real or complex) vector space. It is **normed** if there exists a function $\|\cdot\| : X \rightarrow \mathbb{R}$ (the **norm**) with the following properties:

- ① $\|x\| \geq 0$ with $\|x\| = 0$ iff $x = 0$.
- ② $\|\alpha x\| = |\alpha| \|x\|$.
- ③ $\|x + y\| \leq \|x\| + \|y\|$.


Example 3.3 The most common vector norms are the **p -norms** defined (on \mathbb{C}^n) by

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p},$$

which are norms for $1 \leq p < \infty$. Another common norm is the **infinity-norm**,

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

It can be shown that $\|\cdot\|_\infty = \lim_{p \rightarrow \infty} \|\cdot\|_p$.

 **Exercise 3.1** Show that the p -norms do indeed satisfy the properties of a norm.

Solution 3.1: The positivity and homogeneity are trivial. The triangle inequality is proved below.

Lemma 3.1 (Hölder inequality) Let $p, q > 1$ with $1/p + 1/q = 1$. Then,

$$\left| \sum_{k=1}^n x_k y_k \right| \leq \left(\sum_{k=1}^n |x_k|^p \right)^{1/p} \left(\sum_{k=1}^n |y_k|^q \right)^{1/q}.$$

Proof: From **Young's inequality**¹

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q},$$

follows

$$\frac{\left| \sum_{k=1}^n x_k y_k \right|}{\|x\|_p \|y\|_q} \leq \sum_{k=1}^n \frac{|x_k|}{\|x\|_p} \frac{|y_k|}{\|y\|_q} \leq \sum_{k=1}^n \frac{1}{p} \frac{|x_k|^p}{\|x\|_p^p} + \sum_{k=1}^n \frac{1}{q} \frac{|y_k|^q}{\|y\|_q^q} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

■

Lemma 3.2 (Minkowski inequality) Let $p, q > 1$ with $1/p + 1/q = 1$, then

$$\left(\sum_{k=1}^n |x_k + y_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^n |x_k|^p \right)^{1/p} + \left(\sum_{k=1}^n |y_k|^p \right)^{1/p}.$$

¹Since $\log x$ is a concave function, then for every $a, b > 0$,

$$\log \left(\frac{1}{p} a + \frac{1}{q} b \right) \geq \frac{1}{p} \log a + \frac{1}{q} \log b,$$

i.e.,

$$\frac{a}{p} + \frac{b}{q} \geq a^{1/p} b^{1/q},$$

and it only remains to substitute $a \mapsto a^p$ and $b \mapsto b^q$.

Proof: We write

$$|x_k + y_k|^p \leq |x_k| |x_k + y_k|^{p-1} + |y_k| |x_k + y_k|^{p-1}.$$

Using Hölder's inequality for the first term,

$$\sum_{k=1}^n |x_k| |x_k + y_k|^{p-1} \leq \left(\sum_{k=1}^n |x_k|^p \right)^{1/p} \left(\sum_{k=1}^n |x_k + y_k|^{q(p-1)} \right)^{1/q}.$$

Note that $q(p-1) = p$. Similarly, for the second term

$$\sum_{k=1}^n |y_k| |x_k + y_k|^{p-1} \leq \left(\sum_{k=1}^n |y_k|^p \right)^{1/p} \left(\sum_{k=1}^n |x_k + y_k|^p \right)^{1/q},$$

Summing up,

$$\sum_{k=1}^n |x_k + y_k|^p \leq \left(\sum_{k=1}^n |x_k + y_k|^p \right)^{1/q} (\|x\|_p + \|y\|_p).$$

Dividing by the factor on the right-hand side, and using the fact that $1 - 1/q = 1/p$ we get the required result. \blacksquare

Definition 3.2 (Inner product space) Let X be a (complex) vector space. The function $(\cdot, \cdot) : X \times X \rightarrow \mathbb{C}$ is called an **inner product** if:

- ① $(x, y) = \overline{(y, x)}$.
- ② $(x, y + z) = (x, y) + (x, z)$ (bilinearity).
- ③ $(\alpha x, y) = \alpha(x, y)$.
- ④ $(x, x) \geq 0$ with $(x, x) = 0$ iff $x = 0$.

Example 3.4 For $X = \mathbb{C}^n$ the form

$$(x, y) = \sum_{i=1}^n x_i \bar{y}_i$$

is an inner product.

Lemma 3.3 (Cauchy-Schwarz inequality) The following inequality holds in an inner product space.

$$|(x, y)|^2 \leq (x, x)(y, y).$$

Proof: We have,

$$0 \leq (x - \alpha y, x - \alpha y) = (x, x) - \alpha(y, x) - \bar{\alpha}(x, y) + |\alpha|^2(y, y).$$

Suppose that $(y, x) = r \exp(i\theta)$, then take $\alpha = t \exp(-i\theta)$. For every t ,

$$(x, x) - 2rt + t^2(y, y) \geq 0.$$

Since we have a quadratic inequality valid for all t we must have

$$r^2 - (x, x)(y, y) \leq 0,$$

which completes the proof. ■

Comments:

- ① The Cauchy-Schwarz inequality is a special case of Hölder's inequality.
- ② A third method of proof is from the inequality

$$0 \leq ((y, y)x - (x, y)y, (y, y)x - (x, y)y) = (y, y) \left[(x, x)(y, y) - |(x, y)|^2 \right].$$

Lemma 3.4 In an inner product space $\sqrt{(x, x)}$ is a norm.

Proof: Let $\|x\| = \sqrt{(x, x)}$. The positivity and the homogeneity are immediate. The triangle inequality follows from the Cauchy-Schwarz inequality

$$\begin{aligned} \|x + y\|^2 &= (x + y, x + y) = \|x\|^2 + \|y\|^2 + (x, y) + (y, x) \\ &\leq \|x\|^2 + \|y\|^2 + 2|(x, y)| \leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| = (\|x\| + \|y\|)^2. \end{aligned}$$
■

Definition 3.3 An Hermitian matrix A is called **positive definite** (s.p.d) if

$$x^\dagger Ax > 0$$

for all $x \neq 0$.

Definition 3.4 (Convergence of sequences) Let (x_n) be a sequence in a normed vector space X . It is said to converge to a limit x if $\|x_n - x\| \rightarrow 0$.

In \mathbb{R}^n convergence in norm always implies convergence of each of the component.

Lemma 3.5 The norm $\|\cdot\|$ is a continuous mapping from X to \mathbb{R} .

Proof: This is an immediate consequence of the triangle inequality, for

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|,$$

hence

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Take now $y = x_n$ and the limit $n \rightarrow \infty$. ■

Definition 3.5 Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on X . They are called *equivalent* if there exist constants $c_1, c_2 > 0$ such that

$$c_1\|x\| \leq \|x\|' \leq c_2\|x\|$$

for all $x \in X$.

Theorem 3.1 All norms over a finite dimensional vector space are equivalent.


Proof: Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms. It is sufficient to show the existence of a constant $c > 0$ such that

$$\|x\|' \leq c\|x\|$$

for all x . In fact, it is sufficient to restrict this on the unit ball of the norm $\|\cdot\|^2$. Thus, we need to show that for all x on the unit ball of $\|\cdot\|$, the norm $\|\cdot\|'$ is bounded. This follows from the fact that the norm is a continuous function and that the unit ball of a finite-dimensional vector space is compact. ■

Lemma 3.6 In \mathbb{R}^n the following inequalities hold:

$$\begin{aligned}\|x\|_2 &\leq \|x\|_1 && \leq \sqrt{n}\|x\|_2 \\ \|x\|_\infty &\leq \|x\|_2 && \leq \sqrt{n}\|x\|_\infty \\ \|x\|_\infty &\leq \|x\|_1 && \leq n\|x\|_\infty.\end{aligned}$$

 *Exercise 3.2* Prove the following inequalities for vector norms:

$$\begin{aligned}\|x\|_2 &\leq \|x\|_1 && \leq \sqrt{n}\|x\|_2 \\ \|x\|_\infty &\leq \|x\|_2 && \leq \sqrt{n}\|x\|_\infty \\ \|x\|_\infty &\leq \|x\|_1 && \leq n\|x\|_\infty.\end{aligned}$$

Solution 3.2:

① On the one hand, $\|x\|_2^2 = \sum |x_i|^2 \leq (\sum |x_i|)^2 = \|x\|_1^2$. On the other hand

$$\|x\|_1 = \sum_i |x_i| = \sum_i |x_i| \cdot 1 \leq \left(\sum_i x_i^2 \right)^{1/2} \left(\sum_i 1^2 \right)^{1/2} = \sqrt{n}\|x\|_2,$$

which follows from the Cauchy-Schwarz inequality.

② We have

$$\|x\|_\infty^2 = \max_i |x_i|^2 \leq \sum_i |x_i|^2 = \|x\|_2^2 = \sum_i |x_i|^2 \leq n \times \max_i |x_i|^2 = n\|x\|_\infty^2.$$

③ Similarly,

$$\|x\|_\infty = \max_i |x_i| \leq \sum_i |x_i| = \|x\|_1 = \sum_i |x_i| \leq n \times \max_i |x_i| = n\|x\|_\infty.$$

²If this holds on the unit ball, then for arbitrary $x \in X$,

$$\|x\|' = \|x\| \left\| \frac{x}{\|x\|} \right\|' \leq c\|x\| \left\| \frac{x}{\|x\|} \right\| = c\|x\|.$$

Definition 3.6 (Subordinate matrix norm) Let $\|\cdot\|$ be a norm in $X = \mathbb{R}^n$. For every $A : X \rightarrow X$ (a linear operator on the space) we define the following function $\|\cdot\| : \mathcal{B}(X, X) \rightarrow \mathbb{R}$,

$$\|A\| = \sup_{0 \neq x \in X} \frac{\|Ax\|}{\|x\|}. \quad (3.1)$$

Comments:

- ① By the homogeneity of the norm we have

$$\|A\| = \sup_{0 \neq x \in X} \left\| A \frac{x}{\|x\|} \right\| = \sup_{\|x\|=1} \|Ax\|.$$

- ② Since the norm is continuous and the unit ball is compact then,

$$\|A\| = \max_{\|x\|=1} \|Ax\|,$$

and the latter is always finite.

- ③ By definition, for all A and x ,

$$\|Ax\| \leq \|A\| \|x\|.$$

Theorem 3.2 Eq. (3.1) defines a norm on the space of matrices $\mathbb{R}^n \rightarrow \mathbb{R}^n$, which we call the matrix norm **subordinate** to the vector norm $\|\cdot\|$.

Proof: The positivity and the homogeneity are immediate. It remains to show the triangle inequality:

$$\begin{aligned} \|A + B\| &= \sup_{\|x\|=1} \|(A + B)x\| \\ &\leq \sup_{\|x\|=1} (\|Ax\| + \|Bx\|) \\ &\leq \sup_{\|x\|=1} \|Ax\| + \sup_{\|x\|=1} \|Bx\|. \end{aligned}$$

■


Lemma 3.7 For every two matrices A, B and subordinate norm $\|\cdot\|$,

$$\|AB\| \leq \|A\|\|B\|.$$

In particular,

$$\|A^k\| \leq \|A\|^k.$$

Proof: Obvious. ■

 *Exercise 3.3* Show that for every invertible matrix A and norm $\|\cdot\|$,


$$\|A\|\|A^{-1}\| \geq 1.$$

Solution 3.3: Since the norm of the unit matrix is always one for a subordinate matrix norm,

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|.$$

Example 3.5 (infinity-norm) Consider the infinity norm on vectors. The matrix norm subordinate to the infinity norm is

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \max_i \left| \sum_j a_{i,j} x_j \right| = \max_i \sum_j |a_{i,j}|.$$

 *Exercise 3.4* Prove that the matrix norm subordinate to the vector norm $\|\cdot\|_1$ is

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{i,j}|.$$

Solution 3.4: Note that,

$$\|A\|_1 = \sup_{\|x\|_1=1} \sum_i \left| \sum_j a_{i,j} x_j \right| \leq \sup_{\|x\|_1=1} \sum_i \sum_j |a_{i,j}| |x_j| = \sup_{\|x\|_1=1} \sum_j |x_j| \sum_i |a_{i,j}|,$$

from which we get

$$\|A\|_1 \leq \sup_{\|x\|_1=1} \left(\max_j \sum_i |a_{i,j}| \right) \sum_j |x_j| = \max_j \sum_i |a_{i,j}|.$$

The equality is established by choosing x to be a unit vector with a non-zero component that maximizes $\sum_i |a_{ij}|$.

Example 3.6 (2-norm) Consider now the matrix 2-norm subordinate to the vector 2-norm

$$\|x\|_2 = \sqrt{(x, x)}.$$

By definition,

$$\|A\|_2^2 = \sup_{\|x\|_2=1} (Ax, Ax) = \sup_{\|x\|_2=1} (A^\dagger Ax, x).$$

The matrix $A^\dagger A$ is Hermitian, hence it can be diagonalized $A^\dagger A = Q^\dagger \Lambda Q$, where Q is unitary. Then

$$\|A\|_2^2 = \sup_{\|x\|_2=1} (Q^\dagger \Lambda Qx, x) = \sup_{\|x\|_2=1} (\Lambda Qx, Qx) = \sup_{\|y\|_2=1} (\Lambda y, y),$$


where we have used the fact that $y = Q^{-1}x$ has unit norm. This gives,

$$\|A\|_2^2 = \sup_{\|y\|_2=1} \sum_{i=1}^n \lambda_i |y_i|^2,$$

which is maximized by taking y_i to choose the maximal eigenvalue. Thus,

$$\|A\|_2 = \max_{\lambda \in \Sigma(A^\dagger A)} \sqrt{|\lambda|},$$

where we have used the fact that all the eigenvalue of an Hermitian matrix of the form $A^\dagger A$ are real and positive.

 *Exercise 3.5* ① Let $\|\cdot\|$ be a norm on \mathbb{R}^n , and S be an n -by- n non-singular matrix. Define $\|x\|' = \|Sx\|$, and prove that $\|\cdot\|'$ is a norm on \mathbb{R}^n .

② Let $\|\cdot\|$ be the matrix norm subordinate to the above vector norm. Define $\|A\|' = \|SAS^{-1}\|$, and prove that $\|\cdot\|'$ is the matrix norm subordinate to the corresponding vector norm.


Solution 3.5:

① The homogeneity is trivial. For the positivity $\|0\|' = 0$ and $\|x\|' = 0$ only if $Sx = 0$, but since S is regular it follows that $x = 0$. It remains to verify the triangle inequality.

$$\|x + y\|' = \|S(x + y)\| \leq \|Sx\| + \|Sy\| = \|x\|' + \|y\|'.$$

② By definition

$$\|A\|' = \sup_{x \neq 0} \frac{\|SAx\|}{\|Sx\|} = \sup_{y \neq 0} \frac{\|SAS^{-1}y\|}{\|S^{-1}y\|} = \|SAS^{-1}\|.$$

 **Exercise 3.6** True or false: if $\|\cdot\|$ is a matrix norm subordinate to a vector norm, so is $\|\cdot\|' = \frac{1}{2}\|\cdot\|$ (the question is not just whether $\|\cdot\|'$ satisfies the definition of a norm; the question is whether there exists a vector norm, for which $\|\cdot\|'$ is the subordinate matrix norm!).

Solution 3.6: False because the norm of the identity has to be one.

Neumann series Let A be an n -by- n matrix and consider the infinite series

$$\sum_{k=0}^{\infty} A^k,$$

where $A^0 = I$. Like for numerical series, this series is said to converge to a limit B , if the sequence of partial sums

$$B_n = \sum_{k=0}^n A^k$$

converges to B (in norm). Since all norms on finite dimensional spaces are equivalent, convergence does not depend on the choice of norm. Thus, we may consider any arbitrary norm $\|\cdot\|$.

Recall the root test for the convergence of numerical series. Since it only relies on the completeness of the real numbers, it can be generalized as is for arbitrary complete normed spaces. Thus, if the limit

$$L = \lim_{n \rightarrow \infty} \|A^n\|^{1/n}$$

exists, then $L < 1$ implies the (absolute) convergence of the above series, and $L > 1$ implies that the series does not converge.

Proposition 3.1 If the series converges absolutely then

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$$

(and the right hand side exists). It is called the **Neumann series** of $(I - A)^{-1}$.

Proof: We may perform a term-by-term multiplication

$$(I - A) \sum_{k=0}^{\infty} A^k = \sum_{k=0}^{\infty} (A^k - A^{k+1}) = I - \lim_{k \rightarrow \infty} A^k,$$

but the limit must vanish (in norm) if the series converges. ■

We still need to establish the conditions under which the Neumann series converges. First, we show that the limit L always exists:

Proposition 3.2 The limit $\lim_{n \rightarrow \infty} \|A^n\|^{1/n}$ exists and is independent of the choice of norms. The limit is called the **spectral radius** of A and is denoted by $\text{spr}(A)$.

Proof: Let $a_n = \log \|A^n\|$. Clearly,

$$a_{n+m} = \log \|A^{n+m}\| \leq \log \|A^n\| + \log \|A^m\| = a_n + a_m,$$

i.e., the sequence (a_n) is **sub-additive**. Since the logarithm is a continuous function on the positive reals, we need to show that the limit

$$\lim_{n \rightarrow \infty} \log \|A^n\|^{1/n} = \lim_{n \rightarrow \infty} \frac{a_n}{n}$$

exists. This follows directly from the sub-additivity (the Fekete lemma).

Indeed, set m . Then, any integer n can be written as $n = mq + r$, with $0 \leq r < m$.

We have,

$$\frac{a_n}{n} = \frac{a_{mq+r}}{n} \leq \frac{q}{n} a_m + \frac{r}{n} a_r.$$

Taking $n \rightarrow \infty$, the right hand side converges to a_m/m , hence,

$$\limsup \frac{a_n}{n} \leq \frac{a_m}{m}.$$

Taking then $m \rightarrow \infty$ we have

$$\limsup \frac{a_n}{n} \leq \liminf \frac{a_m}{m}$$

which proves the existence of the limit. The independence on the choice of norm results from the equivalence of norms, as

$$c^{1/n} \|A^n\|^{1/n} \leq (\|A^n\|')^{1/n} \leq C^{1/n} \|A^n\|^{1/n}.$$

■

Corollary 3.1 The Neumann series $\sum_k A^k$ converges if $\text{spr } A < 1$ and diverges if $\text{spr } A > 1$.

Thus, the spectral radius of a matrix is always defined, and is a property that does not depend on the choice of norm. We now relate the spectral radius with the eigenvalues of A . First, a lemma:

Lemma 3.8 Let S be an invertible matrix. Then, $\text{spr } S^{-1}AS = \text{spr } A$.

Proof: This is an immediate consequence of the fact that $\|S^{-1} \cdot S\|$ is a matrix norm and the independence of the spectral radius on the choice of norm. ■

Proposition 3.3 Let $\Sigma(A)$ be the set of eigenvalues of A (the **spectrum**). Then,

$$\text{spr } A = \max_{\lambda \in \Sigma(A)} |\lambda|.$$

Proof: By the previous lemma it is sufficient to consider A in Jordan canonical form. Furthermore, since all power of A remain block diagonal, and we are free to choose, say, the infinity norm, we can consider the spectral radius of a single Jordan block; the spectral radius of A is the maximum over the spectral radii of its Jordan blocks.

Let then A be an m -by- m Jordan block with eigenvalue λ , i.e.,

$$A = \lambda I + D,$$

where D has ones above its main diagonal, i.e., it is nil-potent with $D^m = 0$. Raising this sum to the n -th power ($n > m$) we get

$$A^n = \lambda^n I + n \lambda^{n-1} D + \binom{n}{2} \lambda^{n-2} D^2 + \cdots + \binom{n}{m-1} \lambda^{n-m+1} D^{m-1}.$$

Taking the infinity norm we have

$$|\lambda|^n \leq \|A^n\| \leq m \binom{n}{m-1} |\lambda|^{n-m+1} \max(|\lambda|^{m-1}, 1).$$

Taking the n -th root and going to the limit we obtain that $\text{spr } A = |\lambda|$. ■

Proposition 3.4 For every matrix A ,

$$\text{spr } A \leq \inf_{\|\cdot\|} \|A\|,$$

where the infimum is over all choices of subordinate matrix norms.

Proof: For every eigenvalue λ with (normalized) eigenvector u , and every subordinate matrix norm $\|\cdot\|$,

$$\|A\| \geq \|Au\| = |\lambda| \|u\| = |\lambda|.$$

It remains to take the maximum over all $\lambda \in \Sigma(A)$ and the infimum over all norms. ■

We will now prove that this inequality is in fact an identity. For that we need the following lemma:

Lemma 3.9 Every matrix A can be “almost” diagonalized in the following sense: for every $\epsilon > 0$ there exists a non-singular matrix S such that

$$A = S^{-1}(\Lambda + T)S,$$

where Λ is diagonal with its element coinciding with the eigenvalues of A , and T is strictly upper triangular with $\|T\|_\infty < \epsilon$.

Proof: There exists a transformation into the Jordan canonical form:

$$A = P^{-1}(\Lambda + D)P,$$

where D is nil-potent with ones above its main diagonal. Let now

$$E = \begin{pmatrix} \epsilon & 0 & \cdots & 0 \\ 0 & \epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \epsilon^n \end{pmatrix}.$$

and set $E^{-1}P = S$. Then

$$A = S^{-1}E^{-1}(\Lambda + D)ES = S^{-1}(\Lambda + E^{-1}DE)S,$$

where $T = EDE^{-1}$ is given by

$$T_{i,j} = \sum_{k,l} E_{i,k}^{-1} D_{k,l} E_{l,j} = \epsilon^{j-i} D_{i,j}.$$

But since the only non-zero elements are $D_{i,i+1} = 1$, we have $T^{i,i+1} = \epsilon$, and $\|T\|_\infty = \epsilon$. ■

Theorem 3.3 For every matrix A ,

$$\text{spr } A = \inf_{\|\cdot\|} \|A\|.$$


Proof: We have already proved the less-or-equal relation. It remains to show that for every $\epsilon > 0$ there exists a subordinate matrix norm $\|\cdot\|$ such that

$$\|A\| \leq \text{spr } A + \epsilon.$$

This follows from the fact that every matrix is similar to an almost diagonal matrix, and that the spectral radius is invariant under similarity transformations. Thus, for every ϵ we take S as in the lemma above, and set $\|\cdot\| = \|S^{-1} \cdot S\|_\infty$, hence

$$\|A\| = \|\Lambda + T\|_\infty \leq \|\Lambda\|_\infty + \|T\|_\infty = \text{spr } A + \epsilon.$$

■

 **Exercise 3.7** A matrix is called **normal** if it has a complete set of orthogonal eigenvectors. Show that for normal matrices,


$$\|A\|_2 = \text{spr } A.$$

Solution 3.7: If A has a complete set of orthogonal eigenvectors, then every vector x can be written as $x = \sum_i a_i e_i$, where $Au_i = \lambda_i u_i$ and $(u_i, u_j) = \delta_{ij}$. For $x = \sum_i a_i e_i$ we have $Ax = \sum_i \lambda_i a_i e_i$, and

$$(Ax, Ax) = \sum_i |\lambda_i|^2 |a_i|^2.$$

Now,


$$\|A\|_2^2 = \sup_{x \neq 0} \frac{(Ax, Ax)}{(x, x)} = \sup_{x \neq 0} \frac{\sum_i |\lambda_i|^2 |a_i|^2}{\sum_i |a_i|^2} = \max_i |\lambda_i|^2 = (\text{spr } A)^2.$$

 **Exercise 3.8** Show that $\text{spr } A < 1$ if and only if


$$\lim_{k \rightarrow \infty} A^k x = 0, \quad \forall x.$$

Solution 3.8: If $\text{spr } A < 1$, then there exists a matrix norm for which $\|A\| < 1$, hence $\|A^k\| \leq \|A\|^k \rightarrow 0$. Conversely, let $A^k x \rightarrow 0$ for all x . By contradiction, suppose that $\text{spr } A \geq 1$, which implies the existence of an eigenvalue $|\lambda| \geq 1$. Let u be the corresponding eigenvector, then

$$A^k u = \lambda^k u \not\rightarrow 0.$$

 **Exercise 3.9** True or false: the spectral radius $\text{spr } A$ is a matrix norm.

Solution 3.9: False, because for non-zero nil-potent A , $\text{spr } A = 0$. In fact, the spectral radius is a semi-norm.


 **Exercise 3.10** Is the inequality $\text{spr } AB \leq \text{spr } A \text{ spr } B$ true for all pairs of n -by- n matrices? What about if A and B were upper-triangular? Hint: try to take $B = A^T$ and

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}.$$

Solution 3.10: The general assertion is false. Indeed, take A and B as suggested, then

$$AB = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$


Now, $\text{spr } A = \text{spr } B = \sqrt{2}$, but $\text{spr}(AB) = 4$. For upper-diagonal matrices we have an equality since the eigenvalues are the diagonal values, and the diagonal elements of the product are the product of the diagonal elements.

 **Exercise 3.11** Can you use the Neumann series to approximate the inverse of a matrix A ? Under what conditions will this method converge?

Solution 3.11: Take,

$$A^{-1} = (I - (I - A))^{-1} = \sum_{k=1}^{\infty} (I - A)^k.$$

This method will converge if $\text{spr}(I - A) < 1$.

 **Computer exercise 3.1** Construct a “random” 6-by-6 matrix A . Then plot the 1, 2, and infinity norms of $\|A^n\|^{1/n}$ as function of n with the maximum n large enough so that the three curves are sufficiently close to the expected limit.

Normal operators

Definition 3.7 A matrix A is called **normal** if it commutes with its adjoint, $A^\dagger A = AA^\dagger$.

Lemma 3.10 A is a normal operator if and only if

$$\|Ax\|_2 = \|A^\dagger x\|_2$$

for every $x \in \mathbb{R}^n$.

Proof: Suppose first that A is normal, then for all $x \in \mathbb{R}^n$,

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^\dagger Ax) = (x, AA^\dagger x) = (A^\dagger x, A^\dagger x) = \|A^\dagger x\|_2^2.$$

Conversely, let $\|Ax\|_2 = \|A^\dagger x\|_2$. Then,

$$(x, AA^\dagger x) = (A^\dagger x, A^\dagger x) = (Ax, Ax) = (x, A^\dagger Ax),$$

from which follows that

$$(x, (AA^\dagger - A^\dagger A)x) = 0, \quad \forall x \in \mathbb{R}^n.$$

Since $AA^\dagger - A^\dagger A$ is symmetric then it must be zero (e.g., because all its eigenvalues are zero, and it cannot have any nilpotent part). ■

Lemma 3.11 For every matrix A ,

$$\|A^\dagger A\|_2 = \|A\|_2^2.$$

Proof: Recall that the 2-norm of A is given by

$$\|A\|_2^2 = \text{spr } A^\dagger A.$$

On the other hand, since $A^\dagger A$ is Hermitian, its 2-norm coincides with its largest eigenvalue. ■

Theorem 3.4 If A is a normal operator then

$$\|A^n\|_2 = \|A\|_2^n,$$

and in particular $\text{spr } A = \|A\|_2$.

Proof: Suppose first that A was Hermitian. Then, by the previous lemma

$$\|A^2\|_2 = \|A^\dagger A\|_2 = \|A\|_2^2.$$

Since A^2 is also Hermitian we then have $\|A^4\|_2 = \|A\|_2^4$, and so on for every $n = 2^m$. Suppose then that A is normal (but not necessarily Hermitian), then for every $n = 2^m$,

$$\|A^n\|_2^2 = \|(A^\dagger)^n A^n\|_2 = \|(A^\dagger A)^n\|_2 = \|(A^\dagger A)\|_2^n = \|A\|_2^{2n},$$

hence $\|A_n\|_2 = \|A\|_2^n$. It remains to treat the case of general n . Write then $n = 2^m - r$, $r \geq 0$. We then have

$$\|A\|_2^{n+r} = \|A^{n+r}\|_2 \leq \|A^n\|_2 \|A\|_2^r,$$

hence $\|A\|_2^n \leq \|A^n\|_2$. The reverse inequality is of course trivial, which proves the theorem. ■

3.3 Perturbation theory and condition number

Consider the linear system

$$Ax = b,$$

and a “nearby” linear system

$$(A + \delta A)\hat{x} = (b + \delta b).$$

The question is under what conditions the smallness of δA , δb guarantees the smallness of $\delta x = \hat{x} - x$. If δx is small the problem is well-conditioned, and it is ill-conditioned otherwise.

Subtracting the two equations we have

$$A(\hat{x} - x) + \delta A \hat{x} = \delta b,$$

or,

$$\delta x = A^{-1}(-\delta A \hat{x} + \delta b).$$

Taking norms we obtain an inequality

$$\|\delta x\| \leq \|A^{-1}\| (\|\delta A\| \|\hat{x}\| + \|\delta b\|),$$

which we further rearrange as follows,

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \|A\| \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \|\hat{x}\|} \right).$$

We have thus expressed the relative change in the output as the product of the relative change in the input (we'll look more carefully at the second term later) and the number

$$\kappa(A) = \|A^{-1}\| \|A\|,$$

which is the (relative) **condition number**. When $\kappa(A)$ is large a small perturbation in the input can produce a large perturbation in the output.

In practice, \hat{x} will be the computed solution. Then, provided we have estimates on the “errors” δA , and δb , we can estimate the relative error $\|\delta x\|/\|\hat{x}\|$. From a theoretical point of view, however, it seems “cleaner” to obtain an error bound which is independent of δx (via \hat{x}). This can be achieved as follows. First from

$$(A + \delta A)(x + \delta x) = (b + \delta b) \quad \Rightarrow \quad (A + \delta A)\delta x = (-\delta A x + \delta b)$$

we extract

$$\begin{aligned} \delta x &= (A + \delta A)^{-1}(-\delta A x + \delta b) \\ &= [A(I + A^{-1}\delta A)]^{-1}(-\delta A x + \delta b) \\ &= (I + A^{-1}\delta A)^{-1}A^{-1}(-\delta A x + \delta b). \end{aligned}$$

Taking now norm and applying the standard inequalities we get

$$\frac{\|\delta x\|}{\|x\|} \leq \|(I + A^{-1}\delta A)^{-1}\| \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right).$$

Now, if $\text{spr } A^{-1}\delta A < 1$, we can use the Neumann series to get the following estimate,

$$\|(I + A^{-1}\delta A)^{-1}\| = \left\| \sum_{n=0}^{\infty} (-A^{-1}\delta A)^n \right\| \leq \sum_{n=0}^{\infty} \|A^{-1}\|^n \|\delta A\|^n = \frac{1}{1 - \|A^{-1}\| \|\delta A\|}.$$

Combining with the above,

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \\ &= \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \|x\|} \right) \\ &\leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right), \end{aligned}$$

where we have used the fact that $\|A\| \|x\| \geq \|Ax\| = \|b\|$. In this (cleaner) formulation the condition number is

$$\frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}},$$

which is close to $\kappa(A)$ provided that δA is sufficiently small, and more precisely, that $\kappa(A) \frac{\|\delta A\|}{\|A\|} = \|A^{-1}\| \|\delta A\| < 1$.

We conclude this section by establishing another meaning to the condition number. It is *the reciprocal on the distance to the nearest ill-posed problem*. A large condition number means that the problem is close *in a geometrical sense* to a singular problem.

Theorem 3.5 Let A be non-singular, then

$$\frac{1}{\kappa(A)} = \min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \text{ is singular} \right\},$$

where $\kappa(A)$ is expressed in terms of 2-norm (Euclidean).

Proof: Since $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$, we need to show that

$$\frac{1}{\|A^{-1}\|_2} = \min \{ \|\delta A\|_2 : A + \delta A \text{ is singular} \}.$$

If $\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2}$, then $\|A^{-1}\|_2 \|\delta A\|_2 < 1$, which implies the convergence of the Neumann series

$$\sum_{n=0}^{\infty} (-A^{-1} \delta A)^n = (1 + A^{-1} \delta A)^{-1} = A^{-1} (A + \delta A)^{-1},$$

i.e.,

$$\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2} \quad \Rightarrow \quad A + \delta A \text{ is not singular},$$

or,

$$\min \{ \|\delta A\|_2 : A + \delta A \text{ is singular} \} \geq \frac{1}{\|A^{-1}\|_2}.$$

To show that this is an equality it is sufficient to construct a δA of norm $\frac{1}{\|A^{-1}\|_2}$ so that $A + \delta A$ is singular. By definition, there exists an $x \in \mathbb{R}^n$ on the unit sphere for which $\|A^{-1}x\|_2 = \|A^{-1}\|_2$. Let then $y = \frac{A^{-1}x}{\|A^{-1}x\|_2}$, be another unit vector and construct

$$\delta A = -\frac{xy^T}{\|A^{-1}\|_2}.$$

First note that


$$\|\delta A\|_2 = \frac{1}{\|A^{-1}\|_2} \max_{\|z\|_2=1} \|xy^T z\|_2 = \frac{1}{\|A^{-1}\|_2} \max_{\|z\|_2=1} |y^T z| = \frac{1}{\|A^{-1}\|_2},$$

where we have used the fact that $\|x\|_2 = 1$, and the fact that $|y^T z|$ is maximized for $z = y$. Finally, $A + \delta A$ is singular because

$$(A + \delta A)y = \left(A - \frac{xy^T}{\|A^{-1}\|_2}\right)y = Ay - \frac{x}{\|A^{-1}\|_2} = 0.$$

■

Comment: Note how the theorem relies on the use of the Euclidean norm.

 **Exercise 3.12** The **spectrum** $\Sigma(A)$ of a matrix A is the set of its eigenvalues. The **ϵ -pseudospectrum** of A , which we denote by $\Sigma_\epsilon(A)$, is defined as the set of complex numbers z , for which there exists a matrix δA such that $\|\delta A\|_2 \leq \epsilon$ and z is an eigenvalue of $A + \delta A$. In mathematical notation,

$$\Sigma_\epsilon(A) = \{z \in \mathbb{C} : \exists \delta A, \|\delta A\|_2 \leq \epsilon, z \in \Sigma(A + \delta A)\}.$$

Show that

$$\Sigma_\epsilon(A) = \left\{z \in \mathbb{C} : \|(zI - A)^{-1}\|_2 \geq 1/\epsilon\right\}.$$

Solution 3.12: By definition, $z \in \Sigma_\epsilon(A)$ if and only if

$$\exists \delta A, \|\delta A\|_2 \leq \epsilon, z \in \Sigma(A + \delta A),$$

which in turn holds if and only if

$$\exists \delta A, \|\delta A\|_2 \leq \epsilon, 0 \in \Sigma(A - zI + \delta A).$$

Now, we have shown that

$$\frac{1}{\|(A - zI)^{-1}\|_2} = \min \{\|\delta A\|_2 : 0 \in \Sigma(A - zI + \delta A)\}.$$


This means that there exists such a δA if and only if

$$\epsilon \geq \frac{1}{\|(A - zI)^{-1}\|_2}.$$

I.e., $z \in \Sigma_\epsilon(A)$ if and only if

$$\|(A - zI)^{-1}\|_2 \geq 1/\epsilon,$$

which completes the proof.

 **Exercise 3.13** Let $Ax = b$ and $(A + \delta A)\hat{x} = (b + \delta b)$. We showed in class that $\delta x = \hat{x} - x$ satisfies the inequality,

$$\|\delta x\|_2 \leq \|A^{-1}\|_2 (\|\delta A\|_2 \|\hat{x}\|_2 + \|\delta b\|_2).$$

Show that this is not just an upper bound: that for sufficiently small $\|\delta A\|_2$ there exist non-zero δA , δb such that the above is an equality. (**Hint:** follow the lines of the proof that links the reciprocal of the condition number to the distance to the nearest ill-posed problem.)

3.4 Direct methods for linear systems

Algorithms for solving the linear system $Ax = b$ are divided into two sorts: **direct methods** give, in the absence of roundoff errors, an exact solution after a finite number of steps (of floating point operations); all direct methods are variations of **Gaussian elimination**. In contrast, **iterative methods** compute a sequence of iterates (x_n) , until x_n is sufficiently close to satisfying the equation. Iterative methods may be much more efficient in certain cases, notably when the matrix A is **sparse**.

3.4.1 Matrix factorization

The basic direct method algorithm uses **matrix factorization**—the representation of a matrix A as a product of “simpler” matrices. Suppose that A was **lower-triangular**:

$$\begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Then the system can easily be solved using **forward-substitution**:

Algorithm 3.4.1: FORWARD-SUBSTITUTION(A, b)

```

for  $i = 1$  to  $n$ 
  do  $x_i = (b_i - \sum_{k=1}^{i-1} a_{ik}x_k) / a_{ii}$ 

```

Similarly, if A was **upper-diagonal**,

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Then the system can easily be solved using **backward-substitution**:

Algorithm 3.4.2: BACKWARD-SUBSTITUTION(A, b)

```

for  $i = n$  downto  $1$ 
  do  $x_i = (b_i - \sum_{k=i+1}^n a_{ik}x_k) / a_{ii}$ 

```

Finally, if A is a **permutation matrix**, i.e., an identity matrix with its rows permuted, then the system $Ax = b$ only requires the permutation of the rows of b .

Matrix factorization consists of expressing any non-singular matrix A as a product $A = PLU$, where P is a permutation matrix, L is non-singular lower-triangular and U is non-singular upper-triangular. Then, the system $Ax = b$ is solved as follows:

$$\begin{aligned} LUx &= P^{-1}b = P^Tb && \text{permute the entries of } b \\ Ux &= L^{-1}(P^Tb) && \text{forward substitution} \\ x &= U^{-1}(L^{-1}P^Tb) && \text{backward substitution.} \end{aligned}$$

This is the general idea. We now review these steps in a systematic manner.

Lemma 3.12 Let P, P_1, P_2 be n -by- n permutation matrices and A be an n -by- n matrix. Then,

- ① PA is the same as A with its rows permuted and AP is the same as A with its column permuted.
- ② $P^{-1} = P^T$.
- ③ $\det P = \pm 1$.
- ④ $P_1 P_2$ is also a permutation matrix.

Proof: Let $\pi : [1, n] \rightarrow [1, n]$ be a permutation function (one-to-one and onto). Then, the entries of the matrix P are of the form $P_{ij} = \delta_{\pi^{-1}(i), j}$. Now,

$$(PA)_{i,j} = \sum_{k=1}^n \delta_{\pi^{-1}(i), k} a_{kj} = a_{\pi^{-1}(i), j}$$

$$(AP)_{i,j} = \sum_{k=1}^n a_{ik} \delta_{\pi^{-1}(k), j} = a_{i, \pi(j)},$$

which proves the first assertion. Next,

$$(P^T P)_{i,j} = \sum_{k=1}^n \delta_{\pi^{-1}(i), k} \delta_{k, \pi^{-1}(j)} = \sum_{k=1}^n \delta_{i, \pi(k)} \delta_{\pi(k), j} = \delta_{i,j},$$

which proves the second assertion. The determinant of a permutation matrix is ± 1 because when two rows of a matrix are interchanged the determinant changes sign. Finally, if P_1 and P_2 are permutation matrices with maps π_1 and π_2 , then

$$(P_1 P_2)_{i,j} = \sum_{k=1}^n \delta_{\pi_1^{-1}(i), k} \delta_{\pi_2^{-1}(k), j} = \sum_{k=1}^n \delta_{\pi_1^{-1}(i), k} \delta_{k, \pi_2(j)}$$

$$= \delta_{\pi_1^{-1}(i), \pi_2(j)} = \delta_{\pi_2^{-1}(\pi_1^{-1}(i)), j}.$$

■

Definition 3.8 The m -th **principal sub-matrix** of an n -by- n matrix A is the square matrix with entries a_{ij} , $1 \leq i, j \leq m$.

Definition 3.9 A lower triangular matrix L is called **unit lower triangular** if its diagonal entries are 1.

Theorem 3.6 A matrix A has a unique decomposition $A = LU$ with L unit lower triangular and U non-singular upper triangular if and only if all its principal sub-matrices are non-singular.

Proof: Suppose first that $A = LU$ with the above properties. Then, for every $1 \leq m \leq n$,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & \\ & L_{22} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ & U_{22} \end{pmatrix},$$

where A_{11} is the m -th principal sub-matrix, L_{11} and L_{22} are unit lower triangular and U_{11} and U_{22} are upper triangular. Now,

$$A_{11} = L_{11}U_{11}$$

is non-singular because $\det A_{11} = \det L_{11} \det U_{11} = \prod_{i=1}^m u_{ii} \neq 0$, where the last step is a consequence of U being triangular and non-singular.

Conversely, suppose that all the principal sub-matrices of A are non-singular. We will show the existence of L, U by induction on n . For $n = 1$, $a = 1 \cdot a$. Suppose that the decomposition holds all $(n - 1)$ -by- $(n - 1)$ matrices. Let A' be of the form

$$A' = \begin{pmatrix} A & b \\ c^T & d \end{pmatrix}$$

where b, c are column vectors of length $(n - 1)$ and d is a scalar. By assumption, $A = LU$. Thus, we need to find vectors $l, u \in \mathbb{R}^{n-1}$ and a scalar γ such that

$$\begin{pmatrix} A & b \\ c^T & d \end{pmatrix} = \begin{pmatrix} L & \\ l^T & 1 \end{pmatrix} \begin{pmatrix} U & u \\ & \gamma \end{pmatrix}.$$

Expanding we have

$$\begin{aligned} b &= Lu \\ c^T &= l^T U \\ d &= l^T u + \gamma. \end{aligned}$$

The first and second equation for u, l can be solved because by assumption L and U are invertible. Finally, γ is extracted from the third equation. It must be non-zero otherwise A' would be singular. ■

A matrix A may be regular and yet the LU decomposition may fail. This is where permutations are necessary.

Theorem 3.7 Let A be a non-singular n -by- n matrix. Then there exist permutation matrices P_1, P_2 , a unit lower triangular matrix L and an upper triangular matrix U , such that

$$P_1 A P_2 = LU.$$

Either P_1 or P_2 can be taken to be a unit matrix.

Proof: The proof is by induction. The case $n = 1$ is trivial. Assume this is true for dimension $n - 1$. Let then A be a non-singular matrix. Thus, every row and every column has a non-zero element, and we can find permutation matrices P'_1, P'_2 such that $a_{11} = (P'_1 A P'_2)_{11} \neq 0$ (only one of them is necessary).

Now, we solve the block problem

$$P'_1 A P'_2 = \begin{pmatrix} a_{11} & A_{12}^T \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{A}_{22} \end{pmatrix},$$

where A_{22}, I and \tilde{A}_{22} are $(n - 1)$ -by- $(n - 1)$ matrices, and $A_{12}, A_{21}, L_{21}, U_{12}$ and are $(n - 1)$ -vectors; u_{11} is a scalar. Expanding, we get

$$u_{11} = a_{11}, \quad A_{12} = U_{12}, \quad A_{21} = L_{21} u_{11}, \quad A_{22} = L_{21} U_{12}^T + \tilde{A}_{22}.$$

Since $\det A \neq 0$ and multiplication by a permutation matrix can at most change the sign of the determinant, we have

$$0 \neq \det P'_1 A P'_2 = 1 \cdot u_{11} \cdot \det \tilde{A}_{22},$$

from which we deduce that \tilde{A}_{22} is non-singular. Applying the induction, there exist permutation matrices \tilde{P}_1, \tilde{P}_2 and triangular matrices $\tilde{L}_{22}, \tilde{U}_{22}$ such that

$$\tilde{P}_1 \tilde{A}_{22} \tilde{P}_2 = \tilde{L}_{22} \tilde{U}_{22}.$$

Substituting we get

$$\begin{aligned} P'_1 A P'_2 &= \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{P}_1^T \tilde{L}_{22} \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ L_{21} & \tilde{P}_1^T \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \tilde{P}_1 L_{21} & \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \tilde{P}_2^T \\ 0 & \tilde{U}_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_2^T \end{pmatrix} \end{aligned}$$

The two outer matrices are permutation matrices whereas the two middle matrices satisfy the required conditions. This completes the proof. ■

A practical choice of the permutation matrix, known as Gaussian elimination with partial pivoting (GEPP) is given in the following corollary:

Corollary 3.2 It is possible to choose $P'_2 = I$ and P'_1 so that a_{11} is the largest entry in absolute value in its column.

The PLU with partial pivoting algorithm is implemented as follows:

Algorithm 3.4.3: LU FACTORIZATION(A)

```

for  $i = 1$  to  $n - 1$ 
  /* permute only with rows under  $i$  */
  permute the rows of  $A, L$  such that  $a_{ii} \neq 0$ 
  /* calculate  $L_{21}$  */
  for  $j = i + 1$  to  $n$ 
    do  $l_{ji} = a_{ji}/a_{ii}$ 
  /* calculate  $U_{12}$  */
  for  $j = i$  to  $n$ 
    do  $u_{ij} = a_{ij}$ 
  /* change  $A_{22}$  into  $\tilde{A}_{22}$  */
  for  $j = i + 1$  to  $n$ 
    do for  $k = i + 1$  to  $n$ 
      do  $a_{jk} = a_{jk} - l_{ji}u_{ik}$ 

```


Comments:

- ① It can be checked that once l_{ij} and u_{ij} are computed, the corresponding entries of A are not used anymore. This means that U, L can overwrite A . (No need to keep the diagonal terms of L .)
- ② Since the algorithm involves row permutation, the output must also provide the permutation matrix, which can be represented by a vector.
- ③ In practice, there is no need to actually permute the entries of the matrix. This can be done “logically” only.

Operation count The number of operations needed for LU factorization can be deduced directly from the algorithm:

$$\sum_{i=1}^{n-1} \left(\sum_{j=i+1}^n + \sum_{j=i+1}^n \sum_{k=i+1}^n 2 \right) = \sum_{i=1}^{n-1} [(n-i) + 2(n-i)^2] = \frac{2}{3}n^3 + O(n^2).$$

Since the forward and backward substitution require $O(n^2)$ operations, the number of operations needed to solve the system $Ax = b$ is roughly $\frac{2}{3}n^3$.

 **Exercise 3.14** Show that every matrix of the form


$$\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}$$

$a, b, \neq 0$, has an LU decomposition. Show that even if the diagonal elements of L are 1 the decomposition is not unique.

Solution 3.14: Note that this matrix is singular, hence does not fit to the scope considered above. Yet, setting

$$\begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix} = \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix},$$

we get $u_{11} = 0$, $l_{21}u_{11} = 0$ (which is redundant), $u_{12} = a$, and $l_{21}u_{12} + u_{22} = b$. These constraints can be solved for arbitrary l_{21} .

 **Exercise 3.15** Show that if $A = LU$ is symmetric then the columns of L are proportional to the rows of U .


Solution 3.15: From the symmetry of A follows that

$$LU = A = A^T = U^T L^T.$$


Now $U^T L^T$ is also an LU decomposition of A , except that the lower-triangular matrix is not normalized. Let $S = \text{diag}(u_{11}, u_{22}, \dots)$, then

$$LU = (U^T S^{-1})(S L^T).$$

By the uniqueness of the LU decomposition (for regular matrices), it follows that $L = U^T S^{-1}$, which is what we had to show.

 **Exercise 3.16** Show that every symmetric positive-definite matrix has an LU -decomposition.


Solution 3.16: By a previous theorem, it is sufficient to show that all the principal submatrices are regular. In fact, they are all s.p.d., which implies their regularity.

 **Exercise 3.17** Suppose you want to solve the equation $AX = B$, where A is n -by- n and X, B are n -by- m . One algorithm would factorize $A = PLU$ and then solve the system column after column using forward and backward substitution. The other algorithm would compute A^{-1} using Gaussian elimination and then perform matrix multiplication to get $X = A^{-1}B$. Count the number of operations in each algorithm and determine which is more efficient.

Solution 3.17: The first algorithm requires roughly $\frac{2}{3}n^3$ operations. The second requires about the same number for matrix inversion, but then, $O(n^3)$ more operations for matrix multiplication.

 **Exercise 3.18** Determine the LU factorization of the matrix

$$\begin{pmatrix} 6 & 10 & 0 \\ 12 & 26 & 4 \\ 0 & 9 & 12 \end{pmatrix}.$$

 **Computer exercise 3.2** Construct in Matlab an n -by- n matrix A (its entries are not important, but make sure it is non-singular), and verify how long it takes to perform the operation $B = \text{inv}(A)$; . Repeat the procedure for $n = 10, 100, 1000, 2000$.

3.4.2 Error analysis

The two-step approach for obtaining error bounds is as follows:

- ① Analyze the accumulation of roundoff errors to show that the *algorithm* for solving $Ax = b$ generates the exact solution \hat{x} of the nearby problem $(A + \delta A)\hat{x} = (b + \delta b)$, where $\delta A, \delta b$ (the **backward errors**) are small.
- ② Having obtained estimates for the backward errors, apply perturbation theory to bound the error $\hat{x} - x$.

Note that perturbation theory assumes that $\delta A, \delta b$ are given. In fact, these perturbations are just “backward error estimates” of the roundoff errors present in the computation.

We start with backward error estimates, in the course of which we will get a better understanding of the role of **pivoting** (row permutation). As a demonstration, consider the matrix

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

with an arithmetic device accurate to three decimal digits. Note first that

$$\kappa(A) = \|A\|_{\infty} \|A^{-1}\|_{\infty} \approx 2 \times 2,$$

so that the result is quite insensitive to perturbations in the input. Consider now an LU decomposition, taking into account roundoff errors:

$$\begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \ell_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}.$$

Then,

$$u_{11} = \text{fl}(0.0001/1) = 0.0001$$

$$\ell_{21} = \text{fl}(1/u_{11}) = 10000$$

$$u_{12} = 1$$

$$u_{22} = \text{fl}(1 - \ell_{21}u_{12}) = \text{fl}(1 - 10000 \cdot 1) = -10000.$$

However,

$$\begin{pmatrix} 1 & 0 \\ 10000 & 1 \end{pmatrix} \begin{pmatrix} 0.0001 & 1 \\ 0 & -10000 \end{pmatrix} = \begin{pmatrix} 0.0001 & 1 \\ 1 & 0 \end{pmatrix}.$$

Thus, the a_{22} entry has been completely forgotten! In our terminology, the method is not backward stable because

$$\frac{\|\delta A\|_{\infty}}{\|A\|_{\infty}} = \frac{\|A - LU\|_{\infty}}{\|A\|_{\infty}} = \frac{1}{2}.$$

The relative backward error is large, and combined with the estimated condition number, the relative error in x could be as large as 2.

Had we used GEPP, the order of the rows would have been reversed,

$$\begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \ell_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix},$$

yielding

$$u_{11} = \text{fl}(1/1) = 1$$

$$\ell_{21} = \text{fl}(0.0001/u_{11}) = 0.0001$$

$$u_{12} = \text{fl}(1/1) = 1$$

$$u_{22} = \text{fl}(1 - \ell_{21}u_{12}) = \text{fl}(1 - 0.0001 \cdot 1) = 1,$$

which combined back gives

$$\begin{pmatrix} 1 & 0 \\ 0.0001 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1.0001 \end{pmatrix},$$

and

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} = \frac{\|A - LU\|_\infty}{\|A\|_\infty} = \frac{0.0001}{2}.$$

3.5 Iterative methods

3.5.1 Iterative refinement

Let's start with a complementation of direct methods. Suppose we want to solve the system $Ax = b$, i.e., we want to find the vector $x = A^{-1}b$, but due to roundoff errors (and possible other sources of errors), we obtain instead a vector

$$x_0 = \tilde{A}^{-1}b.$$

Clearly, we can substitute the computed solution back into the linear system, and find out that the **residual**,

$$b - Ax_0 \stackrel{\text{def}}{=} r_0$$

differs from zero. Let $e_0 = x_0 - x$ be the **error**. Subtracting $b - Ax = 0$ from the residual equation, we obtain

$$Ae_0 = r_0.$$

That is, *the error satisfies a linear equation with the same matrix A and the residual vector on its right hand side.*

Thus, we will solve the equation for e_0 , but again we can only do it approximately. The next approximation we get for the solution is

$$x_1 = x_0 + \tilde{A}^{-1}r_0 = x_0 + \tilde{A}^{-1}(b - Ax_0).$$

Once more, we define the residual,

$$r_1 = b - Ax_1,$$

and notice that the error satisfies once again a linear system, $Ae_1 = r_1$, thus the next correction is $x_2 = x_1 + \tilde{A}^{-1}(b - Ax_1)$, and inductively, we get

$$x_{n+1} = x_n + \tilde{A}^{-1}(b - Ax_n). \quad (3.2)$$

The algorithm for iterative refinement is given by

Algorithm 3.5.1: ITERATIVE REFINEMENT(A, b, ϵ)

```

 $x = 0$ 
for  $i = 1$  to  $n$ 
   $r = b - Ax$ 
  if  $\|r\| < \epsilon$ 
    then break
  do
    Solve  $Ae = r$ 
     $x = x + e$ 
return ( $x$ )

```

Of course, if the solver is exact, the refinement procedure ends after one cycle.

Theorem 3.8 If \tilde{A}^{-1} is sufficiently close to A^{-1} in the sense that $\text{spr}(I - A\tilde{A}^{-1}) < 1$, then the iterative refinement procedure converges to the solution x of the system $Ax = b$. (Note that equivalently, we need $\|I - A\tilde{A}^{-1}\|$ in any subordinate matrix norm.)

Proof: We start by showing that

$$x_n = \tilde{A}^{-1} \sum_{k=0}^n (I - A\tilde{A}^{-1})^k b.$$

We do it inductively. For $n = 0$ we have $x_0 = \tilde{A}^{-1}b$. Suppose this was correct for $n - 1$, then

$$\begin{aligned}
 x_n &= x_{n-1} + \tilde{A}^{-1}(b - Ax_{n-1}) \\
 &= \tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k b + \tilde{A}^{-1}b - \tilde{A}^{-1}A\tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k b \\
 &= \tilde{A}^{-1} \left[\sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k + I - A\tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k \right] b \\
 &= \tilde{A}^{-1} \left[I + (I - A\tilde{A}^{-1}) \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k \right] b \\
 &= \tilde{A}^{-1} \sum_{k=0}^n (I - A\tilde{A}^{-1})^k b.
 \end{aligned}$$

We have a Neumann series which converges if and only if $\text{spr}(I - A\tilde{A}^{-1}) < 1$, giving in the limit

$$\lim_{n \rightarrow \infty} x_n = \tilde{A}^{-1}(A\tilde{A}^{-1})^{-1}b = A^{-1}b = x.$$

■

3.5.2 Analysis of iterative methods

Example 3.7 (Jacobi iterations) Consider the following example

$$\begin{aligned} 7x_1 - 6x_2 &= 3 \\ -8x_1 + 9x_2 &= -4, \end{aligned}$$

whose solution is $x = (1/5, -4/15)$. We may try to solve this system by the following iterative procedure:

$$\begin{aligned} x_1^{(n+1)} &= \frac{3 + 6x_2^{(n)}}{7} \\ x_2^{(n+1)} &= \frac{-4 + 8x_1^{(n)}}{9}. \end{aligned}$$

From a matrix point of view this is equivalent to taking the system

$$\begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix},$$

and splitting it as follows,

$$\begin{pmatrix} 7 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n+1)} = - \begin{pmatrix} 0 & -6 \\ -8 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n)} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}.$$

This iterative methods, based on a splitting of the matrix A into its diagonal part and its off-diagonal part is called **Jacobi's method**.

The following table gives a number of iterates:

n	$x_1^{(n)}$	$x_2^{(n)}$
1	0.4286	-0.4444
10	0.1487	-0.1982
20	0.1868	-0.2491
40	0.1991	-0.2655
80	0.2000	-0.2667

Example 3.8 (Gauss-Seidel iterations) Consider now the same system, but with a slightly different iterative method:

$$\begin{aligned}x_1^{(n+1)} &= \frac{3 + 6x_2^{(n)}}{7} \\x_2^{(n+1)} &= \frac{-4 + 8x_1^{(n+1)}}{9}.\end{aligned}$$


The idea here is to use the entries which have already been computed in the present iteration. In matrix notation we have

$$\begin{pmatrix} 7 & 0 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n+1)} = - \begin{pmatrix} 0 & -6 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n)} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}.$$

This iterative method, based on a splitting of the matrix A into its lower-triangular part and the remainder is called the **Gauss-Seidel method**.

The following table gives a number of iterates:

n	$x_1^{(n)}$	$x_2^{(n)}$
1	0.4286	-0.0635
10	0.2198	-0.2491
20	0.2013	-0.2655
40	0.2000	-0.2667
80	0.2000	-0.2667


 *Exercise 3.19* Write an algorithm (i.e., a list of instructions in some pseudocode) that calculates the solution to the linear system, $Ax = b$, by Gauss-Seidel's iterative procedure. The algorithm receives as input the matrix A and the vector b , and returns the solution x . Try to make the algorithm efficient.

Solution 3.19:

Algorithm 3.5.2: GAUSS-SEIDEL(A, b, ϵ, M)

```

 $x = 0$ 
for  $i = 1$  to  $M$ 
     $r = b - Ax$ 
    if  $\|r\| < \epsilon$ 
    do then break
    for  $j = 1$  to  $n$ 
        do  $x_j = (b_j - \sum_{k \neq j} a_{jk}x_k) / a_{jj}$ 
return  $(x)$ 
```

 *Computer exercise 3.3* Solve the system

$$\begin{pmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

using both the Jacobi and the Gauss-Seidel iterations. Plot a graph of the norm of the errors as function of the number of iterations. Use the same graph for both methods for comparison.

We are now ready for a general analysis of iterative methods. Suppose we want to solve the system $Ax = b$. For any non-singular matrix Q we can equivalently write $Qx = (Q - A)x + b$, which leads to the iterative method

$$Qx_{n+1} = (Q - A)x_n + b.$$

Definition 3.10 An iterative method is said to be convergent if it converges for any initial vector x_0 .

The goal is to choose a **splitting matrix** Q such that (1) Q is easy to invert, and (2) the iterations converge fast.

Theorem 3.9 Let A be a non-singular matrix, and Q be such that $\text{spr}(I - Q^{-1}A) < 1$. Then the iterative method is convergent.

Proof: We have

$$x_{n+1} = (I - Q^{-1}A)x_n + Q^{-1}b.$$

It is easy to see by induction that

$$x_n = (I - Q^{-1}A)^n x_0 + \sum_{k=0}^{n-1} (I - Q^{-1}A)^k Q^{-1}b,$$

and as we've already seen, the Neumann series converges iff $\text{spr}(I - Q^{-1}A) < 1$. If it converges, the first term also converges to zero (the initial condition is forgotten). The limit is

$$\lim_{n \rightarrow \infty} x_n = (Q^{-1}A)^{-1} Q^{-1}b = A^{-1}b = x.$$

■

Definition 3.11 A matrix A is called **diagonally dominant** if for any row i ,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

Proposition 3.5 If A is diagonally dominant then Jacobi's method converges.


Proof: For Jacobi's method the matrix Q comprises the diagonal of A , therefore, $Q^{-1}A$ consists of the rows of A divided by the diagonal term, and

$$(I - Q^{-1}A)_{ij} = \begin{cases} 0 & i = j \\ -\frac{a_{ij}}{a_{ii}} & i \neq j \end{cases}.$$

Because A is diagonally dominant,

$$\|I - Q^{-1}A\|_{\infty} = \max_i \sum_j |(I - Q^{-1}A)_{ij}| = \max_i \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1.$$

■

 *Exercise 3.20* Show that the Jacobi iteration converges for 2-by-2 symmetric positive-definite systems.

Hint Suppose that the matrix to be inverted is

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

First, express the positive-definiteness of A as a condition on a, b, c . Then, proceed to write the matrix $(I - Q^{-1}A)$, where Q is the splitting matrix corresponding to the Jacobi iterative procedure. It remains to find a norm in which $\|I - Q^{-1}A\| < 1$ or compute the spectral radius.


Solution 3.20: If A of this form is positive definite, then for every x, y ,

$$p(x, y) = ax^2 + 2bxy + cy^2 \geq 0,$$

For the point $(0, 0)$ to be a minimum of $p(x, y)$ we need $a, c > 0$ and $ac > b^2$. Now,


$$I - Q^{-1}A = I - \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}^{-1} \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & b/a \\ b/c & 1 \end{pmatrix}$$

Thus, $\text{spr}(I - Q^{-1}A) = \sqrt{b^2/ac} < 1$, which proves the convergence of the method.


 **Exercise 3.21** Will Jacobi's iterative method converge for

$$\begin{pmatrix} 10 & 2 & 3 \\ 4 & 50 & 6 \\ 7 & 8 & 90 \end{pmatrix}.$$

Solution 3.21: Yes, because the matrix is diagonally dominant.

 **Exercise 3.22** Explain why at least one eigenvalue of the Gauss-Seidel iterative matrix must be zero.

Solution 3.22: Because the last row of $Q - A$ is zero.

 **Exercise 3.23** Show that if A is strictly diagonally dominant then the Gauss-Seidel iteration converges.

Solution 3.23: The method of Gauss-Seidel reads as follows

$$x_i^{(k+1)} = - \sum_{j<i} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j>i} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}.$$

If x is the solution to this system and $e^{(k)} = x^{(k)} - x$, then


$$e_i^{(k+1)} = - \sum_{j<i} \frac{a_{ij}}{a_{ii}} e_j^{(k+1)} - \sum_{j>i} \frac{a_{ij}}{a_{ii}} e_j^{(k)}.$$

Let $r = \max_i \sum_{j \neq i} |a_{ij}|/|a_{ii}|$, which by assumption is less than 1. It can be shown, by induction on the rows of $e^{(k)}$, that $\|e^{(k+1)}\|_\infty \leq r\|e^{(k)}\|_\infty$, which implies convergence. Indeed, for $i = 1$,

$$|e_1^{(k+1)}| \leq \sum_{j>1} \frac{|a_{1j}|}{|a_{11}|} |e_j^{(k)}| \leq r\|e^{(k)}\|_\infty \leq \|e^{(k)}\|_\infty.$$

Suppose this is true up to row $i - 1$, then,

$$|e_i^{(k+1)}| = \sum_{j<i} \frac{|a_{ij}|}{|a_{ii}|} \|e^{(k)}\|_\infty + \sum_{j>i} \frac{|a_{ij}|}{|a_{ii}|} \|e^{(k)}\|_\infty \leq r\|e^{(k)}\|_\infty.$$

 **Exercise 3.24** What is the explicit form of the iteration matrix $G = (I - Q^{-1}A)$ in the Gauss-Seidel method when

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

Solution 3.24: Do it by inspection:

$$\begin{aligned} 2x_1^{(n+1)} &= x_2^{(n)} + b_1 \\ 2x_2^{(n+1)} &= x_1^{(n+1)} + x_3^{(n+1)} + b_2 \\ 2x_3^{(n+1)} &= x_2^{(n+1)} + x_4^{(n+1)} + b_3, \end{aligned}$$

from which we extract,

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{2}x_2^{(n)} + \cdots \\ x_2^{(n+1)} &= \frac{1}{4}x_2^{(n)} + \frac{1}{2}x_3^{(n)} + \cdots \\ x_3^{(n+1)} &= \frac{1}{8}x_2^{(n)} + \frac{1}{4}x_3^{(n)} + \frac{1}{2}x_4^{(n)} + \cdots, \end{aligned}$$

etc. Thus,

$$I - Q^{-1}A = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ 0 & \frac{1}{4} & \frac{1}{2} & & \\ 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

3.6 Acceleration methods

3.6.1 The extrapolation method

Consider a general iterative method for linear systems

$$x_{n+1} = Gx_n + c.$$

For the system $Ax = b$ we had $G = (I - Q^{-1}A)$ and $c = Q^{-1}b$, but for now this does not matter. We know that the iteration will converge if $\text{spr } G < 1$.

Consider now the one-parameter family of methods,

$$\begin{aligned} x_{n+1} &= \gamma(Gx_n + c) + (1 - \gamma)x_n \\ &= [\gamma G + (1 - \gamma)I]x_n + \gamma c \stackrel{\text{def}}{=} G_\gamma x_n + \gamma c, \end{aligned}$$

$\gamma \in \mathbb{R}$. Can we choose γ such to optimize the rate of convergence, i.e., such to minimize the spectral radius of G_γ ? Note that (1) if the method converges then it converges to the desired solution, and (2) $\gamma = 1$ reduces to the original procedure.

Recall that (1) the spectral radius is the largest eigenvalue (in absolute value), and that (2) if $\lambda \in \Sigma(A)$ and $p(\lambda) \in \Sigma(p(A))$ for any polynomial p . Suppose that we even don't really know the eigenvalues of the original matrix G , but we only know that they are real (true for symmetric or Hermitian matrices) and within the segment $[a, b]$. Then, the spectrum of G_γ lies within

$$\Sigma(G_\gamma) \subseteq \{\gamma z + (1 - \gamma) : z \in [a, b]\}.$$

This means that

$$\text{spr } G_\gamma \leq \max_{a \leq \lambda \leq b} |\gamma \lambda + (1 - \gamma)|.$$

The expression on the right-hand side is the quantity we want to minimize,

$$\gamma^* = \arg \min_{\gamma \in \mathbb{R}} \max_{a \leq z \leq b} |\gamma z + (1 - \gamma)|.$$

Problems of this type are called **min-max problems**. They are very common in optimization.

Theorem 3.10 If $1 \notin [a, b]$, then

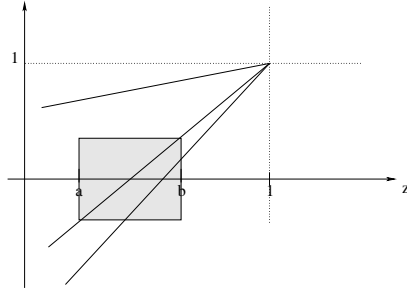
$$\gamma^* = \frac{2}{2 - a - b},$$

and

$$\text{spr } G_\gamma^* \leq 1 - |\gamma^*|d,$$

where $d = \text{dist}(1, [a, b])$.

Proof: Since $1 \notin [a, b]$, then we either have $b < 1$ or $a > 1$. Let's focus on the first case; the second case is treated the same way. The solution to this problem is best viewed graphically:



From the figure we see that the optimal γ is when the absolute values of the two extreme cases coincide, i.e., when

$$\gamma(a - 1) + 1 = -\gamma(b - 1) - 1,$$

from which we readily obtain $2 = (2 - a - b)\gamma^*$. Substituting the value of γ^* into

$$\max_{a \leq z \leq b} |\gamma z + (1 - \gamma)|,$$

whose maximum is attained at either $z = a, b$, we get

$$\text{spr } G_{\gamma^*} \leq \gamma^*(b - 1) + 1 = 1 - |\gamma^*|d,$$

since γ^* is positive and $d = 1 - b$. ■

Example 3.9 The method of extrapolation can be of use even if the original method does not converge, i.e., even if $\text{spr } G > 1$. Consider for example the following iterative method for solving the linear systems $Ax = b$,

$$x_{n+1} = (I - A)x_n + b.$$

It is known as Richardson's method. If we know that A has real eigenvalues ranging between λ_{\min} and λ_{\max} , then in the above notation

$$a = 1 - \lambda_{\max} \quad \text{and} \quad b = 1 - \lambda_{\min}.$$

If $1 \notin [a, b]$, i.e., all the eigenvalues of A have the same sign, then This means that the optimal extrapolation method is

$$x_{n+1} = [\gamma^*(I - A) + (1 - \gamma^*)I] x_n + \gamma^* b,$$

where

$$\gamma^* = \frac{2}{\lambda_{\max} + \lambda_{\min}}.$$

Suppose that $\lambda_{\min} > 0$, then the spectral radius of the resulting iteration matrix is bounded by

$$\text{spr } G_{\gamma^*} \leq 1 - \frac{2\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

It is easy to see that the bounds remains unchanged if $\lambda_{\max} < 0$.

3.6.2 Chebyshev acceleration

Chebyshev's acceleration method takes the idea even further. Suppose we have an iterative method,

$$x_{n+1} = Gx_n + c,$$

and that we have used it to generate the sequence x_0, x_1, \dots, x_n . Can we use this existing sequence to get even closer to the solution? Specifically, consider a linear combination,

$$u_n = \sum_{k=0}^n a_{n,k} x_k.$$

We want to optimize this expression, with respect to the coefficients $a_{n,k}$ such that u_n is as close as possible to the fixed point $x = Gx + c$. Assume that for all n ,

$$\sum_{k=0}^n a_{n,k} = 1.$$

Then,

$$u_n - x = \sum_{k=0}^n a_{n,k} x_k - x = \sum_{k=0}^n a_{n,k} (x_k - x).$$

Now, since $(x_k - x) = (Gx_{k-1} + c) - (Gx + c) = G(x_{k-1} - x)$, repeated application of this recursion gives

$$u_n - x = \sum_{k=0}^n a_{n,k} G^k (x_0 - x) \stackrel{\text{def}}{=} p_n(G)(x_0 - x),$$

where $p_n(z) = \sum_{k=0}^n a_{n,k} z^k$. Optimality will be achieved if we take the coefficients $a_{n,k}$ such to minimize the norm of $p_n(G)$, or instead, its spectral radius. Note that

$$\text{spr } p_n(G) = \max_{z \in \Sigma(p_n(G))} |z| = \max_{z \in \Sigma(G)} |p_n(z)|.$$

Suppose all we knew was that the eigenvalues of G lie in a set S . Then, our goal is to find a polynomial of degree n , satisfying $p_n(1) = 1$, which minimizes

$$\max_{z \in S} |p_n(z)|.$$

That is, we are facing another min-max problem,

$$p_n^* = \arg \min_{p_n} \max_{z \in S} |p_n(z)|.$$

This can be quite a challenging problem. We will solve it again for the case where the spectrum of G is real, and confined to the set $S = [a, b]$.

Definition 3.12 (Chebyshev polynomials) The Chebyshev polynomials, $T_k(x)$, $k = 0, 1, \dots$, are a family of polynomials defined recursively by

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2x T_n(x) - T_{n-1}(x). \end{aligned}$$

Applying the iterative relation we have

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1. \end{aligned}$$

Note that for $y \in [-1, 1]$, we can express y as $\cos x$, in which case

$$\begin{aligned} T_2(y) &= T_2(\cos x) = 2 \cos^2 x - 1 = \cos 2x = \cos(2 \cos^{-1} y) \\ T_3(y) &= T_3(\cos x) = 4 \cos^3 x - 3 \cos x = \cos 3x = \cos(3 \cos^{-1} y), \end{aligned}$$

and so on. This suggests the following relation:

Lemma 3.13 For $x \in [-1, 1]$ the Chebyshev polynomials have the following explicit representation:

$$T_n(x) = \cos(n \cos^{-1} x).$$

Proof: We have the following relations,

$$\begin{aligned} \cos[(n+1)\theta] &= \cos \theta \cos n\theta - \sin \theta \sin n\theta \\ \cos[(n-1)\theta] &= \cos \theta \cos n\theta + \sin \theta \sin n\theta, \end{aligned}$$

which upon addition gives

$$\cos[(n+1)\theta] = 2 \cos \theta \cos n\theta - \cos[(n-1)\theta].$$

Set now $x = \cos \theta$, we get

$$\cos[(n+1)\cos^{-1} x] = 2x \cos[n\cos^{-1} x] - \cos[(n-1)\cos^{-1} x],$$

i.e., the functions $\cos[n\cos^{-1} x]$ satisfy the same recursion relations as the Chebyshev polynomials. It only remains to verify that they are identical for $n = 0, 1$. ■

Properties of the Chebyshev polynomials

- ① $T_n(x)$ is a polynomial of degree n .
- ② $|T_n(x)| \leq 1$ for $x \in [-1, 1]$.
- ③ For $j = 0, 1, \dots, n$,

$$T_n\left(\cos \frac{j\pi}{n}\right) = \cos(j\pi) = (-1)^j.$$

These are the extrema of $T_n(x)$.

- ④ For $j = 1, 2, \dots, n$,

$$T_n\left(\cos \frac{(j - \frac{1}{2})\pi}{n}\right) = \cos\left((j - \frac{1}{2})\pi\right) = 0.$$

That is, the n -th Chebyshev polynomial has n real-valued roots and *all* reside within the segment $[-1, 1]$.

Proposition 3.6 Let $p_n(z)$ be a polynomial of degree n with $p(\tilde{z}) = 1$, $\tilde{z} \notin [-1, 1]$. Then

$$\max_{-1 \leq z \leq 1} |p_n(z)| \geq \frac{1}{|T_n(\tilde{z})|}.$$

Equality is satisfied for $p_n(z) = T_n(z)/T_n(\tilde{z})$.

This proposition states that given that p_n equals one at a point z_n , there is a limit on how small it can be in the interval $[-1, 1]$. The Chebyshev polynomials are optimal, within the class of polynomials of the same degree, in that they can fit within a strip of minimal width.

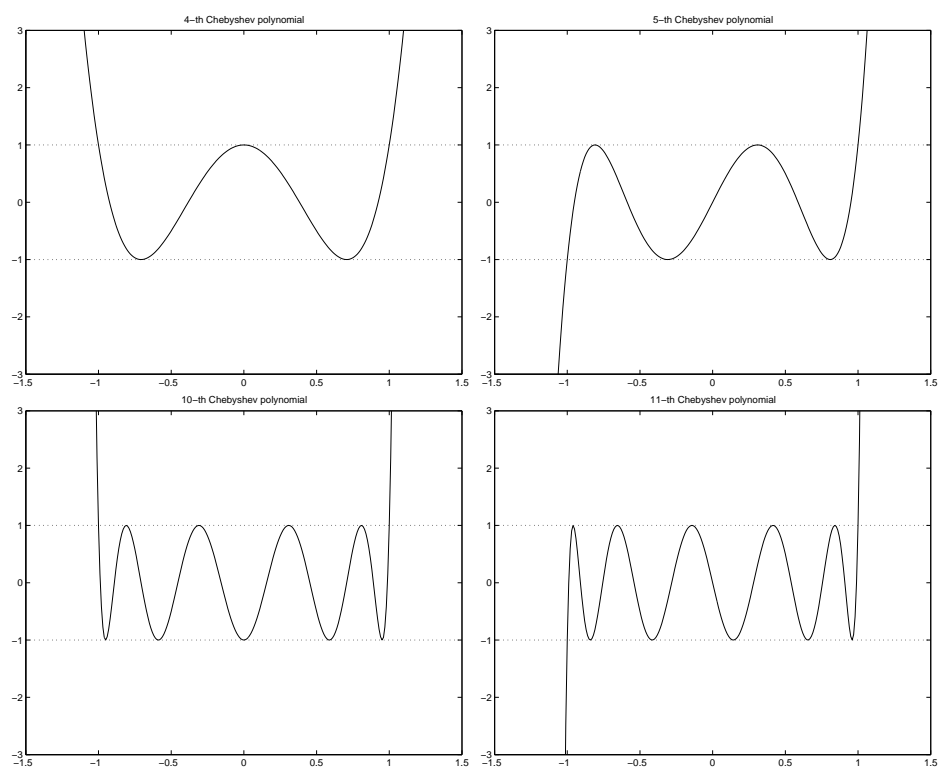


Figure 3.1: The functions $T_4(x)$, $T_5(x)$, $T_{10}(x)$, and $T_{11}(x)$.

Proof: Consider the $n + 1$ points $z_i = \cos(i\pi/n) \in [-1, 1]$, $i = 0, 1, \dots, n$. Recall that these are the extrema of the Chebyshev polynomials, $T_n(z_i) = (-1)^i$.

We now proceed by contradiction, and assume that

$$\max_{-1 \leq z \leq 1} |p_n(z)| < \frac{1}{|T_n(\tilde{z})|}.$$

If this holds, then a-fortiori,

$$|p_n(z_i)| - \frac{1}{|T_n(\tilde{z})|} < 0, \quad i = 0, 1, \dots, n.$$

This can be re-arranged as follows

$$\operatorname{sgn}[T_n(\tilde{z})](-1)^i p_n(z_i) - \frac{(-1)^i T_n(z_i)}{\operatorname{sgn}[T_n(\tilde{z})] T_n(\tilde{z})} < 0,$$

or,

$$\operatorname{sgn}[T_n(\tilde{z})](-1)^i \left[p_n(z_i) - \frac{T_n(z_i)}{T_n(\tilde{z})} \right] < 0.$$

Consider now the function

$$f(z) = p_n(z) - \frac{T_n(z)}{T_n(\tilde{z})}.$$

It is a polynomial of degree at most n ; its sign alternates at the z_i , implying the presence of n roots on the interval $[-1, 1]$; it has a root at $z = \tilde{z}$. This is impossible, contradicting the assumption. ■

Proposition 3.7 Let $p_n(z)$ be a polynomial of degree n , $p_n(1) = 1$, and let a, b be real numbers such that $1 \notin [a, b]$. Then,

$$\max_{a \leq z \leq b} |p_n(z)| \geq \frac{1}{|T_n(w(1))|},$$

where

$$w(z) = \frac{2z - b - a}{b - a}.$$

Equality is obtained for $p_n(z) = T_n(w(z))/T_n(w(1))$.

Note that a polynomial of degree n composed with a linear function is still a polynomial of degree n ,

Proof: Take the case $a < b < 1$. Then,

$$w(1) = \frac{2 - b - a}{b - a} = 1 + 2 \frac{1 - b}{b - a} \stackrel{\text{def}}{=} \tilde{w} > 1.$$

The converse relation is

$$z(w) = \frac{1}{2}[(b - a)w + a + b],$$

and $z(\tilde{w}) = 1$.

Let p_n be a polynomial of degree n satisfying $p_n(1) = 1$, and define $q_n(w) = p_n(z(w))$. We have $q_n(\tilde{w}) = p_n(1) = 1$, hence, by the previous proposition,

$$\max_{-1 \leq w \leq 1} |q_n(w)| \geq \frac{1}{|T_n(\tilde{w})|},$$

Substituting the definition of q_n , this is equivalent to

$$\max_{-1 \leq w \leq 1} |p_n(z(w))| = \max_{a \leq z \leq b} |p_n(z)| \geq \frac{1}{|T_n(\tilde{w})|}.$$

■

We have thus shown that among all polynomials of degree n satisfying $p_n(1) = 1$, the one that minimizes its maximum norm in the interval $[a, b]$ is

$$p_n(z) = \frac{T_n(w(z))}{T_n(w(1))}, \quad \text{with} \quad w(z) = \frac{2z - b - a}{b - a}.$$

What does this have to do with acceleration methods? Recall that we assume the existence of an iterative procedure,

$$x_{n+1} = Gx_n + c,$$

where $\Sigma G \in [a, b]$, and we want to improve it by taking instead

$$u_n = \sum_{k=0}^n a_{n,k} x_k,$$

where $\sum_{k=0}^n a_{n,k} = 1$. We've seen that this amounts to an iterative method with iteration matrix $p_n(G)$, where p_n is the polynomial with coefficients $a_{n,k}$. Thus, what we want is to find the polynomial that minimizes

$$\max_{a \leq z \leq b} |p_n(z)|,$$

and now we know which it is. This will ensure that

$$\text{error}(n) \leq \frac{\text{error}(0)}{|T_n(w(1))|},$$

and the right hand side decays exponentially fast in n . We are still facing a practical problem of implementation. This will be dealt with now.

Lemma 3.14 The family of polynomials $p_n(z) = \frac{T_n(w(z))}{T_n(w(1))}$ can be constructed recursively as follows:

$$\begin{aligned} p_0(z) &= 1 \\ p_1(z) &= \frac{2z - b - a}{2 - b - a} \\ p_n(z) &= \sigma_n p_1(z) p_{n-1}(z) + (1 - \sigma_n) p_{n-2}, \end{aligned}$$

where the constants σ_n are defined by

$$\sigma_1 = 2 \quad \sigma_n = \left(1 - \frac{\sigma_{n-1}}{2[w(1)]^2}\right)^{-1}.$$

Proof: By the recursive property of the Chebyshev polynomials,

$$T_n(w(z)) = 2w(z) T_{n-1}(w(z)) - T_{n-2}(w(z)).$$

Dividing by $T_n(w(1))$, and converting T_k 's into p_k 's:

$$p_n(z) = \frac{2w(1) T_{n-1}(w(1))}{T_n(w(1))} p(z) p_{n-1}(w(z)) - \frac{T_{n-2}(w(1))}{T_n(w(1))} T_{n-2}(w(z)).$$

It remains to show that

$$\rho_n \stackrel{\text{def}}{=} \frac{2w(1) T_{n-1}(w(1))}{T_n(w(1))} = \sigma_n \quad \text{and} \quad -\frac{T_{n-2}(w(1))}{T_n(w(1))} = 1 - \sigma_n.$$

That their sum is indeed one follows from the Chebyshev recursion relation. It is also obvious that $\rho_1 = 2$. Finally,

$$\begin{aligned} \rho_{n-1} &= \frac{2w(1) T_{n-2}(w(1))}{T_{n-1}(w(1))} \\ &= \frac{2w(1) \frac{T_{n-2}(w(1))}{T_n(w(1))} T_n(w(1))}{\frac{2w(1) T_{n-1}(w(1))}{T_n(w(1))} \frac{T_n(w(1))}{2w(1)}} \\ &= -[2w(1)]^2 \frac{1 - \rho_n}{\rho_n}. \end{aligned}$$

It only remains to invert this relation. ■

Theorem 3.11 The sequence (u_n) of Chebyshev's acceleration's method can be constructed as follows:

$$\begin{aligned} u_1 &= \gamma(Gx_0 + c) + (1 - \gamma)x_0 \\ u_n &= \sigma_n [\gamma(Gu_{n-1} + c) + (1 - \gamma)u_{n-1}] + (1 - \sigma_n)u_{n-2}, \end{aligned}$$

where $\gamma = 2/(2 - b - a)$ and the σ_n are as above.

Comments:

- ① The (u_n) are constructed directly without generating the (x_n) .
- ② The first step is extrapolation, and the next ones are “weighted extrapolations”.
- ③ The Chebyshev polynomials are not apparent (they are hiding...).

Proof: Start with $n = 1$,

$$u_1 = a_{1,1}x_1 + a_{1,0}x_0 = a_{1,1}(Gx_0 + c) + a_{1,0}x_0.$$

The coefficients $a_{1,0}$ and $a_{1,1}$ are the coefficients of the polynomial $p_1(z)$. By Lemma 3.14,

$$a_{1,1} = \frac{2}{2 - b - a} = \gamma \quad a_{1,0} = -\frac{a + b}{2 - b - a} = 1 - \gamma.$$

Now to the n -th iterate. Recall that

$$u_n = \sum_{k=0}^n a_{n,k}x_k = x + \sum_{k=0}^n a_{n,k}(x_k - x) = x + p_n(G)(x_0 - x).$$

By Lemma 3.14,

$$p_n(G) = \sigma_n p_1(G) p_{n-1}(G) + (1 - \sigma_n) p_{n-2}(G),$$


and $p_1(G) = \gamma G + (1 - \gamma)I$. Applying this on $x_0 - x$ we get

$$\begin{aligned} u_n - x &= \sigma_n [\gamma G + (1 - \gamma)I] (u_{n-1} - x) + (1 - \sigma_n)(u_{n-2} - x) \\ &= \sigma_n [\gamma Gu_{n-1} + (1 - \gamma)u_{n-1}] - \sigma_n [\gamma Gx + (1 - \gamma)x] \\ &\quad + (1 - \sigma_n)u_{n-2} - (1 - \sigma_n)x. \end{aligned}$$

It remains to gather the terms multiplying x . Since $x = Gx + c$ is a fixed point,

$$-\sigma_n [\gamma Gx + (1 - \gamma)x] - (1 - \sigma_n)x = \sigma_n \gamma c - x.$$

Substituting into the above we get the desired result. ■

 *Computer exercise 3.4* The goal is to solve the system of equations:

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -4 \\ 0 \\ 4 \\ -4 \end{pmatrix}.$$

- ① Write explicitly the Jacobi iterative procedure,

$$x^{k+1} = Gx^k + c.$$

- ② What is the range of eigenvalues of the matrix G ?
 ③ Is the Jacobi iterative procedure convergent?
 ④ Write an algorithm for the Chebyshev acceleration method based on Jacobi iterations.
 ⑤ Implement both procedures and compare their performance.

3.7 The singular value decomposition (SVD)

Relevant, among other things, to the mean-square minimization: find $x \in \mathbb{R}^n$ that minimizes $\|Ax - b\|_2$, where $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$, with $m > n$ (more equations than unknowns). It has many other uses.

Since we are going to consider vectors in \mathbb{R}^m and \mathbb{R}^n , and operators between these two spaces, we will use the notation $\|\cdot\|_m$ and $\|\cdot\|_n$ for the corresponding vector 2-norms. Similarly, we will use $\|\cdot\|_{m \times n}$, etc., for the operator 2-norms. We will also use I_m, I_n to denote the identity operators in the two spaces.

Recall that the norm of an m -by- n matrix (it will always be assumed that $m \geq n$) is defined by

$$\|A\|_{m \times n} = \sup_{\|x\|_n=1} \|Ax\|_m = \sup_{(x,x)_n=1} \sqrt{(Ax, Ax)_m}.$$

A matrix Q is called **orthogonal** if its columns form an orthonormal set. If the matrix is n -by- n , then its columns form a basis in \mathbb{R}^n , and $Q^T Q = I_n$. Since Q is invertible, it immediately follows that $Q^T = Q^{-1}$, hence $QQ^T = I_n$ as well. If Q is an m -by- n orthogonal matrix, then $Q^T Q = I_n$, but the m -by- m matrix QQ^T is not an identity.

Lemma 3.15 Let $x \in \mathbb{R}^n$, and Q be an orthogonal m -by- n matrix, $m \geq n$, then $\|Qx\|_m = \|x\|_n$.

Proof: This is immediate by

$$\|Qx\|_m^2 = (Qx, Qx)_m = (x, Q^T Qx)_n = (x, x)_n = \|x\|_n^2.$$

■

Lemma 3.16 Let A be an n -by- n matrix, V be an orthogonal n -by- n matrix, and U be an orthogonal m -by- n matrix. Then,

$$\|UAV^T\|_{m \times n} = \|A\|_{n \times n}.$$

Proof: By definition,

$$\begin{aligned} \|UAV^T\|_{m \times n}^2 &= \sup_{(x,x)_n=1} (UAV^T x, UAV^T x)_m \\ &= \sup_{(x,x)_n=1} (AV^T x, AV^T x)_n \\ &= \sup_{(y,y)_n=1} (Ay, Ay)_n \\ &= \|A\|_{n \times n}^2, \end{aligned}$$

where we have used the previous lemma in the passage from the first to the second line, and the fact that x on the unit sphere can be expressed as Vy , with y on the unit sphere. ■

Theorem 3.12 (SVD decomposition) Let A be an m -by- n matrix, $m \geq n$. Then, A can be decomposed as

$$A = U\Sigma V^T,$$

where U is an m -by- n orthogonal matrix, V is an n -by- n orthogonal matrix, and Σ is an n -by- n diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$.

The columns of U , u_i , are called the **left singular vectors**, the columns of V , v_i , are called the **right singular vectors**, and the σ_i are called the **singular values**. This theorem states that in some sense “every matrix is diagonal”. Indeed, for every right singular vector v_i ,

$$Av_i = U\Sigma V^T v_i = U\Sigma e_i = \sigma_i Ue_i = \sigma_i u_i.$$

Thus, it is always possible to find an orthogonal basis $\{v_i\}$ in \mathbb{R}^n , and an orthogonal set $\{u_i\}$ in \mathbb{R}^m , such that any $x = \sum_{i=1}^n a_i v_i$ is mapped into $Ax = \sum_{i=1}^n \sigma_i a_i u_i$.

Proof: The proof goes by induction, assuming this can be done for an $(m-1)$ -by- $(n-1)$ matrix. The basis of induction is a column vector, which can always be represented as a normalized column vector, times its norm, times one.

Let then A be given, and set v to be a vector on the unit sphere, $\|v\|_n = 1$, such that $\|Av\|_m = \|A\|_{m \times n}$ (such a vector necessarily exists). Set then $u = Av/\|Av\|_m$, which is a unit vector in \mathbb{R}^m . We have one vector $u \in \mathbb{R}^m$, which we complete (by Gram-Schmidt orthonormalization) into an orthogonal basis $U = (u, \tilde{U}) \in \mathbb{R}^{m \times m}$, $U^T U = U U^T = I_m$. Similarly, we complete $v \in \mathbb{R}^n$ into an orthonormal basis $V = (v, \tilde{V}) \in \mathbb{R}^{n \times n}$. Consider the m -by- n matrix

$$U^T A V = \begin{pmatrix} u^T \\ \tilde{U}^T \end{pmatrix} A \begin{pmatrix} v & \tilde{V} \end{pmatrix} = \begin{pmatrix} u^T A v & u^T A \tilde{V} \\ \tilde{U}^T A v & \tilde{U}^T A \tilde{V} \end{pmatrix}.$$

Note that $u \in \mathbb{R}^m$, $\tilde{U} \in \mathbb{R}^{m \times (m-1)}$, $v \in \mathbb{R}^n$ and $\tilde{V} \in \mathbb{R}^{n \times (n-1)}$. Hence, $u^T A v \in \mathbb{R}$, $u^T A \tilde{V} \in \mathbb{R}^{1 \times (n-1)}$, $\tilde{U}^T A v \in \mathbb{R}^{(m-1) \times 1}$, and $\tilde{U}^T A \tilde{V} \in \mathbb{R}^{(m-1) \times (n-1)}$.

Now,

$$u^T A v = \|Av\|_m u^T u = \|A\|_{m \times n} \stackrel{\text{def}}{=} \sigma,$$

and

$$\tilde{U}^T A v = \|Av\|_m \tilde{U}^T u = 0,$$

due to the orthogonality of u and each of the rows of \tilde{U} . Thus,

$$U^T A V = \begin{pmatrix} \sigma & w^T \\ 0 & A_1 \end{pmatrix},$$

where $w^T = u^T A \tilde{V}$ and $A_1 = \tilde{U}^T A \tilde{V}$. We are going to prove that $w = 0$ as well. On the one hand we have

$$\left\| U^T A V \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 = \left\| \begin{pmatrix} \sigma^2 + w^T w \\ A_1 w \end{pmatrix} \right\|_m^2 \geq (\sigma^2 + w^T w)^2.$$

On the other hand

$$\left\| U^T A V \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 \leq \|U^T A V\|_{m \times n}^2 \left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 = \|A\|_{m \times n}^2 (\sigma^2 + w^T w),$$

where we have used the above lemma for $\|U^T A V\|_{m \times n}^2 = \|A\|_{m \times n}^2$. Since $\|A\|_{m \times n}^2 = \sigma^2$, it follows from these two inequalities that

$$(\sigma^2 + w^T w)^2 \leq \sigma^2 (\sigma^2 + w^T w) \quad \rightarrow \quad w^T w (\sigma^2 + w^T w) \leq 0,$$

i.e., $w = 0$ as claimed.

Thus,

$$U^T AV = \begin{pmatrix} \sigma & 0 \\ 0 & A_1 \end{pmatrix},$$

At this stage, we use the inductive hypothesis for matrices of size $(m-1) \times (n-1)$, and write $A_1 = U_1 \Sigma_1 V_1^T$, which gives,

$$U^T AV = \begin{pmatrix} \sigma & 0 \\ 0 & U_1 \Sigma_1 V_1^T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V_1 \end{pmatrix}^T,$$

hence

$$A = \begin{bmatrix} U & 0 \\ 0 & U_1 \end{bmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \begin{bmatrix} V & 0 \\ 0 & V_1 \end{bmatrix}^T.$$

It remains to show that σ is larger or equal to all the diagonal entries of Σ , but this follows at once from the fact that

$$\sigma = \|A\|_{m \times n} = \left\| \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \right\|_{n \times n} = \max_i \left| \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix}_{ii} \right|.$$

This concludes the proof. ■

Comment: SVD provides an interpretation of the action of A on a vector x :

- ① Rotate (by V^T).
- ② Stretch along axes by σ_i .
- ③ Pad the vector with $m - n$ zeros.
- ④ Rotate (by U).

Having proved the existence of such a decomposition, we turn to prove a number of algebraic properties of SVD.

Theorem 3.13 Let $A = U \Sigma V^T$ be an SVD of the m -by- n matrix A . Then,

- ① If A is square symmetric with eigenvalues λ_i , and orthogonal diagonalizing transformation $U = (u_1, \dots, u_n)$, i.e., $A = U \Lambda U^T$, then an SVD of A is with $\sigma_i = |\lambda_i|$, the same U , and V with columns $v_i = \text{sgn}(\lambda_i) u_i$.
- ② The eigenvalues of the n -by- n (symmetric) matrix $A^T A$ are σ_i^2 , and the corresponding eigenvectors are the right singular vectors v_i .

- ③ The eigenvalues of the m -by- m (symmetric) matrix AA^T are σ_i^2 and $m - n$ zeros. The corresponding eigenvectors are the left singular vectors supplemented with a set of $m - n$ orthogonal vectors.
- ④ If A has full rank (its columns are independent), then the vector $x \in \mathbb{R}^n$ that minimizes $\|Ax - b\|_m$ is $x = V\Sigma^{-1}U^T b$. The matrix

$$V\Sigma^{-1}U^T$$

is called the **pseudo-inverse** of A .

- ⑤ $\|A\|_{m \times n} = \sigma_1$. If, furthermore, A is square and non-singular then $\|A^{-1}\|_{n \times n} = 1/\sigma_n$, hence the condition number is σ_1/σ_n .
- ⑥ Suppose that $\sigma_1 \geq \sigma_n \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$. Then the rank of A is r , and

$$\begin{aligned} \text{null } A &= \text{span}(v_{r+1}, \dots, v_n) \\ \text{range } A &= \text{span}(u_1, \dots, u_r). \end{aligned}$$

- ⑦ Write $V = (v_1, \dots, v_n)$ and $U = (u_1, \dots, u_n)$. Then,

$$A = \sum_{i=1}^n \sigma_i u_i v_i^T,$$

i.e., it is a sum of rank-1 matrices. The matrix of rank $k < n$ that is closest (in norm) to A is

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T,$$

and $\|A - A_k\|_2 = \sigma_{k+1}$. That is, the dyads $u_i v_i^T$ are ranked in “order of importance”. A_k can also be written as

$$A_k = U\Sigma_k V^T,$$

where $\Sigma_k = \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0)$.

Proof:

- ① This is obvious.
- ② We have

$$A^T A = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T,$$

where we have used the fact that $U^T U = I_m$. This is an eigen-decomposition of $A^T A$.

③ First,

$$AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^T U^T.$$

Take an m -by- $(m-n)$ matrix \tilde{U} such that (U, \tilde{U}) is orthogonal (use Gram-Schmidt). Then, we can also write

$$AA^T = (U, \tilde{U}) \begin{pmatrix} \Sigma^T \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U^T \\ \tilde{U}^T \end{pmatrix}.$$

This is precisely an eigen-decomposition of AA^T .

④ We need to minimize $\|Ax - b\|_2 = \|U\Sigma V^T x - b\|_2$. Since A has full rank, so does Σ , hence it is invertible. Let $(U, \tilde{U}) \in \mathbb{R}^{m \times m}$ be as above, then

$$\begin{aligned} \|U\Sigma V^T x - b\|_2^m &= \left\| \begin{pmatrix} U^T \\ \tilde{U}^T \end{pmatrix} (U\Sigma V^T x - b) \right\|_m^2 \\ &= \left\| \begin{pmatrix} \Sigma V^T x - U^T b \\ -\tilde{U}^T b \end{pmatrix} \right\|_m^2 \\ &= \|\Sigma V^T x - U^T b\|_n^2 + \|\tilde{U}^T b\|_{m-n}^2. \end{aligned}$$

The second term does not depend on x , and the first can be made zero by choosing

$$x = V\Sigma^{-1}U^T b.$$

⑤ Since $\|A\|_{m \times n} = \|\Sigma\|_{n \times n}$, the first statement is obvious. If A is invertible, then $A^{-1} = V\Sigma^{-1}U^T$, hence $\|A^{-1}\|_{n \times n} = \|\Sigma^{-1}\|_{n \times n}$, and the second statement is equally obvious.

⑥ Recall that

$$A : \sum_i a_i v_i \mapsto \sum_i a_i \sigma_i u_i.$$

Then, clearly the range of A is the span of all those u_i for which $\sigma_i > 0$ and its null space is the span of all those v_i for which $\sigma_i = 0$.

⑦ A_k has rank k because it is a sum of k rank-1 matrices, and

$$\|A - A_k\|_{m \times n} = \left\| \sum_{i=k+1}^n \sigma_i u_i v_i^T \right\|_{m \times n} = \left\| U \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & \sigma_{k+1} & \\ & & & \ddots \\ & & & & \sigma_n \end{pmatrix} V^T \right\|_{m \times n} = \sigma_{k+1}.$$


We need to show that there is no closer matrix of rank k . Let B be any rank- k matrix, so that its null space has dimension $n - k$. The space spanned by (v_1, \dots, v_{k+1}) has dimension $k + 1$, hence it must have a non-zero intersection with the null space of B . Let x be a unit vector in this intersection,

$$x \in \text{null } B \cap \text{span}(v_1, \dots, v_{k+1}), \quad \|x\|_n = 1.$$

Then,

$$\begin{aligned} \|A - B\|_{m \times n}^2 &\geq \|(A - B)x\|_m^2 = \|Ax\|_m^2 = \|U\Sigma V^T x\|_m^2 \\ &= \|\Sigma V^T x\|_n^2 \geq \sigma_{k+1}^2 \|V^T x\|_n^2 = \sigma_{k+1}^2, \end{aligned}$$

where we have used the fact that $(V^T x)$ has its last $n - k - 1$ entries zero. ■

 **Exercise 3.25** Let $A = U\Sigma V^T$ be an SVD for the m -by- n matrix A . What are the SVDs for the following matrices:

- ① $(A^T A)^{-1}$.
- ② $(A^T A)^{-1} A^T$.
- ③ $A(A^T A)^{-1}$.
- ④ $A(A^T A)^{-1} A^T$.

Solution 3.25:

- ① The matrix $A^T A$ is n -by- n , and has an SVD of the form $A^T A = V\Sigma^2 V^T$. Its inverse is $(A^T A)^{-1} = V\Sigma^{-2} V^T$, which is almost an SVD, except for the singular values being in increasing order. Let P be a permutation matrix that switches the first row with the last, the second with the $(n - 1)$ -st, etc. It is a symmetric matrix, i.e., $P^T = P = P^{-1}$. Then,

$$(A^T A)^{-1} = (VP)(P\Sigma^{-2}P)(VP)^T$$

is an SVD.

- ② The matrix $(A^T A)^{-1} A^T$ is n -by- m . For such matrices, we SVD its transpose,

$$A(A^T A)^{-T} = U\Sigma V^T V\Sigma^{-2} V^T = (UP)(P\Sigma^{-1}P)(VP)^T.$$

- ③ Same as the previous example.
- ④ The matrix $A(A^T A)^{-1} A^T$ is m -by- m :

$$A(A^T A)^{-1} A^T = U\Sigma V^T V\Sigma^{-2} V^T V\Sigma U^T = UI_m U^T.$$

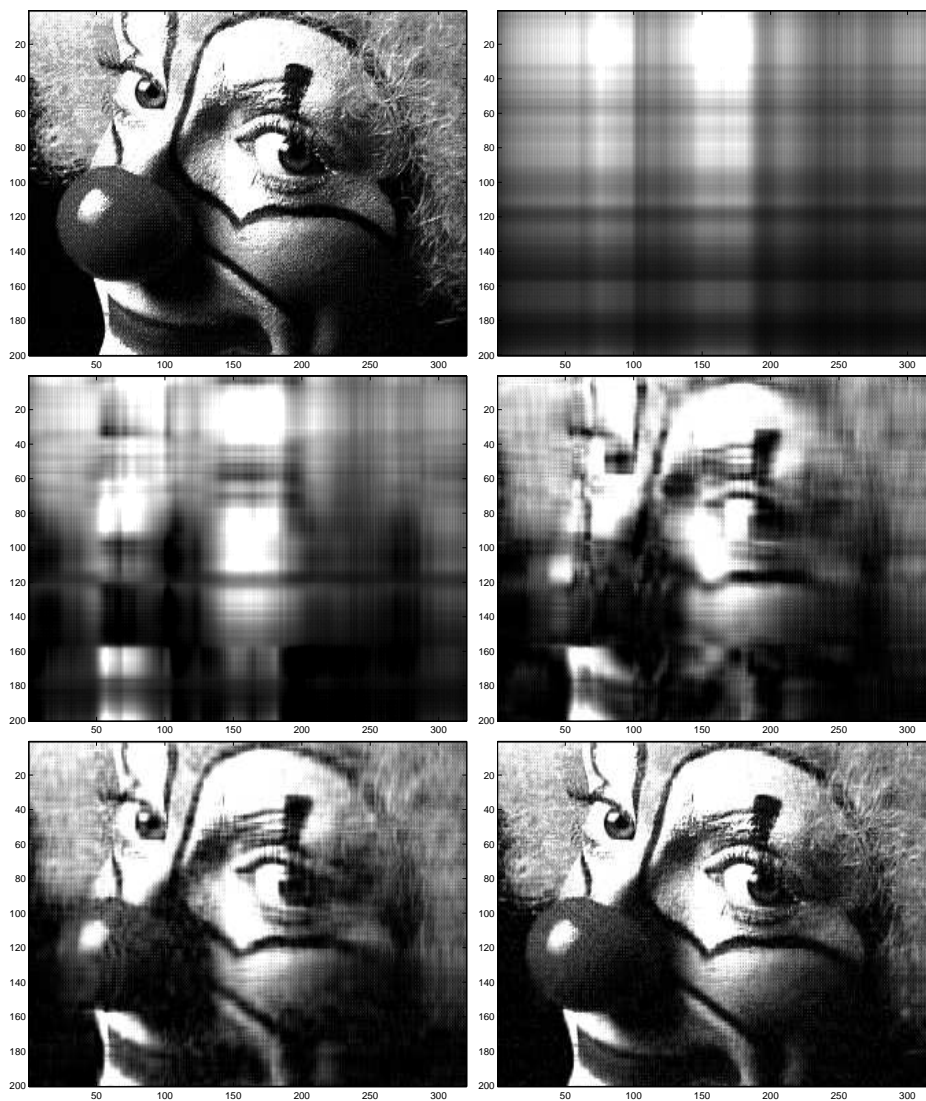





Figure 3.2: (a) Full 320×200 image, (b) $k = 1$, (c) $k = 3$, (d) $k = 10$, (e) $k = 20$, (f) $k = 50$.

 *Exercise 3.26* Decompose the following matrix

$$\begin{pmatrix} 2 & 6 & -4 \\ 6 & 17 & -17 \\ -4 & -17 & -20 \end{pmatrix}$$

into a product of the form LDL^T , where D is diagonal.

 *Exercise 3.27* Write an algorithm for Cholesky factorization (that is, an algorithm that calculates L , so that $LL^T = A$, where A is symmetric, positive-definite).


 *Exercise 3.28* Let $A = I - L - U$ where the matrices L and U are strictly lower and upper diagonal, respectively. Consider the following iterative procedure for solving the linear system $Ax = b$:

$$x_{k+1} = \bar{b} + x_k - (I - U)^{-1}(I - L)^{-1}Ax_k,$$

where

$$\bar{b} = (I - U)^{-1}(I - L)^{-1}b.$$

- (i) Prove that if the procedure converges it converges to the right solution.
- (ii) Explain why this scheme does not require the inversion of $(I - U)$ and $(I - L)$.

 *Exercise 3.29* Prove that if B_n is an approximation to the matrix A^{-1} , i.e., $B_n = A^{-1}(I - E_n)$ and $\|E_n\|$ is “small”, then

$$B_{n+1} = B_n(2I - AB_n)$$

is an even better approximation. How small should $\|E_0\|$ be for the sequence to converge?

Chapter 4

Interpolation

In this chapter we will consider the following question. What is the polynomial of lowest degree that agrees with certain data on its value and the values of its derivatives at given points. Viewing this polynomial as an approximation of a function satisfying the same constraints, we will estimate the error of this approximation.

4.1 Newton's representation of the interpolating polynomial

Suppose we are given $n + 1$ set of points in the plane,

$$\begin{array}{c|c|c|c} x_0 & x_1 & \cdots & x_n \\ \hline y_0 & y_1 & \cdots & y_n \end{array}$$

The goal is to find a polynomial of *least degree* which agrees with this data. Henceforth we will denote the set of polynomials of degree n or less by Π_n .

Theorem 4.1 Let x_0, x_1, \dots, x_n be $n + 1$ distinct points. For every set of values y_0, y_1, \dots, y_n , there exists a unique $p_n \in \Pi_n$ such that $p(x_i) = y_i$, $i = 0, 1, \dots, n$.

Proof: We start by proving uniqueness. Suppose that there exists $p_n, q_n \in \Pi_n$ satisfying

$$p_n(x_i) = q_n(x_i) = y_i, \quad i = 0, \dots, n.$$

Then the polynomial $r_n = p_n - q_n$ is in Π_n and satisfies

$$r_n(x_i) = 0, \quad i = 0, \dots, n,$$

hence it must be identically zero.

We then prove existence using induction on n . For $n = 0$ we choose

$$p_0(x) = y_0.$$

Suppose then the existence of a polynomial $p_{n-1} \in \Pi_{n-1}$ that interpolates the function at the points (x_0, \dots, x_{n-1}) . We take then

$$p_n(x) = p_{n-1}(x) + c(x - x_0)(x - x_1) \dots (x - x_{n-1}).$$

This polynomial is in Π_n , it agrees with p_{n-1} on the first n points. It only remains to require that

$$y_n = p_{n-1}(x_n) + c(x_n - x_0)(x_n - x_1) \dots (x_n - x_{n-1}),$$

i.e., take

$$c = \frac{y_n - p_{n-1}(x_n)}{\prod_{j=0}^{n-1} (x_n - x_j)}.$$

■

This proof is in fact constructive. Given $n + 1$ points $(x_i, y_i)_{i=0}^n$, we construct a sequence of interpolating polynomials:

$$p_0(x) = c_0$$

$$p_1(x) = c_0 + c_1(x - x_0)$$

$$p_2(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1),$$

and in general,

$$p_n(x) = \sum_{i=0}^n c_i \prod_{j=0}^{i-1} (x - x_j),$$

where the coefficients c_i are given by

$$c_i = \frac{y_i - p_{i-1}(x_i)}{\prod_{j=1}^{i-1} (x_i - x_j)}.$$

This representation of the (unique!) interpolating polynomials is known as **Newton's representation**.

The following example is only presented for didactic reasons since we will learn a much more efficient way to calculate the interpolating polynomial.

Example 4.1 Find the interpolating polynomial for the following data:

x	5	-7	-6	0
y	1	-23	-54	-954

4.2 Lagrange's representation

4.3 Divided differences

Recall how we construct Newton's interpolating polynomial: once we have a polynomial $p_{k-1} \in \Pi_{k-1}$ interpolating through the points (x_0, \dots, x_{k-1}) we proceed to construct $p_k \in \Pi_k$ by finding a constant c_k such that

$$y(x_k) = p_{k-1}(x_k) + c_k(x_k - x_{k-1}) \cdots (x_k - x_0).$$

The constant c_k is the coefficient of x^k in $p_k(x)$, which is the interpolating polynomial through the points (x_0, \dots, x_k) . Note that by construction, the constant c_k only depend on the choice of points (x_0, \dots, x_k) and the values of $y(x)$ at these points. We denote this constant by

$$y[x_0, \dots, x_k] \equiv \text{the coefficient of } x^k \text{ in the interpolating polynomial,}$$

hence Newton's interpolation formula can be written as

$$p_n(x) = \sum_{k=1}^n y[x_0, \dots, x_k] \left[\prod_{j=0}^{k-1} (x - x_j) \right].$$

The coefficients $y[x_0, \dots, x_k]$ are called the **divided differences** of $y(x)$. The reason for this name will be seen shortly.

Our goal in this section is to show a simple way of calculating the divided differences. Let us start with $k = 0$. In this case the coefficient of x^0 in the zeroth-order polynomial passing through $(x_0, y(x_0))$ is $y(x_0)$, i.e.,

$$y[x_0] = y(x_0).$$

Now to $k = 1$. The coefficient of x^1 is

$$y[x_0, x_1] = \frac{y(x_1) - y(x_0)}{x_1 - x_0} = \frac{y[x_1] - y[x_0]}{x_1 - x_0}.$$

Next to $k = 2$,

$$y[x_2] = y[x_0] + y[x_0, x_1](x_2 - x_0) + y[x_0, x_1, x_2](x_2 - x_0)(x_2 - x_1),$$

which we rearrange as

$$\begin{aligned} y[x_0, x_1, x_2] &= \frac{y[x_2] - y[x_0] - y[x_0, x_1](x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{y[x_2] - y[x_1] + y[x_1] - y[x_0] - y[x_0, x_1](x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{y[x_1, x_2](x_2 - x_1) + y[x_0, x_1](x_1 - x_0) - y[x_0, x_1](x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{y[x_1, x_2] - y[x_0, x_1]}{(x_2 - x_0)}. \end{aligned}$$

This is generalized into the following theorem:

Theorem 4.2 Divided differences satisfy the following recursive formula,

$$y[x_0, \dots, x_k] = \frac{y[x_1, \dots, x_k] - y[x_0, \dots, x_{k-1}]}{x_k - x_0}.$$

Proof: We know that $y[x_1, \dots, x_k]$ is the coefficient of x^{k-1} in q_{k-1} , which is the interpolating polynomial through (x_1, \dots, x_k) whereas $y[x_0, \dots, x_{k-1}]$ is the coefficient of x^{k-1} in p_{k-1} , which is the interpolating polynomial through (x_0, \dots, x_{k-1}) . Now, it is easily verified that

$$p_k(x) = q_{k-1}(x) + \frac{x - x_k}{x_k - x_0}[q_{k-1}(x) - p_{k-1}(x)].$$

This completes the proof. ■

AND NOW SHOW HOW TO CALCULATE.


Example 4.2 Find the interpolating polynomial for the following data

x	5	-7	-6	0
y	1	-23	-54	-954

using divided differences.

4.4 Error estimates

4.5 Hermite interpolation

 **Exercise 4.1** Write an algorithm that gets two vectors, (x_0, x_1, \dots, x_n) and (y_0, y_1, \dots, y_n) , and a number x , and returns $p(x)$, where p is the interpolating polynomials through the $n + 1$ points (x_i, y_i) .

Solution 4.1: The first step is to compute the coefficients C_i of Newton's representation. The most efficient way is to use divided differences:

Algorithm 4.5.1: DIVIDED-DIFFERENCES(X, Y)

```

for  $i = 0$  to  $n$ 
  do  $M_{i,0} = Y_i$ 
for  $j = 1$  to  $n$ 
  do for  $i = 0$  to  $n - j$ 
    do  $M_{i,j} = (M_{i+1,j-1} - M_{i,j-1}) / (X_{i+j} - X_i)$ 
for  $i = 0$  to  $n$ 
  do  $C_i = M_{0,i}$ 
return ( $C$ )

```


Once that the coefficients are known, we use nested multiplication to evaluate $p(x)$:

Algorithm 4.5.2: NESTED-MULTIPLICATION(x, X, C)

```

 $p = C_n$ 
for  $i = n - 1$  downto  $0$ 
  do  $p = (x - X_i)p + C_i$ 
return ( $p$ )

```

 **Exercise 4.2** Apply Lagrange's interpolation formula to the set of equally spaced pairs:

x	h	$2h$	$3h$
y	y_0	y_1	y_2

to obtain an approximation for $y(x)$ at $x = 0$.

Solution 4.2: The Lagrange interpolation formula in this case is

$$p(x) = y_0 \frac{(x-2h)(x-3h)}{2h^2} - y_1 \frac{(x-h)(x-3h)}{h^2} + y_2 \frac{(x-h)(x-2h)}{2h^2}.$$

Substituting $x = 0$ we get

$$p(0) = 3y_0 - 3y_1 + y_2.$$

 *Exercise 4.3* Let $\ell_i(x)$ be the Lagrange polynomials for the set of point x_0, \dots, x_n , and let $C_i = \ell_i(0)$. Show that

$$\sum_{i=0}^n C_i x_i^j = \begin{cases} 1 & j = 0 \\ 0 & j = 1, \dots, n \\ (-1)^n x_0 x_1 \cdots x_n & j = n+1, \end{cases}$$

and that

$$\sum_{i=0}^n \ell_i(x) = 1.$$

Solution 4.3: Each of the polynomials x^j , $j = 0, 1, \dots, n$, coincides (by uniqueness) with its interpolating polynomial through the $n+1$ given points. Thus,

$$x^j = \sum_{i=0}^n \ell_i(x) x_i^j.$$


Substituting $x = 0$ we get the first two lines. For the third line, we note that

$$x^{n+1} - (x-x_0)(x-x_1)\cdots(x-x_n)$$

is a polynomial of degree n , hence it coincides with its interpolating polynomial:

$$x^{n+1} - (x-x_0)(x-x_1)\cdots(x-x_n) = \sum_{i=0}^n \ell_i(x) x_i^{n+1}.$$

Substituting $x = 0$ we get the desired result.

 *Exercise 4.4* Suppose that $p(x)$ is the interpolation polynomial of the data:


x	3	7	1	2
y	10	146	2	1

Find a simple expression, in terms of $p(x)$, for the interpolation polynomial of the data:

x	3	7	1	2
y	12	146	2	1


Solution 4.4: Since the only difference is in the first data point, we use the Lagrange representation to write

$$p(x) + (12 - 10)\ell_0(x).$$

 *Exercise 4.5* Show that the divided differences are linear maps on functions. That is, prove the equation


$$(\alpha f + \beta g)[x_0, x_1, \dots, x_n] = \alpha f[x_0, x_1, \dots, x_n] + \beta g[x_0, x_1, \dots, x_n].$$

Solution 4.5: This is immediate by induction.

 *Exercise 4.6* The divided difference $f[x_0, x_1]$ is analogous to a first derivative. Does it have a property analogous to $(fg)' = f'g + fg'$?

Solution 4.6: By definition


$$\begin{aligned} (fg)[x_1, x_2] &= \frac{f[x_2]g[x_2] - f[x_1]g[x_1]}{x_2 - x_1} \\ &= \frac{f[x_2]g[x_2] - f[x_1]g[x_2]}{x_2 - x_1} + \frac{f[x_1]g[x_2] - f[x_1]g[x_1]}{x_2 - x_1} \\ &= f[x_1, x_2]g[x_2] + f[x_1]g[x_1, x_2]. \end{aligned}$$


 *Exercise 4.7* Prove the Leibnitz formula:

$$(fg)[x_0, x_1, \dots, x_n] = \sum_{k=0}^n f[x_0, x_1, \dots, x_k]g[x_k, x_{k+1}, \dots, x_n].$$

Solution 4.7: We use induction on n . We have seen this to be correct for $n = 1$. Suppose this is correct for any n interpolation points. Then,


$$\begin{aligned}
 (fg)[x_0, x_1, \dots, x_n] &= \frac{(fg)[x_1, \dots, x_n] - (fg)[x_0, \dots, x_{n-1}]}{x_n - x_0} \\
 &= \frac{1}{x_n - x_0} \sum_{k=1}^n f[x_1, \dots, x_k] g[x_k, \dots, x_n] \\
 &\quad - \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} f[x_0, \dots, x_k] g[x_k, \dots, x_{n-1}] \\
 &= \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} f[x_1, \dots, x_{k+1}] g[x_{k+1}, \dots, x_n] \\
 &\quad - \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} f[x_0, \dots, x_k] g[x_k, \dots, x_{n-1}] \\
 &\quad \pm \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} f[x_0, \dots, x_k] g[x_{k+1}, \dots, x_n] \\
 &= \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} (x_{k+1} - x_0) f[x_0, \dots, x_{k+1}] g[x_{k+1}, \dots, x_n] \\
 &\quad + \frac{1}{x_n - x_0} \sum_{k=0}^{n-1} (x_n - x_k) f[x_0, \dots, x_k] g[x_k, \dots, x_n] \\
 &= \sum_{k=0}^n f[x_0, \dots, x_k] g[x_k, \dots, x_n].
 \end{aligned}$$

 **Exercise 4.8** Compare the efficiency of the divided difference algorithm to the original procedure we learned in class for computing the coefficients of a Newton interpolating polynomial.

 **Exercise 4.9** Find Newton's interpolating polynomial for the following data:

x	1	3/2	0	2
$f(x)$	3	13/4	3	5/3

Use divided differences to calculate the coefficients.

 **Exercise 4.10** Find an explicit form of the Hermite interpolating polynomial for $k = 2$ (two interpolation points) and $m_1 = m_2 = m$ ($p^{(k)}(x_i) = f^{(k)}(x_i)$ for $k = 0, 1, 2, \dots, m - 1$).

Solution 4.10: There are two interpolation points, in each of which we have m pieces of data. By the Lagrange approach, let's solve the problem for homogeneous data at the point x_2 , i.e., $p^{(k)}(x_2) = 0$. The interpolating polynomials can be written in the form

$$p(x) = \ell_1^m(x) \left[c_0 + c_1 \ell_2(x) + \cdots + c_{m-1} \ell_2^{m-1}(x) \right],$$

where $\ell_i(x)$ are the Lagrange basis polynomials. In the presence of just two points,

$$\ell_1(x) = \frac{x - x_2}{x_1 - x_2} \quad \ell_2(x) = \frac{x - x_1}{x_2 - x_1},$$


and

$$\ell_1'(x) = \frac{1}{x_1 - x_2} \equiv \alpha = -\ell_2'(x).$$

Now,

$$\begin{aligned} p(x_1) &= c_0 \\ p'(x_1) &= \alpha(m c_0 - c_1) \\ p''(x_1) &= \alpha^2(m(m-1)c_0 - 2m c_1 + 2c_2), \end{aligned}$$

ad so on.

 **Exercise 4.11** Find the Hermite interpolating polynomial in the case $m_1 = m_2 = \cdots = m_k = 2$.

Hint: try

$$p(x) = \sum_{i=1}^k h_i(x) f(x_i) + \sum_{i=1}^k g_i(x) f'(x_i)$$

with h_i and g_i polynomial of degrees up to $2k - 1$ which satisfy:

$$\begin{aligned} h_i(x_j) &= \delta_{i,j} & g_i(x_j) &= 0 \\ h_i'(x_j) &= 0 & g_i'(x_j) &= \delta_{i,j}. \end{aligned}$$

Solution 4.11: The proposed polynomial satisfies the requirement, but we need to construct the polynomials h, g . The g 's are easy,

$$g_i(x) = (x - x_i) \ell_i^2(x),$$

where the $\ell_i(x)$ are the Lagrange basis polynomials. For the h 's we look for


$$h_i(x) = \ell_i^2(x)(1 + b(x - x_i)).$$


Differentiating and substituting $x = x_i$ we get

$$h_i(x_i) = 2\ell_i(x_i)\ell_i'(x_i) + b\ell_i(x_i) = 0,$$

hence $b = -2\ell'_i(x_i)$, and

$$h_i(x) = \ell_i^2(x) [1 - 2\ell'_i(x_i)(x - x_i)].$$


 **Exercise 4.12** Suppose that a function $f(x)$ is interpolated on the interval $[a, b]$ by a polynomial $P_n(x)$ whose degree does not exceed n . Suppose further that $f(x)$ is arbitrarily often differentiable on $[a, b]$ and that there exists an M such that $|f^{(i)}(x)| \leq M$ for $i = 0, 1, \dots$ and for any $x \in [a, b]$. Can it be shown without further hypotheses that $P_n(x)$ converges uniformly on $[a, b]$ to $f(x)$ as $n \rightarrow \infty$?

 **Exercise 4.13** Assume a set of $n + 1$ equidistant interpolation points, $x_i = x_0 + ih$, $i = 1, \dots, n$. Prove that the divided difference, $f[x_0, \dots, x_n]$, reduces to

$$f[x_0, \dots, x_n] = \frac{1}{h^n n!} \sum_{k=0}^n (-1)^{n-k} \binom{n}{k} f(x_k).$$

Hint: you may have to use the identity

$$\binom{m}{j-1} + \binom{m}{j} = \binom{m+1}{j}$$

 **Exercise 4.14** Prove that if f is a polynomial of degree k then for $n > k$ the divided difference $f[x_0, \dots, x_n]$ vanishes identically for all choices of interpolation points, (x_0, \dots, x_n) .

Chapter 5

Approximation theory

5.1 Weierstrass' approximation theorem

Theorem 5.1 Let $[a, b]$ be a bounded domain. For every continuous function $f(x)$ on $[a, b]$ and $\epsilon > 0$ there exists a polynomial $p(x)$ such that

$$\|f - p\|_{\infty} = \sup_{a \leq x \leq b} |f(x) - p(x)| \leq \epsilon.$$

This theorem states that continuous functions on bounded domains can be uniformly approximated by polynomials. Equivalently, it states that the space of polynomials is dense in the space of continuous functions in the topology induced by the sup-norm. Note that the theorem says nothing about the degree of the polynomial. Since polynomials depend continuously on their coefficients, this theorem remains valid if we restrict the polynomials to rational coefficients. This means that the space $C[a, b]$ as a dense subspace which is countable; we say then that the space of continuous functions endowed with the sup-norm topology is **separable**.

Proof: It is sufficient to restrict the discussion to functions on $[0, 1]$, for polynomials of linear transformations remain polynomials. Thus, we need to prove that for any function $f \in C[0, 1]$ and $\epsilon > 0$ we can find a polynomial p such that

$$\|f - p\|_{\infty} \leq \epsilon.$$

Equivalently, that we can construct a sequence of polynomials p_n such that

$$\lim_{n \rightarrow \infty} \|f - p_n\|_{\infty} = 0.$$

We introduce now a operator B_n which maps functions in $f \in C[0, 1]$ into polynomials $B_n f \in \Pi_n$:

$$B_n f(x) = \sum_{k=0}^n \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$

We note the following properties of B_n :

① Linearity:

$$B_n(\alpha f + \beta g) = \alpha B_n f + \beta B_n g.$$

② Positivity: if $f(x) \geq 0$ then $B_n f(x) \geq 0$.

③ For $f(x) = 1$,

$$B_n f(x) = \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} = (x + 1 - x)^n = 1,$$

i.e., $B_n f(x) = f(x)$.

④ For $f(x) = x$,

$$\begin{aligned} B_n f(x) &= \sum_{k=0}^n \binom{n}{k} \frac{k}{n} x^k (1-x)^{n-k} \\ &= \sum_{k=1}^n \binom{n-1}{k-1} x^k (1-x)^{n-k} \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k} x^{k+1} (1-x)^{n-1-k} \\ &= x, \end{aligned}$$

so that again $B_n f(x) = f(x)$.

⑤ For $f(x) = x^2$,

$$\begin{aligned} B_n f(x) &= \sum_{k=0}^n \binom{n}{k} \frac{k^2}{n^2} x^k (1-x)^{n-k} \\ &= \frac{n-1}{n} x^2 + \frac{1}{n} x, \end{aligned}$$

so that $\|B_n f(x) - f(x)\|_\infty \rightarrow 0$.

We claim that this is sufficient to conclude that for any continuous f , the sequence of polynomials $B_n f$ converges uniformly to f . This is established in the next theorem. ■

Theorem 5.2 (Bohman-Korovkin) Let L_n be a sequence of operators in $C[a, b]$ that are linear, positive, and satisfy

$$\lim_{n \rightarrow \infty} \|L_n f - f\|_\infty = 0 \quad \text{for} \quad f = 1, x, x^2.$$

Then $\|L_n f - f\|_\infty \rightarrow 0$ for all $f \in C[a, b]$.

Proof: The operators L_n are linear and positive. Therefore, if $f(x) \geq g(x)$ for all x , then

$$L_n f(x) - L_n g(x) = L_n(f - g)(x) \geq 0,$$

i.e., $L_n f(x) \geq L_n g(x)$ for all x . In particular, since $\pm f(x) \leq |f(x)|$ it follows that $\pm L_n f(x) \leq L_n |f|(x)$, or

$$|L_n f(x)| \leq L_n |f|(x). \quad (5.1)$$

Let $f \in C[a, b]$ be given as well as $\epsilon > 0$. Since f is continuous on a bounded domain, it is uniformly continuous: there exists a $\delta > 0$ such that for every x, y such that $|x - y| \leq \delta$, $|f(x) - f(y)| \leq \epsilon$. On the other hand, if $|x - y| > \delta$ then $|f(x) - f(y)| \leq 2\|f\|_\infty \leq 2\|f\|_\infty(x - y)^2/\delta^2$. In either case, there exists a constant C , which depends on f and ϵ , such that

$$|f(x) - f(y)| \leq C_\epsilon(x - y)^2 + \epsilon.$$

View now this inequality as an inequality between functions of y with x being a parameter. By (5.1),

$$\begin{aligned} |L_n(f(x) - f)(y)| &= |f(x)L_n 1(y) - L_n f(y)| \\ &\leq L_n |f(x) - f|(y) \\ &\leq C_\epsilon(x^2 - 2xL_n y + L_n y^2) + \epsilon L_n 1(y). \end{aligned}$$

In particular, this should hold for $y = x$, hence

$$|f(x)L_n 1(x) - L_n f(x)| \leq C_\epsilon(x^2 - 2xL_n x + L_n x^2) + \epsilon L_n 1(x).$$

Since we eventually want to bound $f(x) - L_n f(x)$, we write

$$\begin{aligned} |f(x) - L_n f(x)| &\leq |f(x)L_n 1(x) - L_n f(x)| + |f(x) - f(x)L_n 1(x)| \\ &\leq C_\epsilon(x^2 - 2xL_n x + L_n x^2) + \epsilon L_n 1(x) + \|f\|_\infty |1 - L_n 1(x)|. \end{aligned}$$

Since the assumptions of this theorem are that for every $\eta > 0$ there exists an N such that for every $n > N$

$$|L_n 1 - 1|_\infty \leq \eta \quad |L_n x - x|_\infty \leq \eta \quad |L_n x^2 - x^2|_\infty \leq \eta,$$

it follows that

$$|f(x) - L_n f(x)| \leq C_\epsilon(2|x|\eta + \eta) + \epsilon(1 + \eta) + \|f\|_\infty \eta.$$

By taking η sufficiently small we can make the right hand side smaller than, say, 2ϵ , which concludes the proof. ■

5.2 Existence of best approximation

Consider the following general problem. We are given a function f on some interval $[a, b]$. The function could be continuous, differentiable, piecewise-smooth, square-integrable, or belong to any other family of functions. For given $n \in \mathbb{N}$, we would like to find the polynomial $p_n \in \Pi_n$ that **best approximates** f , i.e., that minimizes the difference $\|f - p_n\|$:

$$p_n = \arg \min_g \|f - g\|.$$

Three questions arise:

- ① Which norm should be used?
- ② Does a best approximation exist? Is it unique? Does the existence and uniqueness depend on the choice of norms.
- ③ If it does, how to find it?

The answer to the first question is that the choice is arbitrary, or more precisely, depends on one's needs. The answer to the second question is “yes”, independently of the choice of norms. There is no general answer to the third question. We will see below how to find the best approximation for a specific norm, the L^2 norm.

But first, the existence of a best approximation follows from the following theorem:

Theorem 5.3 Let $(X, \|\cdot\|)$ be a normed space, and let $G \subset X$ be a finite-dimensional subspace. For every $f \in X$ there exists at least one best approximation within G . That is, there exist a $g \in G$, such that

$$\|f - g\| \leq \|f - h\|$$

for all $h \in G$.

Proof: Let $f \in X$ be given, and look at the subset of G :

$$K = \{g \in G : \|f - g\| \leq \|f\|\}.$$

This set is non-empty (since it contains the zero vector), bounded, since every $g \in K$ satisfies,

$$\|g\| \leq \|g - f\| + \|f\| \leq 2\|f\|,$$

and closed, i.e., K is compact with respect to the norm topology. Consider now the real-valued function $a : K \mapsto \mathbb{R}^+$:

$$a(g) = \|f - g\|.$$

It is continuous (by the continuity of the norm), and therefore reaches its minimum in K . ■

5.3 Approximation in inner-product spaces

We are now going to examine the problem of determining the best approximation in inner-product spaces. To avoid measure-theoretic issues, we will consider the space of continuous functions $X = C[a, b]$ endowed with an inner product:

$$(f, g) = \int_a^b f(x)g(x)w(x) dx,$$

where $w(x)$ is called a **weight function**, and must be strictly positive everywhere in $[a, b]$. The corresponding norm is the weighted- L^2 norm:

$$\|f\| = \sqrt{(f, f)}.$$

For $f \in C[a, b]$, we will be looking for $g \in \Pi_n$ that minimized the difference:

$$\|f - g\|^2 = \int (f(x) - g(x))w(x) dx.$$

The choice $w(x) = 1$ reduces to the standard L^2 norm.

Recall also the Cauchy-Schwarz inequality, valid for all inner-product spaces,

$$(f, g) \leq \|f\| \|g\|,$$

and the parallelepiped identity:

$$\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2.$$

Definition 5.1 The vectors f, g are called **orthogonal** (denoted $f \perp g$) if $(f, g) = 0$. f is called orthogonal to the set $G \subset X$ if $f \perp g$ for all $g \in G$.

The theory of best approximation in inner-product spaces hinges on the following theorem:

Theorem 5.4 Let X be an inner-product space and $G \subset X$ a finite-dimensional subspace. Let $f \in X$. Then $g \in G$ is the best approximation of f in G iff $f - g \perp G$.

Proof: Suppose first that $f - g \perp G$. We need to show that g is a best approximation in the sense that

$$\|f - g\| \leq \|f - h\|$$

for all $h \in G$. Now,

$$\|f - h\|^2 = \|f - g + g - h\|^2 = \|f - g\|^2 + \|g - h\|^2 \geq \|f - g\|^2,$$

where we have used the assumption that $f - g \perp g - h$.

Conversely, suppose that g is a best approximation and let $h \in G$. For all $\alpha > 0$,

$$0 \leq \|f - g + \alpha h\|^2 - \|f - g\|^2 = \alpha(f - g, h) + \alpha^2\|h\|^2.$$

Dividing by α :

$$(f - g, h) + \alpha\|h\|^2 \leq 0,$$

and taking $\alpha \rightarrow 0$ we get that $f - g \perp h$. Since this holds for all $h \in G$, this proves the claim. ■

Corollary 5.1 There exists a unique best approximation.

Proof: Let $g, h \in G$ be best approximations, then

$$(g - h, g - h) = (f - h, g - h) - (f - g, g - h) = 0,$$

since $g - h \in G$. ■

Example 5.1 Let $X = C[-1, 1]$ with the standard inner product, and $G = \text{span}\{g_1, g_2, g_3\} = \text{span}\{x, x^3, x^5\}$. Take $f = \sin x$. The best approximation in G is of the form

$$g(x) = c_1 g_1(x) + c_2 g_2(x) + c_3 g_3(x).$$

The best approximation is set by the orthogonality conditions,

$$(f - g, g_i) = 0 \quad \text{or} \quad (g, g_i) = (f, g_i), \quad i = 1, 2, 3.$$

This results in the following linear system

$$\begin{pmatrix} (g_1, g_1) & (g_2, g_1) & (g_3, g_1) \\ (g_1, g_2) & (g_2, g_2) & (g_3, g_2) \\ (g_1, g_3) & (g_2, g_3) & (g_3, g_3) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} (f, g_1) \\ (f, g_2) \\ (f, g_3) \end{pmatrix}.$$

The matrix of coefficients is called the Gram matrix. Computing these integrals we get

$$\begin{pmatrix} \frac{2}{3} & \frac{2}{5} & \frac{2}{7} \\ \frac{2}{5} & \frac{2}{7} & \frac{2}{9} \\ \frac{2}{7} & \frac{2}{9} & \frac{2}{11} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \sin 1 - \cos 1 \\ -3 \sin 1 + 5 \cos 1 \\ 65 \sin 1 - 101 \cos 1 \end{pmatrix}.$$

Orthonormal systems Life becomes even simpler if we span the subspace G with an orthonormal basis $\{g_i\}_{i=1}^n$. Then, every $g \in G$ has a representation

$$g = \sum_{i=1}^n \alpha_i g_i, \quad (g_i, g_j) = \delta_{ij}.$$

Theorem 5.5 Let $G = \text{span}\{g_1, \dots, g_n\} \subset X$. The best approximation $g \in G$ of a vector $f \in X$ is

$$g = \sum_{i=1}^n (f, g_i) g_i.$$

Proof: For all $j = 1, \dots, n$,

$$(f - g, g_j) = (f, g_j) - \sum_{i=1}^n (f, g_i)(g_j, g_i) = 0,$$

hence $f - g \perp G$. ■

Example 5.2 Let's return to the previous example of $X = C[-1, 1]$ and G span x, x^3, x^5 . It can be checked that the following vectors,

$$\begin{aligned} g_1(x) &= \sqrt{\frac{3}{2}}x \\ g_2(x) &= \sqrt{\frac{7}{2}}(5x^3 - 3x) \\ g_3(x) &= \sqrt{\frac{11}{2}}(63x^5 - 70x^3 + 15x) \end{aligned}$$

form an orthonormal basis in G (the Legendre polynomials). Then, for $f(x) = \sin x$,

$$g(x) = c_1 g_1(x) + c_2 g_2(x) + c_3 g_3(x),$$

with

$$\begin{aligned} c_1 &= \sqrt{\frac{3}{2}} \int x \sin x \, dx \\ c_2 &= \sqrt{\frac{7}{2}} \int (5x^3 - 3x) \sin x \, dx \\ c_3 &= \sqrt{\frac{11}{2}} \int (63x^5 - 70x^3 + 15x) \sin x \, dx. \end{aligned}$$

Lemma 5.1 (Generalized Pythagoras lemma) Let $\{g_i\}_{i=1}^n$ be an orthonormal set, then

$$\left\| \sum_{i=1}^n \alpha_i g_i \right\|^2 = \sum_{i=1}^n \alpha_i^2.$$

Proof: By induction on n . ■

Lemma 5.2 (Bessel inequality) Let $\{g_i\}_{i=1}^n$ be an orthonormal set then for every $f \in X$:

$$\sum_{i=1}^n |(f, g_i)|^2 \leq \|f\|^2.$$


Proof: Set

$$h = \sum_{i=1}^n (f, g_i) g_i,$$

which is the best approximation of f within the span of the g_i 's. Now,

$$\|f\|^2 = \|f - h + h\|^2 = \|f - h\|^2 + \|h\|^2 \geq \|h\|^2 = \sum_{i=1}^n |(f, g_i)|^2,$$

where we have used the fact that $f - h \perp h$. ■

 *Exercise 5.1* Show that the set of polynomials,

$$\phi_0(x) = \frac{1}{\sqrt{\pi}} \quad \phi_k(x) = \frac{2}{\sqrt{\pi}} T_k(x), \quad k = 1, 2, \dots,$$

where $T_k(x)$ are the Chebyshev polynomials, form an orthonormal basis on the segment $[-1, 1]$ with respect to the inner product,

$$(f, g) = \int_{-1}^1 f(x) g(x) \frac{dx}{\sqrt{1-x^2}}.$$

Derive an expression for the best approximation of continuous functions in the interval $[-1, 1]$ with respect to the norm


$$\|f\|^2 = \int_{-1}^1 \frac{f^2(x)}{\sqrt{1-x^2}} dx,$$


where the approximating function is a polynomial of degree less or equal n .

 *Exercise 5.2* Consider the space $C[-1, 1]$ endowed with inner product

$$(f, g) = \int_{-1}^1 f(x) g(x) dx.$$


Use the Gram-Schmidt orthonormalization procedure to construct a basis for $\text{span}\{1, x, x^2, x^3\}$.

 *Exercise 5.3* Let X be an inner product space, and G a subspace spanned by the orthonormal vectors $\{g_1, g_2, \dots, g_n\}$. For every $f \in X$ denote by Pf the best L^2 -approximation of f by an element of G . Find an explicit formula for $\|f - Pf\|$.

 **Exercise 5.4** Suppose that we want to approximate an even function f by a polynomial $p_n \in \Pi_n$, using the norm

$$\|f\| = \left(\int_{-1}^1 f^2(x) dx \right)^{1/2}.$$

Prove that p_n is also even.

 **Exercise 5.5** Let $\{p_n(x)\}$ be a sequence of polynomials that are orthonormal with respect to the weight function $w(x)$ in $[a, b]$, i.e.,


$$\int_a^b p_n(x) p_m(x) w(x) dx = \delta_{m,n}.$$

Let $P_{n-1}(x)$ be the Lagrange interpolation polynomial agreeing with $f(x)$ at the zeros of p_n . Show that

$$\lim_{n \rightarrow \infty} \int_a^b w(x) [P_{n-1}(x) - f(x)]^2 dx = 0.$$


Hint: Let B_{n-1} be the Bernstein polynomial of degree $n-1$ for $f(x)$. Estimate the right-hand side of the inequality

$$\int w [P_{n-1} - f]^2 dx \leq 2 \int w [P_{n-1} - B_{n-1}]^2 dx + 2 \int w [B_{n-1} - f]^2 dx.$$

 **Exercise 5.6** Find the Bernsteins polynomials, $B_1(x)$ and $B_2(x)$ for the function $f(x) = x^3$. Use this result to obtain the Weierstrass polynomials of first and second degree for $f(y) = \frac{1}{8}(y+1)^3$ on the interval $-1 \leq y \leq 1$.


Chapter 6

Numerical integration


 *Exercise 6.1* Approximate


$$\int_0^1 e^{-x^2} dx$$

to three decimal places.


 *Exercise 6.2* Prove that if $f \in C^2[a, b]$ then there exists an $\bar{x} \in (a, b)$ such that the error of the trapezoidal rule is

$$\int_a^b f(x) dx - \frac{1}{2}(b-a)(f(a) + f(b)) = -\frac{1}{12}(b-a)^3 f''(\bar{x}).$$

 *Exercise 6.3* Determine the interval width h and the number m so that Simpson's rule for $2m$ intervals can be used to compute the approximate numerical value of the integral $\int_0^\pi \cos x dx$ with an accuracy of $\pm 5 \cdot 10^{-8}$.

 *Exercise 6.4* By construction, the n 'th Newton-Cotes formula yields the exact value of the integral for integrands which are polynomials of degree at most n . Show that for even values of n , polynomials of degree $n + 1$ are also integrated exactly. Hint: consider the integral of x^{n+1} in the interval $[-k, k]$, with $n = 2k + 1$.

 *Exercise 6.5* Derive the Newton-Cotes formula for $\int_0^1 f(x) dx$ based on the nodes $0, \frac{1}{3}, \frac{2}{3}$ and 1 .

 *Exercise 6.6* Approximate the following integral:


$$\int_{-2}^2 \frac{dx}{1+x^2}$$

using Gaussian quadrature with $n = 2$.

Chapter 7

More questions

7.1 Preliminaries


 *Exercise 7.1* What is the rate of convergence of the sequence

$$a_n = \frac{1}{n 2^n}.$$

What is the rate of convergence of the sequence


$$b_n = e^{-3^n}.$$

7.2 Nonlinear equations

 *Exercise 7.2* Let $f(x)$ be a continuous differentiable function on the line, which has a root at the point \hat{x} . Consider the following iterative procedure:


$$x_{n+1} = x_n - f(x_n),$$


Determine conditions on f and on the initial point x_0 that guarantee the convergence of the sequence (x_n) to \hat{x} .

 *Exercise 7.3* Let $\Phi : \mathbb{R}^5 \mapsto \mathbb{R}^5$ be an iteration function with fixed point ζ . Suppose that there exists a neighborhood of ζ in which

$$\|\Phi(x) - x\| \leq 45\|x - \zeta\|^{7/3},$$

and the norm is the infinity-norm for vectors. Prove or disprove: there exists a neighborhood of ζ such that for every x_0 in this neighborhood the sequence (x_n) converges to ζ .


 *Exercise 7.4* Let f be twice differentiable with $f(\zeta) = 0$ and $f'(\zeta) \neq 0$. Prove that Newton's method for root finding is locally second order.

 *Exercise 7.5* True or false: the iteration


$$x_{n+1} = 1 + x_n - \frac{1}{4}x_n^2$$

converges to the fixed point $x = 2$ for all $x_0 \in [1, 3]$.


7.3 Linear algebra

 *Exercise 7.6* Let $\|\cdot\|$ be a vector norm in \mathbb{R}^n , and let $\|\cdot\|$ denote also the subordinate matrix norm.

- ① For $x \in \mathbb{R}^n$ define $\|x\|' = \frac{1}{2}\|x\|$. Is $\|\cdot\|'$ a vector norm?
- ② For $A \in \mathbb{R}^{n \times n}$ define $\|A\|' = \frac{1}{2}\|A\|$. Is $\|\cdot\|'$ a matrix norm subordinate to some vector norm?


 *Exercise 7.7* ① Prove that all the diagonal terms of a symmetric positive definite matrix are positive.

- ② Prove that all the principal submatrices of an spd matrix are spd.

 *Exercise 7.8* Prove by an explicit calculation that the 1- and 2-norms in \mathbb{R}^n are equivalent: find constants c_1, c_2 , such that


$$c_1\|x\|_2 \leq \|x\|_1 \leq c_2\|x\|_2$$

for all $x \in \mathbb{R}^n$.


 *Exercise 7.9* Let $\|\cdot\|$ be a vector norm in \mathbb{R}^n . Prove that the real-valued function on matrices $A \in \mathbb{R}^{n \times n}$,

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

satisfies the properties of a norm.

 *Exercise 7.10* Let $\|\cdot\|$ denote a vector norm in \mathbb{R}^n and its subordinate matrix norm.

- ① Let $x = (1, 0, 0, \dots, 0)^T$. Is it necessarily true that $\|x\| = 1$?
- ② Let I be the unit n -by- n matrix. Is it necessarily true that $\|I\| = 1$?

 *Exercise 7.11* Derive an explicit expression for the matrix norm subordinate to the 1-norm for vectors.


 *Exercise 7.12* Prove that the spectral radius, which can be defined by


$$\text{spr } A = \max_{\lambda \in \Sigma(A)} |\lambda|,$$

satisfies

$$\text{spr } A = \inf_{\|\cdot\|} \|A\|.$$


 *Exercise 7.13* What is the spectral radius of an upper triangular matrix?


 *Exercise 7.14* Can you use the Neumann series to approximate the inverse of a matrix A ? Under what conditions will this method converge?

 *Exercise 7.15* Let A be a non-singular matrix, and let B satisfy

$$\|B\|_2 < \frac{1}{\|A^{-1}\|_2}.$$

Prove that the matrix $A + B$ is not singular.

 *Exercise 7.16* Show that every symmetric positive-definite matrix has an LU-decomposition. Justify every step in the proof.

 *Exercise 7.17* Consider the iterative method

$$x_{n+1} = x_n + B(b - Ax_n),$$

with $x_1 = 0$. Show that if $\text{spr}(I - AB) < 1$, then the method converges to the solution of the linear system $Ax = b$.


7.4 Interpolation

 *Exercise 7.18* Consider the function


$$f(x) = \frac{1}{1+x}$$

on the interval $[0, 1]$. Let p_n be its interpolating polynomial with uniformly spaced interpolation points $x_i = i/n$, $i = 0, 1, \dots, n$. Prove or disprove:

$$\lim_{n \rightarrow \infty} \|p_n - f\|_{\infty} = 0.$$

 *Exercise 7.19* Let f be interpolated on $[a, b]$ by a polynomial $p_n \in \Pi_n$. Suppose that f is infinitely differentiable and that $|f^{(k)}(x)| \leq M$ for all $x \in [a, b]$. Can we conclude, without further information, that

$$\lim_{n \rightarrow \infty} \|p_n - f\|_{\infty} = 0.$$

 *Exercise 7.20* Compute the Hermite interpolating polynomial for the data $f(0) = f'(0) = f''(0) = 0$ and $f(1) = 1$.

7.5 Approximation theory

Index

- backward-substitution, 57
- Cauchy-Schwarz inequality, 41
- Chebyshev
 - acceleration, 74
 - polynomials, 75
- forward-substitution, 57
- Hölder inequality, 39
- inequality
 - Cauchy-Schwarz, 41
 - Hölder, 39
 - Minkowski, 39
 - Young, 39
- inner product, 40
- Matrix
 - norm, 43
 - positive-definite, 41
- matrix
 - permutation, 58
- Minkowski inequality, 39
- Neumann series, 46
- norm
 - p -norms, 39
 - equivalence, 42
 - Matrix norm, 43
 - vector, 38
- permutation matrix, 58
- Singular value decomposition, 83
- Spectral radius, 47
- spectrum, 48
- Young inequality, 39

