

LANDMARKS IN THE HISTORY OF ITERATIVE METHODS

MARTIN J. GANDER, PHILIPPE HENRY AND GERHARD WANNER

Abstract. “One of the ways to help make computer science respectable is to show that it is deeply rooted in history (...).” (Donald E. Knuth [106, p.671]). A great deal of the “respectable” modern numerical methods proceed iteratively, for which we give an overview in the last Section 11. Also teaching and learning science on a historical basis leads to a “respectable” deeper understanding. The first problems requiring iterative processes were square root calculations in Babylon, Greece and India. More complicated problems such as sine tables in the Arabic, Indian and Medieval calculations including Kepler’s Problem were performed with Fixed Point iterations. With Newton, Raphson and Simpson we enter “respectable” discussions of methods based on derivatives, with Mourraille and Cayley their geometric properties in \mathbf{R} and \mathbf{C} and with Fourier, Cauchy and Kantorovich rigorous error estimations. Surprisingly, also linear problems became interesting for very large dimensions, beginning with Gauss, Seidel, SOR, Richardson, Krylov, Domain Decomposition and Multigrid Methods. We explain all of them and illustrate them on the same “Montreal test problem”.

Contents

1. Ancient Calculations of Square Roots	2
1.1 The Babylonian Tablet YBC 7289	2
1.2 Heron’s Method	3
1.3 The Bakhshālī Manuscript	6
1.4 A Problem by Leonardo Pisano	7
1.5 Bombelli’s Continued Fraction for \sqrt{N}	8
2 Early Fixed Point Iterations	9
2.1 Al-Kāshī’s and Nityānanda’s Calculation of $\sin 1^\circ$	9
2.2 Theon of Smyrna’s Iteration	11
2.3 Jost Bürgi’s Calculation of his Sine Table	12
2.4 Indian Astronomical Tables and Ḥabash’s Problem	13
2.5 Kepler’s Laws and Kepler’s Problem	14
2.6 Fourier’s zigzags	16
2.7 Euler’s iterated exponentials	17
3 Emergence of Newton’s Method	18
3.1 Newton’s Famous Example	19
3.2 Newton and Wren’s Graphical Solution of Kepler’s Problem	20
3.3 Newton’s method for m th roots	21
3.4 Raphson’s contribution	22
3.5 Simpson’s fluxion approach	23
4 Geometry of Newton’s Method	26
4.1 Mourraille’s Geometrical Interpretation	26
4.2 Cayley’s “Newton-Fourier Imaginary Problem”	27
4.3 Basins of Attraction	31
5 Error Estimates for Newton’s Method	35
5.1 Fourier’s Estimate	35
5.2 Cauchy’s Estimate	36
5.3 Systems of Equations; Banach and Kantorovich	38

6	Stationary Iterative Methods for Large Linear Problems	41
6.1	A Letter of Gauss	43
6.2	The Method of Jacobi	43
6.3	The Method of Seidel	44
6.4	Back to Gauss' letter	47
6.5	Successive Overrelaxation Method (SOR) of David Young	48
7	Non-Stationary Extrapolation Methods	50
7.1	Richardson's 1911 Paper	50
7.2	John von Neumann's Letter	53
7.3	Golub's Modified Chebyshev Semi-iterative Method	55
7.4	Cabay and Jackson: Modified Polynomial Extrapolation	56
8	Krylov Methods	59
8.1	Relaxation and the Method of Steepest Descent	60
8.2	The Conjugate Gradient Method	63
8.3	General Krylov Methods	68
8.4	Preconditioning	68
9	Domain Decomposition Methods	70
9.1	Schwarz Methods	70
9.2	Dirichlet-Neumann Methods	72
9.3	Neumann-Neumann Methods	73
9.4	FETI (Finite Element Tearing and Interconnect)	74
9.5	Optimized Schwarz Methods	74
10	Multigrid Methods	75
10.1	Two Grid Method	76
10.2	Multi Grid Method	77
11	Current research and outlook	78

1. Ancient Calculations of Square Roots.

“(...) et diuidatur recta mo. ad punctum .q. in duo equa, et erit (...)”

(Fibonacci, [139, II, p. 250], the main idea of this section, see Fig. 1.2)

It is widely believed that the oldest mathematical problem requiring iterative methods, somehow urged by the application of Pythagoras' Theorem, was the calculation of the square root of a given number N .

1.1. The Babylonian Tablet YBC 7289. This clay tablet (first third of the second millennium BC, [128, pp. 42–43]) shows a square with both diagonals and contains, in base 60, the numbers

$$1; 24, 51, 10 = 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} \simeq \sqrt{2} \quad (1.1)$$

and

$$42; 25, 35 = 42 + \frac{25}{60} + \frac{35}{60^2} \simeq 30\sqrt{2},$$

with all hexadecimal digits correct (see Fig. 1.1). The latter is the length of the diagonal of the square of side length 30. The value (1.1) improves the Babylonian

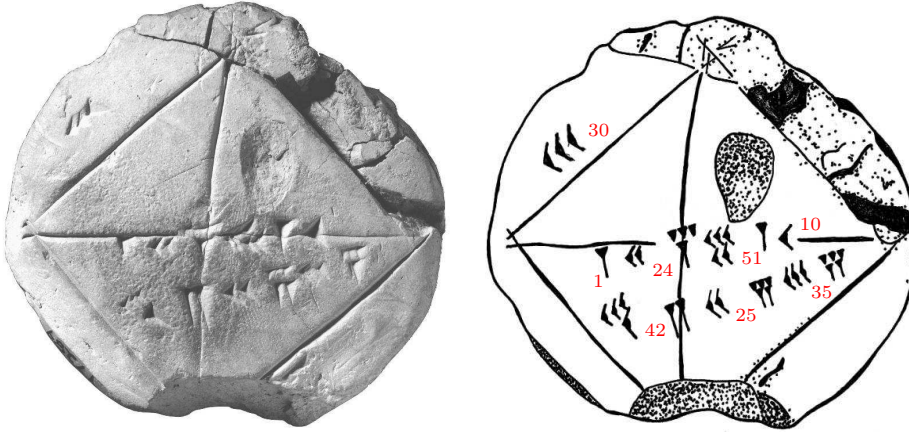


FIG. 1.1. The tablet YBC 7289 (diameter about 7 cm, ©Yale Babylonian Collection) and a drawing by the authors.

approximation $\sqrt{2} \simeq 1;25$ and also occurs in an other tablet containing a list of coefficients [128, p. 136]. As the “decimal point” of the numbers is not indicated by the scribe, the interpretation $0;30 = \frac{1}{2}$ for the side of the square is also possible and thus

$$0;42,25,35 = \frac{42}{60} + \frac{25}{60^2} + \frac{35}{60^3} \simeq \frac{30}{60} \sqrt{2} = \frac{1}{\sqrt{2}},$$

so that the tablet gives a pair of reciprocal numbers, an important matter of interest in Babylonian mathematics, because it allows to replace divisions by multiplications.

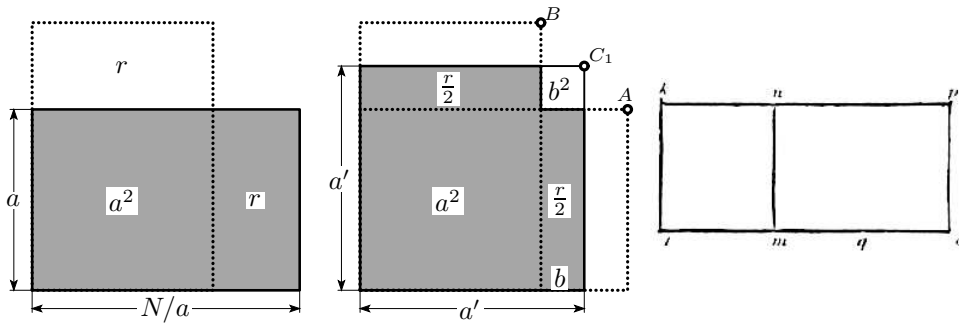
Unfortunately, it is not known how the accurate value (1.1) was calculated. Neugebauer and Sachs [128, p. 43] propose a general procedure attested by another tablet.

1.2. Heron’s Method. For a given N , we search \sqrt{N} , the side length of the square with area N . We let a be a first approximation and consider the rectangle with sides a and N/a which has the desired area but is not necessarily a square (gray in Fig. 1.2, left). We therefore divide (in a very similar way to Eucl. II.5) the surplus of this rectangle beyond the square a^2 , i.e., the *residual* $r = N - a^2 = a(\frac{N}{a} - a)$, into two equal parts at the arithmetic mean of a , N/a and display them symmetrically around a^2 (Fig. 1.2, middle). This creates approximately a square with sides

$$a' = a + b = a + \frac{r}{2a} = \frac{1}{2} \left(a + \frac{N}{a} \right). \quad (1.2)$$

One of the oldest documents attesting the use of the formula (1.2) is the demotic papyrus BM 10520 kept at the British Museum (probably dating of the early Roman period of Egypt, i.e., 30 BC to 641 AD). The scribe uses it to obtain $\sqrt{10}$ with $a = 3$ and $b = \frac{10-3^2}{2 \cdot 3} = \frac{1}{6}$ (see Fig. 1.3). Here is a translation of this text¹:

¹After these few lines the scribe also calculates using the same formula $\sqrt{\frac{1}{2}}$ under the form $\sqrt{\frac{18}{36}} = \frac{\sqrt{18}}{6} = \frac{\sqrt{4^2+2}}{6} = \frac{1}{6} \left(4 + \frac{1}{4} \right) = \frac{2}{3} + \frac{1}{24}$. Unfortunately, “the scribe was very careless or did not understand what he was writing, and so made numerous errors (...)” [137, p. 70]



“Cause that 10 reduce to its square root.

You shall reckon 3 times 3: result 9, remainder 1; $\frac{1}{2}$ (of 1): result $\frac{1}{2}$.

You shall cause that $\frac{1}{2}$ make part of 3: result $\frac{1}{6}$.

You shall add $\frac{1}{6}$ to 3: result $3\frac{1}{6}$. It is the square root.

Causing knowing it. Viz.

You shall reckon $3 \frac{1}{6}$ times $3 \frac{1}{6}$: result $10 \frac{1}{36}$.

Its difference with the square root $\frac{1}{36}$.” [137, p. 69]

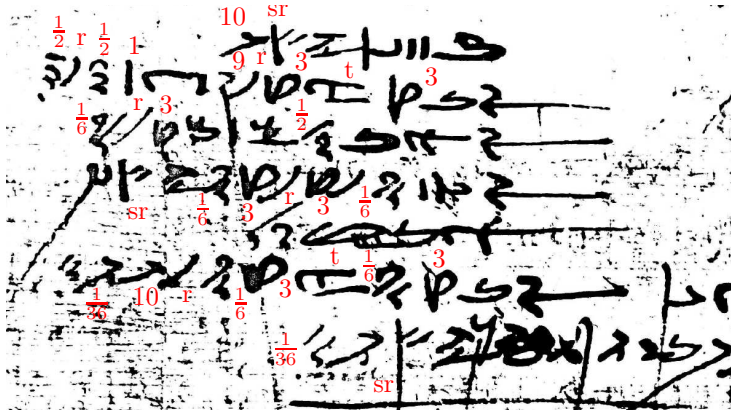


FIG. 1.3. *Reproduction of part of the papyrus BM 10520 where the scribe extracts the square root of 10 with arabics numbers and words “square root” (sr), “times” (t), “result” (r) [137, Plate 23].*

We remark that the same geometric intuition behind formula (1.2) allowed Al-Khwārizmī (*Al-jabr w'al muqābala*, 830 AD) to solve $x^2 + 10x = 39$, his first quadratic equation (Fig. 1.4). Indeed, treating the rectangle $10 \times x$ in the same way as above allows to conclude that $(x + 5)^2 = 39 + 25 = 64$, hence $x + 5 = 8$. Also Leonardo Pisano's solution (c. 1175–1250, also known as Fibonacci) of equation (1.5) below is the same algorithm.

The first ancient author to more explicitly explain this method was Heron of Alexandria (1st century AD) in his *Metrica* discovered in 1896 (see [161, p. 18]). Heron presents as example the computation of $\sqrt{720}$ as follows:

“But as 720 has not a rational square root, we shall take the root with a very small difference. As the square lying next to 720 is 729 and has the square root 27, divide 720 by 27, result 26 and two thirds. Add 27, result 53 and two thirds. One half of these, result 26 $\frac{1}{2}$ $\frac{1}{3}$. Thus the square

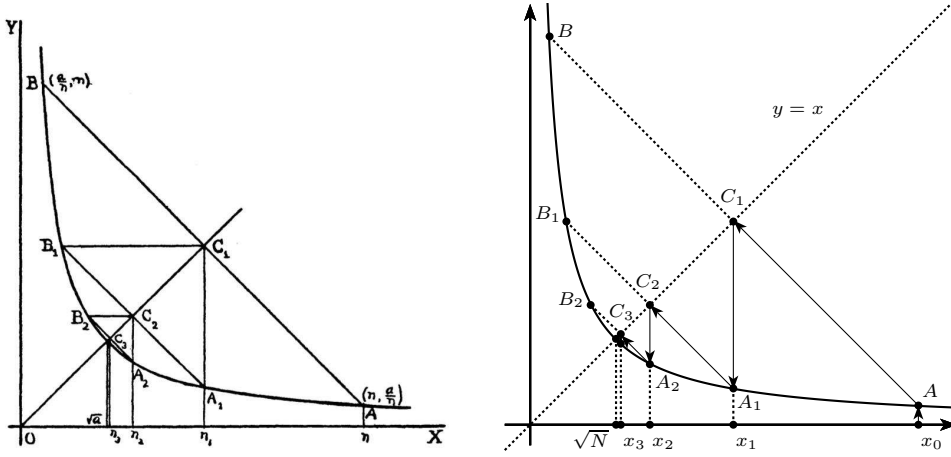


FIG. 1.5. Bouton's geometrical interpretation for the convergence of (1.2).

Another geometric interpretation. In 1909, Charles Leonard Bouton (1869–1922) still feels the need to revisit the convergence of the sequence (1.2) to \sqrt{N} [12]. He also gives a nice picture for illustrating the convergence of method (1.2) by drawing the hyperbola $y = \frac{N}{x}$ (see Fig. 1.5). The points A and B of Figure 1.2 lie symmetrically on this hyperbola. If C_1 , which is their mid point, is projected back to this hyperbola, vertically and horizontally, we obtain the next iterates A_1 , B_1 and so on. The rapid convergence of the method then becomes clearly visible.

1.3. The Bakhshālī Manuscript. Another appearance of the same algorithm is contained in an ancient Indian collection of mathematical fragments written on birch bark and found in 1881 near the village of Bakhshālī, not far from Peshawar (now in Pakistan). It is probably the oldest surviving document of Indian mathematics and seems to be a copy of the second half of the first millennium of an older text².

Among the seventy folios containing problems and prose commentary, we find on folio 56r the following method to obtain a square root:

“In the case of a number whose square root is to be found, divide it by the approximate root (the root of the nearest square number) ; multiply the denominator of the resulting *śeṣa* (the ratio of the remainder to the divisor) by two ; square it (the fraction just obtained) ; halve it ; divide it by the composite fraction (the first approximation) ; subtract (from the composite fraction) ; the result is the refined root.” [29, p. 121]

This sequence of steps

$$\frac{N}{a} = \frac{a^2 + r}{a} = a + \frac{r}{a}, \quad \frac{r}{2a}, \quad \left(\frac{r}{2a}\right)^2, \quad \frac{1}{2}\left(\frac{r}{2a}\right)^2, \quad \frac{\left(\frac{r}{2a}\right)^2}{2\left(a + \frac{r}{2a}\right)}, \quad a + \frac{r}{2a} - \frac{\left(\frac{r}{2a}\right)^2}{2\left(a + \frac{r}{2a}\right)}$$

yields precisely the two iterations of formula (1.2), i.e., formula (1.3). The folio 57v proposes the example (see the red numbers in Fig. 1.6)

$$\sqrt{41} = \sqrt{36 + 5} \simeq 6 + \frac{5}{12} - \frac{\frac{25}{144}}{2\left(6 + \frac{5}{12}\right)}.$$

²About recent debates regarding the dating of the manuscript and the use of zero as a “number in its own right” or as “placeholder” see [141].

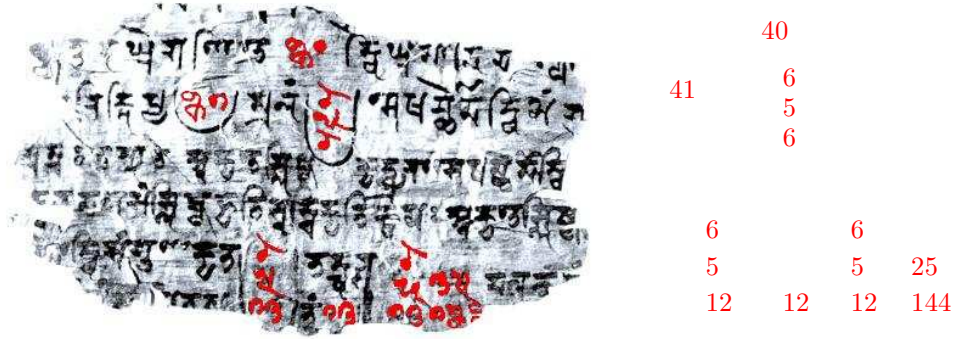


FIG. 1.6. The folio 57v of the Bakhshālī manuscript (Bodleian Library, Oxford, [98, p. 148 & Pl. XXXIX]): for easier understanding, we have coloured the numbers in red.

1.4. A Problem by Leonardo Pisano. In a passage of his *Epistola ad magistrum Theodorum*, Leonardo Pisano (c. 1175–1250), also known as Fibonacci, solves the following problem [139, II, pp. 249–250]: Inscribe an equilateral pentagon in an isosceles triangle with base 12 and legs 10, so that one vertex of the pentagon is at the apex of the triangle, one on each leg and two on the base (see Fig. 1.7).

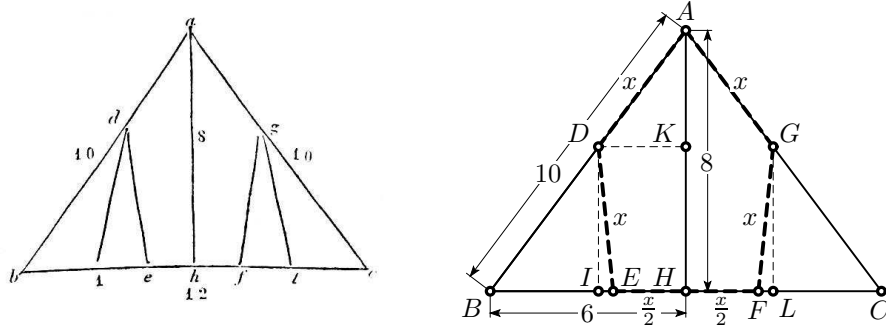


FIG. 1.7. Reproduction of Leonardo’s figure in [139, II] (left).

Let x be the side of the pentagon and draw the perpendiculars DI , GL and DK (Leonardo’s picture, or that of his graphic artist, is not very precise). The triangles BHA and DKA are similar to the Pythagorean triangle 3, 4, 5. This gives us $DK = \frac{3x}{5}$ and $KA = \frac{4x}{5}$. Computing DE by Pythagoras (“angulus .d.i.e est rectus”) we have

$$x^2 = \left(\frac{3x}{5} - \frac{x}{2}\right)^2 + \left(8 - \frac{4x}{5}\right)^2 \Rightarrow x^2 + \left(36 + \frac{4}{7}\right)x = 182 + \frac{6}{7}. \quad (1.5)$$

Here, the square x^2 together with the rectangle $\left(36 + \frac{4}{7}\right) \times x$ is a given quantity. Exactly as in the first problem of Al-Khwārizmī (“et sic reducta est questio ad unam ex regulis algebre”), Leonardo divides the rectangle “in duo equa”, so that all three parts together with the square $\left(18 + \frac{2}{7}\right)^2$ can be assembled to a complete square (see Figs. 1.2 and 1.4). We thus get

$$\left(x + 18 + \frac{2}{7}\right)^2 = 182 + \frac{6}{7} + \left(18 + \frac{2}{7}\right)^2 = 517 + \frac{11}{49}. \quad (1.6)$$

22; 00, 00
 22; 45, 18, 22, 02, 26, 56, 19, 35, 30, 36, 44, 04, 53, 52, 39, 11, 01, 13, 28, 09, 47, 45, 18, 22, 02, 26, 56, 19, 35, 30, 36
 22; 44, 33, 15, 51, 48, 26, 49, 04, 42, 20, 47, 19, 56, 15, 58, 04, 01, 49, 22, 49, 33, 41, 53, 26, 37, 32, 50, 29, 42, 22, 26
 22; 44, 33, 15, 07, 05, 00, 29, 25, 54, 05, 16, 37, 22, 37, 16, 25, 49, 58, 53, 42, 38, 32, 20, 25, 09, 24, 31, 14, 05, 20, 41
 22; 44, 33, 15, 07, 05, 00, 29, 25, 54, 05, 16, 37, 22, 37, 15, 43, 19, 01, 04, 16, 29, 54, 20, 51, 43, 13, 39, 42, 30, 29, 27

(1.7)

We thus have to find x from $\sqrt{N} = \sqrt{a^2 + r} = a + x$. After expanding $(a + x)^2$ and simplification, this leads to

Bombelli's algorithm corresponds to solve the equation to the right for x iteratively, starting from $x_0 = 0$. We obtain

Bombelli uses this method with $\sqrt{13} = \sqrt{3^2 + 4}$ and obtains so the approximations³

$\frac{2}{3}$, ma uolendo più prossimo, si aggiunga il rotto al 6 fa $6\frac{2}{3}$, e con esso si parta pur il 4, ne uiene $\frac{2}{3}\frac{2}{3}$, e questo si aggiunga, come si è fatto di sopra al 3 fa $3\frac{2}{3}\frac{2}{3}$, ch'è l'altro numero più prossimo, perche il suo quadrato è $13\frac{2}{3}\frac{2}{3}$, ch'è troppo $\frac{1}{3}\frac{2}{3}\frac{2}{3}$, e

We show in Figure 1.8 how Bombelli explains the computation of $\frac{20}{33}$ by “aggionga” the “rotto” $\frac{3}{5}$ to 6 and divide 4 by that. Then the “quadrato” of $3\frac{20}{33}$ becomes $13\frac{4}{1089}$ which is declared to be “più prossimo”.

8

According to C. Brezinski this is “the birth” of the theory of continued fractions⁴. The golden age of such expansions will culminate during the next centuries in the work of Euler, Legendre and Lambert. We will see in the next section other examples of equations solved by iterations.

2. Early Fixed Point Iterations.

“(…) the lines of development for these methods within each textual tradition, and the potential links between them, are still very poorly understood.”
(K. Plofker, [140, p. 259])

Solving equations of the type

$$x = f(x) \quad \text{by iterating} \quad x_{n+1} = f(x_n) \quad (2.1)$$

is rooted in Arab mathematical practice as well as in Indian one, but the relationships between the two are not so clear. Today, standard error analysis assures convergence, if in a neighbourhood of the solution point $|f'(x)| \leq q < 1$.

For example, for Bombelli’s computation of $\sqrt{13}$ (see equation (1.10) above) we have

$$f(x) = \frac{4}{6+x}, \quad x = \sqrt{13} - 3 = 0.605551\dots, \quad f'(x) = -\frac{4}{(6+x)^2} = -0.09167\dots$$

which explains the (for continued fractions typical) alternating convergence and the fact that each iteration improves the result by approximately one digit:

n	p	q	$3 + \frac{p}{q}$	$(3 + \frac{p}{q})^2$
1	2	3	3.666666666667	13.444444444444
2	3	5	3.600000000000	12.960000000000
3	20	33	3.606060606060	13.0036730945822
4	66	109	3.6055045871560	12.9996633280027
5	109	180	3.6055555555556	13.0000308641975
6	720	1189	3.6055508830950	12.9999971705874
7	2378	3927	3.6055513114337	13.0000002593810
8	3927	6485	3.6055512721665	12.9999999762217
9	25940	42837	3.6055512757663	13.0000000021798
10	85674	141481	3.6055512754363	12.999999998002

2.1. Al-Kāshī’s and Nityānanda’s Calculation of $\sin 1^\circ$. The trisection of a given angle was one of the great problems which the ancient Greeks could not resolve precisely. In particular, the value of $\sin 1^\circ$ was required as first step for the calculation of any table of sines. The value of $\sin 3^\circ = \sin(18^\circ - 15^\circ)$ can be obtained from the dimensions of the regular pentagon and hexagon as

$$\sin 3^\circ = \frac{1}{16} \left(\sqrt{2}(\sqrt{5}-1)(\sqrt{3}+1) - 2\sqrt{5+\sqrt{5}}(\sqrt{3}-1) \right) = 0; 3, 8, 24, 33, 59, 34, 28, 14\dots \quad (2.2)$$

and computed to any precision. But the calculation of $\sin 1^\circ$ is more difficult. Ptolemy (c. 100–c. 170) obtained from $\sin 3^\circ$ by repeated halving of the angle and linear interpolation $\sin 1^\circ = 0; 1, 2, 50$ (see for example Chapter 4 in [2]).

⁴About the history of continued fractions see [15], especially pp. 61–70 about the work of Bombelli and Pietro Cataldi (1548–1626).

Al-Kāshī. A tremendous progress for more precision was made by the *Treatise on the Chord and the Sine* by the Iranian astronomer Jamshīd Ghīāth ud-Dīn al-Kāshī (c.1380–1429), director of the Ulugh Beg Observatory in Samarkand. The paper [147] by Rosenfeld and Hogendijk gives a presentation of the surviving manuscripts and begins with “One of the highlights of the medieval Islamic mathematical tradition is (...)”. An earlier presentation of the underlying idea is the paper [1] by Asger Aaboe (1922–2007), which was based on [153], a translation of a text by Mīrim Chelebī, a grandson of the Turkish astronomer Ṣalāḥ al-Dīn Mūsā Qāḍī-Zādeh al-Rūmī (1360–1437, teacher of Ulugh Beg).

Al-Kāshī’s, and thus Chelebī’s, text starts with Ptolemy’s Lemma, which leads to the addition properties of the sine functions. By applying addition formulas to $\sin(\alpha + 2\alpha)$ we obtain the triple angle formula for sine,

$$\sin 3\alpha = 3 \sin \alpha - 4 \sin^3 \alpha \quad \text{hence} \quad x = \frac{\sin 3^\circ}{3} + \frac{4}{3}x^3, \quad (2.3)$$

a third degree equation to be solved for $x = \sin 1^\circ$.

Al-Kāshī computes from this equation the sexagesimal digits of $x = 0; a_1, a_2, a_3, \dots$ one after the other. Suppose that we have already $x_2 = 0; 1, 2$ and want to find $x_3 = 0; 1, 2, a_3$. Inserting this into (2.3) and using (2.2) gives for a_3 the condition

$$0; 1, 2, a_3 = 0; 1, 2, 48, 11, \dots + \frac{4}{3}(0; 1, 2, a_3)^3 \approx 0; 1, 2, 48, 11, \dots + \frac{4}{3}(0; 1, 2)^3.$$

The error committed to the right by neglecting a_3 is of size $4 \cdot (0; 1, 2)^2 \cdot a_3 \cdot 60^{-3} \approx 60^{-5}$. So a_3 can just be computed by subtracting $0; 1, 2$ on both sides and evaluating the leading sexagesimal digit. The general algorithm is thus the following

$$a_{n+1} = \text{Int.part} \left[60^{n+1} \cdot \left(\frac{\sin 3^\circ}{3} + \frac{4}{3}x_n^3 - x_n \right) \right]. \quad (2.4)$$

We have coded this 600 year old algorithm in a 6-line Maple procedure in high precision, which produces for $\sin 1^\circ$ the following sequence of sexagesimal digits⁵:

0; 01,02,49,43,11,14, 44,16,26,18,28,49, 20,26,50,41,13,06, 46,25,26,26,34,06,
40,18,50,31,06,35, 20,44,06,39,18,05, 38,58,02,00,05,04, 33,59,11,35,33,50,
34,07,56,43,38,30, 15,49,36,42,06,43, 10,38,45,53,15,59, 07,19,46,22,23,42,
12,01,52,27,42,58, 47,42,58,28,56,53, 07,50,58,27,11,17, 14,38,29,51,35,44,
52,04,34,18,19,41, 56,39,49,23,38,33, 32,52,06,36,13,45, 25,03,20,47,44,22,
02,18,30,28,22,07, 33,05,15,38,54,20, 07,05,27,28,18,52, 25,41,16,56,20,33,
24,34,27,28,10,01, 02,31,08,34,40,34, 16,20,17,15,34,13, 01,34,47,48,03,04,
27,05,18,40,40,16, 13,02,45,22,42,05, 38,34,51,38,34,31, 42,56,46,59,38,01, ...

Nityānanda. The Hindu astronomer Nityānanda described in his monumental astronomical treatise *Sarvasiddhāntarāja* (1639) first a method similar to Al Kāshī’s, and then continued (Verses 63 (second half)–64 (first half), [124, p. 14]):

“And I will explain this by another method (prakāra). [Considering] the Sine of three degrees [and] removing one third of it, [may one put it down] separately. The cube of this is divided by three. When this result [obtained thus] is divided by the square of the Radius [and] added to [the initial result that was stored], [and the process] is repeated, half of it (i.e. the iterated value) is the Sine of one degree.”

⁵The formula, as it stands, produces a few times an $a_{n+1} > 59$. An Arabic scribe, doing his calculations by hand, would then recognize that the previous a_n must be increased by 1 and the new a_{n+1} decreased by 60. We, too, have made the same corrections in our list.

The “Radius” in this text is 60, the standard radius back then for expressing the sine values. And “half of it” means that Nityānanda was computing $y = 2 \cdot 60 \sin 1^\circ$, for which equation (2.3) becomes $y = \frac{2}{3} 60 \sin 3^\circ + \frac{y^3}{3 \cdot 60^2}$. Nityānanda thus describes the iteration

$$y_{n+1} = \frac{2}{3} 60 \sin 3^\circ + \frac{y_n^3}{3 \cdot 60^2} \quad \text{or, for (2.3),} \quad x_{n+1} = \frac{\sin 3^\circ}{3} + \frac{4}{3} x_n^3. \quad (2.5)$$

Starting with $y_0 = x_0 = 0$, these formulas produce the following results:

Nityānanda’s method	Iterations for $\sin 1^\circ$
2.093438249717753308884745184363	0.017445318747647944240706209870
2.094287736663828315042153435267	0.017452397805531902625351278627
2.094288771212046996374290307845	0.017452406426767058303119085899
2.094288772472484016744106641437	0.017452406437270700139534222012
2.094288772474019665097061718702	0.017452406437283497209142180989
2.094288772474021536048031298080	0.017452406437283512800400260817
2.094288772474021538327496856609	0.017452406437283512819395807138
2.094288772474021538330274034267	0.017452406437283512819418950286
2.094288772474021538330277417831	0.017452406437283512819418978482
2.094288772474021538330277421953	0.017452406437283512819418978516

Since $f'(x) = 4x^2 \approx 0.0012$, we understand why at each iteration we gain approximately three additional digits.

2.2. Theon of Smyrna’s Iteration. Theon of Smyrna (c. 70–c. 135) writes in his *Mathematics Useful for Understanding Plato*:

“Let two units be laid out, of which we take one as the side and the other as the diagonal (...); add to the side the diagonal and to the diagonal two sides (...). (...) the side is now 2 (...) [and] the diagonal 3. Again, to the side 2 add the diagonal 3 (...) [and] to the diagonal 3 twice the side (...) so that the diagonal is now 7 and the side 5 (...). And so on by continuing the addition. The ratio alternates: the square on the diagonal being now one more now one less than twice the square on the side (...). Therefore the squares of all the diagonals are the double of the squares of all the sides (...); indeed, what is missing in the preceding diagonal is found in excess in the following diagonal.” ([41, pp. 71–75], transl. [170, p. 675])

Starting with $a_1 = d_1 = 1$, Theon thus describes the recursion⁶

$$\begin{aligned} a_{n+1} &= a_n + d_n \\ d_{n+1} &= 2a_n + d_n \end{aligned} \Rightarrow \begin{pmatrix} a_n \\ d_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 7 \end{pmatrix}, \begin{pmatrix} 12 \\ 17 \end{pmatrix}, \begin{pmatrix} 29 \\ 41 \end{pmatrix}, \begin{pmatrix} 70 \\ 99 \end{pmatrix}, \dots \quad (2.6)$$

Expanding the squares a_{n+1}^2 and d_{n+1}^2 we see that always

$$(d_{n+1}^2 - 2a_{n+1}^2) = -(d_n^2 - 2a_n^2), \quad (2.7)$$

thus, as Theon claimed, $d_n^2 = 2a_n^2 \pm 1$. Therefore, according to Heath⁷, “no one familiar with the truth of the proposition stated by Theon could have failed to observe that,

⁶The sequence (a_i) is today known as Pell numbers and the sequence $(2d_i)$ as Pell–Lucas numbers.

⁷[85, I, p. 399]. At the beginning of the quoted passage, Theon writes: “As the unity is the principle of all figures, according to the highest and generating reason, so also is the ratio of the diagonal to the side found in the unit” [41, p. 71]. So as $a_1 = d_1 = 1$, Vedova suggests that Theon, a known Pythagorean, is still trying, some 600 years after the death of Pythagoras, to defend the Master’s doctrine that the unit is the constituent element of all things [170, p. 677].

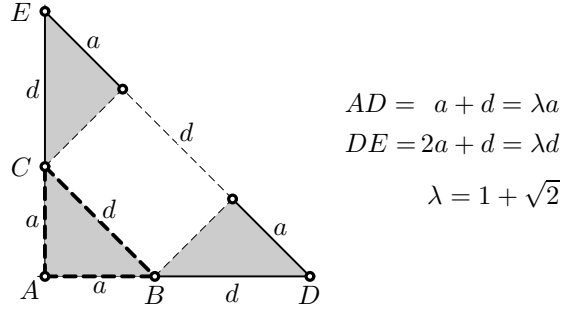


FIG. 2.1. Eigenvector of Theon's algorithm.

(...) the successive fractions d_n/a_n would give nearer and nearer approximations to the value $\sqrt{2}$ ".

A geometric interpretation. Theon's use of the words "side" and "diagonal" indicates that he may have had some geometric intuition in mind. Indeed, if we start with half a square ABC of sides a and diagonal d (see Fig. 2.1) and twist it twice around B and C respectively by 135° , then $ECBD$ is part of a regular octagon and ADE is similar to ABC with similarity factor $1 + \sqrt{2}$. Thus

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix} \text{ has dominant eigenvector } \begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix} \text{ and dominant eigenvalue } 1 + \sqrt{2}$$

and Theon's method (2.6) can be seen as the oldest use of what is known today as the *Power Iteration Method* for an eigenvalue problem.

2.3. Jost Bürgi's Calculation of his Sine Table. One and a half millennia later, but still four centuries before the power iteration method in von Mises [123] (see [91] and [166] for more information on the history of the power method), Jost Bürgi (1552–1632) applied a similar algorithm around 1584 to linear maps $s^n \mapsto s^{n+1}$ for computing the sine values at m equidistant points between 0° and 90° . This algorithm (Kunstweg) is defined as follows: Given approximate values for the sine values s_1^n, \dots, s_m^n on an equidistant grid, Bürgi computes approximations for the cosines $c_{1/2}^n, \dots, c_{(2m-1)/2}^n$ on the staggered grid as (here written for $m = 3$, see Fig. 2.2, neglecting the constant $K = 2 \sin \delta$)

$$\begin{aligned} \cos(\alpha - \delta) - \cos(\alpha + \delta) &= K \sin \alpha \\ \sin(\alpha + \delta) - \sin(\alpha - \delta) &= K \cos \alpha \end{aligned} \Rightarrow \begin{aligned} &0 \quad c_{1/2}^n - c_{3/2}^n = s_1^n \quad 0 \\ s_1^n &\swarrow \quad \quad \quad \swarrow s_1^{n+1} = c_{1/2}^n \\ &s_2^n \quad c_{3/2}^n - c_{5/2}^n = s_2^n \quad s_2^{n+1} - s_1^{n+1} = c_{3/2}^n \\ &\swarrow \quad \quad \quad \swarrow s_3^{n+1} - s_2^{n+1} = c_{5/2}^n \\ &s_3^n \quad c_{5/2}^n = \frac{1}{2} s_3^n \end{aligned}$$

Here, the new cosine values are calculated from bottom to top. Bürgi carefully replaced the starting condition $\cos 90^\circ = 0$ (which is not located on the staggered grid) as $c_{7/2}^n = -c_{5/2}^n$, which leads to $2c_{5/2}^n = s_3^n$. In the second step, we compute new sine values from top to bottom, starting from the boundary condition $s_0^{n+1} = 0$. Bürgi calculated his table by using $m = 90$. We demonstrate here its convergence for $m = 3$:

0°	0	0	0	0	0	0	0	0	0
30°	2	7	7	26	26	97	97	362	362
		5	12	19	45	71	168	265	627
60°	3	2	14	7	52	26	194	97	724
90°	4								

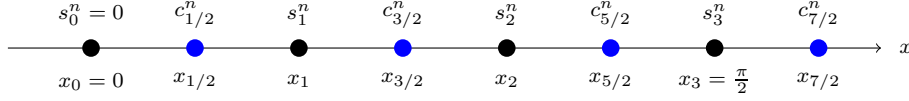


FIG. 2.2. Variables in Bürgi's algorithm in a staggered finite difference setting.

The obtained sine values are correct up to a common factor and must be scaled to achieve $\sin 90^\circ = 1$. Due to the clever choice of the initial guess (inspired by Bürgi himself), the results for $\sin 30^\circ = \frac{1}{2}$ are correct for all n . The result for $\sin 60^\circ \approx \frac{2340}{2702}$, after 5 iterations, is correct to 6 digits⁸.

Written in matrix notation, the two steps of Bürgi's algorithm become

$$\begin{pmatrix} 1 & -1 & \\ & 1 & -1 \\ & & 2 \end{pmatrix} \begin{pmatrix} c_{1/2}^n \\ c_{3/2}^n \\ c_{5/2}^n \end{pmatrix} = \begin{pmatrix} s_1^n \\ s_2^n \\ s_3^n \end{pmatrix}, \quad \begin{pmatrix} 1 & & \\ -1 & 1 & \\ & -1 & 1 \end{pmatrix} \begin{pmatrix} s_1^{n+1} \\ s_2^{n+1} \\ s_3^{n+1} \end{pmatrix} = \begin{pmatrix} c_{1/2}^{n+1} \\ c_{3/2}^{n+1} \\ c_{5/2}^{n+1} \end{pmatrix}. \quad (2.8)$$

Inserting the second expression of (2.8) into the first we obtain

$$\begin{pmatrix} 1 & -1 & \\ & 1 & -1 \\ & & 2 \end{pmatrix} \begin{pmatrix} 1 & \\ -1 & 1 \\ & -1 & 1 \end{pmatrix} \begin{pmatrix} s_1^{n+1} \\ s_2^{n+1} \\ s_3^{n+1} \end{pmatrix} = \begin{pmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -2 & 2 \end{pmatrix} \begin{pmatrix} s_1^{n+1} \\ s_2^{n+1} \\ s_3^{n+1} \end{pmatrix} = \begin{pmatrix} s_1^n \\ s_2^n \\ s_3^n \end{pmatrix}. \quad (2.9)$$

We thus recognize the finite difference Laplacian with a Neumann condition at the right boundary. Bürgi's algorithm can therefore be interpreted in modern terms as an inverse power iteration on the discrete Laplacian and thus converges to the lowest eigenmode, which is the sine function $\sin x$.

We admire in Bürgi's work the systematic use of symmetric formulas on, as we say today, a staggered grid ("Doch etwas umb eine halbe Zall erhoben"). Unfortunately, he did not show it to competent readers as Kepler, but offered it in 1592 to the emperor Rudolph II, who discarded it, so that it was only rediscovered four centuries later by Menso Folkerts (see [55], [88, p. 145]). Bürgi's algorithm is a discrete version of an algorithm of Joh. Bernoulli relating to iterated involutes (see [86]) which is also the subject of a manuscript of Lagrange (see [88]).

2.4. Indian Astronomical Tables and Ḥabash's Problem. Not all processes representing smooth periodic phenomena observed by ancient Indian and Arabic astronomers behaved like a nice sine function with the maximal value precisely in the middle of $[0, \pi]$. Therefore they started to develop tables of "sine" functions whose maximum has been moved out of this mid point. A manuscript by Ḥabash al-Ḥāsib al-Marwazī (Baghdad, 9th century), explaining such a procedure, has been kept in Istanbul and analysed by Edward Stewart Kennedy (1912–2009) (see [103], [101] and [102]).

We reproduce from [101, p. 52] the following values (in base 60) of Ḥabash compared to the values of al-Khwārizmī:

	30°	60°	90°	120°	150°	
Ḥabash	1; 10, 56	1; 35, 28	1; 28, 52	1; 6, 12	0; 35, 8	
al-Khwārizmī	1; 11, 5	1; 35, 27	1; 28, 51	1; 6, 1	0; 34, 51	

⁸This, together with $\sin 60^\circ = \frac{\sqrt{3}}{2} = \frac{3}{2\sqrt{3}}$, leads to $\sqrt{3} \approx \frac{1351}{780}$, a value known to Archimedes.

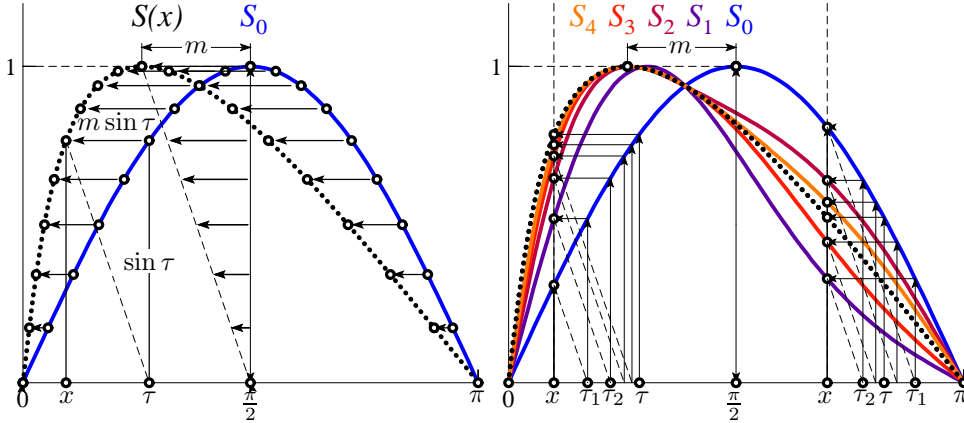


FIG. 2.3. The shear transformed $S(x)$ (dotted) of $S_0(x) = \sin(x)$ for $m = 0.75$ (left); the convergence $\tau_i \rightarrow \tau$ and the functions $S_i(x) = \sin(\tau_i(x))$ for $i = 1, 2, 3, 4$ (right).

Apparently, in this function, the highest point has been moved from $\frac{\pi}{2}$ to a point $\frac{\pi}{2} - m$ with the help of a shear transformation applied to a symmetric sine function. For simplicity, we take the maximal value equal to 1 and move every point of $y = \sin \tau$ horizontally from τ to x by an amount depending linearly on y (see Fig. 2.3, left):

$$x = \tau - m \cdot \sin \tau. \quad (2.10)$$

The problem is thus, for a given x , to find the value of τ satisfying (2.10), in order to have $S(x) = \sin \tau$. According to Kennedy, “Ḥabash (...) gives a neat recursion relation. (...) in all the sources the processes are written out as verbal statements” [103, p. 82]. This “neat recursion relation” was the fixed point iteration

$$\tau_0 = x, \quad \tau_{i+1} = x + m \cdot \sin \tau_i \quad (i = 0, 1, 2, \dots). \quad (2.11)$$

Geometrically, this means that the lines connecting $(x, \sin \tau_i)$ with $(\tau_{i+1}, 0)$ are all parallel with slope $-\frac{1}{m}$ (see Fig. 2.3, right). The algorithm is thus a movement along zig zag lines and convergence can nicely be observed, as long as $m < 1$. If we set $S_i(x) = \sin \tau_i$, we obtain a sequence of functions converging to the required $S(x)$.

Recomputing the historical values. Most of the old tables used the value of $m = 24^\circ$, which is close to the obliquity of the ecliptic, so that the maximal value of S moves to $x = 66^\circ$. They also multiplied the sine values by $k = 1;36$, which is $4m$. So we use $m = \frac{24\pi}{180} = 0.41887902$ and $k = 1.6$ and represent the final values in base 60 with one additional digit in Table 2.1. We can clearly observe that Ḥabash computed precisely four iterations, while the values of al-Khwārizmī, calculated by trigonometric interpolation, are more precise. We see also that towards the borders of $[0, \pi]$, where $f'(\tau) = m \cdot \cos \tau$ is largest, the convergence slows down.

According to Kennedy “Ḥabash does not claim to have originated the algorithm [sic], and there is reason to think that it came to him along with many other techniques used by the astronomers of India” [102, p. 248] because “the techniques in which this section of Ḥabash’s work is embedded are demonstrably Hindu” [103, p. 83].

2.5. Kepler’s Laws and Kepler’s Problem. One of the great moments in the history of science was when Johannes Kepler (1571–1630) extracted from the precise

It.	30°	60°	90°	120°	150°
S_0	0; 47, 59, 60	1; 23, 8, 18	1; 36, 0, 0	1; 23, 8, 18	0; 47, 59, 60
S_1	1; 4, 14, 12	1; 34, 45, 39	1; 27, 42, 1	1; 0, 41, 41	0; 29, 39, 56
S_2	1; 9, 7, 31	1; 35, 25, 6	1; 29, 3, 24	1; 7, 40, 34	0; 36, 52, 1
S_3	1; 10, 31, 50	1; 35, 26, 54	1; 28, 50, 34	1; 5, 34, 16	0; 34, 3, 51
S_4	1; 10, 55, 43	1; 35, 26, 59	1; 28, 52, 36	1; 6, 12, 44	0; 35, 9, 33
S_5	1; 11, 2, 27	1; 35, 26, 59	1; 28, 52, 17	1; 6, 1, 3	0; 34, 43, 55
S_6	1; 11, 4, 21	1; 35, 26, 59	1; 28, 52, 20	1; 6, 4, 36	0; 34, 53, 55
S_7	1; 11, 4, 53	1; 35, 26, 59	1; 28, 52, 20	1; 6, 3, 31	0; 34, 50, 1
S_8	1; 11, 5, 2	1; 35, 26, 59	1; 28, 52, 20	1; 6, 3, 51	0; 34, 51, 32
S_9	1; 11, 5, 4	1; 35, 26, 59	1; 28, 52, 20	1; 6, 3, 45	0; 34, 50, 57
S_{10}	1; 11, 5, 5	1; 35, 26, 59	1; 28, 52, 20	1; 6, 3, 47	0; 34, 51, 11

TABLE 2.1
Recomputing the historical values with $m = 24^\circ$ and $k = 1; 36$.

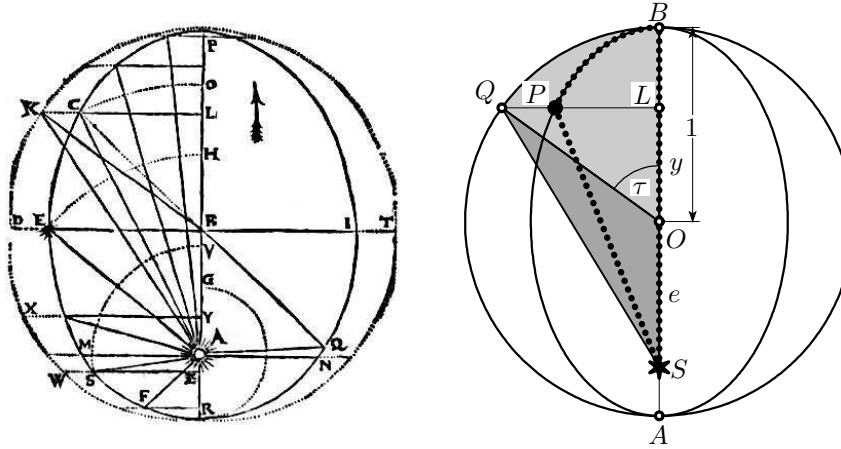


FIG. 2.4. Kepler's Laws (left: Kepler's Picture from [104, p. 676]).

measurements of the orbit of Mars⁹ by Tycho Brahe (1546–1601) the two laws (see Fig. 2.4): • Planets move on elliptic orbits with the Sun in one of the foci [104, Chap. 56];

• The surface of the area enclosed by the dotted line (see Fig. 2.4 right) is proportional to the time [104, Chap. 40].

We take the semi-major axis a as 1¹⁰, then $OS = e$ is the *eccentricity*. The angle τ is called the *eccentric anomaly* and determines the position P of the planet. By homothety $P \mapsto Q$, the dotted area is proportional to the area dashed in grey, which is the union of the circle sector OBQ of area $\frac{\tau}{2}$ and the triangle SOQ of area $\frac{e}{2} \sin \tau$. Therefore, if t denotes the time, we have

$$\text{Const. } t = \tau + e \sin \tau. \quad (2.12)$$

⁹Mars was the planet with the largest eccentricity and refused to obey all the ancient theories of circular movements.

¹⁰Kepler, in his calculations, did nearly the same, he used $a = 10^5$ in order to avoid decimal fractions.

If $0 \leq t \leq T$ runs through an entire period of the orbit, τ performs an entire revolution 2π , while $\sin \tau$ returns to 0, hence the constant in (2.12) is $\text{Const.} = \frac{2\pi}{T}$.

Kepler's Problem is now the question: *For a given time t , find the position of the planet, i.e. find the eccentric anomaly τ , such that equation (2.12) is satisfied.* It is a nice coincidence that this equation is precisely the same as Ĥabash's equation (2.10), and that Kepler solves his equation with the same iteration (2.11), i.e.

$$\tau_{i+1} = \text{Const.} \cdot t - e \sin \tau_i. \quad (2.13)$$

Kepler's numerical example [105, pp. 695–696]. In his calculations, Kepler used degrees for the angles and normalised the time t to angles, also measured in degrees. We thus multiply equation (2.13) by $\frac{180}{\pi}$, insert 0.09265 for the eccentricity of Mars and use Kepler's value $50^\circ 9' 10''$ for the normalised time t . This gives the recursion¹¹

$$\tau_{i+1} = 50^\circ 9' 10'' - 11910'' \sin \tau_i \quad (2.14)$$

with starting approximation (“prima positione”) $\tau_0 = 44^\circ 25'$, and leads to the results

	from the recursion	Kepler's values
τ_0	$44^\circ 25' 00''$	$44^\circ 25'$
τ_1	$47^\circ 50' 15''$	$47^\circ 50'$
τ_2	$47^\circ 42' 02''$	$47^\circ 42' 17''$

and the next following correct values from (2.14) are $\tau_3 = 47^\circ 42' 21''$ and $\tau_4 = 47^\circ 42' 20''$.

2.6. Fourier's zigzags. A better understanding of solving equations by iteration began with the work of Leonhard Euler (1707–1783) and Joseph Fourier (1768–1830).

In his monumental treatise [60] on heat conduction, Fourier discussed in Chapter V the propagation of heat in a solid sphere. By his method of separation of variables, adjusting mixed boundary conditions at the surface leads for the eigenvalues ε to the equation

$$\tan \varepsilon = \frac{\varepsilon}{\lambda}, \quad (2.15)$$

where λ is a fixed constant $0 < \lambda < 1$. We have thus to find the intersections of the curve $u = \tan \varepsilon$ with the straight line $u = \frac{\varepsilon}{\lambda}$ of slope > 1 . This problem has apparently “an infinite number of real roots” for each of the branches of $\tan \varepsilon$. Fourier explains his method on the first non-trivial root $0 < \varepsilon < \frac{\pi}{2}$ as follows ([60, §286], see Fig. 2.5, left and his Fig. 13): start from an ε , $u = \tan \varepsilon$ above the solution point, move down vertically to our line, i.e., $u' = \frac{\varepsilon}{\lambda}$ and move horizontally back to the curve, i.e., $\varepsilon' = \arctan u'$. Apparently, the sequence of values $\varepsilon, \varepsilon', \varepsilon'', \varepsilon''', \dots$ would converge from above to the solution. Similarly, by starting from below ([60, §287], see Fig. 2.5, right, his Fig. 14) the same sequence converges from below to the solution. In our notation, we thus have the iteration procedure

$$\varepsilon_{n+1} = \arctan \frac{\varepsilon_n}{\lambda} \quad (2.16)$$

¹¹Unfortunately, Kepler had the misprint 11910'' instead of the correct 19110''; the computed values are therefore mathematically correct, but astronomically wrong.

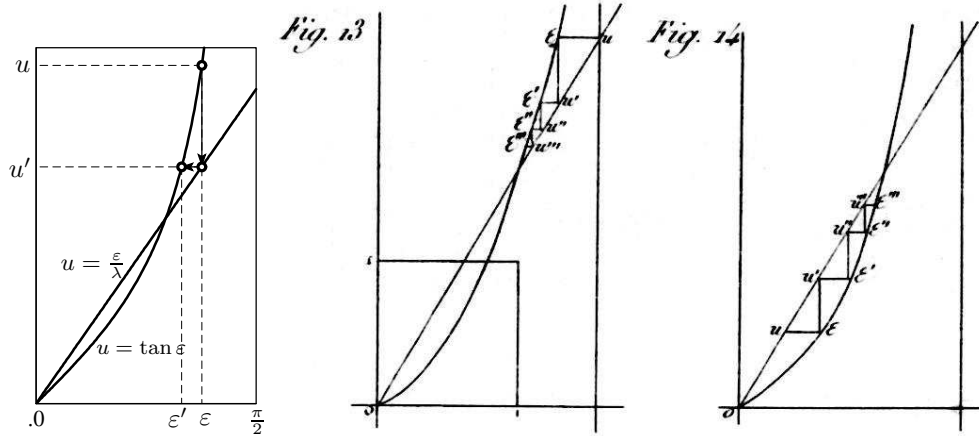


FIG. 2.5. Fourier's figure for computing the first positive root of (2.15), starting with an overestimate (middle) and an underestimate (right).

converging for any initial value $0 < \varepsilon_0 < \frac{\pi}{2}$.

Fourier concludes by saying that he has explained “this approximation procedure, because it is based on a remarkable construction and can be useful in many other cases” and makes visible at once “the nature and the limits of the roots” [60, §288].

2.7. Euler's iterated exponentials. Many phenomena of fixed point iterations were discovered by Euler in [49, E489] by studying his *exponentialibus replicatis*. Inspired by the *profundissimas speculationes* of *Illustr. Marchio de Condorcet*, Euler wanted to understand the effects of repeating infinitely often the exponential function. The problem is thus: for a given constant r and initial value α , study the behaviour of

$$r^\alpha = \beta, \quad r^{r^\alpha} = r^\beta = \gamma, \quad \text{and simili modo } r^\gamma = \delta, \quad r^\delta = \varepsilon, \quad r^\varepsilon = \zeta \text{ etc.} \quad (2.17)$$

For $r = 2$ and $\alpha = 2$ the iterates are 2, 4, 16, 65536, and already the next term *ex 19729 figuris*, a sequence which *multo rapidius in immensum esse excreturam*. Hence, let us choose a smaller value of r and *examinemus casum* $r = \frac{3}{2}$. For calculating numerically powers like $\beta = r^\alpha$, taking *logarithmos* turns the exponential into a product, and a second *similique modo* turns the product into a sum. So Euler computed, by adding up 7-digit logarithms,

$$1.5000 \mapsto 1.8371 \mapsto 2.1062 \mapsto 2.3490 \mapsto 2.5920 \mapsto 2.8604 \mapsto 3.1893 \mapsto 3.6443. \quad (2.18)$$

For this slowly increasing sequence (*lente increscunt*) we have doubts as to its convergence (*dubitare queamus (...) convergant*).

For a better understanding, Euler made numerous numerical tests and tried to represent the algorithm using a picture (*Solutio geometrica eiusdem problematis*). We represent here, in Fig. 2.6, some of Euler's calculations in “Fourier style” for various values of r : for $r = \frac{3}{2}$ there is no solution to $r^x = x$ (this happens for $r > e^{\frac{1}{e}} = 1.44467$, [49, §8]), see Figure 2.6 (top left); for $r = \sqrt{2}$ there are two solutions $\varphi = 2$ (stable) and $\psi = 4$ (unstable), see Figure 2.6 (top right); for $r = \frac{1}{2}$ the derivative f' becomes negative and we have alternating convergence, see Figure 2.6 (bottom left); finally for $r = \frac{1}{20}$ the derivative $f' < -1$ (this happens for $r < \frac{1}{e^e} = \frac{1}{15.154}$, [49, §39]), thus the algorithm converges alternatingly to two values, see Figure 2.6 (bottom right).

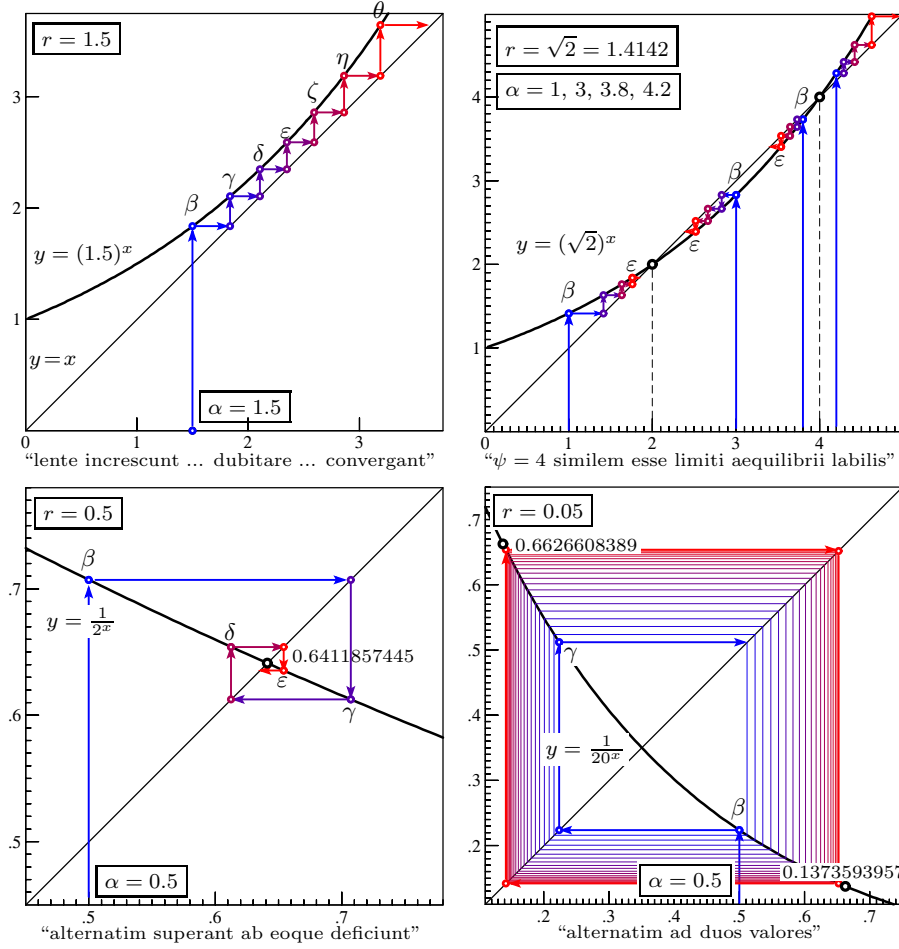


FIG. 2.6. Representation of Euler's calculations for various values of r (from [49]).

3. Emergence of Newton's Method.

"Isaac Newton was not a pleasant man. His relations with other academics were notorious, with most of his later life spent embroiled in heated disputes." (S. Hawking, *A Brief History of Time*, 1988, p. 181)

"Il est facheux que Mr. NEWTON se soit contenté d'étaler ses découvertes sans y joindre les Démonstrations, & qu'il ait préféré le plaisir de se faire adminer à celui d'instruire."

(G. Cramer, *Introduction à l'analyse Éc.*, 1750, pp. XIII–IX)

Like for the fixed point iterations, also Newton's method, today written as

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad \text{for solving a nonlinear equation } f(x) = 0, \quad (3.1)$$

emerged slowly over the centuries. Ypma [184, p. 76] traced back the influence of François Viète (1540–1603) and William Oughtred (1574–1660) on Newton's interest for solving nonlinear equations. Although the origin of the methods in the body of knowledge at Newton's disposal presents some gray areas, it is clear that already the Babylonian algorithm in Section 1.1 as well as the Indian version in Section 1.3 can

$y^3 - 2y - 5 = 0$		$+ 2,10000000$ $- 0,00544852$ $+ 2,09455148, \&c. = y$
$2 + p = y$	$+ y^3$ $- 2y$ $- 5$	$+ 8 + 12p + 6p^2 + p^3$ $- 4 - 2p$ $- 5$
SOMME.		$- 1 + 10p + 6p^2 + p^3$
$0,1 + q = p$	$+ p^3$ $+ 6p^2$ $+ 10p$ $- 1$	$+ 0,001 + 0,03q + 0,3q^2 + q^3$ $+ 0,06 + 1,2 + 6$ $+ 1, + 10,$ $- 1,$
SOMME.		$+ 0,061 + 11,23q + 6,3q^2 + q^3$
$- 0,0054 + r = q$	$+ q^3$ $+ 6,3 q^2$ $+ 11,23 q$ $+ 0,061$	$- 0,00000157488 + 0,000087487 - 0,000021^2 + r^3$ $+ 0,000183708 - 0,000087488 + 0,000021^2$ $- 0,0000642 + 11,23$ $+ 0,061$
SOMME.		$+ 0,0000416 + 11,162r$
$- 0,00004852 + s = r$		

FIG. 3.1. Newton's calculations to find the root $\simeq 2.09455148$ of $x^3 - 2x - 5 = 0$ [130, p. 7].

be seen of being the same as (3.1) with

$$f(a) = a^2 - N, \quad f'(a) = 2a,$$

for which the rapid convergence (1.4), (1.7) is typical.

3.1. Newton's Famous Example. In his book *The Method of Fluxions and Infinite Series* (manuscript of 1671, published "translated from the author's Latin original not yet made publick", London 1736), Newton explains how to solve $x^3 - 2x - 5 = 0$ (see Fig. 3.1; the same array was published by Wallis fifty years earlier [176, p. 338] and also appears in Newton's (unpublished) *De analysi per aequationes infinitas*, 1669).

By computing some values of this polynomial, one sees that $x = 2$ is a good approximation to the wanted root. Newton thus defines $x = 2 + p$ (with p small); inserted into $x^3 - 2x - 5 = 0$, this gives the "SOMME" (see the third box in Fig. 3.1)

$$-1 + 10p + 6p^2 + p^3 = 0. \quad (3.2)$$

Usually in Algebra, the higher the degree of the polynomial is, the harder the solution process becomes. But here p is small, thus the higher powers p^2 and p^3 can be neglected and we obtain $p = \frac{1}{10}$ ¹².

The next step is to put $p = \frac{1}{10} + q$ (with q even smaller than p), which inserted into (3.2) gives (Fig. 3.1, fifth box)

$$0.061 + 11.23q + 6.3q^2 + q^3 \Rightarrow q = -\frac{0.061}{11.23} = -0.0054 \quad (3.3)$$

and finally with $q = -0.0054 + r$ we obtain (Fig. 3.1, seventh box)

$$0.0005416 + 11.162r + \dots \Rightarrow r = -\frac{0.0005416}{11.162} = -0.00004852. \quad (3.4)$$

¹²Newton does not tell the reader that -1 is the value of this polynomial at $p = 0$ and 10 the corresponding *Fluxion*, so that $\frac{1}{10}$ is the correction in (3.1).

“But since the description of this curve [the contracted cycloid] is difficult, a solution by approximation will be preferable.” [Cæterum ob difficultatem describendi hanc curvam præstat constructiones vero proximas in praxi Mechanica adhibere.] [129, p. 109] or [131, p. 157]

$$E = \frac{N - AOQ + D}{1 - e \cos(AOQ)} \quad \text{and similar further corrections } G \text{ and } I. \quad (3.6)$$

“(...) the infinite series $AOQ + E + G + I + \&c.$ converges so very fast, that it will be scarcely ever needful to proceed beyond the second term E .”

$$\tau_1 - \tau_0 = \frac{\tilde{t} - \tau_0 - e \sin \tau_0}{1 + e \cos \tau_0}, \quad (3.7)$$
$$\begin{aligned} & 44^\circ, 25, 00 \\ & 47^\circ, 42, 06, 45, 55, 18, 39, 15, 12, 48, 01, 24, 34, 50, 47, 08, 59, 50, 12, 07, 22, 48, 02, 14, 14, 59, 34 \\ & 47^\circ, 42, 20, 12, 30, 08, 26, 30, 53, 11, 16, 27, 48, 11, 14, 27, 02, 05, 55, 45, 47, 59, 59, 56, 07, 43, 12 \\ & 47^\circ, 42, 20, 12, 30, 12, 19, 55, 12, 03, 17, 56, 39, 42, 55, 52, 28, 10, 08, 22, 29, 43, 58, 58, 55, 53, 02 \\ & 47^\circ, 42, 20, 12, 30, 12, 19, 55, 12, 03, 17, 56, 59, 15, 40, 46, 11, 13, 28, 22, 54, 32, 28, 42, 48, 29, 31 \end{aligned} \quad (3.8)$$

Let us thus compute here “in numbers” a famous ancient problem, the cube root $\sqrt[3]{2}$, necessary for ‘doubling the cube’ in the Oracle of Delphi, which gives¹⁴

$$f(x) = x^3 - 2 \Rightarrow x_{n+1} = x_n - \frac{x_n^3 - 2}{3x_n^2}$$

$$= \frac{1}{3}(2x_n + \frac{2}{x_n^2})$$

21

P R O P. II.

Proponatur $ba - aaa = c$

Sumatur (g) quantitas quaecunq; minor (a). Dico proximam (g) (per methodum nostram) enatam, semper maiorem esse praecedenti, minorem vero quam (a), ac proinde ad verum convergere,

Ex hypothesi $g + z = a$. Erit $bg - ggg + b - 3gg \cdot z - 3gzz - zzz = ba - aaa = c$
Ergo $b - 3gg \cdot z - 3gzz - zzz = c + ggg - bg$. Ergo $+z = +x + \frac{3gzz + zzz}{b - 3gg} = \frac{c + ggg - bg}{b - 3gg} + x$

Seu Theoremati convergenti, inde. $+z = +x + \frac{3gzz + zzz}{b - 3gg}$ utrique parti addatur (g)

proveniet $g + z = x + g + \frac{3gzz + zzz}{b - 3gg}$ Sed (g) nova $= g + x$ maior praecedenti,

quantitate (x,) minor vero (a,) quantitate $\frac{3gzz + zzz}{b - 3gg}$, pars suo toto. Q. E. D.

FIG. 3.3. Raphson's method for solving equation (3.9) [142, p. 8]. (In line 9, read $g + z = a = \dots$)

3.4. Raphson's contribution. Very often, Newton's method is called "Method of Newton-Raphson". The reason for this is the book entitled *Analysis æquationum universalis* [142] (first published as a tract in 1690) written by Joseph Raphson (c. 1648–c. 1715), which contains thirty-four problems that require solving polynomial equations [142, pp. 9–39]. Raphson knows Newton's approach¹⁵ but, according to him, his method "is not of the same origin" and does not have "the same development" [184, p. 548].

One of the problems treated by Raphson was the *trisection of the angle* 60° in a circle of radius 10, for which the chord becomes $a = 20 \sin 10^\circ$. If we insert this into equation (2.3), we obtain the equation

$$300a - a^3 = 1000 \quad \text{which Raphson writes generally as} \quad ba - a^3 = c, \quad (3.9)$$

a cubic equation to solve. We see in Fig. 3.3 that Raphson takes g as a first approximation and searches a correction z such that $g + z = a$. Inserting this into (3.9) we have

$$bg - g^3 + (b - 3g^2) \cdot z - 3gz^2 - z^3 = c.$$

Newton would here neglect z^2 and z^3 and conclude that $z = \frac{c + g^3 - bg}{b - 3g^2}$. Raphson, since his method "is not of the same origin", called this quantity x and developed in three additional lines of calculations the formula containing the additional error term¹⁶

$$z = x + \frac{3gz^2 + z^3}{b - 3g^2}. \quad (3.10)$$

For a numerical application [142, p. 20], Raphson starts with $g = 3$ and obtains the iterates 3.4, 3.472, 3.4729636, 3.472963553338607 with all 15 digits correct (see Fig. 3.4).

The end of Raphson's book contains a set of tables giving the appropriate formulas of the correction x for a list of sixty-four polynomial equations up to degree four [142,

¹⁵He also solves Newton's equation $x^3 - 2x - 5 = 0$ in the tract of 1690, see [184, p. 546] or [8, p. 151]. This second source gives an original reproduction of the calculations. Raphson's final solution is 2.0945514815427104141 (where the last 7 digits are incorrect).

¹⁶This error term can however not really be used effectively, so Raphson's method is identical to Newton's method.

$$\begin{array}{r}
\begin{array}{r}
g = 3 \\
c + ggg - bg = + 127 \\
b - 3gg = + 273 \quad + 127.0 \quad (+.4 = x \\
\hline
3. \\
+ .4 \\
\hline
\end{array} \\
\begin{array}{r}
g = 3.4 \\
c + ggg - bg = + 19.304 \\
b - 3gg = 265.32 \quad + 19.3040 \quad (+.072 = x \\
\hline
3.4 \\
+ .072 \\
\hline
\end{array} \\
\begin{array}{r}
g = 3.472 \\
c + ggg - bg = + 254210048. \\
b - 3gg = 263.835648 \quad + 2542100480 \quad (+.0009636 = x \\
\hline
3.472 \\
+ .0009636 \\
\hline
\end{array} \\
\begin{array}{r}
g = 3.4729636 \\
c + ggg - bg = -.0000123100020899. \\
b - 3gg = 263.8155715 \quad -.0000123100020899 \quad (-.00000046661393 = x \\
\hline
3.4729636 \\
-.000000046661393 \\
\hline
a = 3.472963553338607
\end{array}
\end{array}$$

FIG. 3.4. Raphson's method to calculate $20 \sin 10^\circ = 3.472963553338606977 \dots$

pp. 41–45]. However, Raphson does not give a general expression involving a derivative as in (3.1). He will never associate the calculus with his iterative technique and he never extended it to problems other than polynomial equations¹⁷.

3.5. Simpson's fluxion approach. In the sixth essay of his book [160, pp. 81–86], published in 1740, Thomas Simpson (1710–1761) explains “a new method for the solution of all kinds of algebraical equations in numbers”. According to him, “as it is more general than any hitherto given, [this method] cannot but be of considerable use¹⁸”. Simpson is the first to use explicitly the (then) modern calculus¹⁹, but was afraid that “(...) it perhaps may be objected, that the Method of Fluxions, whereon it is founded, being a more exalted Branch of the Mathematicks, cannot be so properly applied to what belongs to common Algebra²⁰”. Here is the explanation of his method:

“Take the Fluxion of the given Equation (be it what it will) supposing, x , the unknown, to be the variable Quantity ; and having divided the whole by \dot{x} , let the Quotient be represented by A . Estimate the Value of x pretty near the Truth, substituting the same in the Equation, as also in the Value of A , and let the Error, or resulting Number in the former, be divided by this numerical Value of A , and let the Quotient be subtracted from the said former Value of x ; and from thence will arise a new Value of that Quantity much nearer to the Truth than the former, wherewith proceeding as before, another new Value may be had, and so another, &c. 'till we arrive to any Degree of Accuracy desired.” [160, p. 81]

¹⁷[107, pp. 349–350], [184, p. 547].

¹⁸[160, p. vii].

¹⁹Thereby Kollerstrom [107, p. 351] suggests to consider Simpson as the inventor of this method!

²⁰[160, p. vii].

We see how cumbersome the statement of Formula (3.1) was without proper notation. The text “the Fluxion of the given Equation (...) having divided the whole by \dot{x} , let the Quotient be represented by A ” is Newton’s notation $\frac{\dot{f}}{\dot{x}} = A$ (whereas Leibniz wrote $\frac{df}{dx}$ and Lagrange $f'(x)$ for the same quantity).

Simpson then presented two examples, the first being the same as Raphson’s computation in Fig. 3.4, giving after three iterations from the starting value 3.5 the solution $x = 3.47296351$ “which is true, at least, to 7 or 8 Places”. The second example is the solution of

$$\sqrt{1-x} + \sqrt{1-2x^2} + \sqrt{1-3x^3} = 2$$

for which, starting from $x_0 = 0.5$, he obtained $x_1 = 0.557$ and $x_2 = 0.5516$ (correct value 0.551586152497047117247685268).

Systems of equations. While the origin of Newton’s method in one variable can be debated, the first use of this method for *systems of nonlinear equations*

$$f(x, y) = 0, \quad g(x, y) = 0 \quad (3.11)$$

is clearly in this work of Simpson. He stated the method as follows:

“Take the Fluxions of both the Equations, considering x and y as variable, and in the former collect all the Terms, affected with \dot{x} , under their proper Signs, and having divided by \dot{x} , put the Quotient = A ; and let the remaining Terms, divided by \dot{y} , be represented by B : In like manner, having divided the Terms in the latter, affected with \dot{x} , by \dot{x} , let the Quotient be put = a , and the rest, divided by \dot{y} , = b . Assume the Values of x and y pretty near the Truth, and substitute in both the Equations, marking the Error in each, and let these Errors, whether positive or negative, be signified by R and r respectively : Substitute likewise in the Values of A, B, a, b , and let $\frac{Br-bR}{Ab-aB}$ and $\frac{aR-Ar}{Ab-aB}$ be converted into Numbers, and respectively added to the former Values of x and y ; and thereby new Values of those Quantities will be obtained ; from whence, by repeating the Operation, the true Values may be approximated *ad libitum*.” [160, p. 82]

Here, “the former” is $f(x, y)$, so that (“divided by \dot{x} ”) $A = \frac{\partial f}{\partial x}$ and (“divided by \dot{y} ”) $B = \frac{\partial f}{\partial y}$; “the latter” is $g(x, y)$, so that $a = \frac{\partial g}{\partial x}$ and $b = \frac{\partial g}{\partial y}$; “and let these errors...” $R = f(x_0, y_0)$, $r = g(x_0, y_0)$. With these notations, Simpson’s algorithm

$$x_1 = x_0 + \frac{Br - bR}{Ab - aB}, \quad y_1 = y_0 + \frac{aR - Ar}{Ab - aB} \quad (3.12)$$

is

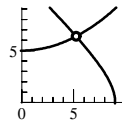
$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \frac{1}{Ab - Ba} \begin{pmatrix} b & -B \\ -a & A \end{pmatrix} \begin{pmatrix} R \\ r \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \left(\begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix} \right)^{-1} \begin{pmatrix} f_0 \\ g_0 \end{pmatrix}, \quad (3.13)$$

which is the same as (3.1) with $\frac{1}{f'(x)}$ replaced by the inverse of the Jacobian J .

Moreover, Simpson observes that the method converges quadratically: “when x and y are near the Truth, (the method) doubles the Number of Places at each Operation (...)” [160, p. 83]. Let us present in detail three examples given by Simpson.

Example 1. This is the very first solution by Newton’s method of a nonlinear system ever published. Solve

$$\begin{cases} y + \sqrt{y^2 - x^2} - 10 = 0, \\ x + \sqrt{y^2 + x} - 12 = 0. \end{cases} \quad (3.14)$$



The Fluxions here being $\dot{y} + \frac{\dot{y}\dot{x} - x\dot{x}}{\sqrt{yy - xx}}$ and $\dot{x} + \frac{\dot{y}\dot{x} + \frac{1}{2}\dot{x}}{\sqrt{yy + x}}$
or $\dot{y} + \frac{\dot{y}\dot{x}}{\sqrt{yy - xx}} - \frac{x\dot{x}}{\sqrt{yy - xx}}$, and $\dot{x} + \frac{\frac{1}{2}\dot{x}}{\sqrt{yy + x}} + \frac{\dot{y}\dot{x}}{\sqrt{yy + x}}$,
we have A equal $-\frac{x}{\sqrt{yy - xx}}$, B equal $1 + \frac{y}{\sqrt{yy + x}}$, $a = \left(\begin{array}{cc} \frac{-x}{\sqrt{y^2 - x^2}} & 1 + \frac{y}{\sqrt{y^2 - x^2}} \\ 1 + \frac{1}{2\sqrt{y^2 + x}} & \frac{y}{\sqrt{y^2 + x}} \end{array} \right)$
 $1 + \frac{\frac{1}{2}}{\sqrt{yy + x}}$, and $b = \frac{y}{\sqrt{yy + x}}$ (*Cafe II.*)

FIG. 3.5. Simpson's calculation of the Jacobian for the system (3.14) (with 1 printing error).

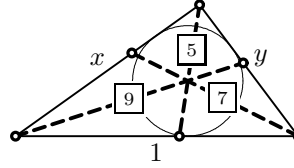
See in Fig. 3.5 how Simpson computed the Jacobian of the system. He then performed two iterations using Formula (3.12) from the initial values (5, 6) as

$$\mathbf{x}^{(0)} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} 5.23 \\ 6.37 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 5.23263 \\ 6.36898 \end{pmatrix}, \quad \text{correct: } \mathbf{x} = \begin{pmatrix} 5.2326413353 \\ 6.3690267672 \end{pmatrix}.$$

He added that the equations “had been much easier solved, had they been first reduced (...) to $20y - xx - 100 = 0$, and $yy - xx + 25x - 144$ equal 0”.

Example 2. Solve

$$\begin{cases} 49\left(x - \frac{x}{(x+y)^2}\right) - 25\left(1 - \frac{x^2}{(1+y)^2}\right) = 0, \\ 81\left(1 - \frac{x^2}{(1+y)^2}\right) - 49\left(\frac{x}{y} - \frac{xy}{(1+x)^2}\right) = 0. \end{cases}$$



Here, Simpson states the Jacobian

$$J = \begin{pmatrix} 49\left(1 + \frac{x-y}{(x+y)^3}\right) + \frac{50x}{(1+y)^2} & \frac{98x}{(x+y)^3} - \frac{50x^2}{(1+y)^3} \\ -\frac{162x}{(1+y)^2} + 49\left(\frac{y}{(1+x)^2} - \frac{1}{y} - \frac{2xy}{(1+x)^3}\right) & \frac{162x^2}{(1+y)^3} + 49x\left(\frac{1}{y^2} + \frac{1}{(1+x)^2}\right) \end{pmatrix}$$

without comment and writes that an initial guess (0.8, 0.6) is facilitated by knowing “the Nature of the *Problem* from whence those Equations are derived”, namely “when it is known, that 1, x , and y , are the Sides of a Plain Triangle, wherein Lines, drawn to bisect each Angle and terminate in those Sides, are to one another, respectively, as 5, 7, and 9”. Then he obtains with formula (3.12)

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} 0.799 \\ 0.582 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 0.79912 \\ 0.58138 \end{pmatrix}, \quad \text{correct } \mathbf{x} = \begin{pmatrix} 0.799178525272 \\ 0.581448098856 \end{pmatrix}.$$

Example 3. Solve

$$\begin{cases} x^x + y^y - 1000 & = & 0 \\ x^y + y^x - 100 & = & 0. \end{cases}$$

We leave it to the reader to decipher and check Simpson's formulas for the Jacobian $A = \overline{1 + L : x} \times x^x$, $B = \overline{1 + L : y} \times y^y$, $a = \frac{y}{x} \times x^y + y^x L : y$, $b = \frac{x}{y} \times y^x + x^y L : x$. The solutions are symmetric for $x \leftrightarrow y$; we take x to be the greater number. Since, from the first equation, x^x should be close to 1000, we take x as 4.5; “and from the first and second together, that the Difference of x and y must be pretty large”. “I therefore take $y = 2.5$ ”. This leads with formula (3.12) to the sequence of values

$$\mathbf{x}^{(0)} = \begin{pmatrix} 4.5 \\ 2.5 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} 4.55 \\ 2.45 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 4.5519 \\ 2.4495 \end{pmatrix}, \quad \text{correct } \mathbf{x} = \begin{pmatrix} 4.5519514195775 \\ 2.4496246279281 \end{pmatrix}.$$

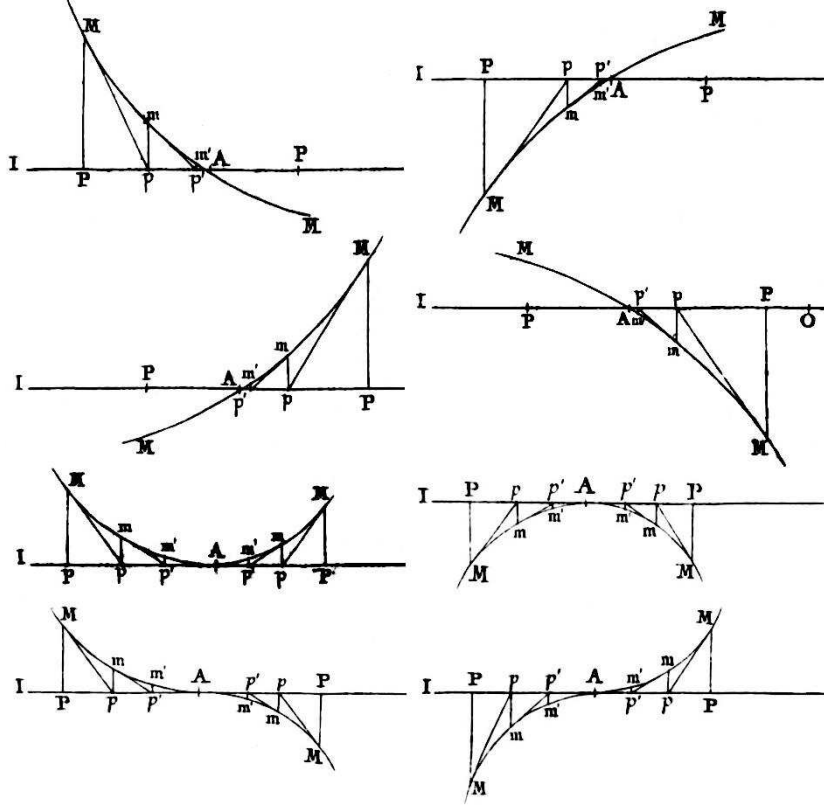


FIG. 4.1. Mourraille's illustrations for the convergence of Newton's method [125, Fig. 41–48].

4. Geometry of Newton's Method.

"Needless to say, computer based observations do not provide a substitute for actual proofs (...) but mathematical study of my observations on the \mathcal{M}^* set turned out to be fruitful and useful, witness Douady and Hubbard 1982 and forthcoming works. Needless to say (again) this mathematical study could not have been undertaken without my computer based observations."
(B. Mandelbrot, [117, p. 231])

4.1. Mourraille's Geometrical Interpretation. In his *Traité de la résolution des équations* from 1768 [125], Jean Raymond Mourraille²¹ (1721–1808, Mayor of Marseille 1791–1793) writes (§103, p. 346): "NEWTON & les autres Auteurs" ("Newton and the other authors") did, perhaps, not pay attention that the terms (...) are the successive tangents of the curve described by the Equation". Mourraille was the first to illustrate the convergence of Newton's method by pictures which are nowadays in many books, not only for simple roots, but also for double and triple roots (see Fig. 4.1).

He then went on to demonstrate the difficulties of the method ("Inconvénients de cette Méthode") with a less trivial example (his "Fig. 54", our Fig. 4.2): a polynomial of degree 4 with two real roots A and B close together. If the starting point x_0 moves

²¹According to Cajori, this book "has remained quite unnoticed by mathematicians" [21, p. 133]. For biographical information about Mourraille, see the references in [21].

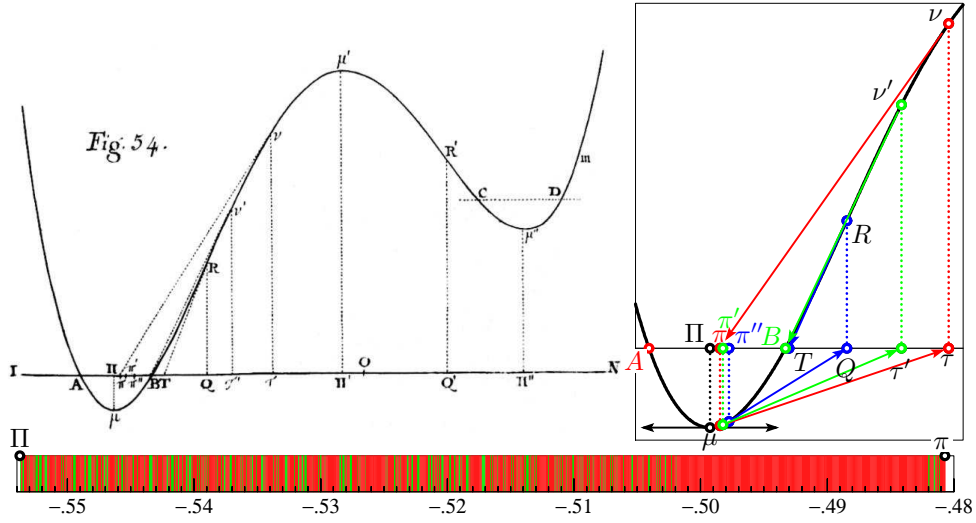


FIG. 4.2. Mourraille's "Fig. 54" to explain different types of behaviour for Newton's method; below: ultimate convergence to A (red) and B (green) for 1000 equidistant initial values x_0 between Π and π for the equation $(x+1)x((x-4)^2+0.2)=0$.

from B to A, then the correction x_1 moves from B to N "d'un mouvement très-accélééré". When x_0 crosses the point Π , where the polynomial is minimal between A and B, the correction x_1 would disappear "à une distance infinie du point B" and come back from the other side "à une distance presque immense" [125, p. 341]. Then come three points π (in red), π' (in green) and π'' (in blue), extremely close together, with totally different convergence behavior:

- The iteration starting in π (red) would, via τ and ν , come back to the point π where it has started ("dans le même point d'où l'on est parti", [125, p. 343]) and thus represents a two-cycle.
- The iteration starting in π' (green) would, via τ' and ν' , lead directly after two steps to the exact solution in B.
- The iteration starting in π'' (blue) would, via Q, come to R, the inflection point; as a consequence, the error of x_2 at T has a local maximum: neighbouring points on either side of the inflection point R would lead to an x_2 with smaller error.

Mourraille claims that the point π is "the limit between the convergence and the divergence of the sequence" and therefore that the initial points between Π and π lead to divergence [125, p. 343], which is not true (see Fig. 4.2, below). Indeed, this is an interval where the basin of attraction is of fractal nature (see also Fig. 4.11).

4.2. Cayley's "Newton-Fourier Imaginary Problem". The first to propose the application of "the Newtonian method as completed by Fourier²²" in the complex plane was Arthur Cayley (1821–1895). Apparently, he judged the subject sufficiently important for six publications: beginning with the "Smith's Prize Examination, Jan. 28, 1879", then in a section "Desiderata and Suggestions" of the *American Journal of Mathematics* (March 3d, 1879), then with proofs in the articles [25], [26], [27] and, more than ten years later, in the note [28] for the French public.

²²See Section 5.1 below.

“(…) throwing aside the restrictions as to reality, we have what I call the Newton-Fourier Imaginary Problem, as follows.

Take $f(u)$, a given rational and integral function of u , with real or imaginary coefficients; ξ , a given real or imaginary value, and from this derive ξ_1 by the formula $\xi_1 = \xi - \frac{f(\xi)}{f'(\xi)}$, and thence $\xi_1, \xi_2, \xi_3, \dots$ each from the preceding one by the like formula.

A given imaginary quantity $x + iy$ may be represented by a point the coordinates of which are (x, y) : the roots of the equation are thus represented by given points A, B, C, \dots and the values ξ, ξ_1, ξ_2, \dots by points P, P_1, P_2, \dots the first of which is assumed at pleasure, and the others each from the preceding one by the like given geometrical construction.” [26, p. 97]

The problem is to determine the behaviour of the iterations in the different regions of the plane. In the case of the quadratic equation $f(z) = z^2 - a^2$ ($a \in \mathbb{C}$), “the solution is easy and elegant” [26, p. 97] and “the division into regions is made without difficulty” [27, p. 232].

In order to simplify the calculations, we suppose with Cayley $a = 1$. For a given $z_0 \in \mathbb{C}$, one step of Newton’s method “as completed by Fourier” is then

$$z_1 = z_0 - \frac{z_0^2 - 1}{2z_0} = \frac{z_0^2 + 1}{2z_0} = \frac{1}{2} \left(z_0 + \frac{1}{z_0} \right), \quad (4.1)$$

(remember that the last expression is Heron’s (1.2)) “and the question is, under what conditions do we thus approximate to one determinate root (selected out of the two roots at pleasure), say $a[+1]$, of the given equation²³.”

We immediately see that for a purely imaginary starting point $z_0 = iy_0$ we obtain $z_1 = \frac{i}{2}(y_0 - \frac{1}{y_0})$, again purely imaginary, so that “all the points $[z_1, z_2, \dots]$ will also be on the y axis so that we will approach neither the point $A [+1]$ nor the point $B [-1]$ ” [28, p. 217]. In the same way, the negative as well as the positive half-planes remain invariant.

Cayley’s main tool for studying the precise movement is the *ratio of the distances of a point z from the roots 1 and -1 respectively*, i.e., the ratio of $z - 1$ and $z - (-1) = z + 1$, as a new complex variable

$$q = \frac{z - 1}{z + 1} \quad \Leftrightarrow \quad z = \frac{1 + q}{1 - q}. \quad (4.2)$$

By construction, $k = |q| < 1$ if and only if z is closer to 1 than to -1 , i.e., $\Re z > 0$ (green in Fig. 4.3). Furthermore, a Newton step (4.1) in the z -plane corresponds in the q -plane to

$$q_1 = \frac{z_1 - 1}{z_1 + 1} = \frac{z_0^2 + 1 - 2z_0}{z_0^2 + 1 + 2z_0} = \frac{(z_0 - 1)^2}{(z_0 + 1)^2} = q_0^2, \quad (4.3)$$

just taking the square. Thus, whenever q_0 moves on a circle of radius $k < 1$, q_1 moves twice as fast on the circle of radius k^2 and all subsequent q_2, q_3, \dots of moduli k^4, k^8, \dots tend to 0. This proves that the corresponding Newton iterates z_2, z_3, \dots converge to 1, “and that very rapidly”.

A simple geometrical construction. If q_0 moves, with $|q_0| = k$, on a circle with diameter $[-k, k]$, the corresponding z_0 moves, using (4.2), on a circle with diameter

²³[25, p. 179].

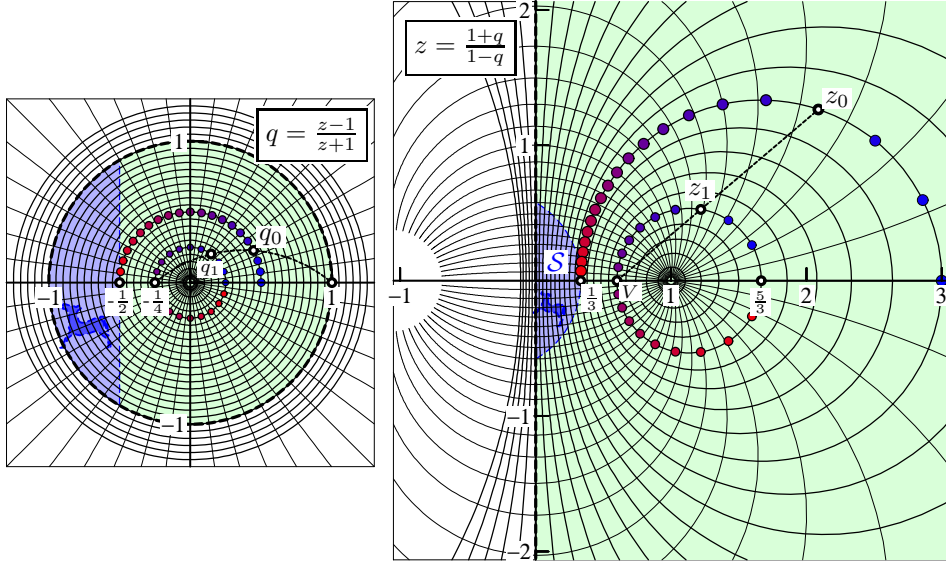


FIG. 4.3. The map $z \mapsto q$ of (4.2); a sequence of q_0 with $k = \frac{1}{2}$, the corresponding z_0 moving on $\mathcal{C}_{\frac{1}{2}}$ with corresponding z_1 moving on $\mathcal{C}_{\frac{1}{4}}$, the red ones approach the limit of inequality (4.5), while the corresponding q_0 approach (4.6); the “segment of unfitness” S in blue.

$\left[\frac{1-k}{1+k}, \frac{1+k}{1-k}\right]$ ²⁴. We denote this circle by \mathcal{C}_k . Consequently, the corresponding values z_1 move on the second circle \mathcal{C}_{k^2} . With the help of $V = \frac{1-k^2}{1+k^2}$, the left intersection of \mathcal{C}_{k^2} with the real axis, Cayley shows how “to get z_1 from z_0 ”:

$$\text{the points } z_0, z_1 \text{ and } V \text{ “are in a right line” [25, p. 181]} \quad (4.4)$$

(see Fig. 4.3 and the lines $z_0 = A_1, V_2, z_1 = A_2$ in Fig. 4.4). This means in the q -plane, since the point 1 goes to ∞ , that the points $-k^2, q_1, q_0, 1$ are concyclic.

Proof of (4.4). Denote the complex numbers $q_1 = k^2 e^{2i\phi}$, $q_0 = k e^{i\phi}$, $-k^2 = a$, $1 = b$ (see Fig. 4.5, right). The angles ω_1 and ω_2 are parallel and thus equal. The sides of the triangle $q_0, 0, a$ are in ratio $k : k^2 = 1 : k$ with exterior angle ϕ , hence $\omega_2 = \arctan \frac{k \sin \phi}{1+k \cos \phi}$. However the complex argument ω_3 of the ratio

$$\frac{q_1 - b}{q_0 - b} = \frac{k^2 e^{2i\phi} - 1}{k e^{i\phi} - 1} = 1 + k e^{i\phi}$$

has the same value. Hence the equality of ω_1 and ω_3 , the peripheral angles at a and b to the arc $q_0 q_1$, implies that the four points are concyclic.

Regular and irregular steps. Cayley distinguishes carefully between regular and irregular steps. A step $z_i \mapsto z_{i+1}$ is *irregular* if the new correction z_{i+1} is *not* closer to the solution than z_i ,

$$|z_{i+1} - 1| \geq |z_i - 1| \quad (4.5)$$

²⁴Mappings like (4.2) are so-called Möbius transforms and are known to map circles (including straight lines) to circles (including straight lines).

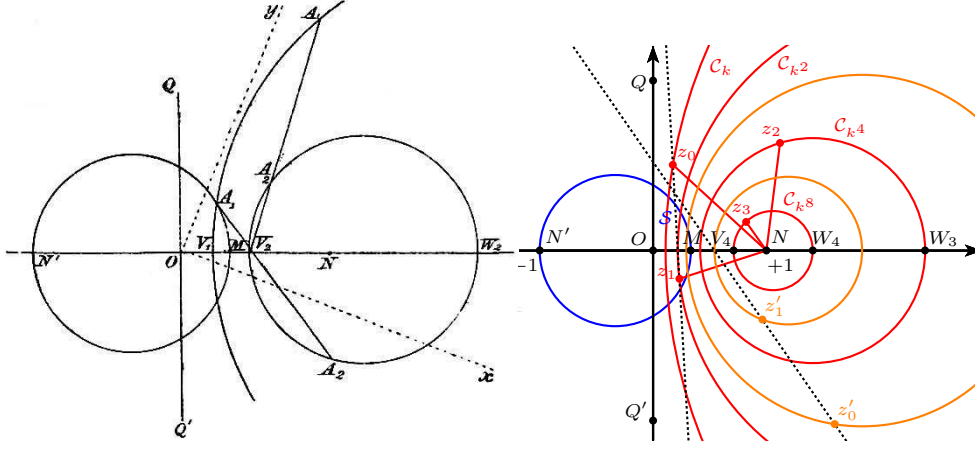


FIG. 4.4. Cayley's figure for the "Newton-Fourier imaginary problem" [25, fig. 48]. In the right picture, the point z_0 is not in S , but the point z_1 is. Hence the second step is not regular. Only the sequence z_2, z_3, z_4, \dots (in red) as well as z'_0, z'_1, \dots (in orange) are "regular from the beginning".

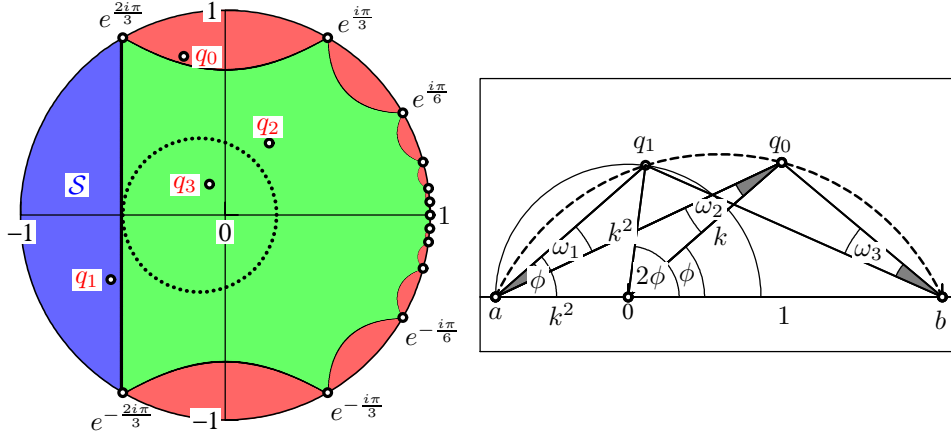


FIG. 4.5. Left: region for initial values of q_0 (in green) for which the corresponding sequence z_0, z_1, z_2, \dots "is regular from the beginning"; dotted: Cayley's "safe circle" in the q -plane; q_0, q_1, q_2, q_3 the points for z_0, z_1, z_2, z_3 of Fig. 4.4. Right: proof of property (4.4).

and *regular* otherwise. Their characterization becomes simpler in the q -plane because $z_i - 1 = \frac{1+q_i}{1-q_i} - 1 = \frac{2q_i}{1-q_i}$ and thus (4.5) becomes with (4.3)

$$\left| \frac{2q_i^2}{1-q_i^2} \right| \geq \left| \frac{2q_i}{1-q_i} \right| \Rightarrow \left| \frac{q_i}{1+q_i} \right| \geq 1 \Rightarrow \Re q_i \leq -\frac{1}{2}. \quad (4.6)$$

Cayley calls the intersection of $\Re q \leq -\frac{1}{2}$ with the unit circle, i.e. in the z -plane the intersection of the circle with diameter $[-1, \frac{1}{3}]$ and the positive half-plane, the "segment of unfitness" (the set S in blue in Fig. 4.3). Cayley claimed that for any z_0 inside "the safe circle" $|z - 1| < \frac{2}{3}$ the sequence z_0, z_1, z_2, \dots "is regular from the beginning". The corresponding circle in the q -plane is compared in Fig. 4.5 (left, dotted) to the green set of "regular-from-the-beginning" initial values and can be seen to be far from optimal.

It is nice to observe the dynamics of the map $q \mapsto q^2$ in this figure: Initial values

with k very close to 1 might be in one of the little red lunes²⁵. These are then mapped again and again to the next larger lune to the left, finally ending up in the large lune \mathcal{S} for an irregular step. Thereafter, the border of \mathcal{S} is mapped to the entire arc between $e^{-\frac{2i\pi}{3}}$ and $e^{\frac{2i\pi}{3}}$ and the value of q , in dependence of the actual size of k , might begin another lune walk towards another irregular step or spiral regularly inside the green set towards the solution.

In 1890, when Cayley again presented the previous results, but this time to the French public, he had hope of generalizing them to degree 3 [28, p. 218]:

“J’espère appliquer cette théorie au cas d’une équation cubique, mais les calculs sont beaucoup plus difficiles.” (“I hope to apply this theory to the case of a cubic equation, but the calculations are much more difficult.”)

Cayley’s dream had to wait for several decades, until more progress in the notions for “set of points” was available, and the French public was the right choice.

4.3. Basins of Attraction.

“The problem is to determine the regions of the plane, such that P being taken at pleasure anywhere within one region we arrive ultimately at the point A ; anywhere within another region at the point B ; and so for the several points representing the roots of the equation. (...) [The] (...) case of the cubic equation appears to present considerable difficulty.”

(Cayley, [26, p. 97])

“(...) it is anything but obvious what the division is, and the author had not succeeded in finding it.” (Cayley, speaking about himself, [27, p. 232])

“If Ahlfors, the creator of one of the main tools (...) needed pictures to come to terms with the subject, what can one say of lesser mortals? (...) It took me some time to discover that no one knew, and even longer to understand that the question really meant: what do the basins of the roots look like? (...) Adrien Douady and I poured over these pictures, and eventually got a glimpse of how to understand some of them, more particularly Newton’s method for $z^3 - 1$ and $z^3 - z$.”

(John Hubbard, [113, pp. xiii-xiv]).

In 1976, John Hubbard needed to teach undergraduate analysis to students at Orsay university and wanted to include some numerical content. So he decided to teach Newton’s method and the question of choosing the first estimate did not fail to arise (Hubbard cited by Douady, [138, p. 170]):

“Now, for equations of, say, degree three, the situation seems more complicated. I will think of it and tell you next week.”

Hubbard explained that “he assumed that although he didn’t know where to start, the experts surely did” and added that “it took some time to discover that no one knew anything about the global behavior of Newton’s method” [92, p. 233].

An early important paper for understanding the structure of these sets, i.e., explaining “l’insuccès de la tentative de Cayley” (the failure of Cayley’s attempt [95, p. 158]) was a memoir written for the *Grand prix des sciences mathématiques* in 1918 [95] by Gaston Julia (1893–1878). One of his tools for understanding the structures created by iteration processes was to study *repeated inverse images* (“antécédents successifs” [95, p. 160]) of the method.

We explain this method for the simplest of all cubic equations $z^3 - z = 0$ (see Fig. 4.6) with three real roots -1 (in blue), 0 (in green) and 1 (in red).

²⁵The authors had used ‘moon’ for the German ‘Mond’, but an anonymous reviewer said the correct term in English is the French ‘lune’.

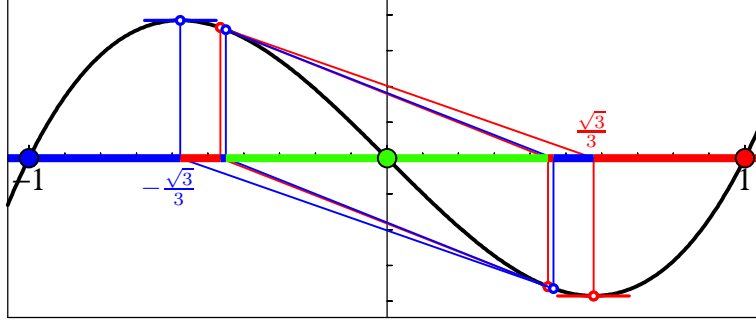


FIG. 4.6. Basins of attraction in the real for the three roots of $x^3 - x = 0$.

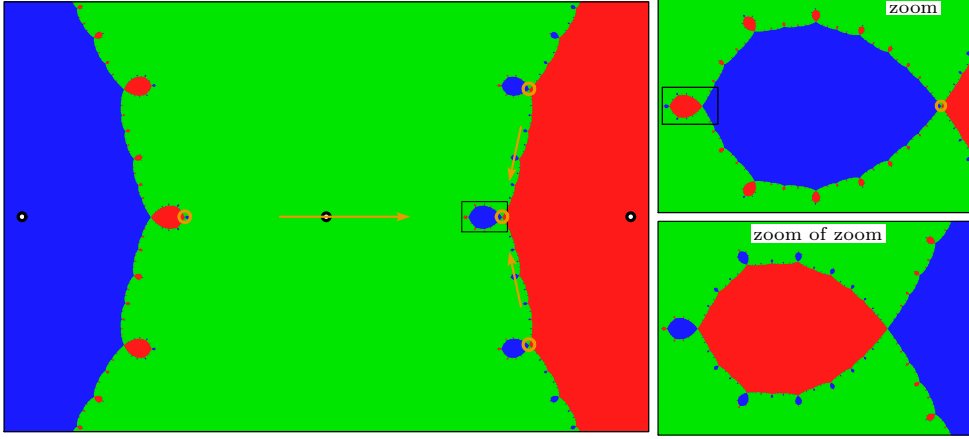


FIG. 4.7. Basins of attraction in the complex plane for the three roots of $z^3 - z = 0$ with zoom and double zoom. In orange: the Mourraille point $\sqrt{3}/3$ and its three preimage Newton points.

The real case. Already on the real line, Cayley’s “regions of the plane”, nowadays called *basins of attraction*, are complicated: the separation between the blue root and the red root happens at the points²⁶ $\pm \frac{\sqrt{3}}{3}$, where the tangent is horizontal. We compute from there repeatedly Newton’s method backwards

$$z_{n+1} = z_n - \frac{z_n^3 - z_n}{3z_n^2 - 1} = \frac{2z_n^3}{3z_n^2 - 1} \Rightarrow z_n^3 - \frac{3}{2}z_{n+1}z_n^2 + \frac{1}{2}z_{n+1} = 0. \quad (4.7)$$

Starting from the point $\frac{\sqrt{3}}{3}$ (in red), we compute backwards one, two, etc. preimages, which lead to an infinite red spiral, inside of which lies the blue spiral starting at $-\frac{\sqrt{3}}{3}$. These spirals create an infinity of alternating red and blue regions (only the first two are visible with this resolution). They converge to a 2-periodic orbit $\{\pm \frac{\sqrt{5}}{5}\}$, inside of which is the basin of attraction for the green root 0.

The complex case. Here, the three basins of attraction become particularly spectacular (see Fig. 4.7). The reason for this complexity is the fact that the inverse formula in (4.7) (right) gives in the complex plane *three* preimages z_n from each z_{n+1} . If, for example, we choose for z_1 the Mourraille point $\frac{\sqrt{3}}{3}$, we get for z_0 three possible

²⁶They correspond to Mourraille’s point Π in Fig. 4.2.

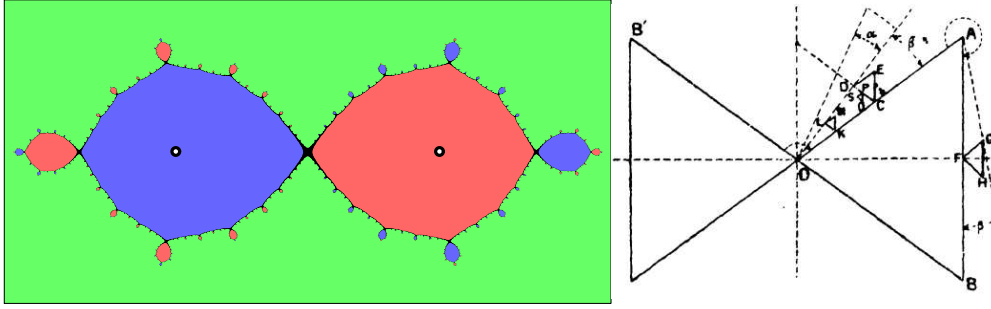


FIG. 4.8. The image of Fig. 4.7 by $z \mapsto \frac{1}{z}$ (left) and Julia's schema to represent the Julia set (“which is obviously just a schema”, p. 170).

preimages (drawn in orange in Fig. 4.7). These preimages have together 9 preimages to be mapped to z_1 after two iterations, then 27 for three iterations, and so on. It is understandable that Cayley found their analysis “anything but obvious”.

Precisely the same problem had been studied “à fond” (17 pages) by Julia [95, pp. 158-175]. He thereby brought the point ∞ to zero by setting $z \mapsto w = \frac{1}{z}$, such that (4.7) becomes

$$w_{n+1} = \frac{3w_n - w_n^3}{2} \quad (4.8)$$

(see Fig. 4.8, left). Julia observed that “the total domain of convergence towards a root consists of infinitely many separate areas” [95, p. 158], and that the points for which there is no convergence form “a continuous closed curve having double points everywhere dense on itself” [95, p. 52], today called a “Julia set”:

“It is quite difficult to imagine exactly what this continuous E' can be. But it is not impossible to get an idea of it using a constructive process which I will now expose and which will show that there is no impossibility, for example, that an area R_∞ , simply connected, is limited by a linear continuum E' having multiple points everywhere dense on it, a continuum which divides the plane into an infinity of regions, each of which touches R_∞ through its entire boundary; the frontier of each small region being moreover a simple curve (...) I found the idea in a beautiful paper by M. Helge von Koch, *Acta Mathematica* de 1906 [171]” [95, pp. 169–170].

His “constructive process” consists of starting from two equilateral triangles and constantly adding smaller triangles on the sides (see Fig. 4.8, right). This allows him to explain that “the structure of the whole E' is the same as that of any part \mathcal{E}' of E' ” [95, p. 99]: “On voit bien de quelle étrange espèce est ce continu (...)” [95, p. 168].

The equation $z^3 - 1 = 0$. This is the other “simplest of all cubic equations”, and it was studied in 1983 by J.-P. Eckmann [45] and the same year by J.H. Curry, L. Garnett and D. Sullivan [38] as particular case of a more general equation. Here, Newton's algorithm (see Sect. 3.3) reads as

$$z_{n+1} = \frac{1}{3} \left(2z_n + \frac{1}{z_n^2} \right) \quad \text{or equivalently} \quad z_n^3 - \frac{3}{2} z_{n+1} z_n^2 + \frac{1}{2} = 0. \quad (4.9)$$

The only singular point in Newton's iteration is the point $z = 0$. Fig. 4.9 (left) shows the original picture of Eckmann. As before, the complexity of the basins is best

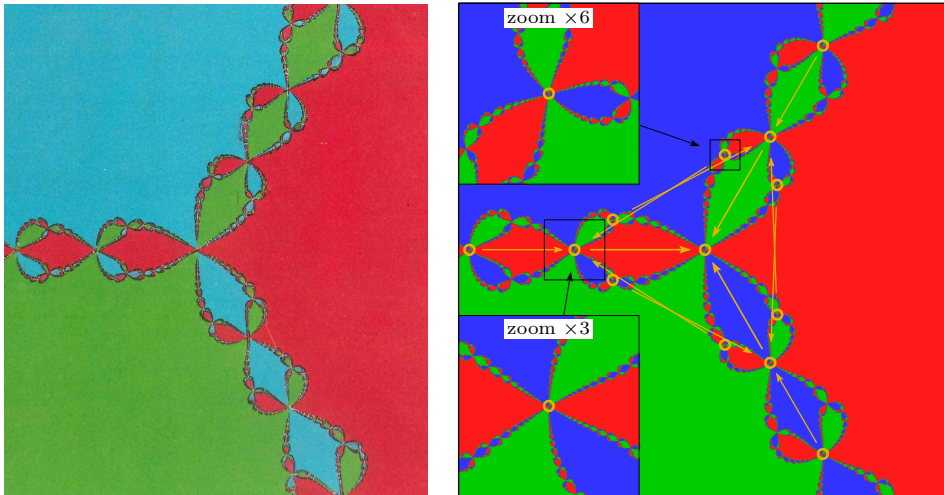


FIG. 4.9. The equation $z^3 - 1 = 0$: original picture of Eckmann [45] (left) together with the first and second preimages of $z = 0$, two of them with a zoom of their neighbourhood (right).

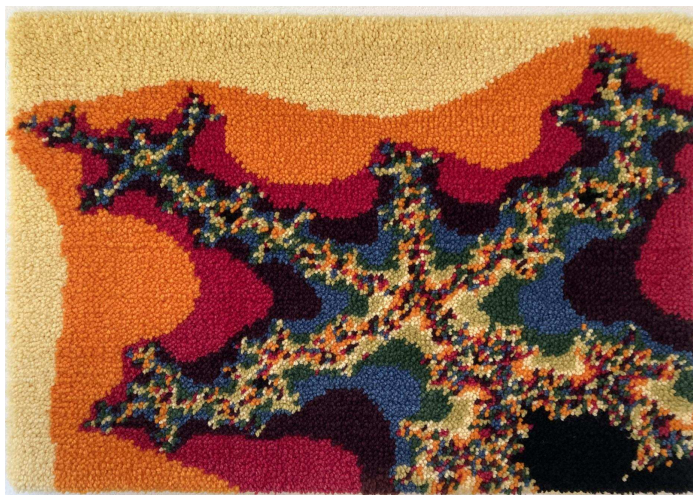


FIG. 4.10. Wall Carpet for the Mandelbrot set computed by Martin Hairer on his Macintosh II.

understood by looking at the 3 preimages z_n for $z_{n+1} = 0$ by solving (4.9) (right), then the next following 9 points which attain 0 after two iterations (drawn in orange in Fig. 4.9, right) and so on. The Newton steps are holomorphic maps, and thus angle preserving, between the neighbourhoods of these points of which we draw two of them as a zoom.

The Mandelbrot set. On March 1, 1980 Benoit Mandelbrot discovered that already the simplest “Cayley map” (4.3) $q \mapsto q^2$ becomes a highly interesting (and since become very famous) object, if one adds a complex parameter c as $q \mapsto q^2 + c$ and discusses convergence (or the lack of it), starting with $q = 0$, in dependence of c . We present in Fig. 4.10 the picture of a wall carpet for which young Martin Hairer, when he was still in school, had computed on his Macintosh II computer, pixel by pixel, the

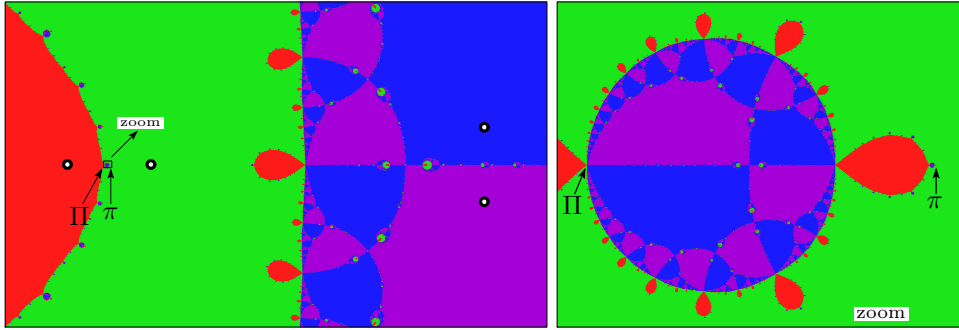


FIG. 4.11. Basins of attraction of Mourraille's equation "Fig. 54" with zoom on the interval $[\Pi, \pi]$.

corresponding colours.

Mourraille's equation. By looking at the basins of attraction of Mourraille's equation of Fig. 4.2 in the complex plane (in four colours), the exceptional situation of the interval $[\Pi, \pi]$ becomes clearly visible (see Fig. 4.11). This is again an occurrence of Julia's maxim that Mandelbrot repeatedly heard during his lectures at École Polytechnique: "To simplify, you should complexify" [118, p. 272].

5. Error Estimates for Newton's Method.

"On voit qu'il est nécessaire d'assigner un caractère certain, d'après lequel on puisse toujours distinguer si les limites sont assez voisines pour que l'application de la règle donne nécessairement des résultats convergens."

(J. Fourier, [58, p. 62])

While Newton, Raphson and Simpson "knew" how fast their method converges, the first rigorous proofs were given by Fourier and Cauchy. These proofs are based on Lagrange's formula for the error of a Taylor series (see [109]; see also [61, p. 22], [23, Huitième Leçon])

$$\begin{aligned}
 (1) \quad & f(x) = f(x_0) + (x - x_0)f'(\xi) && (\text{for } f \text{ in } C^1) \\
 (2) \quad & f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(\xi) && (\text{for } f \text{ in } C^2) \\
 & \text{etc.} &&
 \end{aligned} \tag{5.1}$$

where ξ is an unknown intermediate value between x_0 and x .

5.1. Fourier's Estimate. Joseph Fourier (1768-1830) published his results in a paper submitted to the Société Philomatique in 1818 [58] and later, in a much more elaborated version, in his last work [61], published posthumously in 1831. This book is divided into two *Livres*, the first is discussing at length the position and separation of real roots of a polynomial based on sign changes of all its derivatives²⁷. *Livre II* (pages 157–248) is dedicated to the *Calcul des Racines*, giving neat rigorous estimates, once a root has been isolated by the methods of *Livre I*. We thus suppose to have a polynomial $f(x)$ and an interval $[a, b]$ knowing that

- $f(x)$ has exactly one root inside $[a, b]$;

²⁷In 1820, Fourier published the paper [59] about this problem which was analysed by Darboux (Œuvres de Fourier II, pp. 310–314). He emphasizes that Fourier brings "absolutely new views" on this question (p. v). It is interesting to note that this part was at the origin of Charles Sturm's discovery of *Sturm sequences* [164].

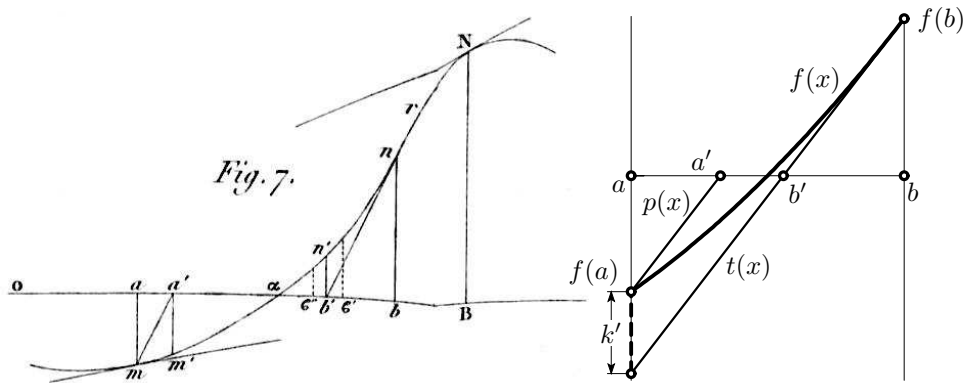


FIG. 5.1. Fourier's Error Estimation (left: from Fourier [61, p. 258]).

- $f'(x)$ and $f''(x)$ have constant signs on $[a, b]$.

To search “la mesure de la convergence” [61, p. 169], Fourier treats in detail the case where $f'(x) > 0$ and $f''(x) > 0$, and computes (see Fig. 5.1)

- b' with a Newton step using $t(x)$, the tangent in b with slope $f'(b)$;
- a' using the parallel $p(x)$ through $(a, f(a))$, also with slope $f'(b)$.

Because of $f''(x) > 0$, we have $f'(x) < f'(b) = t'(x) = p'(x)$ inside $[a, b]$. Hence using Formula (1) in (5.1) twice, we have $t(x) < f(x) < p(x)$ inside $[a, b]$. As a consequence, the interval $[a', b']$ is a new pair of rigorous bounds for the root. Evaluating $f(a) - t(a)$ from Formula (2) in (5.1), we obtain the difference k'

$$k' = \frac{(b-a)^2}{2} f''(\xi), \quad \text{hence} \quad (b' - a') = (b-a)^2 \cdot \frac{f''(\xi)}{2f'(b)}, \quad (5.2)$$

because $k' = f'(b) \cdot (b' - a')$ [61, p. 201]. This is the famous (“très remarqué” [61, p. 170]) *quadratic convergence*, and repeating the algorithm with $[a', b']$ etc., leads to a sequence of rapidly shrinking *nested intervals*. As an application, Fourier computes in [61, pp. 209–217] the root of Newton’s example $x^3 - 2x - 5 = 0$ as $x = 2.09455148154232659148238654057930$, where all digits are declared to be correct (they are).

5.2. Cauchy’s Estimate. Following Fourier’s article [58], Augustin-Louis Cauchy (1789–1857) appended the “Note” [23] to his *Calcul Différentiel* published in 1829. His emphasis was not only algebraic equations, but also transcendental ones. As a consequence, there is no *Livre I* giving nice intervals to start with. So Cauchy begins with a starting point x_0 and a first Newton correction x_1 with

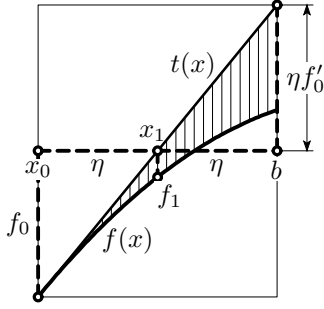
$$x_1 = x_0 + \eta, \quad \text{where} \quad \eta = -\frac{f(x_0)}{f'(x_0)}$$

(we suppose here for simplicity $f(x_0) < 0$, $f'(x_0) > 0$ and $\eta > 0$; see Fig. 5.2). We then choose the interval $I = [x_0, x_0 + 2\eta] = [x_0, b]$ of length 2η , for which x_1 is the mid point, and suppose that $f'(x) \neq 0$ on I . The stripes in Fig. 5.2 represent by Formula (2) of (5.1) the error term $f(x) - t(x) = \frac{(x-x_0)^2}{2!} f''(\xi)$, so we obtain for $x = x_0 + 2\eta$

$$|f(b) - t(b)| \leq 2\eta^2 K \quad \text{if we define} \quad K = \max\{|f''(x)|; x \in I\}. \quad (5.3)$$

Using $t(b) = \eta f'(x_0)$ one direction of this inequality implies

$$\eta(f'(x_0) - 2\eta K) \leq f(b).$$



$2\rho, 2\rho\varepsilon, 2\rho\varepsilon^3, 2\rho\varepsilon^7, \text{ etc....}$

FIG. 5.2. Cauchy's Error Estimation; right: maximal intervals for x_0, x_1, x_2, x_3 , etc. ([23, p. 263]; Cauchy uses ρ instead of η).

Therefore, under the condition

$$f'(x_0) - 2\eta K \geq 0, \quad \text{or equivalently} \quad \frac{2\eta K}{f'(x_0)} \leq 1, \quad (5.4)$$

we have $f(b) \geq 0$, hence $f(x_0)$ and $f(b)$ have opposite signs. This leads to Cauchy's 2.^e Théorème: Under condition (5.4) the equation $f(x) = 0$ has exactly one root in $[x_0, x_0 + 2\eta]$.

Cauchy continues on several pages with another theorem and a corollary for proving the other estimates. We here shortcut this part by using another of Cauchy's inventions in another of his works [22, Chap. VI, §1, p. 126] ("la convergence de la série est assurée"), that of "Cauchy sequences". Besides (5.3) we need also a bound for $f'(x)$ on the entire interval²⁸:

$$B = \max \left\{ \frac{1}{|f'(x)|}; x \in I \right\} \quad \text{and set the abbreviation} \quad \varepsilon = \frac{\eta KB}{2}. \quad (5.5)$$

Then we have alternately, starting with $|x_1 - x_0| = \eta$,

$$\begin{aligned} & \text{from (5.1) (2) and (5.3)} & \text{from (5.5)} \\ |f(x_1)| & \leq \frac{\eta^2}{2} K & |x_2 - x_1| = \frac{|f(x_1)|}{|f'(x_1)|} \leq \frac{\eta^2 KB}{2} = \eta \varepsilon \\ |f(x_2)| & \leq \frac{\eta^2 \varepsilon^2}{2} K & |x_3 - x_2| = \frac{|f(x_2)|}{|f'(x_2)|} \leq \frac{\eta^2 \varepsilon^2 KB}{2} = \eta \varepsilon^3 \\ |f(x_3)| & \leq \frac{\eta^2 \varepsilon^6}{2} K & |x_4 - x_3| = \frac{|f(x_3)|}{|f'(x_3)|} \leq \frac{\eta^2 \varepsilon^6 KB}{2} = \eta \varepsilon^7 \end{aligned} \quad (5.6)$$

etc, see Cauchy's rigorous error intervals displayed in Fig. 5.2, right. These estimates remain valid, as long as the points x_2, x_3, \dots remain in I . To assure this, we estimate

$$|x_1 - x_2| + |x_2 - x_3| + |x_3 - x_4| + \dots \leq \eta \cdot (\varepsilon + \varepsilon^3 + \varepsilon^7 + \dots) < \frac{\eta \varepsilon}{1 - \varepsilon^2}, \quad (5.7)$$

and have that the condition

$$\frac{\varepsilon}{1 - \varepsilon^2} \leq 1 \quad \text{or} \quad \varepsilon \leq \frac{\sqrt{5} - 1}{2} \quad \text{or} \quad \eta KB \leq \sqrt{5} - 1 \quad (5.8)$$

²⁸Cauchy used A for our $\frac{1}{B}$, B for our K ; the notations used here are those of Kantorovich (see Section 5.3).

implies $|x_1 - x_k| < \eta$ for all k . We then have a Cauchy sequence, whose x_k remain in I and converge to an x^* with $f(x^*) = 0$ and with maximal error

$$|x_k - x^*| \leq |x_k - x_{k+1}| + |x_{k+1} - x_{k+2}| + \dots \leq \eta \cdot (\varepsilon^{2^k - 1} + \varepsilon^{2^{k+1} - 1} + \dots) < \frac{\eta \varepsilon^{2^k - 1}}{1 - \varepsilon^{2^k}}. \quad (5.9)$$

5.3. Systems of Equations; Banach and Kantorovich. The above study of Cauchy is not valid for systems of equations, since the results are based on the study of sign changes for $f(x)$. During the nineteenth century, more and more applications of iteration processes to more general problems arose such as ordinary differential equations (Picard, Lindelöf), which are described in Section I.8 of [84]. In particular the study of integral equations (Volterra, Fredholm, Hilbert) and variational calculus (Hadamard) led to the emergence of functional analysis and the notion of abstract spaces, due largely to Peano, Volterra, Fréchet and above all Stefan Banach (1892-1945) [5], [6]. Treating n -tuples of numbers, infinite sequences, or functions as “points” in such a space, renders proofs about these subjects nearly as simple as in one dimension. These “Banach” spaces are equipped with a vector space structure, a *norm* $|x|$ with corresponding *operator norm* $|A|$ and must be *complete*, i.e., any Cauchy sequence converges to an x in the space. While the proof for completeness is relatively simple for spaces of continuous functions with the max-norm, it is more difficult for integral norms, such that [6] starts right away with “We assume that the reader knows measure theory and the Lebesgue integral”.

Banach’s Fixed Point Theorems. After having defined his space X , Banach states²⁹ that *any map $f : X \rightarrow X$ which is contractive, i.e., $|f(x) - f(y)| \leq q|x - y|$, with a fixed $q < 1$, possesses a unique fixed point x with $x = f(x)$ and the iterations $x_{n+1} = f(x_n)$ converge to $x \in X$ for any starting value $x_0 \in X$* . This extends our observations from Section 2 (where $|f'(x)| \leq q < 1$ had been sufficient for convergence) to any dimension. We illustrate in Fig. 5.3 (left) this result for a two-dimensional map arising from Fourier’s iteration (2.16) $\varepsilon_{n+1} = \arctan \frac{\varepsilon_n}{\lambda}$, when it is extended to the complex plane. We show the four iterations from blue to red, which Fourier had drawn in Fig. 2.5, now with various initial values ε_0 running through a mathematical animal³⁰.

In the next theorem³¹ Banach treats the important particular case where f is a contractive linear map followed by a shift, i.e., the equation $x = Hx + y$ with $|H| = q < 1$. We illustrate this in Fig. 5.3 (right) by our cat which is mapped by $Hx + y$ to its own picture stuck to its nose, and the iteration $x_{n+1} = Hx_n + y$ reproduces it infinitely often, converging to a point which is $x = (I - H)^{-1}y$. We observe in the picture that $|x| \leq |y| + q|y| + q^2|y| + \dots = \frac{|y|}{1-q}$. Thus

$$\text{If } |H| = q < 1 \text{ then } I - H \text{ is invertible and } |(I - H)^{-1}| \leq \frac{1}{1-q}. \quad (5.10)$$

Kantorovich’s Theorem for Newton’s Method. The classical paper is due to Leonid Kantorovich (1912-1986) [96] in 1948. Assume we need to solve in the space X an equation $f(x) = 0$. We start from an x_0 and apply a first Newton step (see

²⁹[5] Théorème 6, part II, §2, p. 160, *Travaux de Stefan Banach* p. 330

³⁰which frequently bears the name “Arnold’s cat”

³¹[5] Théorème 7, part II, §2, p. 161, *Travaux de Stefan Banach* p. 331; [6], Chap. X, Théorème 17, p. 158; Kantorovich [96], Glava I, §3, Teorema 3, p. 102

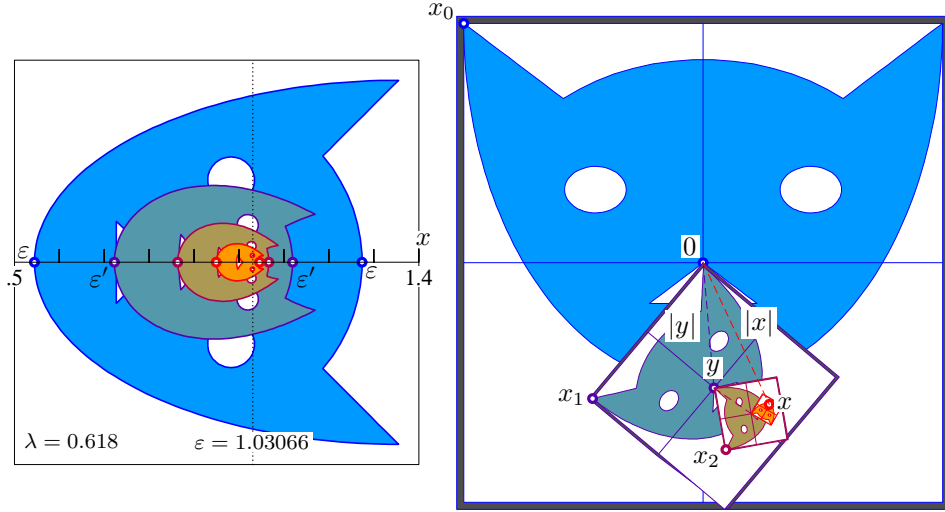


FIG. 5.3. Banachs Fixed Point Theorem; nonlinear (left), linear plus shift (right)

(3.13))

$$x_1 = x_0 - f'(x_0)^{-1}f(x_0) \quad (= \text{root of } t(x) = f(x_0) + f'(x_0)(x - x_0)). \quad (5.11)$$

We suppose x_0 to lie inside a ball I and need an extension of definition (5.3). An easy access to this is to suppose a Lipschitz condition for the matrix $f'(x)$ (see Fenyő [54]),

$$|f'(x) - f'(y)| \leq K|x - y| \quad \text{for } x, y \in I. \quad (5.12)$$

We then have with $x_1 - x_0 = \Delta x$, because of $f(x_0) = t(x_0)$,

$$\begin{aligned} f(x_1) - t(x_1) &= \int_0^1 \frac{d}{d\theta} (f(x_0 + \theta\Delta x) - t(x_0 + \theta\Delta x)) d\theta \\ &= \int_0^1 (f'(x_0 + \theta\Delta x) - f'(x_0)) \Delta x d\theta. \end{aligned}$$

Taking norms, moving them inside the integral, using the Lipschitz condition (5.12) and $\int_0^1 \theta d\theta = \frac{1}{2}$, this gives

$$|f(x_1) - t(x_1)| \leq \frac{1}{2} K |\Delta x|^2, \quad (5.13)$$

exactly as in (5.6). Kantorovich proves this with the help of *formuly Teĭlora*.

The next tool we need is an analogue of (5.5). Computing the maximum of $|f'(x)^{-1}|$ over an entire domain I is not practical, since it requires the inversion of infinitely many matrices. Therefore Kantorovich defines only

$$B_0 = |f'(x_0)^{-1}| \quad (5.14)$$

and expands for other points x (see [96], p. 171)

$$\begin{aligned} f'(x)^{-1} &= [f'(x_0) - (f'(x_0) - f'(x))]^{-1} = \left(f'(x_0) [I - f'(x_0)^{-1}(f'(x_0) - f'(x))] \right)^{-1} \\ &= [I - f'(x_0)^{-1}(f'(x_0) - f'(x))]^{-1} f'(x_0)^{-1} \end{aligned}$$

using, as he says, *ravenstvom* $(AB)^{-1} = B^{-1}A^{-1}$. Here it is required to invert a matrix as in (5.10) with $H = f'(x_0)^{-1}(f'(x_0) - f'(x))$. By using (5.14) and (5.12) and reading this expansion backwards, we see that

$$\text{If } B_0 K |x - x_0| < 1 \text{ then } f'(x) \text{ is invertible and } |f'(x)^{-1}| \leq \frac{B_0}{1 - B_0 K |x - x_0|}. \quad (5.15)$$

Now, starting from $|x_1 - x_0| \leq \eta_0$, we obtain by combining (5.13) and (5.15) that

$$|x_2 - x_1| = |f'(x_1)^{-1} f'(x_1)| \leq \frac{B_0}{1 - B_0 K \eta_0} \cdot \frac{K \eta_0^2}{2} = \frac{1}{2} \frac{h_0 \eta_0}{1 - h_0} = \eta_1,$$

where we have introduced $h_0 = B_0 \eta_0 K$.

For the next round, we prepare

$$B_1 = \frac{B_0}{1 - h_0} \quad \text{and} \quad h_1 = B_1 \eta_1 K = \frac{B_0}{1 - h_0} \frac{1}{2} \frac{h_0 \eta_0}{1 - h_0} K = \frac{1}{2} \frac{h_0^2}{(1 - h_0)^2},$$

and we see that we have in general $|x_{n+1} - x_n| \leq \eta_n$ if we define recursively

$$B_{n+1} = \frac{B_n}{1 - h_n}, \quad \eta_{n+1} = \frac{1}{2} \frac{h_n \eta_n}{1 - h_n}, \quad h_{n+1} = \frac{1}{2} \frac{h_n^2}{(1 - h_n)^2}. \quad (5.16)$$

The recursion for h_n is a fixed point iteration to which we apply what we have learned from Fourier in Section 2.6 and Euler in Section 2.7 (see Fig. 5.4). There is a (unstable) fixed point $h_0 = \frac{1}{2}$ for which we get $\eta_n = 2^{-n} \eta_0$. If $h_0 < \frac{1}{2}$, the h_n exhibit a rapid quadratic convergence to 0 together with the η_n . For $h_0 > \frac{1}{2}$ we have divergence with chaotic dynamics around the point 2 (see Fig. 5.4, below), so we must require

$$h_0 \leq \frac{1}{2} \quad \text{or} \quad B_0 \eta_0 K \leq \frac{1}{2}, \quad (5.17)$$

the famous *Kantorovich condition*. We then have to estimate expressions like

$$|x_0 - x_1| + |x_1 - x_2| + |x_2 - x_3| + \dots \leq \eta_0 + \eta_1 + \eta_2 + \dots \quad (5.18)$$

For this, Kantorovich states a nice identity and proves it in two lines, but its discovery seems not trivial. We may use $1 - 2h_{n+1} = 1 - \left(\frac{h_n}{1 - h_n}\right)^2 = \left(1 + \frac{h_n}{1 - h_n}\right)\left(1 - \frac{h_n}{1 - h_n}\right)$ which gives

$$\sqrt{1 - 2h_{n+1}} = \frac{\sqrt{1 - 2h_n}}{1 - h_n} \quad \text{hence} \quad 1 - \sqrt{1 - 2h_{n+1}} = \frac{1 - h_n - \sqrt{1 - 2h_n}}{1 - h_n}.$$

We finally multiply this on the left and right by $\frac{\eta_{n+1}}{h_{n+1}} = (1 - h_n) \frac{\eta_n}{h_n}$ and obtain

$$\eta_n = \eta_n \cdot \frac{1 - \sqrt{1 - 2h_n}}{h_n} - \eta_{n+1} \cdot \frac{1 - \sqrt{1 - 2h_{n+1}}}{h_{n+1}}. \quad (5.19)$$

From here it is easy to formulate all the theorems and error results, similar to (5.7) and (5.9), because an estimation like (5.18) with (5.19) becomes just a telescoping series.

Later, in his book with Akilov [97], an entire chapter was devoted to the proof of convergence of Newton's method, see also Ortega [133] and Deulhard [42] for later improvements. The article [37] by Ciarlet and Mardare gives a nice overview on these developments together with a new version of the theorem "with only one constant" and a substantially simpler proof.

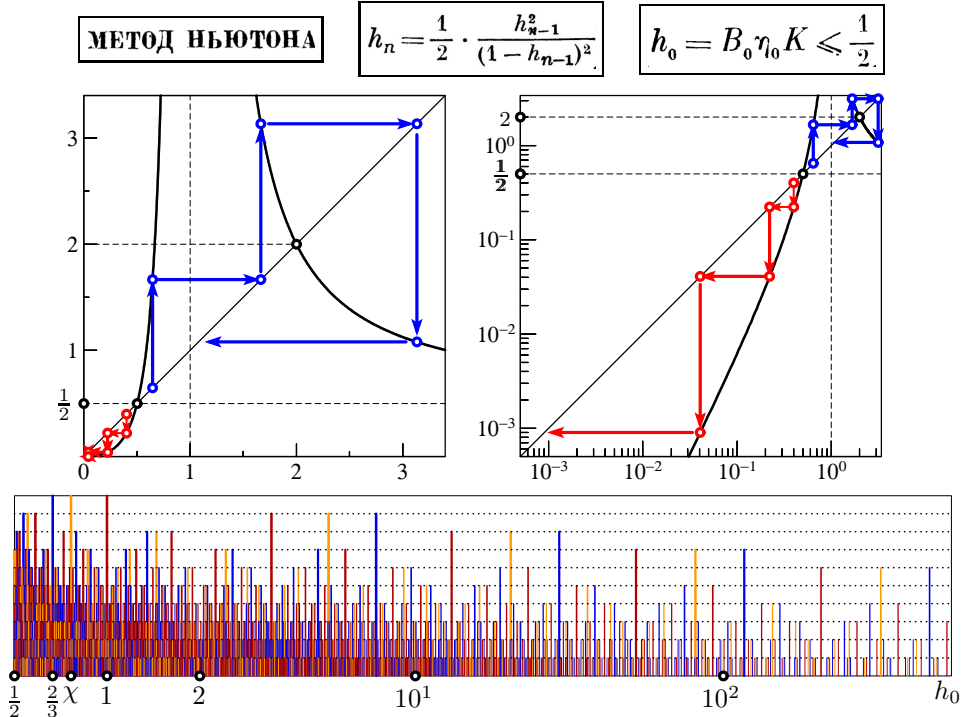


FIG. 5.4. Top: Beginning of “Glava IV” (METOD NIUTONA) in [96, p. 163] as well as the recursion for h_n and the famous condition for h_0 [96, p. 170,172]. Middle: The iterations for h_n (right: in double logarithmic scale). Below: initial values for h_0 which, after 1, 2, 3, ..., 9 iterations, end up at $h_n = 1$ (and then infinity, in red), at $h_n = \frac{2}{3}$ (and then to the fixed point $h_n = 2$, in blue); at $h_n = \chi = 3 - \sqrt{5}$ (giving rise to the periodic two-cycle $3 \pm \sqrt{5}$, in yellow). There exist periodic m -cycles for any $m = 3, 4, 5, \dots$, for example $h_0 = 0.8520440955209$ ($m = 3$), $h_0 = 0.9155259248665$ ($m = 4$), $h_0 = 0.9545792886408$ ($m = 5$).

6. Stationary Iterative Methods for Large Linear Problems.

“Solving linear algebraic equations can be interesting.”

(George E. Forsythe, Bull. Amer. Soc. 59(4), 299-329 (1953); beginning with: “The subject of this talk is mathematically a lowly one.”).

During four millennia, iterative methods were used for nonlinear problems. Linear equations, apparently a “lowly” subject, were simply solved, already centuries before Gauss, by eliminating one variable after the other, or other suitable tricks³². Only when, at first human, later electronic, computers started to treat larger and larger problems, the importance of iterative methods was discovered (Gauss 1823: “You will hardly ever again eliminate directly, at least not when you have more than 2 unknowns”, translation by Forsythe [57] of [78]). This becomes particularly important, when linear systems arise from the approximation of partial differential equations, and such systems are sparse. While one could theoretically still use Gaussian elimination to solve such systems, the substitution process makes the sparse system more and more densely populated. In this case, the cost of Gaussian elimination can rapidly become prohibitive³³.

³²See for example the Section 6 in [168].

³³For a comprehensive journey through the history of numerical linear algebra, see the recent monograph [16].

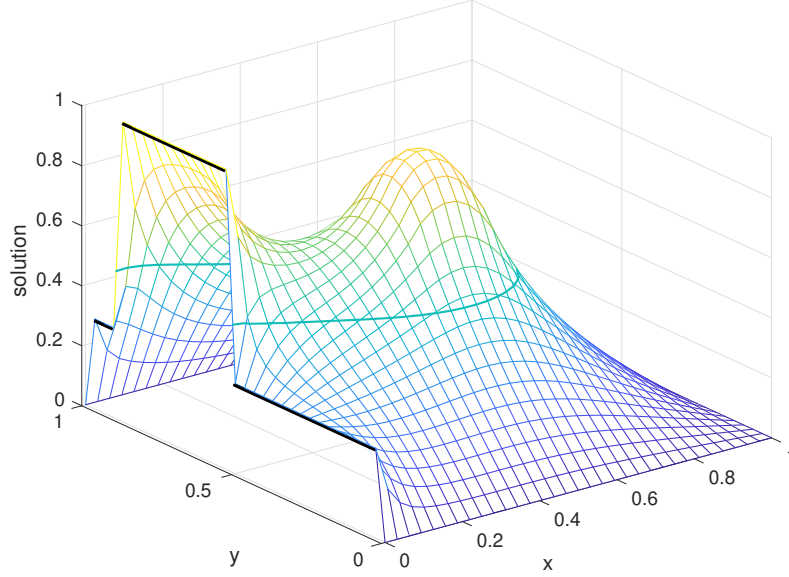


FIG. 6.1. Temperature distribution in a room in Montreal, with a symbolic door on the left and a wall at a fixed temperature, the other walls having the fixed temperature zero, and a stove in the middle that also heats the room a bit. The level set of temperature $\frac{1}{2}$ is also indicated for later comparisons with iterative approximations.

We start to follow in this section the detailed historical development of these methods, using as our model problem to illustrate them the simple example of a temperature distribution in a room in two spatial dimensions we call the “Montreal problem”³⁴, modeled by the Poisson equation

$$-\Delta u = f \quad \text{in } \Omega := (0, 1)^2, \quad u = g \quad \text{on } \partial\Omega, \quad (6.1)$$

which we discretize by a standard centered finite difference method, see e.g. [69, Chapter 2], to obtain the corresponding linear system of equations

$$Au = f. \quad (6.2)$$

The solution we want to compute then looks like shown in Figure 6.1, where we used the mesh size $h = \frac{1}{32}$. We thus have $m = 31$ interior mesh points in each direction, and the matrix A is of size $m^2 \times m^2 = 961 \times 961$. We chose a zero temperature on three sides of the room, and on the left side a door for $y \in (0.5, 0.9)$ with temperature equal to 1, and along the wall on the left the temperature 0.3. In the middle of the room we put a stove, which is modeled by a non-zero source function $f = 50$ in the zone $(0.4, 0.6) \times (0.4, 0.6)$, and leads to the bump in the temperature in the middle of the room. We also show the level curve of temperature $\frac{1}{2}$ in Fig. 6.1.

The first iterative methods for linear systems were stationary iterative methods, i.e. methods which perform the same steps at every iteration, going back to Gauss and Jacobi.

³⁴Which other example would you expect from one of the authors, who has lived five Canadian winters in Montreal?

6.1. A Letter of Gauss. Carl Friedrich Gauss (1777–1855) loved to travel, from 1818 on, for months through the Kingdom of Hannover, surveying the geodesic survey of this country. He computed during long evenings myriads of linear systems arising from the least squares approximation for the positions of triangulation points which, as he wrote [78], was “against the monotony of the surveying business, (...) always a pleasant entertainment” and “can be done half asleep”. But he never published any details.

6.2. The Method of Jacobi. Unaware of the Gauss procedure, Carl Gustav Jacobi (1804-1851) presented in [93] “eine neue” (a new) method for solving the linear system (6.2) $A\mathbf{u} = \mathbf{f}$, in particular systems arising from the method of least squares. He acknowledges the computations that were performed by his friend Dr. Seidel³⁵. Motivated by the fact that the problems Jacobi encountered were *diagonally dominant*, he moved the off-diagonal terms to the right to obtain

$$u_i = \sum_{j \neq i} b_{ij} u_j + g_i \quad \text{or} \quad \mathbf{u} = B\mathbf{u} + \mathbf{g} \quad \text{where} \quad b_{ij} := -\frac{a_{ij}}{a_{ii}}, \quad g_i := \frac{f_i}{a_{ii}}, \quad (6.3)$$

and solved (6.3) by fixed-point iteration,

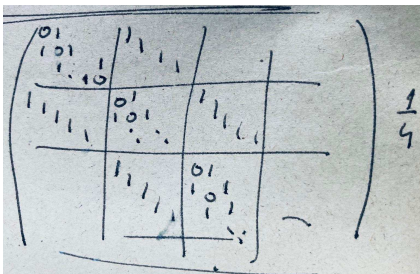
$$u_i^{n+1} = \sum_{j \neq i} b_{ij} u_j^n + g_i \quad \text{or} \quad \mathbf{u}^{n+1} = B\mathbf{u}^n + \mathbf{g} \quad \text{with, say, } \mathbf{u}^0 = \mathbf{0}. \quad (6.4)$$

Starting with a zero initial guess, this leads to the approximate solution

$$\mathbf{u}^n = (B^{n-1} + B^{n-2} + \dots + B + I)\mathbf{g}. \quad (6.5)$$

Example. For the system (6.2), corresponding to our model problem (6.1), Gerhard Wanner wrote the matrix for his students by hand some 30 years ago,

$B =$



$\frac{1}{4}.$

(6.6)

Remark. Jacobi denoted a_{ij} by (ij) , $(i, j = 0, 1, 2, \dots)$ and the corrections $B^i \mathbf{g}$ by Δ^i ; which led to the formulas for their recursive calculation shown in Fig. 6.2.

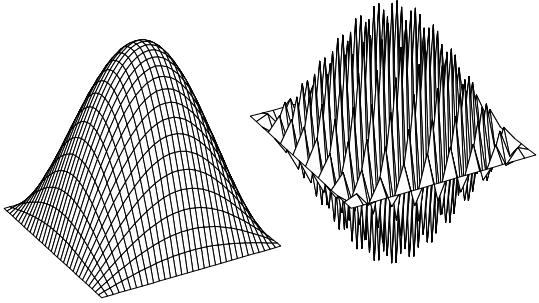
Convergence. For studying the convergence of Jacobi’s method, we have to compute the eigenvalues of the matrix B in (6.6). Such eigenvalues, in one dimension and without calling them so, have been obtained by Lagrange [40]. Extending his result

³⁵ “Man wird dort aus den von einem meiner gelehrten Freunde, Herrn Dr. Seidl in München, mit grosser Sorgfalt geführten Rechnungen ersehen, dass ...” [93, p. 304], and Jacobi used indeed the spelling Seidl for Seidel.

$$\begin{aligned}
(00) \Delta^{i+1} &= -\{(01) \Delta_1^i + (02) \Delta_2^i \text{ etc.}\}, \\
(11) \Delta_1^{i+1} &= -\{(10) \Delta^i + (12) \Delta_2^i \text{ etc.}\}, \\
(22) \Delta_2^{i+1} &= -\{(20) \Delta^i + (21) \Delta_1^i + (23) \Delta_3^i \text{ etc.}\}, \\
&\text{etc.} \qquad \qquad \qquad \text{etc.}
\end{aligned}$$

FIG. 6.2. Jacobi's recursive formulas for the corrections [93, p. 297].

to two dimensions, we obtain the first and last eigenvector as follows:

$$\begin{aligned}
u_{ij} &= \sin\left(\frac{i\pi}{m}\right) \cdot \sin\left(\frac{j\pi}{m}\right) \\
\lambda_{\max} &= \cos\left(\frac{\pi}{m}\right),
\end{aligned}$$

(6.7)

For large m , $\lambda_{\max} \approx 1 - \frac{\pi^2}{2m^2}$ is close to 1 and requires $\mathcal{O}(m^2)$ iterations in order to attain some accuracy.

Numerical Example. If we apply this method to the Montreal problem in Fig. 6.1, starting with a zero initial guess, we obtain the sequence of approximations in Figure 6.3, shown for iteration numbers n that are powers of two. Comparing with the level set at $\frac{1}{2}$ in the solution in Figure 6.1, we see that indeed an accurate solution is only reached around iteration 1024, which is for the $m = 31$ we used here indeed $\mathcal{O}(m^2)$ like predicted by our eigenvalue analysis.

Jacobi's "preconditioner". Realizing that the method can be slow or even fail if the system is not diagonally dominant enough, Jacobi then presents the groundbreaking idea of preconditioning using Jacobi rotations³⁶, see Fig. 6.4:

“As an example we use the method for the equations from Theoria motus p. 219. The original equations are (see Fig. 6.4). If we remove the coefficient 6 in front of q in the first equation, the angle of rotation is $\alpha = 22^0 30'$, and the new equations are...” (translated from [93, p. 304]).

After preconditioning, it takes then only three Jacobi iterations to obtain three accurate digits, which one would nowadays call textbook multi grid convergence! A year after his seminal publication [93], Jacobi showed that his rotations can also be used to compute eigenvalues and eigenvectors [94].

6.3. The Method of Seidel. Philipp Ludwig von Seidel (1821–1896) was a student of Jacobi in Königsberg and had the “honour” (“ich habe noch als Studierenden die Ehre gehabt, für ihn dazu die numerischen Rechnungen auszuführen.” [154, p. 86] of performing painful numerical calculations for Jacobi. Later, as professor in Munich, he made various applications of linear systems (of dimension up to 72) to scientific research, in particular the determination of the luminosity of fixed stars. He thereby realised that it was preferable, once the first component u_1^{n+1} was calculated,

³⁶We will get back to preconditioning in Section 8.4.

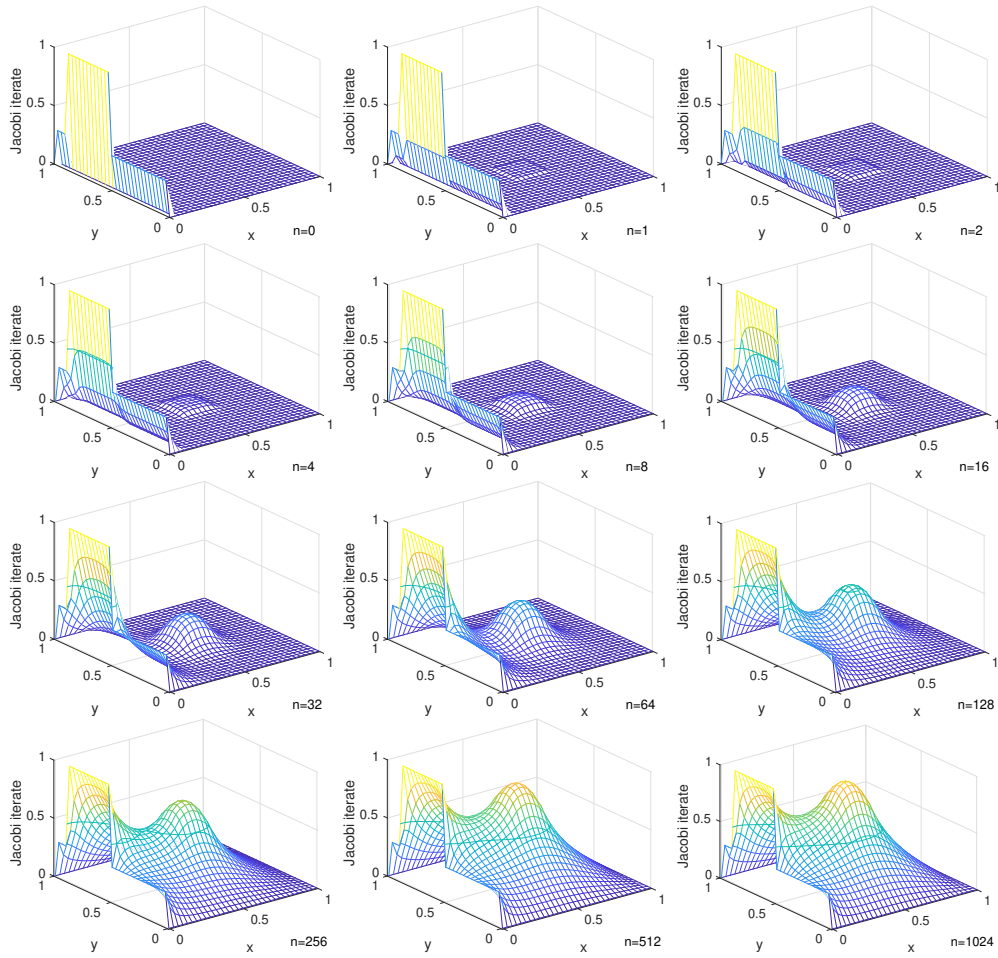


FIG. 6.3. *Jacobi iterations for the Montreal problem.*

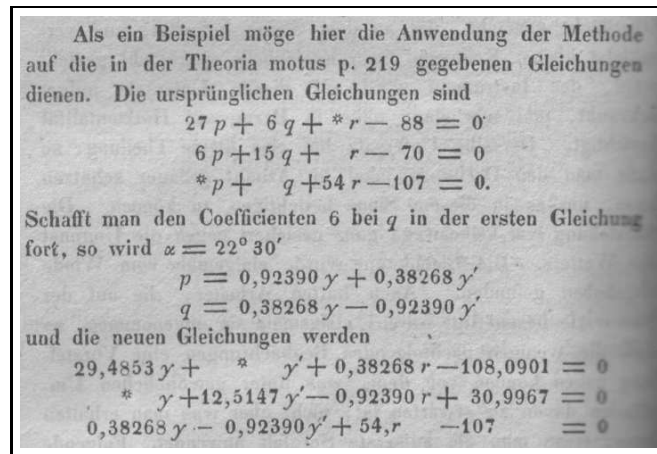


FIG. 6.4. *Jacobi's idea of preconditioning the linear system using Jacobi rotations [93, p. 304].*

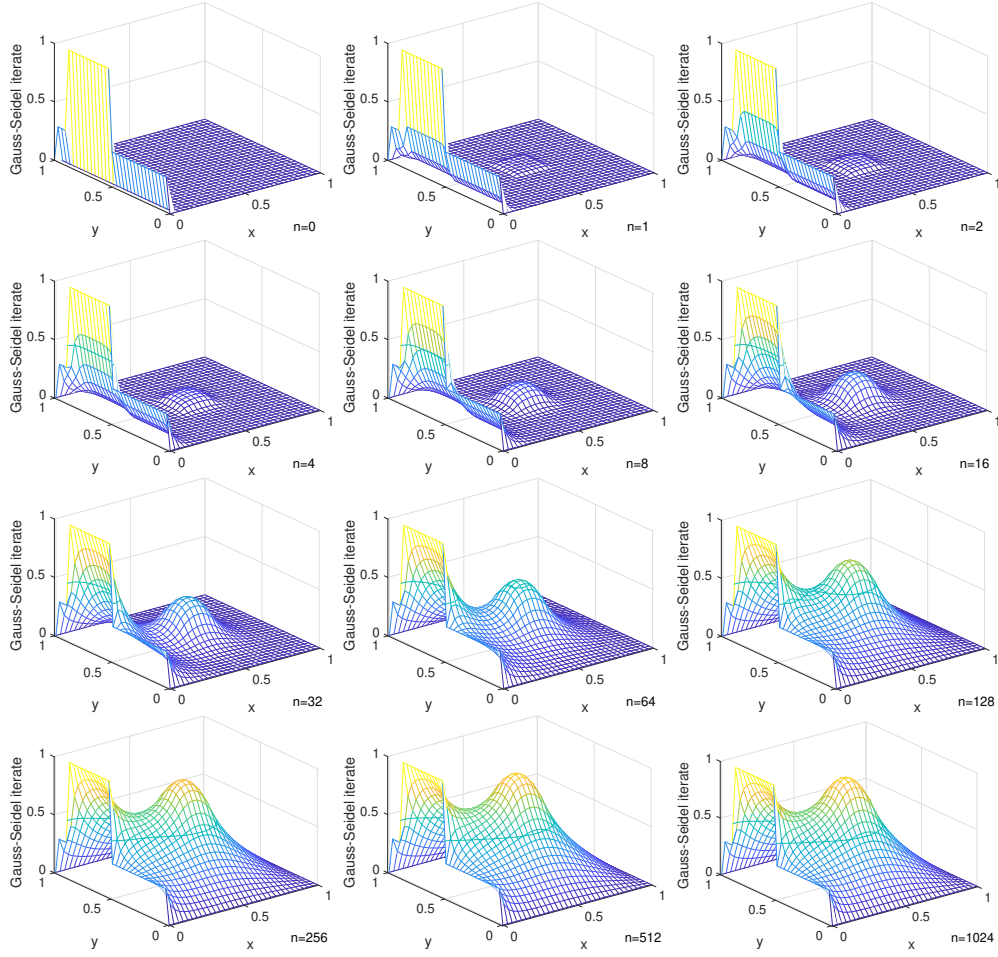


FIG. 6.5. Gauss-Seidel iterations for the Montreal problem.

to use this new value for the determination of u_2^{n+1} etc. We thus have, instead of (6.4),

$$u_i^{n+1} = \sum_{j<i} b_{ij}u_j^{n+1} + \sum_{j>i} b_{ij}u_j^n + g_i \quad \text{or} \quad \mathbf{u}^{n+1} = L\mathbf{u}^{n+1} + U\mathbf{u}^n + \mathbf{g}, \quad (6.8)$$

where

$$L = \begin{pmatrix} 0 & & & \\ b_{21} & 0 & & \\ \vdots & \vdots & \ddots & \\ b_{m1} & b_{m2} & \ddots & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & b_{12} & \cdots & b_{1m} \\ & \ddots & \vdots & \vdots \\ & & 0 & b_{m-1,m} \\ & & & 0 \end{pmatrix}. \quad (6.9)$$

Numerical Example. We now perform the same calculations as in Fig. 6.3 with the Gauss-Seidel method (see Fig. 6.5). Observe that $2n$ iterations of Jacobi produce an equivalent result as n iterations of Gauss-Seidel.

A general description of the method was given by Seidel in [154], who also proved convergence of the method for the case of the normal equations, proposed to do the

so fahre ich fort bis nichts mehr zu corrigiren ist. Von dieser ganzen Rechnung schreibe ich aber in der Wirklichkeit bloss folgendes Schema

	$d = -201$	$b = +92$	$a = -60$	$c = +12$	$a = +5$	$b = -2$	$a = -1$
+6	+5232	+4036	+16	-320	+15	+41	-26
-7558	-6352	-4	+776	+176	+111	-27	-14
-14604	+1074	-3526	-1846	+26	-114	-14	+14
+22156	+46	-506	+1054	+118	-12	0	+26

In sofern ich die Rechnung nur auf das nächste $\frac{1}{1000}$ Sec. führe, sehe ich dafs jetzt nichts mehr zu corrigiren ist. Ich samme daher

$$\begin{array}{rclcl}
 a = -60 & b = +92 & c = +12 & d = -201 \\
 +5 & -2 & & \\
 \hline
 -56 & +90 & +12 & -201
 \end{array}$$

FIG. 6.6. The calculations of Gauss in his letter for Gerling (Cod. MS. Gauss Briefe B : Gerling, Nr 65 / 1 & 2, SUB Göttingen, [177]).

relaxations cyclically, and also to distribute them to two computers (humans) to do parallel computing (“... sich unter zwei Rechner so vertheilen lässt ...” [154, p. 101]). Seidel also considered block variants and using variable precision.

6.4. Back to Gauss’ letter. After the death of Gauss in 1855, hundreds of notes were discovered in his desk and all his correspondence was collected. The enormous work of editing and publishing all these discoveries required half a century. Only in 1903 volume 9 of his *Werke* was published, containing a letter to Christian Ludwig Gerling (1788–1864) from 1823 [78]. In this letter, Gauss explained that solving angle measurements between the locations Berger Warte, Johannisberg, Taufstein and Milseburg, required to solve the system

$$\begin{array}{rcl}
 0 = & +6 & +67a - 13b - 28c - 26d \\
 0 = & -7558 & -13a + 69b - 50c - 6d \\
 0 = & -14604 & -28a - 50b + 156c - 78d \\
 0 = & +22156 & -26a - 6b - 78c + 110d
 \end{array}$$

Gauss then continues (translation by Forsythe [57]):

“In order to eliminate indirectly, I note that, if 3 of the quantities a, b, c, d are set to 0, the fourth gets the largest value when d is chosen as the fourth. Naturally, every quantity must be determined from its own equation, and hence d from the fourth. I therefore set $d = -201$ and substitute this value. The absolute terms then become: +5232, -6352, +1074, +46; the other terms remain the same”

(see the second column of Fig. 6.6). The next value to be corrected would be $b := b + 92$ (third column) and so on. After 7 iterations he then presents the solutions $a = -60 + 5 - 1 = -56$, $b = +92 - 2 = +90$, $c = +12$ and $d = -201$.

We see that Gauss’ method was the same as Seidel’s, with the difference that Seidel just cycled though all variables, while Gauss chose at every step the variable

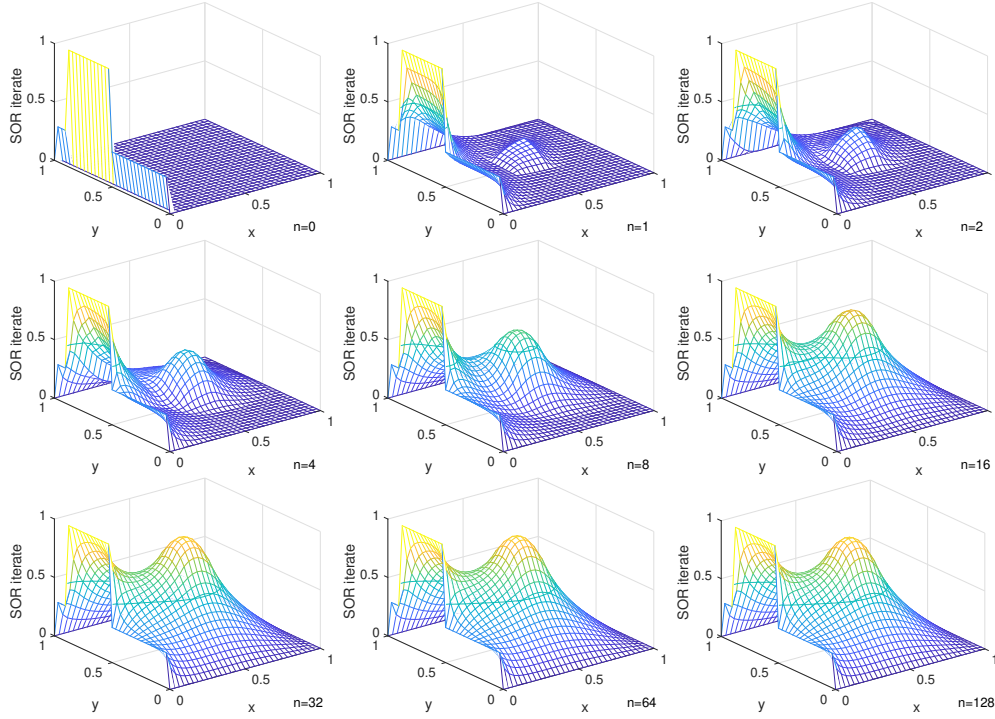


FIG. 6.7. *SOR iterations for the Montreal problem. Note that we only show up to iteration $n = 128$, since there is no visible change any more from $n = 64$ onward!*

with the maximal correction. Therefore method (6.8) is nowadays called *the Gauss-Seidel method*.

6.5. Successive Overrelaxation Method (SOR) of David Young. In his PhD thesis in 1950 [183], David Young (1923-2008) discovered a drastic improvement of the Gauss-Seidel method by applying Richardson’s idea of overrelaxation (see equation (7.2) below) cleverly component-by-component:

“The successive overrelaxation method combines the use of successive displacements and the use of systematic overrelaxation proposed by Richardson [9] as early as 1910” [182, p. 94].

His method, which he called successive overrelaxation (SOR), converges for our model problem as shown in Fig. 6.7. The improvement over the Gauss-Seidel method for the same cost is spectacular, the method provides already very accurate approximations after 32 iterations, whereas Gauss-Seidel needs over 512!

The idea is to **overrelax** Formula (6.8) by introducing an overrelaxation factor $\omega > 1$ and increase for each component the displacement $u_i^{n+1} - u_i^n$ by this factor ω (see also Frankel and the earlier work of Southwell as described in the historical review [151]). In this way, we obtain from (6.8) the SOR algorithm

$$\begin{aligned}
 u_1^{n+1} - u_1^n &= \omega \cdot \begin{pmatrix} -u_1^n & +b_{12}u_2^n & +b_{13}u_3^n & +\dots & +g_1 \end{pmatrix} \\
 u_2^{n+1} - u_2^n &= \omega \cdot \begin{pmatrix} b_{21}u_1^{n+1} & -u_2^n & +b_{23}u_3^n & +\dots & +g_2 \end{pmatrix} \\
 u_3^{n+1} - u_3^n &= \omega \cdot \begin{pmatrix} b_{31}u_1^{n+1} & +b_{32}u_2^{n+1} & -u_3^n & +\dots & +g_3 \end{pmatrix} \\
 &\text{etc.} \qquad \qquad \qquad \text{etc.}
 \end{aligned} \tag{6.10}$$

or, when written in matrix form,

$$\left(\frac{1}{\omega}I - L\right)\mathbf{u}^{n+1} = U\mathbf{u}^n - \frac{\omega - 1}{\omega}\mathbf{u}^n + \mathbf{g}. \quad (6.11)$$

Convergence. For studying the speed of convergence of the algorithm (6.11), we have to compute the maximal eigenvalue μ of the generalized eigenvalue problem

$$\left(\frac{1}{\omega}I - L\right) \cdot \mu \mathbf{x} = U\mathbf{x} - \frac{\omega - 1}{\omega}\mathbf{x} \quad (6.12)$$

or

$$(\mu L + U)\mathbf{x} = \frac{\omega - 1 + \mu}{\omega}\mathbf{x}. \quad (6.13)$$

The idea is to divide this equation by $\sqrt{\mu}$:

$$(\sqrt{\mu}L + \frac{1}{\sqrt{\mu}}U)\mathbf{x} = \frac{\omega - 1 + \mu}{\omega\sqrt{\mu}}\mathbf{x}. \quad (6.14)$$

The “Property A”. In order to treat this seemingly hopeless task, D. Young had another great idea: *We say that the matrix $B = L + U$ has Property A, if for every $\alpha \neq 0$ the matrix $B_\alpha = \alpha L + \frac{1}{\alpha}U$ has the same eigenvalues as B .*

Examples.

$$\begin{pmatrix} 0 & 1 & \\ 1 & 0 & 1 \\ & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & \frac{1}{\alpha} & \\ \alpha & 0 & \frac{1}{\alpha} \\ & \alpha & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ \alpha x_2 \\ \alpha^2 x_3 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ \alpha x_2 \\ \alpha^2 x_3 \end{pmatrix},$$

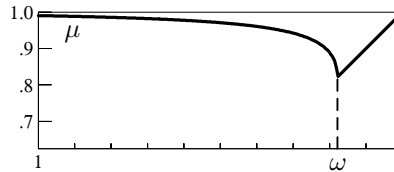
If λ is eigenvalue for the left matrix, it is also eigenvalue for the right one and the coefficients of the eigenvector to the right are scaled by powers of α . The same is true for more general matrices, in particular for the matrix B in (6.6), which is seen as follows:

$$\begin{pmatrix} 0 & 1 & 1 & \\ 1 & 0 & & 1 \\ 1 & & 0 & 1 \\ & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & \frac{1}{\alpha} & \frac{1}{\alpha} & \\ \alpha & 0 & & \frac{1}{\alpha} \\ \alpha & & 0 & \frac{1}{\alpha} \\ & \alpha & \alpha & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ \alpha x_2 \\ \alpha x_3 \\ \alpha^2 x_4 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ \alpha x_2 \\ \alpha x_3 \\ \alpha^2 x_4 \end{pmatrix}.$$

We thus see at once from (6.14): *If B has Property A, the maximal eigenvalue λ of B and the maximal eigenvalue μ for the SOR method (6.11) are connected by*

$$\lambda = \frac{\omega - 1 + \mu}{\omega\sqrt{\mu}} \quad (6.15)$$

or $\nu^2 - \lambda\omega\nu + \omega - 1 = 0$.



The second line in (6.15) is a quadratic equation for $\sqrt{\mu} = \nu$. We have plotted the values of μ as function of ω for the value of $\lambda = \cos \frac{\pi}{32} = 0.9952$, which corresponds to all the Figs. 6.3, 6.5 and 6.7. Apparently, the best value for ω appears where this quadratic equation has double roots, which happens for

$$\omega = \frac{2}{1 + \sqrt{1 - \lambda^2}} \quad \text{for which} \quad \mu_{\text{best}} = \omega - 1 = \frac{1 - \sqrt{1 - \lambda^2}}{1 + \sqrt{1 - \lambda^2}}. \quad (6.16)$$

In our example $\mu_{\text{best}} = 0.8215$ for $\omega = 1.8215$, which explains the drastic improvement of the method observed in Fig. 6.7.

If ω is 1, which corresponds to Gauss-Seidel, we have $\mu = \lambda^2$, hence, under Property A, Gauss-Seidel converges twice as fast as Jacobi, as we have observed by comparing the results in Figures 6.3 and 6.5³⁷. An advantage however of the Jacobi method is that all steps can be executed in parallel.

For more information see [35, Section 2.7], including a very elegant result of Kahan from 1958 which shows that the relaxation parameter must lie in the interval $(0, 2)$, otherwise the SOR method cannot converge! Note that Ostrowski and Reich had previously shown (1947) that if A is SPD, the condition is also sufficient.

7. Non-Stationary Extrapolation Methods. The stationary iterative methods we have seen so far perform the same relaxation step at each iteration, their action is stationary. Non-stationary iterative methods vary in what they do precisely from one step to the next, and this section is devoted to such non-stationary iterative methods based on the principle of extrapolation. We start by describing a spectacular paper of Lewis Fry Richardson (1881–1953) [143], where he treats the complete numerical solution process from modeling using differential equations, over their discretization to the solution of the resulting discrete equations. In addition to Richardson’s now widely known h^2 -extrapolation, he also introduced an important new extrapolation idea for the iterative solution of linear systems, leaving the domain of stationary iterative methods. Richardson’s approach was brought to perfection in 1959 by Gene Golub (1932–2007) in his PhD thesis [80], based on earlier work by John von Neumann (1903–1957) [172, 173], as we will see.

The seminal underlying fundamental idea for Richardson’s iterative method for solving linear systems lies in the extrapolation of a sequence of converging numerical values $S_1, S_2, S_3, \dots, S_n, \dots$ for $n \rightarrow \infty$. Originally initiated in 1926 by Aitken [4], generalized in 1955 by Shanks [155], this led to the famous ε -algorithm by Wynn [180] from 1956, and was subsequently generalized by Wynn to vector sequences [181]. Once one can accelerate vector sequences, approximate solution sequences for linear systems can be accelerated, which led to the minimal polynomial extrapolation method of Cabay and Jackson [18], and the methods in the seminal early book by Brezinski [14]. We mention fundamental analyses and algorithmic improvements by Sidi [158] for minimal polynomial extrapolation and also reduced rank extrapolation, which was invented by Eddy [47], and is a direct generalization to vector sequences of the classical Aitken acceleration. These modern extrapolation methods have strong connections with the Krylov methods we will see in Section 8, and Anderson acceleration also falls into this category of extrapolation methods [50, 174]. It is interesting that all these methods can also be used to accelerate non-linear iterative solvers, see [121] for a recent overview.

7.1. Richardson’s 1911 Paper.

“It is known (W.F.SHEPPARD, “Central-Difference Formulæ,” ‘Proc. Lond. Math. Soc.’ vol. xxxi. (1899)) that (...) the error of the representation of any differential expression by central differences is of the form $h^2 F_2(x, y, z) + h^4 F_4(x, y, z) + \text{terms in higher powers of } h^2, \dots$ ”. (Richardson, [143, p. 310])

Richardson’s h^2 -extrapolation. Based on this observation of Sheppard (see quotation),

³⁷This is not always the case. For an example where Jacobi converges but Gauss-Seidel does not, see [35, Section 2.6]

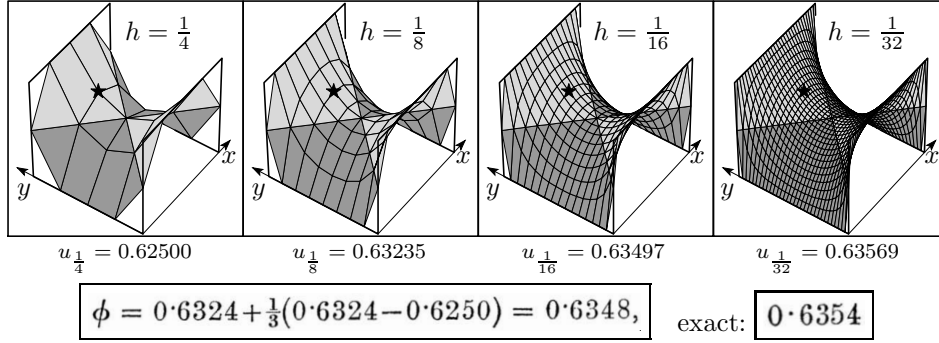


FIG. 7.1. Richardson's h^2 -extrapolation for a Dirichlet Problem (\star is at the point $(0.5, 0.75)$).

Richardson expected that also the errors of numerical discretization results with varying grid-sizes h would behave, at the same point, like $C_2 h^2 + C_4 h^4 + \dots$. Hence, if two numerical values u_h and u_{2h} have been computed for a certain point, one can eliminate the constant C_2 from $u = u_h + h^2 C_2$ and $u = u_{2h} + (2h)^2 C_2$ to obtain $u = u_h + \frac{1}{3}(u_h - u_{2h})$, a much better numerical value at nearly no additional cost; because “numerical cost” was important to Richardson:

“[one has to pay] about $\frac{m}{18}$ pence per co-ordinate point, m being the number of digits [...] the quickest boys averaged 2000 relaxations of Δ_h ³⁸ per week with 3 digits, those done wrong being discounted” [143, p. 325].

Richardson [143, §3.1, p. 315] illustrates this method for the Laplace equation $\Delta u = 0$ on the unit square with 0, 1, 0, 1-constant boundary conditions on the four edges (see Fig. 7.1). For $h = \frac{1}{8}$, due to symmetry, only six unknown values have to be computed, which “was accomplished in an hour”. Of special interest is the value $u_{\frac{1}{8}} = 0.63235$ at the point $(0.5, 0.75)$, since this point is the only remaining grid-point for the grid $h = \frac{1}{4}$ giving $u_{\frac{1}{4}} = 0.62500$. This allows Richardson to extrapolate (Fig. 7.1) to $u = 0.6348$ “and this is only $\frac{1}{10}$ th per cent. in error” from the value “by infinitesimals” $u = 0.6354$ ³⁹.

Initial value problems. The entire §2 of Richardson’s paper treats problems where “the Conditions allow the Integral to be Marched out from a Part of the Boundary”. He chooses the example of the heat equation $\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}$ with initial values $u = 1$ for $t = 0$ and boundary values $u = 0$ for $x = \pm \frac{1}{2}$. Although he refers to classical work by Runge (1895), Heun (1900), Kutta (1901) and Ganz (1903) (see [143, p. 311-312] for references) with results “of remarkable accuracy”, he turns to methods “less accurate, but simpler”: “In satisfying the equation, we must be careful to equate values of⁴⁰”

$$\frac{u_{n-1,m} - 2u_{nm} + u_{n+1,m}}{\Delta x^2} = \frac{u_{n,m+1} - u_{n,m-1}}{2\Delta t},$$

“which are centered at the same point. This causes a little difficulty at starting” [143, p. 313]. In fact, one needs two columns of values for time t_{m-1} and t_m for obtaining a column for time t_{m+1} . Therefore for the first step, Richardson used the implicit

³⁸This is the discretized Laplacian matrix A .

³⁹Computed by adding up Fourier series during 3 hours; in fact, the correct solution is 0.63594, so the extrapolation error is actually only $\frac{1}{6}$ th per cent.

⁴⁰Richardson had a particular notation for these differences which are no longer in use.

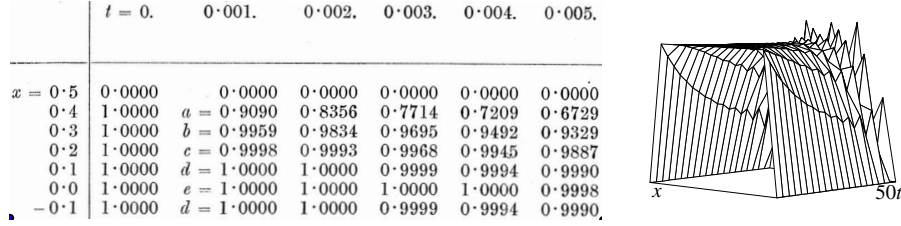


FIG. 7.2. Richardson's solution of Fourier's problem [143, p. 313]; right: results for 20 steps.

mid-point rule centered at $\frac{\Delta t}{2}$. He also knew, by referring to his §3.2.1, that the step-size Δt must be small, therefore he chose $\Delta t = 0.001$ for $\Delta x = 0.1$ and presented the numerical values for 5 steps given in Fig. 7.2. Had he asked his boys to compute 20 steps, he would have seen that *any* $\Delta t > 0$ creates instability (Dahlquist [39]).

Richardson's non-stationary relaxation method. In §3.2, Richardson starts to discuss his method of “Successive Approximation to the Integrals”, i.e. solving equation (6.2) iteratively. For its motivation, let us suppose that $\text{diag} A = \alpha I$ with all entries equal. Then the quantities in equation (6.3) become $B = I - \frac{A}{\alpha}$ and $g = \frac{f}{\alpha}$, and hence the Jacobi iteration (6.4) becomes

$$u^{n+1} - u^n = \frac{1}{\alpha}(f - Au^n). \quad (7.1)$$

We recognize on the left the correction from u^n to the new u^{n+1} and on the right the residual $f - Au^n$ of u^n . The idea is now to vary the factor $\frac{1}{\alpha}$ from one iteration step to the next in order to possibly over- or under-relax the correction and thus accelerate convergence, i.e., to set

$$u^{n+1} = u^n + \frac{1}{\alpha_{n+1}}(f - Au^n), \quad (7.2)$$

where α_{n+1} is a parameter chosen differently at each iteration.

Study of convergence. We subtract equation (7.2) from (6.2) and obtain for the error $e^n := u - u^n$ the equation

$$e^{n+1} = e^n - \frac{1}{\alpha_{n+1}}Ae^n. \quad (7.3)$$

Richardson assumes now an expansion of e^0 in terms of the eigenvectors v_k of A with eigenvalues λ_k as⁴¹

$$e^0 = \sum_k e_k^0 v_k \Rightarrow e^1 = \sum_k \left(1 - \frac{\lambda_k}{\alpha_1}\right) e_k^0 v_k \Rightarrow e^2 = \sum_k \left(1 - \frac{\lambda_k}{\alpha_2}\right) \left(1 - \frac{\lambda_k}{\alpha_1}\right) e_k^0 v_k \dots$$

or, in general,

$$e^n = \sum_k P_n(\lambda_k) e_k^0 v_k \quad \text{where} \quad P_n(\lambda) = \left(1 - \frac{\lambda}{\alpha_1}\right) \left(1 - \frac{\lambda}{\alpha_2}\right) \dots \left(1 - \frac{\lambda}{\alpha_n}\right), \quad (7.4)$$

and “a diagram of the kind shown in figs. 1 and 2 is a great help” (here Fig. 7.3).

⁴¹Richardson wrote originally “ $\phi_1 - \phi_u = \sum_k A_k P_k$ ” and called the P_k the “Principal modes of Vibration” and attributed them to “POCKELS in his book ‘Über die Gleichung, $\Delta^2 u + k^2 u = 0$.’ See also RAYLEIGH, ‘Sound,’ vol. I. chap. IV.”; following Pockels, he wrote λ_k^2 for our λ_k [143, p. 319 and 351].

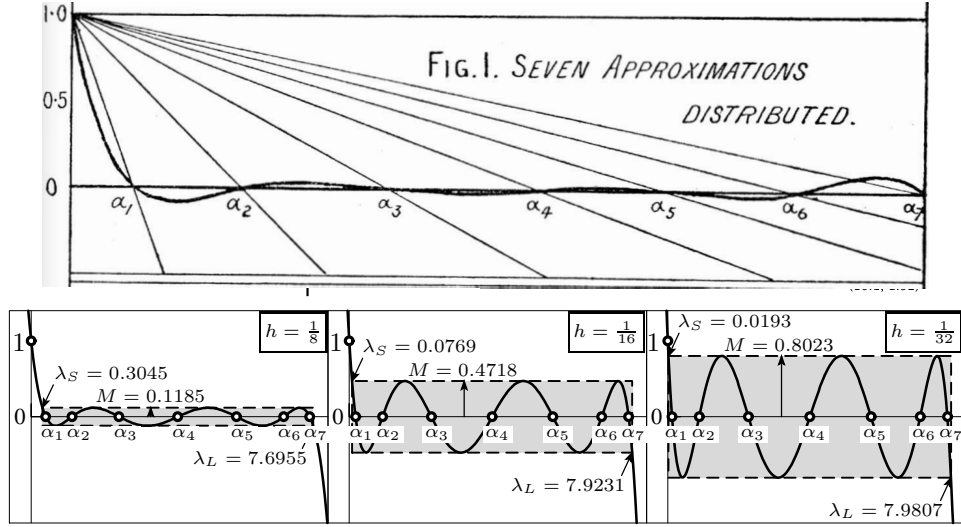
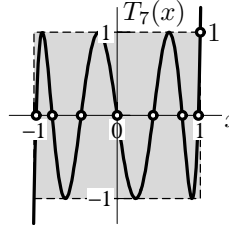


FIG. 7.3. Top: Richardson's "Fig. I" for P_7 in dependence of $\alpha_1, \dots, \alpha_7$; below: The Chebyshev distributions for three grid-sizes on the unit square with $|P|_{\max}$ for the λ 's, i.e. $M = 1/T_7(\mu)$.

Choice of the Parameters. The goal is clearly to render the product P_n as small as possible for all eigenvalues of A . For this it is required to have a rough estimation of an interval (λ_S, λ_L) containing these eigenvalues⁴². For rectangular domains the eigenvectors are exactly known (see (6.7)), and for other domains they have to be estimated. Then, Richardson distributed the (α) 's somehow uniformly over this interval⁴³ (see Fig. 7.3, above). We show in Fig. 7.4 how Richardson's method performs on the Montreal problem when his seven α 's are distributed in the best possible way as indicated in the bottom part of Fig. 7.3, see the next subsection for more details. Convergence of Richardson's method is very good, but not as good as SOR in Fig. 6.7: Richardson's iterate $n = 32$ is comparable to SOR iterate $n = 8$, for essentially the same cost per iteration.

7.2. John von Neumann's Letter. Richardson did not search for the optimal choice of the zeros in order to minimize $|P_n|$ over (λ_S, λ_L) . The solution became clear later, in particular by a letter of von Neumann [172, 173] to the authors of [10], written "with his characteristic brilliance and clarity of style" [10, p. 5]. The main ingredients are the Chebyshev polynomials

$$\begin{aligned}
 T_n(x) &= \cos n\theta \quad \text{with } x = \cos \theta \\
 \text{zeros: } x_i &= \cos(\pi(n-i+\frac{1}{2})/n) \\
 \text{recursion: } T_0(x) &= 1, \quad T_1(x) = x, \\
 T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \quad \text{since} \\
 \cos(n+1)\theta + \cos(n-1)\theta &= 2\cos n\theta \cos \theta.
 \end{aligned}
 \tag{7.5}$$



They were originally published 1853 in the paper [34] on steam engines⁴⁴. Over the years their theory was simplified and extended mainly by Russian authors. Von

⁴²The λ_L is Richardson's own notation (without the square); the λ_S is analogous by replacing "large" by "small".

⁴³"This graph was arrived at by trial."

⁴⁴It is less known that T_2, T_3, T_4 were found by Jakob Bernoulli in 1699, see [87, formula (26)].

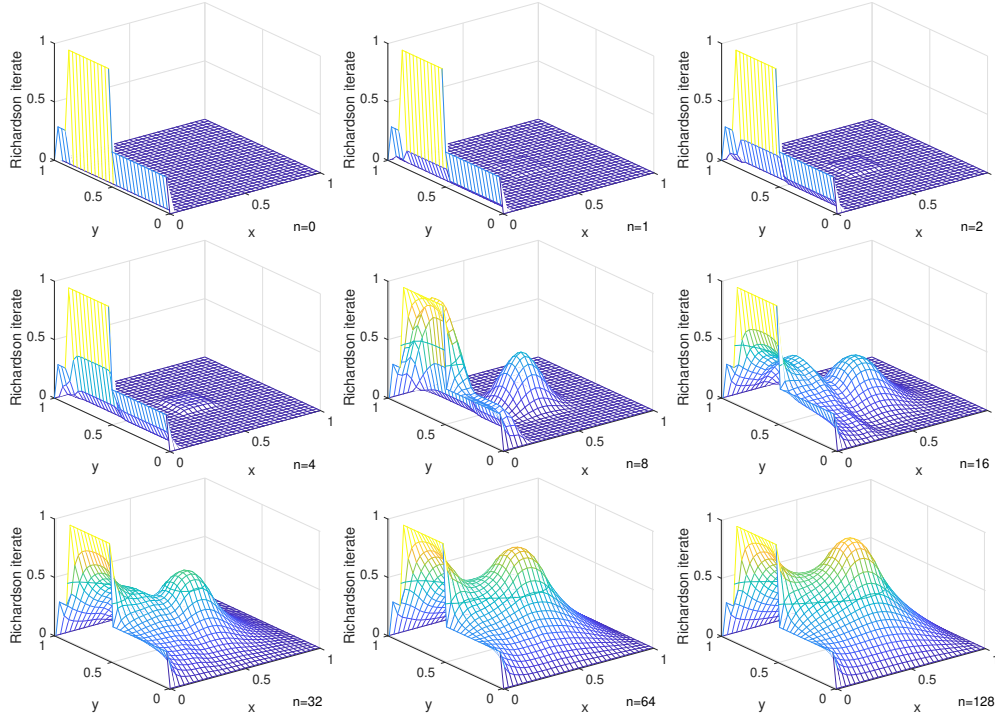


FIG. 7.4. *Richardson iterations for the Montreal problem with seven optimized parameters.*

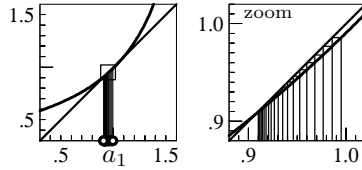
Neumann, who carefully explained how the three-term recursion formula in (7.5) is obtained from the addition theorem of $\cos \theta$, gave as reference: “L. Bernstein: Leçons sur les propriétés extrémales etc., Gauthier-Villars, Paris, 1926 – pp. 7,8”⁴⁵. The optimal $P_n(\lambda)$ is then obtained by shifting T_n from the interval $(1, -1)$ (in this order) to (λ_S, λ_L) and scaling such that $P_n(0) = 1$. This leads to

$$P_n(\lambda) = \frac{T_n(\mu - \mu\nu\lambda)}{T_n(\mu)} \quad \text{with} \quad \mu = \frac{\lambda_L + \lambda_S}{\lambda_L - \lambda_S}, \quad \nu = \frac{2}{\lambda_L + \lambda_S}. \quad (7.6)$$

In Fig. 7.3, we show these curves for three different grid sizes on the unit square. The pictures for the Montreal Problem in Fig. 7.4 were computed with $h = \frac{1}{32}$ for which the reduction factor is $M := 1/T_7(\mu) = 0.8023$, similarly spectacular as the value 0.8215 for SOR in (6.15).

Use of the Recursion Formula. The idea of John von Neumann is now, instead of realizing the polynomial $P_n(\lambda)$ in (7.4) via the iterates for \mathbf{u}_n in (7.2), to use the three-term recursion formula in (7.5). We use this recursion for both the numerators and denominators of P_n in (7.6). In the latter case, we divide the formula by $T_n(\mu)$, after which it becomes the two-term recursion formula

$$\begin{aligned} a_{n+1} &= \frac{1}{2\mu - a_n}, \quad a_1 = \frac{1}{\mu} \\ \text{for } a_n &= \frac{T_{n-1}(\mu)}{T_n(\mu)}. \end{aligned} \quad (7.7)$$



⁴⁵The “L” was wrong, it was Sergei Natanovich Bernstein; by the way: both famous Bernstein’s originated from Jewish families living in Ukraine.

These govern the maximal error after n iterations,

$$M_n = \frac{1}{T_n(\mu)} = \frac{T_{n-1}(\mu)}{T_n(\mu)} \cdot \frac{1}{T_{n-1}(\mu)} = \dots = a_n a_{n-1} \dots a_3 a_2 a_1, \quad (7.8)$$

and we see in (7.7) the enormous gain in precision where, for $h = \frac{1}{32}$, $\mu = 1.004838$, hence $a_1 = 0.99518$, $a_2, a_3 \dots$ continuously descend towards $a_\infty = \mu - \sqrt{\mu^2 - 1} = 0.906348$. The recursion for the numerators then gives

$$P_{n+1}(\lambda) = 2\mu(1 - \nu\lambda)a_{n+1}P_n(\lambda) - a_{n+1}a_nP_{n-1}(\lambda).$$

Here, von Neumann writes “it is convenient to introduce $b_n = \mu a_n$ ” [172, p. 175], and we obtain, after using $-\frac{1}{\mu^2}b_{n+1}b_n = 1 - 2b_{n+1}$, the final formulas

$$P_{n+1}(\lambda) = 2b_{n+1}[(1 - \nu\lambda)P_n(\lambda) - P_{n-1}(\lambda)] + P_{n-1}(\lambda), \quad b_{n+1} = \frac{1}{2 - \frac{1}{\mu^2}b_n}, \quad (7.9)$$

with $P_0 = 1, P_1(\lambda) = 1 - \nu\lambda$ and $b_1 = 1$. Based on these formulas, with the audacity of a great mind, he now defines the

$$\text{Algorithm: } \mathbf{u}^{n+1} = 2b_{n+1}[(\mathbf{u}^n + \nu(\mathbf{b} - A\mathbf{u}^n) - \mathbf{u}^{n-1})] + \mathbf{u}^{n-1}, \quad (7.10)$$

with \mathbf{u}^0 arbitrary; $\mathbf{u}^1 = \mathbf{u}^0 + \nu(\mathbf{b} - A\mathbf{u}^0)$. In order to see that it works, we subtract on the left and right the exact solution \mathbf{u} , for which $\mathbf{b} - A\mathbf{u} = 0$, and obtain for the

$$\text{Error: } \mathbf{e}^{n+1} = 2b_{n+1}[(\mathbf{e}^n - \nu(A\mathbf{e}^n) - \mathbf{e}^{n-1})] + \mathbf{e}^{n-1}, \quad (7.11)$$

with \mathbf{e}^0 arbitrary; $\mathbf{e}^1 = \mathbf{e}^0 - \nu A\mathbf{e}^0$. If we now expand, as above, $\mathbf{e}^0 = \sum_k e_k^0 \mathbf{v}_k$ in the space of eigenvectors, the formulas (7.11) become, for each component, the same as (7.9), hence we obtain exactly the same result as (7.4). For a fixed n , both methods are thus equivalent, but the advantage here is that we have for *all* n the optimal solution. If, for example, in order to double the precision after 7 Richardson iterations, one has to redo 7 new steps, 3 additional von Neumann iterations suffice to double the precision. A numerical illustration of the algorithm (7.10) for the Montreal problem is given in Fig. 7.5. We see that iteration 8 is slightly better than iteration 8 of the Richardson method in Fig. 7.4, iteration 7 would have been identical. From then on the Chebyshev iterations are however much much better, but SOR is still a bit better, as one can see from comparing iterate 32 for example with Fig. 6.7, where the bump in the middle is a bit higher already.

7.3. Golub’s Modified Chebyshev Semi-iterative Method. The authors of [10] wrote at the end of their “Preface”: “The conversion of these methods into a computational form has gone through several phases and has occupied several of us over a much longer period”. Eventually, one of them, Abraham Haskel Taub (1911–1999) from the University of Illinois, suggested this subject to one of his brilliant students, Gene Golub (1932–2007), and the acceleration of stationary iterative methods became his PhD thesis [80]: instead of comparing the performance of the best Chebyshev polynomial at each step in Fig. 7.5 with SOR in Fig. 6.7 which is still a bit better, why not try to combine the power of these two approaches? Gene carefully studied this and provided the essential insight in his PhD thesis⁴⁶

⁴⁶According to Dianne O’Leary [135], when Golub was finishing his thesis, Richard Varga visited and told Golub’s advisor, Abraham Taub, that he had independently obtained very similar results. Taub then told Gene that if Varga publishes first, he will have to write a new thesis. Golub went to see Varga and the latter agreed to joint publication [81], thereby allowing Golub to receive his degree.

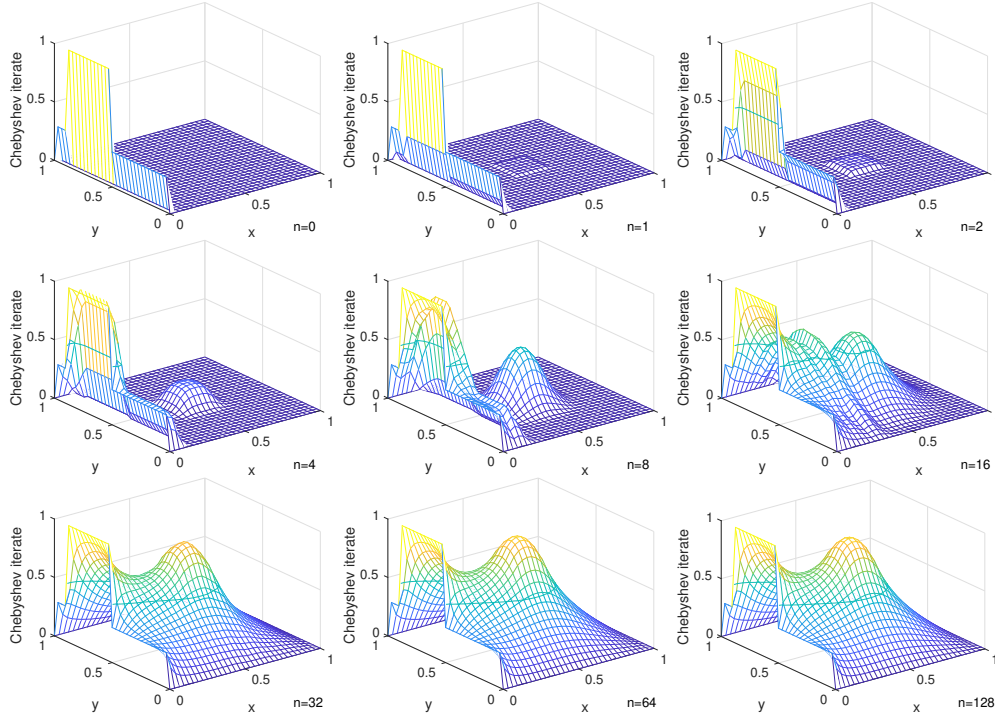


FIG. 7.5. *Chebyshev iterations for the Montreal problem.*

“If $c_m = b_m$ [...] then (1.4) describes the modified Chebyshev method, and if $c_m = b$, then (1.4) describes the successive over-relaxation method” [80, p. 4].

Gene had discovered that one cannot just accelerate the simple Richardson iteration with $\alpha_n = \alpha$ constant using the Chebyshev polynomials and their recursion to find an optimized residual polynomial, one can in fact accelerate any stationary iteration of the form (7.15) that satisfies certain assumptions. Writing this Chebyshev acceleration for SOR led to his quote above, the method he calls modified Chebyshev Semi-iterative Method, and allowed him to prove that then again the accelerated version converges much faster than SOR alone. With this, Gene laid the foundations of modern preconditioning techniques we will see in Subsection 8.4: the Chebyshev acceleration will be replaced by a Krylov method which is even better as we will see in Section 8, and then preconditioned by SOR or any other even more effective stationary iterative method like domain decomposition or multigrid covered in Sections 9 and 10.

7.4. Cabay and Jackson: Modified Polynomial Extrapolation. The main drawback of the Chebyshev Semi-iterative Method is that one needs to know the spectrum of the iteration operator in order to estimate the interval (λ_S, λ_L) on which the residual polynomial needs to be small. In 1976, Cabay and Jackson [18] proposed a new idea for an extrapolation method for accelerating the convergence of vector sequences \mathbf{u}^n , for example generated by the stationary iterative method SOR (6.11), but in contrast to the Chebyshev acceleration in the previous subsection, without the need of knowing either the iteration matrix and an estimate of its spectrum, or the right hand side:

“The extrapolation method is intended primarily for those problems where

the iterates $\mathbf{u}^0, \mathbf{u}^1, \mathbf{u}^2, \dots$ only are known" [18, p. 735].

The method can thus even be used for non-linear stationary iterations, only the iterates are needed! The underlying idea is as for the Chebyshev Semi-iterative Method: one tries to form a linear combination of the iterates, $\mathbf{v}^n := \sum_{j=0}^n \gamma_j \mathbf{u}^j$, $\sum_{j=0}^n \gamma_j = 1$, that converges more quickly to the desired solution \mathbf{u} . If we introduce the error $\mathbf{e}^n := \mathbf{u} - \mathbf{u}^n$, we obtain by isolating \mathbf{u} , multiplying by γ_n and summing

$$\sum_{j=0}^n \gamma_j \mathbf{u}^j = \mathbf{u} \sum_{j=0}^n \gamma_j - \sum_{j=0}^n \gamma_j \mathbf{e}^j, \quad (7.12)$$

and since the sum of the γ_j equals 1

$$\mathbf{v}^n = \mathbf{u} - \sum_{j=0}^n \gamma_j \mathbf{e}^j. \quad (7.13)$$

A good choice of the coefficients γ_j should thus make the sum of the errors as small as possible, to make \mathbf{v}^n as close as possible to the solution \mathbf{u} . To solve however a least squares problem

$$\sum_{j=0}^n \gamma_j \mathbf{e}^j \longrightarrow \min, \quad (7.14)$$

one would need to know the errors \mathbf{e}^j , which is not possible without knowing the solution. To overcome this problem, we need to find a related computable quantity that can be made small, and we show now that the differences of iterates $\mathbf{d}^n := \mathbf{u}^{n+1} - \mathbf{u}^n$ can play this role: in the case of a general stationary iteration for the linear system $A\mathbf{u} = \mathbf{f}$ using a matrix splitting $A = M - N$ ⁴⁷,

$$\mathbf{u}^n = M^{-1}N\mathbf{u}^{n-1} + M^{-1}\mathbf{f}, \quad (7.15)$$

the exact solution satisfies

$$\mathbf{u} = M^{-1}N\mathbf{u} + M^{-1}\mathbf{f},$$

and subtracting the stationary iteration (7.15), we obtain for the error

$$\mathbf{e}^n = \mathbf{u} - \mathbf{u}^n = M^{-1}N(\mathbf{u} - \mathbf{u}^{n-1}) = M^{-1}N\mathbf{e}^{n-1}. \quad (7.16)$$

By induction, denoting the iteration matrix by $G := M^{-1}N$, we thus obtain

$$\mathbf{e}^n = G^n \mathbf{e}^0.$$

Hence the least squares problem (7.14) becomes

$$\sum_{j=0}^n \gamma_j \mathbf{e}^j = \sum_{j=0}^n \gamma_j G^j \mathbf{e}^0 = p_n(G) \mathbf{e}^0 \longrightarrow \min, \quad (7.17)$$

which means we have to find a polynomial $p_n(G)$ in the iteration matrix that makes the sum small. If $p_n(G)$ was the minimal polynomial of the matrix G for the vector \mathbf{e}^0

⁴⁷For example $M = \text{diag}(A)$ for Jacobi.

we would in fact obtain zero, and thus the exact solution at step n of this procedure. Taking the difference of (7.15) at two consecutive iterates, we find for the difference $\mathbf{d}^n := \mathbf{u}^{n+1} - \mathbf{u}^n$ the same iteration formula as for the error,

$$\mathbf{d}^n := G\mathbf{d}^{n-1} = G^n \mathbf{d}^0.$$

The idea of modified polynomial extrapolation (MPE) is to make the quantity

$$\sum_{j=0}^n \gamma_j \mathbf{d}^j = \sum_{j=0}^n \gamma_j G^j \mathbf{d}^0 = p_n(G) \mathbf{d}^0 \longrightarrow \min, \quad (7.18)$$

because the \mathbf{d}^j can be computed, in contrast to the \mathbf{e}^j in (7.17). This is further motivated by the fact that if $p_n(G)\mathbf{e}^0 = 0$, then also $p_n(G)\mathbf{e}^m = 0$ and $p_n(G)\mathbf{d}^m = 0$ for $m \geq 0$: the first result holds since polynomials in G commute, and we thus have

$$0 = G^m p_n(G) \mathbf{e}^0 = p_n(G) G^m \mathbf{e}^0 = p_n(G) \mathbf{e}^m.$$

The second result uses that $\mathbf{d}^m = \mathbf{u}^{m+1} - \mathbf{u}^m = \mathbf{u} - \mathbf{u}^m + \mathbf{u}^{m+1} - \mathbf{u} = (I - G)\mathbf{e}^m$, which implies for $m \geq 0$

$$0 = (I - G)p_n(G)\mathbf{e}^m = p_n(G)(I - G)\mathbf{e}^m = p_n(G)\mathbf{d}^m.$$

In MPE, one solves the minimization problem (7.18) approximately by solving the least squares problem

$$\sum_{j=0}^{n-1} \tilde{\gamma}_j \mathbf{d}^j = -\mathbf{d}^n,$$

and the resulting polynomial

$$p_n(G) = \tilde{\gamma}_0 + \tilde{\gamma}_1 G + \dots + \tilde{\gamma}_{n-1} G^{n-1} + G^n$$

is then an approximation of the minimal polynomial, and MPE gives the exact solution as soon as $p_n(G)$ is the minimal polynomial! The normalized coefficients for the extrapolation are simply obtained by

$$\gamma_j = \frac{\tilde{\gamma}_j}{\sum_{j=0}^m \tilde{\gamma}_j}.$$

We show in Fig. 7.6 the results of MPE for the Montreal problem. We see that MPE converges rapidly to the solution⁴⁸, even a bit faster than the Chebyshev Semi-iterative Method (compare the bump at iteration 32 which is a bit higher for MPE), and we will see in Section 8 on Krylov methods that MPE is in fact equivalent to the conjugate gradient method, and thus in general faster than the Chebyshev Semi-iterative Method, and this without knowing the spectrum of the iteration operator.

MPE is not the only possibility to solve the minimization problem (7.18) approximately. Eddy [46] and Mesina [122] proposed a different variant called reduced rank extrapolation (RRE). Brezinski focused in his early book [14] on the ε -algorithm which is a further extrapolation technique:

⁴⁸We used for the stationary iteration (7.15) the scaled identity $M = \frac{4}{h^2}I$. Using an unscaled identity would work as well in theory, but produces rapidly growing iterates which leads to problems in floating point arithmetic.

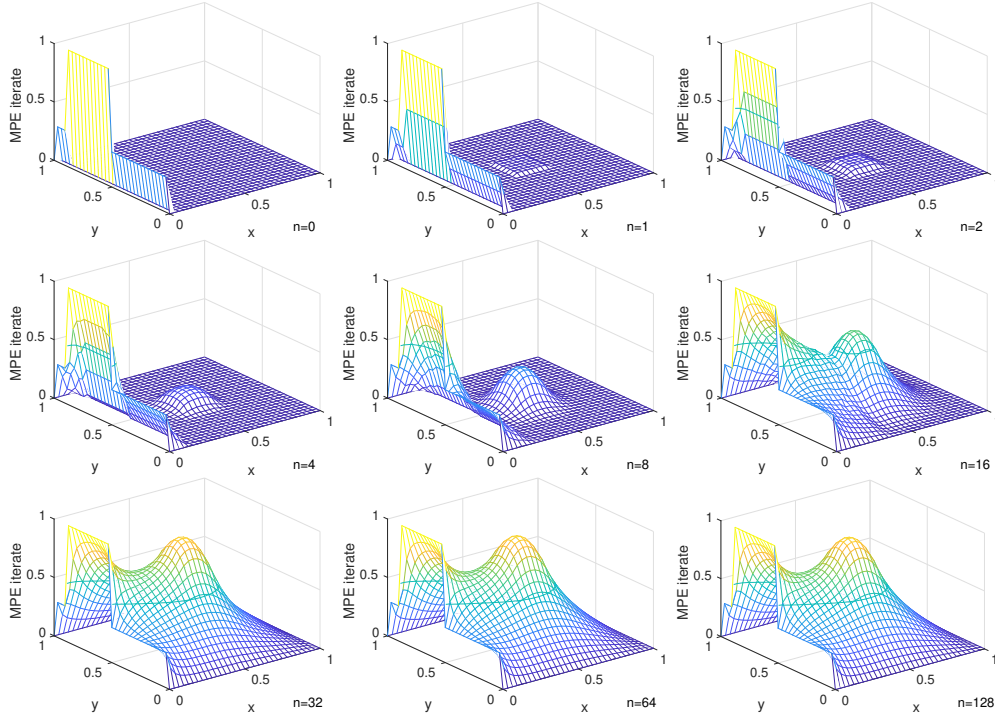


FIG. 7.6. *MPE iterations for the Montreal problem.*

“Nous effectuerons toujours ces transformations à l’aide de modifications appropriées de l’ ε -algorithme et cela pour deux raisons:

- l’ ε -algorithme est le plus puissant de tous les algorithmes que nous avons étudiés pour des suites de nombres.
- Il n’existe de résultats théoriques sur les suites non scalaires que pour l’ ε -algorithme” [14, p. 136].

Sidi studied both MPE and RRE in great detail [156] and also compared them to the vector ε -algorithm of Wynn [181] and the topological ε -algorithm from the seminal work of Brezinski [14], and observed that MPE and RRE have similar convergence properties and are in general more efficient than the ε -algorithm based vector acceleration techniques. Sidi also established a fundamental relation to Krylov methods [157] which have become the mainstream iterative solvers for linear systems of equations, for more details, see [77, Section 11.6 and 11.7], and in particular [77, Theorem 11.32] for the precise equivalence of MPE with the Krylov method FOM. This implies that MPE for symmetric positive definite problems like our model problem is equivalent to the Conjugate Gradient method, which is the first Krylov method we will see in the next section.

8. Krylov Methods. Krylov methods were invented in the early 1950’s independently by Lanczos [111] based on work on an algorithm to compute eigenvalues [110], Forsythe, Hestenes and Rosser [56], and Stiefel [163], in the form of the conjugate gradient method (CG) for symmetric positive definite systems. The structure they use, namely the Krylov space, goes back to work by Krylov [108] in 1931, not related to the solution of linear systems however, but to the concept of characteristic polynomials for the solution of systems of second order ordinary differential equa-

tions describing vibrations. The conjugate gradient method was so successful in the 1970's that many variants were developed for indefinite symmetric and general non-symmetric problems, and their relation to extrapolation methods was discovered, see Subsection 7.4. Since Krylov methods can still exhibit slow convergence, the research field of preconditioning developed, which builds a natural connection between the stationary methods in Section 6 and Krylov methods, and is still currently an intensively researched field in numerical analysis, as we will explain at the end of this section.

8.1. Relaxation and the Method of Steepest Descent.

“On peut tirer des principes ici exposés un parti très-avantageux pour la détermination de l'orbite d'un astre [...]”

(Cauchy, [24, p. 538])

In analogy to Riemann [145], who reduced a hopeless differential equation to a hopefully solvable Minimization Problem (see also [74, Eq. (2.7)]),

$$\boxed{-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f} \iff \iint_{\Omega} \left(\frac{1}{2} \left(\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right) - u \cdot f \right) dx dy \rightarrow \min, \quad (8.1)$$

we can also reduce the corresponding “hopeless” system (6.2) of myriads of linear equations to a Minimization Problem for **one** real function,

$$\boxed{A\mathbf{u} = \mathbf{f}} \iff F(\mathbf{u}) := \frac{1}{2} \mathbf{u}^T A \mathbf{u} - \mathbf{u}^T \mathbf{f} \rightarrow \min, \quad (8.2)$$

where our matrix A in (6.2) is symmetric and positive definite. The equivalence in (8.2) is seen by computing the *gradient*

$$F'(\mathbf{u}) = \frac{1}{2} A \mathbf{u} + \frac{1}{2} A^T \mathbf{u} - \mathbf{f} = A \mathbf{u} - \mathbf{f}. \quad (8.3)$$

Since the gradient $F'(\mathbf{u})$ is the vector indicating the direction in which $F(\mathbf{u})$ increases fastest, we denote, for a given position \mathbf{u} , by

$$\mathbf{r} := -F'(\mathbf{u}) = \mathbf{f} - A \mathbf{u} \quad (\text{the “residual”}) \quad (8.4)$$

the vector indicating the direction “of steepest descent”. Both, $F'(\mathbf{u})$ as well as \mathbf{r} , are perpendicular to the level set of F through the point \mathbf{u} .

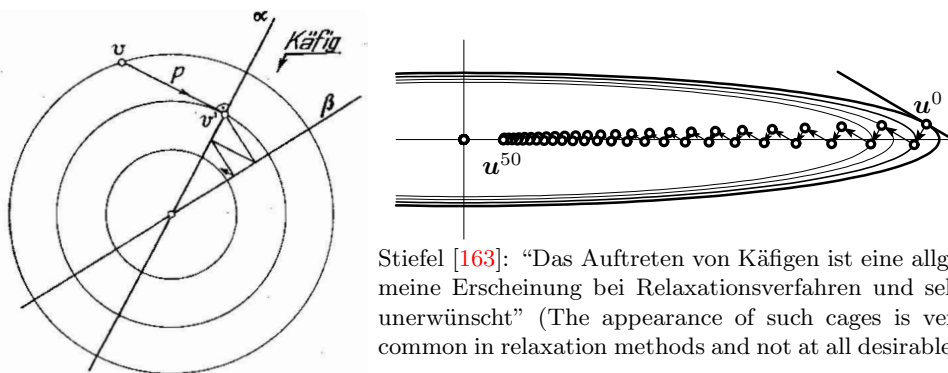
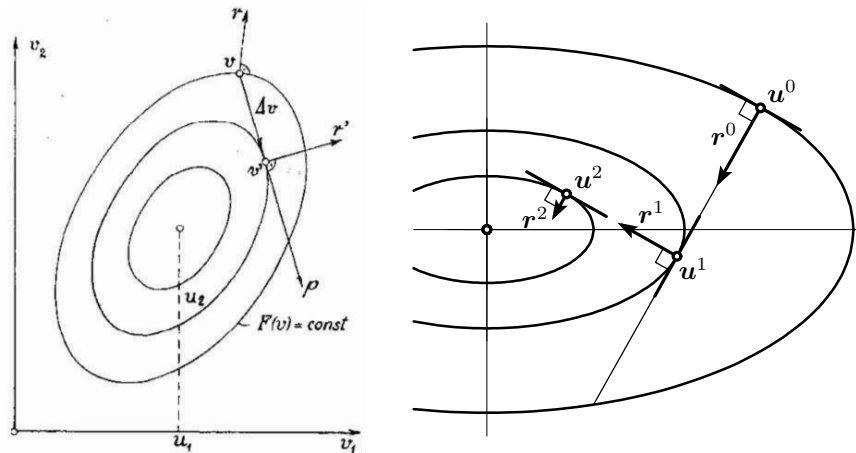
General relaxation step. Given a position \mathbf{u}^n and a search direction \mathbf{p}^n , which is often the residual \mathbf{r}^n from (8.4), we search a new point

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \alpha_n \mathbf{p}^n. \quad (8.5)$$

This method was suggested and used several times during the centuries, often in order to minimize Least Squares errors. An early paper was Cauchy [24], written during the years of enthusiasm for the discovery of new asteroids as well as of Neptune in 1846 (see the quotation). Also one step of Richardson's method (7.2) as well as Jacobi's iteration in the form of equation (7.1) are the same as (8.5) with (8.4), where the α 's have a different meaning. A complete discussion of this method and its generalizations is given in the wonderful paper by Stiefel [163].

Determination of α_n . The standard determination of the free parameter α_n at each step, already suggested by Cauchy⁴⁹, is such that along the straight line $\mathbf{u}^n + \alpha_n \mathbf{p}^n$

⁴⁹Stiefel calls this idea of optimizing on a subspace “der Ritzsche Gedanke” (idea of Ritz).



the function $F(\mathbf{u})$ becomes minimal, i.e., at \mathbf{u}^{n+1} this line with direction \mathbf{p}^n is tangent to the level set, hence the residual vector \mathbf{r}^{n+1} (using (8.4) and (8.5))

the function $F(\mathbf{u})$ becomes minimal, i.e., at \mathbf{u}^{n+1} this line with direction \mathbf{p}^n is tangent to the level set, hence the residual vector \mathbf{r}^{n+1} (using (8.4) and (8.5))

the function $F(\mathbf{u})$ becomes minimal, i.e., at \mathbf{u}^{n+1} this line with direction \mathbf{p}^n is tangent to the level set, hence the residual vector \mathbf{r}^{n+1} (using (8.4) and (8.5))

$$\mathbf{r}^{n+1} = \mathbf{f} - A\mathbf{u}^{n+1} = \mathbf{f} - A\mathbf{u}^n - \alpha_n A\mathbf{p}^n = \mathbf{r}^n - \alpha_n A\mathbf{p}^n, \quad (8.6)$$

must be perpendicular to \mathbf{p}^n (this is “Satz 2” in Stiefel [163]). Solving the condition $(\mathbf{p}^n)^T \mathbf{r}^{n+1} = 0$ for α_n , we get for the optimized distance

$$\alpha_n = \frac{(\mathbf{p}^n)^T \mathbf{r}^n}{(\mathbf{p}^n)^T A \mathbf{p}^n}. \quad (8.7)$$

If $\mathbf{p}_n = \mathbf{r}_n$ for all n , we have the steepest descent method. Illustrations for dimension 2 are given in Fig. 8.1.

Stiefel’s “Käfig”. We now present in Fig. 8.2 the steepest descent method applied to a matrix A whose eigenvalues have a large ratio, i.e., where the corresponding ellipses become narrow. Here, Stiefel imagined the computing “animal” trapped in a cage (“Käfig”), running up and down, far away from the desired solution. Also in Fig. 8.3,

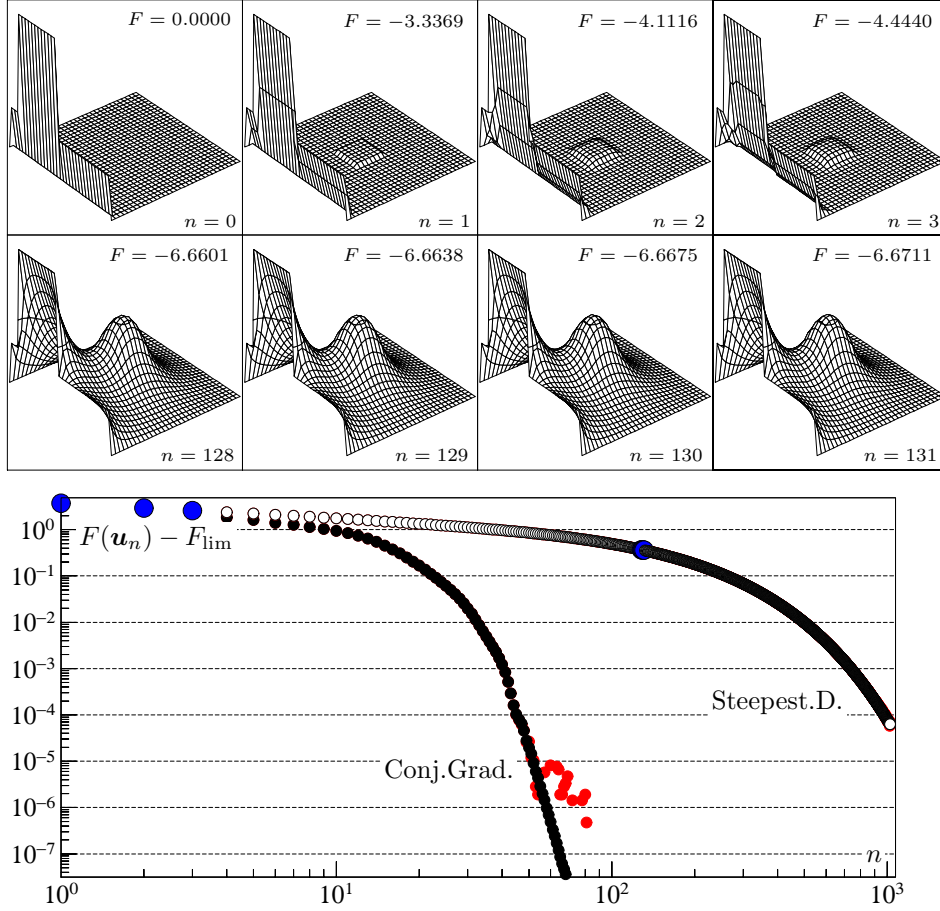


FIG. 8.3. Steepest descent for the Montreal Problem struggling in Stiefel’s “Käfig”; below: the errors of $F(u_n)$ compared to $F_{\text{lim}} = -7.0291170905542$ in double logarithmic scale; in blue: the values for $n = 1-3, 128-131$ of the above figures; compared to the values of the Conjugate Gradient Method; in red: results for single precision.

where the Montreal Problem, of dimension $31^2 = 961$, has been calculated with the standard steepest descent method, we observe that our “animal” has problems to get rid of the “Käfig”.

Ways out of the “Käfig”. Stiefel then proposes in [163] several innovative remedies for this failure:

- Block relaxation, which is a simultaneous relaxation of several equations by the same averaged amount, and thus corresponds to a block Jacobi method, which can be interpreted as a Schwarz domain decomposition method, see Subsection 9.1, and [35, Chapter 4 and Problem 55] for more details.
- An approach Stiefel calls “Scheibenrelaxation”, either choosing search directions related to eigenfunctions on subdomains, or solving directly small subproblems for low modes by relaxation. Stiefel recommends for a given operator to precompute such directions:

“Es ist zweckmässig, für einen gegebenen Operator eine Sammlung

von Scheiben anzulegen⁵⁰.”

He thus already foresaw the coarse correction in multigrid methods, and more specifically the new coarse spaces designed in recent domain decomposition methods like GenEO and SDEM, see [127, 162, 71, 70], precisely by eigenfunction computations; for a historical introduction, see [66].

- The method of conjugate search directions: in this case, one can eliminate completely error components in the direction of each \mathbf{p}^n , independently of the other directions.

While the first ideas with their many possibilities of choosing good procedures were thought for “ein geübter Rechner” (experienced (human) calculator), Stiefel claimed that the steepest descent and conjugate gradient methods “die zwangsläufig fortschreiten” (with precisely prescribed operations) were “suitable for use on sequence-controlled computing machines”, in particular the “programmgesteuerte Zuse-Rechenmaschine in Zürich”, the very first freely programmable computer working in binary floating point arithmetic⁵¹.

We show in Figure 8.4 for the Montreal Problem the results of Steepest Descent for the same iteration steps as we did for all the earlier iterative methods. We see that steepest descent is comparable to Jacobi in Figure 6.3, a very slow method (the level set indicates at iteration 256 that it is slightly faster). We see that even doing locally the best, which is called a good tactic, is not enough for obtaining an overall globally good strategy (according to Stiefel, who was a colonel in the Swiss army). A better strategy is needed for iteratively solving linear systems.

8.2. The Conjugate Gradient Method.

“Rosser (...) returned to INA (Institute for Numerical Analysis) in the summer of 1951 to pursue his studies of solutions of linear equations and to attend a conference on “Solutions of Linear Equations and the Determination of Eigenvalues” to be held at INA in August 1951. In June or July 1951, after almost two years of studying algorithms for solving systems of linear equations, we finally “hit” upon a conjugate-gradient method. I had the privilege of first formulating this new method. However, it was an outgrowth of my discussions with my colleagues at INA. In particular, my conversations with George Forsythe had a great influence on me. During the month of July 1951, I wrote an INA report on this new development. When E. Stiefel arrived at INA in August to attend the conference on solutions of linear equations, he was given a copy of my paper. Shortly thereafter he came to my office and said about the paper, “This is my talk.” (...) Accordingly, I invited Stiefel to remain at UCLA and INA for one semester so that we could write an extensive paper on this subject (the paper [90]). In the meantime, Lanczos observed that the conjugate-gradient method could be derived from his algorithm for finding eigenvalues of matrices.”

(M. R. Hestenes [89, pp. 173–174])

Magnus R. Hestenes (1906–1991), originally expert in optimization and variational calculus, was invited in 1949 by J. Barkley Rosser (1907–1989) to join his group at INA for studying the solution of large linear equations (see [89] and the above quote). After two years of struggle, the conjugate-gradient method (CG) came finally out from discussions among this group. An early unpublished paper of Hestenes was certainly

⁵⁰It is very useful for a given operator to precompute a collection of such directions.

⁵¹Stiefel [163, p. 29]: “[...] dauerte ein Zyklus etwa 2 h 20 m” (a cycle took approximately 2 h 20 min).

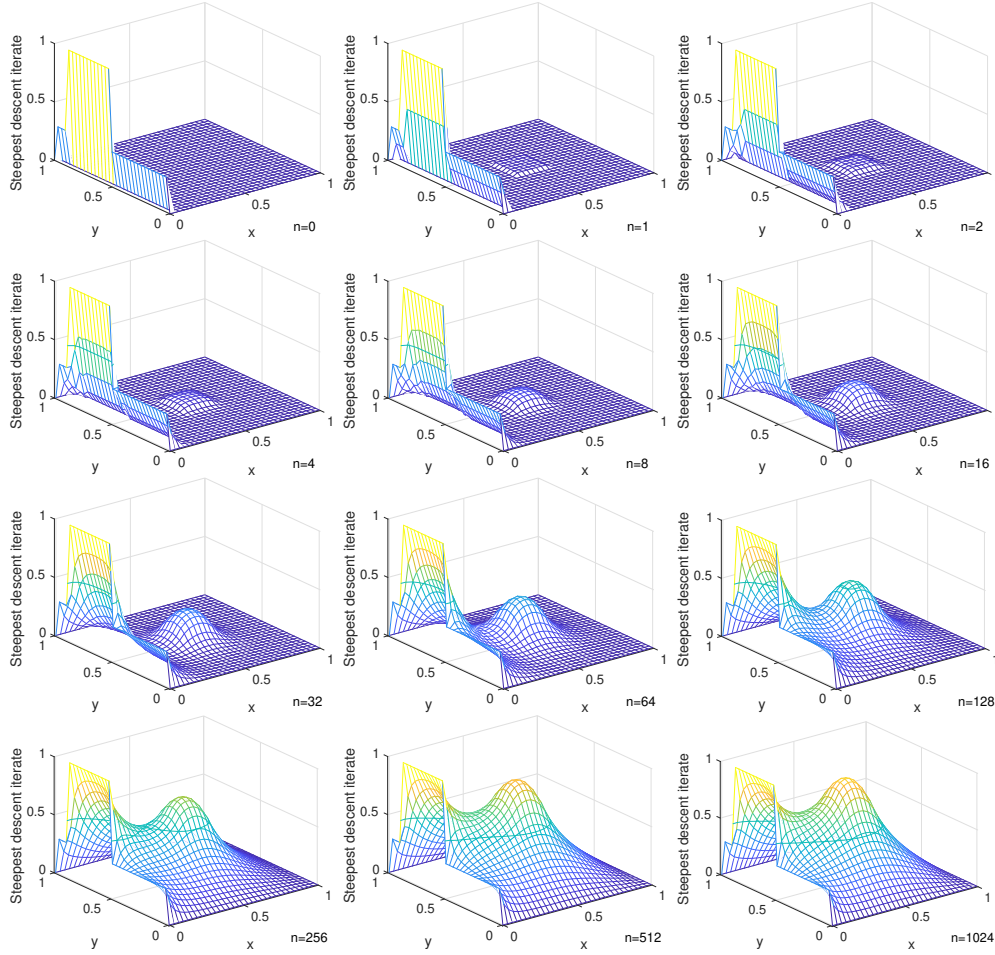


FIG. 8.4. *Steepest Descent iterations for the Montreal Problem. Note that we plot again up to iteration 1024 because of slow convergence.*

helpful⁵².

The surprise was that E. Stiefel (1909–1978), from ETH Zürich, had invented independently the same method. Stiefel had sent already (July 14, 1951) his manuscript [163] to the editors of ZAMP and acknowledged (p. 23) in a “Zusatz bei der Korrektur” the independent discovery of Hestenes.

Simultaneous Relaxation. Following Stiefel [163, §3], we start by presenting the idea of a simultaneous relaxation from a given point \mathbf{u}^0 along *several* directions, we take here \mathbf{p}^0 and \mathbf{p}^1 , as

$$\mathbf{u}^2 = \mathbf{u}^0 + \alpha_0 \mathbf{p}^0 + \alpha_1 \mathbf{p}^1 \quad (8.8)$$

⁵²“It is interesting to recall that in 1936, I developed an algorithm for constructing a set of mutually conjugate directions in Euclidean space for the purpose of studying quadric surfaces. I showed my results to Professor Graustein, a geometer at Harvard University. His reaction was that it was too obvious to merit publication” [89, p. 168].



FIG. 8.5. *Eduard Stiefel 1909-1978 (left); Magnus R. Hestenes 1906-1991 (right).*



FIG. 8.6. *For such a Zuse machine Stiefel developed CG (admired by Gerhard Wanner, photographed by Peter Deuflhard, left); Dr. h.c. (for half an hour) Konrad Zuse, congratulated by Walter and Martin Gander at ETH (right).*

(instead of (8.5)) with the new residual

$$\mathbf{r}^2 = \mathbf{f} - A\mathbf{u}^2 = \mathbf{f} - A\mathbf{u}^0 - \alpha_0 A\mathbf{p}^0 - \alpha_1 A\mathbf{p}^1 = \mathbf{r}^0 - \alpha_0 A\mathbf{p}^0 - \alpha_1 A\mathbf{p}^1, \quad (8.9)$$

similar to (8.6). Restricted to the plane through \mathbf{u}^0 spanned by \mathbf{p}^0 and \mathbf{p}^1 as in (8.8), the function $F(\mathbf{u})$ represents an elliptic paraboloid, whose minimum \mathbf{u}^2 , the “Ritzsche Gedanke” again, we are looking for. At this minimum, the residual \mathbf{r}^2 , in generalization to (8.7), must be perpendicular to *both* \mathbf{p}^0 and \mathbf{p}^1 . These conditions now become with (8.9) the following system for α_0 and α_1 (see [163, equation (34)]):

$$\begin{pmatrix} (\mathbf{p}^0)^T A\mathbf{p}^0 & (\mathbf{p}^0)^T A\mathbf{p}^1 \\ (\mathbf{p}^1)^T A\mathbf{p}^0 & (\mathbf{p}^1)^T A\mathbf{p}^1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} (\mathbf{p}^0)^T \mathbf{r}^0 \\ (\mathbf{p}^1)^T \mathbf{r}^0 \end{pmatrix}. \quad (8.10)$$

Here comes the great idea of Siefel: If the directions \mathbf{p}^0 and \mathbf{p}^1 were chosen in such a way that (see [163, Section 3.b])

$$(\mathbf{p}^0)^T A\mathbf{p}^1 = 0 \quad \text{and, by symmetry,} \quad (\mathbf{p}^1)^T A\mathbf{p}^0 = 0, \quad (8.11)$$

then the system (8.10) becomes diagonal and leads directly to the solutions

$$\alpha_0 = \frac{(\mathbf{p}^0)^T \mathbf{r}^0}{(\mathbf{p}^0)^T A\mathbf{p}^0}, \quad \alpha_1 = \frac{(\mathbf{p}^1)^T \mathbf{r}^0}{(\mathbf{p}^1)^T A\mathbf{p}^1}, \quad (8.12)$$

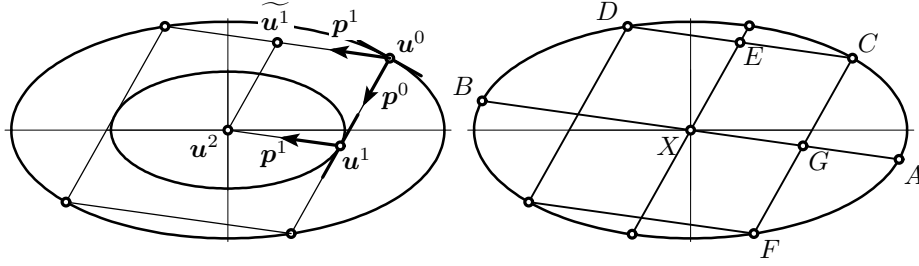
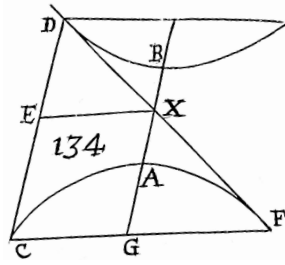


FIG. 8.7. Left: Simultaneous relaxation with conjugate directions; Right: Apollonius' theorem drawn for an ellipse.



Prop. XXXVII.

Si oppositas sectiones (A, B) secet recta linea (CD), non transiens per centrum X, quæ (EX) ab ipsius medio (E) ad centrum ducitur, oppositarum sectionum diameter erit, quæ recta appellatur: transversa verò diameter, ipsi conjugata, est ea (AB), quæ per centrum ducitur æquidistans lineæ bisectæ (CD).

Ducatur DX sectioni^a occurrens in F, & connectatur FC, & producat B A G. Atque ob CE^b = ED, & FX^c = XD^c erit FC ad XE parallela, & FG^d = (XE^e) GC. ^fquare FC tangenti ad A parallela est: & ergo AB, & EX sunt conjugatæ diametri. Q.E.D.

FIG. 8.8. Conjugate directions by Apollonius drawn for the hyperbola (from Isaacus Barrow, *Apollonii Conica : Methodo nova illustrata, & succinctè demonstrata*, London, 1675, p. 50).

which both are the same as (8.7).

We now have a situation as in Fig. 8.7, left: Since the function $F(\mathbf{u})$, restricted to the straight lines $\mathbf{u}^0 + \alpha_0 \mathbf{p}^0$ respectively $\mathbf{u}^0 + \alpha_1 \mathbf{p}^1$, represents parabolas, the minimal points \mathbf{u}^1 respectively \mathbf{u}^2 lie in the *middle* between the points on the same level ellipse. We thus arrive at a 22 century old theorem (Apollonius, end of IIIrd cent. B.C.; see Fig. 8.7, right and Fig. 8.8), where each of the two conjugate diameters intersects the parallels to the other in the middle (“ $CE = ED$ & ... $FG = GC$ ergo AB, & EX sunt conjugatæ diametri”). Unlike as in Fig. 8.2, the line search departing at \mathbf{u}^1 in direction of a conjugate direction \mathbf{p}^1 leads directly to the minimum \mathbf{u}^2 . Based on this idea, Stiefel thus defined in [163] the following algorithm: Beginning at \mathbf{u}^0 , define search directions⁵³

$$\begin{aligned} \mathbf{p}^0 &= \mathbf{r}^0 \\ \mathbf{p}^n &= \mathbf{r}^n + \varepsilon_{n-1} \mathbf{p}^{n-1} \quad n = 1, 2, 3, \dots \end{aligned} \quad (8.13)$$

where the coefficients ε_{n-1} are chosen such that \mathbf{p}^n and \mathbf{p}^{n-1} are conjugate, i.e. from (8.11), $(\mathbf{p}^{n-1})^T A \mathbf{p}^n = 0$, which leads to

$$\varepsilon_{n-1} = -\frac{(\mathbf{p}^{n-1})^T A \mathbf{r}^n}{(\mathbf{p}^{n-1})^T A \mathbf{p}^{n-1}}. \quad (8.14)$$

This is then followed by a standard line search (8.5) $\mathbf{u}^n \mapsto \mathbf{u}^{n+1}$ in direction of \mathbf{p}^n using (8.7). An induction proof then shows that *all* search directions $\mathbf{p}^0, \dots, \mathbf{p}^{n-1}$ are

⁵³Stiefel used ε for the distance to go along the search direction.

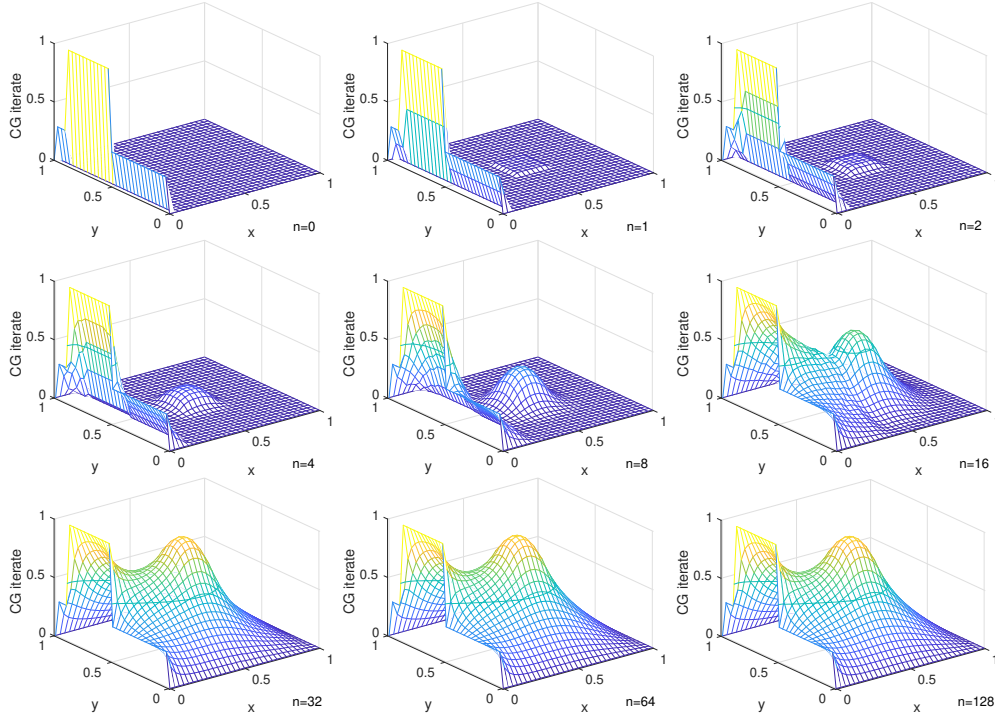


FIG. 8.9. *CG iterations for the Montreal problem.*

pairwise conjugate and the algorithm leads after m iterations to the exact solution, where m is the size of the matrix⁵⁴. We see the immense progress of this method, if we compare the results with those for the steepest descent in Fig. 8.3.

In modern notation, to solve approximately $A\mathbf{u} = \mathbf{f}$, where A is a symmetric and positive definite matrix, CG starts with an initial guess \mathbf{u}^0 with initial residual $\mathbf{r}^0 := \mathbf{f} - A\mathbf{u}^0$, and finds at step n in the affine Krylov space

$$\mathcal{K}_n(A, \mathbf{r}^0) := \mathbf{u}^0 + \text{span}\{\mathbf{r}^0, A\mathbf{r}^0, \dots, A^{n-1}\mathbf{r}^0\}$$

an approximate solution \mathbf{u}_n which satisfies

$$\|\mathbf{u} - \mathbf{u}_n\|_A \longrightarrow \min, \quad (8.15)$$

where the norm $\|\mathbf{u}\|_A := \sqrt{\mathbf{u}^T A \mathbf{u}}$ is the natural energy norm associated with the problem matrix A , see [35, Section 3.2]. CG thus finds the best possible solution in the Krylov space in this norm, it is not possible to do better than this, and CG does it iteratively, without the need to globally minimize, a truly optimal iterative approach.

We show in Fig. 8.9 the results of Conjugate Gradients. Comparing CG with steepest descent in Fig. 8.4, we see that they are worlds apart: steepest descent converges about as slowly as Jacobi in Fig. 6.3, whereas CG converges precisely as MPE in Fig. 7.6, the fastest we have seen so far, and in exact arithmetic, they are equivalent, but CG is a more robust formulation, since it uses the structure of the problem, while MPE just accelerates a sequence of vectors of potentially unknown origin.

⁵⁴More precisely m is the degree of the minimum polynomial of the matrix, and the statement holds only in exact arithmetic.

8.3. General Krylov Methods. The conjugate gradient method became so successful that intensive research efforts were undertaken to develop such methods for more general linear systems, not just symmetric and positive definite ones. Since such matrices however do not define an inner product and associated norm any more, the key ideas of CG above can not be used any more. Inspired by other important properties satisfied by CG, two main classes of such more general Krylov methods were developed.

Minimum residual methods (MR): as stated in (8.15), CG is minimizing the error in the A norm, and we have for symmetric positive definite A

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}_n\|_A^2 &= (\mathbf{u} - \mathbf{u}_n)^T A (\mathbf{u} - \mathbf{u}_n) = (\mathbf{u} - \mathbf{u}_n)^T A A^{-1} A (\mathbf{u} - \mathbf{u}_n) \\ &= (A\mathbf{u} - A\mathbf{u}_n)^T A^{-1} (A\mathbf{u} - A\mathbf{u}_n) = \|\mathbf{f} - A\mathbf{u}_n\|_{A^{-1}}^2.\end{aligned}$$

CG thus also minimizes the residual in the A^{-1} norm. For a general matrix A , one can thus try to also minimize the residual, simply in the 2-norm,

$$\|\mathbf{f} - A\mathbf{u}^n\|_2 \longrightarrow \min. \quad (8.16)$$

This leads for symmetric but not necessarily positive definite matrices A to the MIN-RES method [136] invented by Paige and Saunders in 1975, which runs at a computational cost comparable to CG. For arbitrary matrices A , this gives GMRES [150] proposed by Saad and Schultz in 1986, which however needs to store and work with the entire Krylov space and thus is computationally more expensive than CG. Approximations were therefore also developed that solve the minimum residual problem (8.16) only approximately, like the quasi minimum residual method (QMR) from 1991 by Freund and Nachtigal in [62], with computational cost comparable to CG.

Methods based on orthogonalization (OR): since in CG, the new residual is orthogonal to the current Krylov space (see e.g. [35, Remark 1, Section 3.2]), for a general matrix A one can also determine the new iterate $\mathbf{u}^n \in \mathbf{u}_0 + \mathcal{K}_n(A, \mathbf{r}^0)$ such that the new residual satisfies this orthogonality property

$$\mathbf{f} - A\mathbf{u}^n \perp \mathcal{K}_n(A, \mathbf{r}^0). \quad (8.17)$$

This leads for symmetric but not necessarily positive definite matrices A to the SymmLQ method [136] invented by Paige and Saunders in 1975, whose computational cost is comparable to CG. For general matrices A one obtains the Full Orthogonalization Method (FOM) [148] introduced by Saad in 1981, which requires however again to store the complete Krylov space and is thus computationally more expensive than CG. There are however also very successive methods using this orthogonality condition in different ways to keep computational costs comparable to CG, for example BiCGstab [167] by Van Der Vorst in 1992.

8.4. Preconditioning. For normal matrices, i.e. when $A^T A = A A^T$, all these Krylov methods converge well, if the spectrum of the matrix A is clustered around 1, because they can be interpreted as polynomial approximation problems. This can be seen for CG as follows: since the iterate \mathbf{u}^n of CG lies in the affine Krylov space,

$$\mathbf{u}^n \in \mathbf{u}^0 + \text{span}\{\mathbf{r}^0, A\mathbf{r}^0, \dots, A^{n-1}\mathbf{r}^0\},$$

there exist coefficients γ_j such that

$$\mathbf{u}^n = \mathbf{u}^0 + \sum_{j=0}^{n-1} \gamma_j A^j \mathbf{r}^0 = \mathbf{u}^0 + \sum_{j=0}^{n-1} \gamma_j A^j (\mathbf{f} - A\mathbf{u}^0) = \mathbf{u}^0 + \sum_{j=1}^n \gamma_{j-1} A^j (\mathbf{u} - \mathbf{u}^0).$$

The quantity minimized by CG in the A-norm is therefore

$$\mathbf{u} - \mathbf{u}^n = \mathbf{u} - \mathbf{u}^0 - \sum_{j=1}^n \gamma_{j-1} A^j (\mathbf{u} - \mathbf{u}^0) = p_n(A)(\mathbf{u} - \mathbf{u}^0),$$

where $p_n(A)$ is a so-called residual polynomial of degree n with $p_n(0) = 1$. CG chooses this polynomial such that

$$\|\mathbf{u} - \mathbf{u}^n\|_A = \|p_n(A)(\mathbf{u} - \mathbf{u}^0)\|_A \rightarrow \min,$$

and thus a bound can be obtained by using the eigendecomposition of A (see e.g. [35, Section 3.2, Theorem 25])

$$\|\mathbf{u} - \mathbf{u}^n\|_A \leq \max_j |p_n(\lambda_j(A))| \|\mathbf{u} - \mathbf{u}^0\|_A,$$

where $\lambda_j(A)$ are the eigenvalues of A . CG thus finds at each iteration n a residual polynomial $p_n(\lambda)$ with $p_n(0) = 1$ which is small on the spectrum of A . For symmetric positive definite matrices A their spectrum lies in an interval on the positive real axis, and the shifted and scaled Chebyshev polynomials are the smallest polynomials on such intervals, equaling 1 at zero, see Subsection 7.2, and [35, Section 3.2] for more details. They become especially small when the spectrum is clustered around 1.

For more general matrices A , residual polynomials found by Krylov methods again often become small especially if the spectrum of A is clustered around 1; but what can one do if the system matrix A does not at all have eigenvalues that cluster around 1, and an optimized residual polynomial obtained by a Krylov method does not become small on the spectrum of A , even after many iterations?

This is where preconditioning, which is a very active field of research, comes in: instead of solving the original system, one solves a preconditioned system with preconditioner M ,

$$A\mathbf{u} = \mathbf{f} \implies M^{-1}A\mathbf{u} = M^{-1}\mathbf{f},$$

where M should be close to A such that $M^{-1}A$ has a spectrum clustered at 1⁵⁵, but M^{-1} should be inexpensive, since linear solves with M have to be performed at each iteration of the Krylov method, that now works with the preconditioned Krylov space

$$\mathcal{K}_n(M^{-1}A, \mathbf{r}^0) := \mathbf{u}^0 + \text{span}\{\mathbf{r}^0, M^{-1}A\mathbf{r}^0, \dots, (M^{-1}A)^{n-1}\mathbf{r}^0\},$$

with the preconditioned initial residual $\mathbf{r}^0 := M^{-1}(\mathbf{f} - A\mathbf{u}^0)$. The main insight for preconditioning is that all stationary iterative methods we have seen in Section 6 (and we will see in the upcoming sections!) can be written using the matrix splitting $A = M - N$, i.e.

$$M\mathbf{u}^{n+1} = N\mathbf{u}^n + \mathbf{f} \iff \mathbf{u}^{n+1} = \mathbf{u}^n + M^{-1}(\mathbf{f} - A\mathbf{u}^n).$$

For example for Jacobi, $M = D$, the diagonal of A . Now the stationary iterative method converges well, if the spectral radius of the iteration matrix $M^{-1}N$ is small,

⁵⁵Note that there are also other preconditioners, see for example [126] where the preconditioned system should have a minimum polynomial of low degree, or [7] where the preconditioner is tuned to obtain two distinct tight clusters.

i.e. the eigenvalues of $M^{-1}N$ are all close to zero. This means however that the eigenvalues of the associated preconditioned matrix

$$M^{-1}A = M^{-1}(M - N) = I - M^{-1}N$$

are all close to 1, i.e. clustered at 1. We thus see that rapidly converging stationary iterative methods have spectra of the preconditioned system matrix $M^{-1}A$ clustered at 1 which is often very good also for Krylov methods. Now if the stationary iterative method already converges fast, why bother with a Krylov method? This is because the residual polynomial of the stationary iterative method is the simple polynomial $q_n(M^{-1}A) = (I - M^{-1}A)^n$, since

$$\begin{aligned} \mathbf{u} - \mathbf{u}^n &= \mathbf{u} - \mathbf{u}^{n-1} - M^{-1}(\mathbf{f} - A\mathbf{u}^{n-1}) \\ &= \mathbf{u} - \mathbf{u}^{n-1} - M^{-1}A(\mathbf{u} - \mathbf{u}^{n-1}) = (I - M^{-1}A)^n(\mathbf{u} - \mathbf{u}^0), \end{aligned}$$

whereas the Krylov method applied to the preconditioned system finds a residual polynomial $p_n(M^{-1}A)$ which is tuned to be small on the specific spectrum of $M^{-1}A$ and thus in general much smaller than $q_n(M^{-1}A)$ there. All stationary iterative methods should thus in practice never be used as such, but as preconditioners for Krylov methods: like Extrapolation methods, Krylov methods are accelerators for stationary methods; they find in general much better residual polynomials and thus give much faster convergence than the stationary iteration. This is why even the best stationary iterative methods for PDEs like domain decomposition and multigrid, which we will see in the next sections, are always used as preconditioners for Krylov methods in practice, even though their design and analysis is often simpler and more transparent as stationary iterations.

9. Domain Decomposition Methods. All the methods we have seen so far were based on the solution of a sparse linear system arising from the discretization of partial differential equations, but the most successful iterative methods tackle directly the underlying partial differential equation using their structure. Domain decomposition methods following the principle of divide and conquer are doing this, and they are in addition naturally parallel.

9.1. Schwarz Methods. Domain decomposition methods go back to 1869, when the alternating Schwarz method was invented [152]. Earlier, Bernhard Riemann (1826–1866) after his PhD thesis [144] had coined the expression *Dirichlet Principle* by claiming that he had heard it in Dirichlet’s lectures. Riemann also claimed in his PhD thesis that any Dirichlet Problem, e.g.

$$\Delta u = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega, \tag{9.1}$$

always had a solution, just by finding the function $u(x, y)$ which minimizes the Dirichlet Integral

$$\iint_{\Omega} \left(\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right) dx dy = \min!, \quad u = g \text{ on } \partial\Omega. \tag{9.2}$$

In the subsequent years, more and more mathematicians had doubts about Riemann’s claim, in particular Karl Weierstrass (1815–1897) with the counter-example [178] $\int_{-1}^{+1} (x \cdot y')^2 dx = \min!$, $y(-1) = a$, $y(+1) = b$, whose “solution” was discontinuous at $x = 0$ if $a \neq b$, see also [74, Sec. 2.2]. As a consequence, the Dirichlet Problem

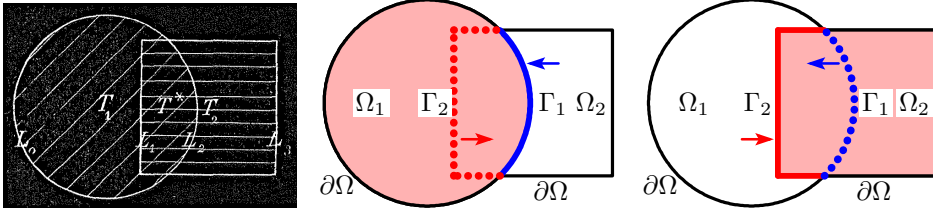


FIG. 9.1. Original example of Schwarz (left: from [152]).

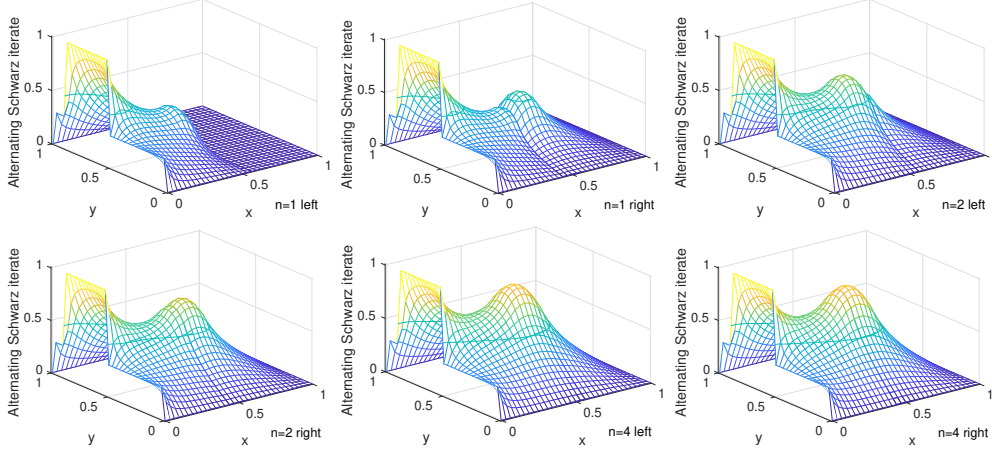


FIG. 9.2. Alternating Schwarz method for the Montreal problem ($\Omega_1 := (0, 0.5625) \times (0, 1)$ and $\Omega_2 := (0.4375, 1) \times (0, 1)$, overlap on the narrow strip between grid points number 14 and 18 out of 33 (including the boundary points) in the x -direction).

remained only solvable for rectangular domains and circular disks (using the methods introduced by Fourier [60]).

As an escape from this dilemma, Hermann Amandus Schwarz (1843-1921) invented his *alternating method* [152] by choosing as domain Ω the union of a disk Ω_1 and an overlapping rectangle Ω_2 (see Figure 9.1). He then computed alternately on the disk and on the rectangle the Dirichlet problem by always carrying the newly computed solution on the dotted boundary curves Γ_1 or Γ_2 over to the other side as new boundary condition for the next round:

$$\begin{aligned} \Delta u_1^n &= 0 & \text{in } \Omega_1, & \quad \Delta u_2^n = 0 & \text{in } \Omega_2, \\ u_1^n &= g & \text{on } \partial\Omega \cap \overline{\Omega}_1, & \quad u_2^n = g & \text{on } \partial\Omega \cap \overline{\Omega}_2, \\ u_1^n &= u_2^{n-1} & \text{on } \Gamma_1, & \quad u_2^n = u_1^n & \text{on } \Gamma_2. \end{aligned}$$

In [152] Schwarz proved convergence (the convergence rate depends on the size of the overlap) using the maximum principle. In Fig. 9.2 we show the iterates 1, 2 and 4 on the left and right subdomain of the alternating Schwarz method when applied to the Montreal problem, i.e. with a non-zero right hand side f from the heater⁵⁶. We see that convergence of this PDE based method is very fast, compared to the

⁵⁶In order to obtain these plots, it suffices to discretize the alternating Schwarz method, which would then correspond to a block Gauss-Seidel iteration applied to the overall discretized problem; for more information on the relation between continuous and discrete Schwarz methods, see [64].

point-wise iterative methods like Jacobi or Gauss-Seidel. However, at each iteration smaller subdomain problems have to be solved which are of the same type as the original problem. It is therefore of interest to divide the domain into more than two subdomains, so that the subdomain solves become cheap, and then also to solve the subdomain problems in parallel, as proposed by Lions in [115], i.e. to replace the transmission condition on Γ_2 by $u_2^n = u_1^{n-1}$. One can show that Schwarz methods converge independently of the mesh size, provided the overlap does not depend on the mesh size, see e.g. [63, Section 6.1]. An additive variant of the Schwarz alternating method due to Dryja and Widlund [44] led to substantial further development and the abstract Schwarz framework [165], but the additive Schwarz method is a preconditioner for the conjugate gradient method, and does not converge when used as a stationary iterative method [64]. Of course, convergence can be obtained by means of sufficient damping.

9.2. Dirichlet-Neumann Methods. In Schwarz methods, Dirichlet conditions are used to transmit information from one subdomain to the next, and overlap is needed for the methods to converge, since without overlap, the initial guess along the interface would never change. The idea of the Dirichlet-Neumann method is to also use Neumann transmission conditions, so that the method can converge without overlap. The Dirichlet-Neumann method was proposed by Bjørstad and Widlund in 1986 [9]. For the same model problem (9.1), including a possibly non-zero right hand side f , and two now non-overlapping subdomains Ω_1 and Ω_2 , $\Omega_1 \cap \Omega_2 = \emptyset$ with interface Γ , the Dirichlet-Neumann method starts with an initial guess λ^0 for the Dirichlet trace of the solution on the interface Γ , and then computes for iteration index $n = 1, 2, \dots$

$$\begin{aligned} \Delta u_1^n &= f & \text{in } \Omega_1, & \quad \Delta u_2^n &= f & \text{in } \Omega_2, \\ u_1^n &= g & \text{on } \partial\Omega \cap \overline{\Omega}_1, & \quad u_2^n &= g & \text{on } \partial\Omega \cap \overline{\Omega}_2, \\ u_1^n &= \lambda^{n-1} & \text{on } \Gamma, & \quad \partial_{n_2} u_2^n &= \partial_{n_2} u_1^n & \text{on } \Gamma, \end{aligned}$$

where ∂_{n_2} denotes the outward normal derivative for Ω_2 along the interface Γ . The new Dirichlet condition for subdomain Ω_1 is then computed on the interface Γ using a relaxation factor θ ,

$$\lambda^n := \theta u_2^n + (1 - \theta) \lambda^{n-1}.$$

As example we show the Dirichlet-Neumann method applied to the Montreal problem in Fig. 9.3 for a symmetric decomposition into two subdomains and relaxation parameter $\theta = 0.4$. We observe that the method converges also rapidly, comparable to the alternating Schwarz method in Fig. 9.2. It is interesting to know that for a perfectly symmetric problem and domain decomposition, the Dirichlet-Neumann method converges in two iterations, provided that the relaxation parameter is chosen to be $\theta = \frac{1}{2}$; the method becomes a direct solver! Now in a diffusion problem, when the mesh is refined, high frequency components of the error are damped rapidly by the diffusion operator and thus do not travel far. This means that high frequencies only 'see' the neighborhood of the interface in a symmetric way. In practice one thus should always choose $\theta = \frac{1}{2}$, and then obtains convergence of the Dirichlet-Neumann method independent of the mesh parameter, like in the Schwarz method when the overlap does not depend on the mesh parameter.

For three subdomains, it is also possible to choose relaxation parameters to obtain a direct solver, but for more than three subdomains this is only possible in certain special situations [30]. The Dirichlet-Neumann method however retains good convergence

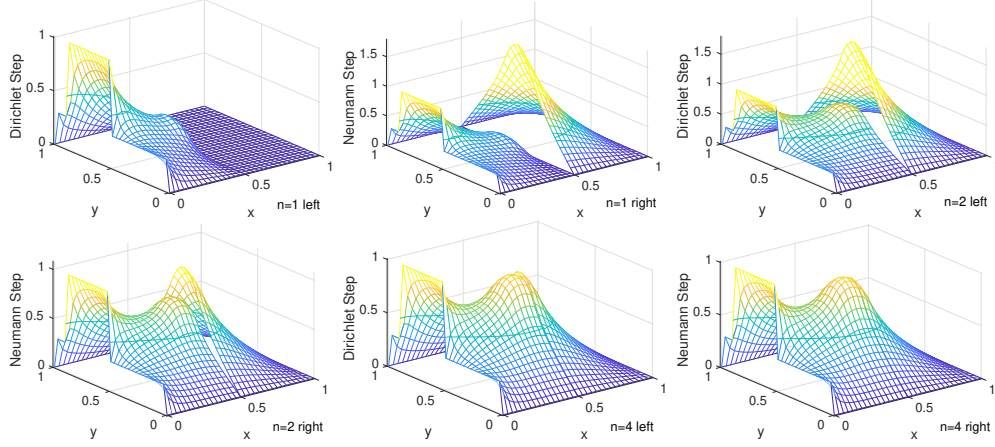


FIG. 9.3. *Dirichlet-Neumann method applied to our model problem for a symmetric domain decomposition and $\theta = 0.4$.*

properties (only logarithmic dependence on the mesh size) for general decompositions with cross points, but only when used as a preconditioner for a Krylov method, otherwise the method is violently divergent at cross points [33]. A disadvantage of the Dirichlet-Neumann method for general decompositions is that one has to choose for each interface which domain will use Dirichlet and which domain will use Neumann conditions.

9.3. Neumann-Neumann Methods. The Neumann-Neumann method was introduced by Bourgat, Glowinski, Le Tallec, and Vidrascu in 1989 [112]. In contrast to what one could expect from its name, it first solves Dirichlet problems on each subdomain, followed by Neumann problems on each subdomain, i.e. it removes the cumbersome choice between Dirichlet and Neumann in the Dirichlet-Neumann method. For the model problem (9.1) with a possibly non-zero right hand side f , and two non-overlapping subdomains Ω_1 and Ω_2 , $\Omega_1 \cap \Omega_2 = \emptyset$ with interface Γ , the Neumann-Neumann method starts with an initial guess λ^0 for the Dirichlet trace of the solution on the interface Γ , and then solves for iteration index $n = 1, 2, \dots$ first Dirichlet subdomain problems

$$\begin{aligned} \Delta u_1^n &= f & \text{in } \Omega_1, & \quad \Delta u_2^n = f & \text{in } \Omega_2, \\ u_1^n &= g & \text{on } \partial\Omega \cap \overline{\Omega}_1, & \quad u_2^n = g & \text{on } \partial\Omega \cap \overline{\Omega}_2, \\ u_1^n &= \lambda^{n-1} & \text{on } \Gamma, & \quad u_2^n = \lambda^{n-1} & \text{on } \Gamma, \end{aligned}$$

followed by Neumann correction problems, i.e problems with f and g equal to zero,

$$\begin{aligned} \Delta \psi_1^n &= 0 & \text{in } \Omega_1, & \quad \Delta \psi_2^n = 0 & \text{in } \Omega_2, \\ \psi_1^n &= 0 & \text{on } \partial\Omega \cap \overline{\Omega}_1, & \quad \psi_2^n = 0 & \text{on } \partial\Omega \cap \overline{\Omega}_2, \\ \partial_{n_1} \psi_1^n &= \frac{\partial_{n_1} u_1^n + \partial_{n_2} u_2^n}{2} & \text{on } \Gamma, & \quad \partial_{n_2} \psi_2^n = \frac{\partial_{n_1} u_1^n + \partial_{n_2} u_2^n}{2} & \text{on } \Gamma, \end{aligned}$$

where ∂_{n_j} denotes the outward normal derivative for Ω_j along the interface Γ . The new Dirichlet condition for subdomain Ω_1 is then computed on the interface Γ using a relaxation factor θ ,

$$\lambda^n := \lambda^{n-1} - \theta(\psi_1^n + \psi_2^n).$$

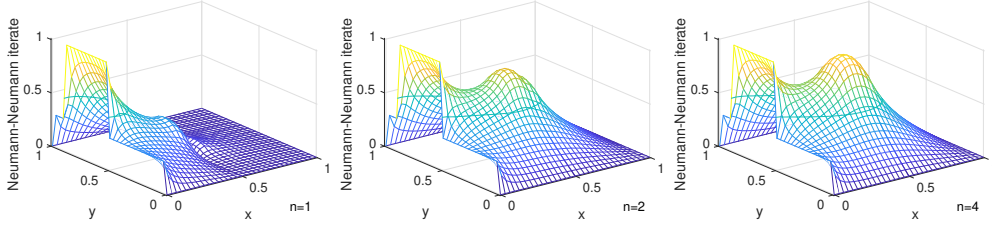


FIG. 9.4. *Neumann-Neumann method applied to our model problem for a symmetric domain decomposition and $\theta = 0.2$.*

As example we show the Neumann-Neumann method applied to the Montreal problem in Fig. 9.4 for a symmetric decomposition into two subdomains and $\theta = 0.2$. We observe that the method converges also rapidly, and the Neumann-Neumann method for a symmetric problem and domain decomposition becomes a direct solver for the special choice $\theta = \frac{1}{4}$. This is also the choice in practice even for non-symmetric domain decompositions, because like for Dirichlet-Neumann, the method then converges independently of the mesh size. Convergence in general is twice as fast as for Dirichlet-Neumann, but one also solves two subdomain problems on each subdomain in each iteration, so the two methods are comparable and the convergence mechanism is the same.

For more than two subdomains, one cannot obtain a direct solver any more, and the Neumann-Neumann method often becomes divergent, see e.g. [31]. It only retains its good convergence properties when used as a preconditioner for a Krylov method. Like the Dirichlet-Neumann method, the Neumann-Neumann method is violently divergent at cross-points, see e.g. [32].

9.4. FETI (Finite Element Tearing and Interconnect). FETI domain decomposition methods were introduced by Farhat and Roux in [51] based on the minimization interpretation (9.2) of the Dirichlet problem (9.1). The resulting methods are very much related to the Neumann-Neumann methods, only the order of the subdomain solves is interchanged: one first solves Neumann problems, and then Dirichlet problems as correction problems in each subdomain. Their convergence properties are therefore as for the Neumann-Neumann method.

9.5. Optimized Schwarz Methods. Research on optimized Schwarz methods was launched when Lions designed a Schwarz method that converges without overlap. To achieve this, he replaced in [116] the Dirichlet transmission conditions with Robin transmission conditions in the Schwarz algorithm,

$$\begin{aligned} \Delta u_1^n &= f & \text{in } \Omega_1, & & \Delta u_2^n &= f & \text{in } \Omega_2, \\ u_1^n &= g & \text{on } \partial\Omega \cap \bar{\Omega}_1, & & u_2^n &= g & \text{on } \partial\Omega \cap \bar{\Omega}_2, \\ (\partial_{n_1} + p_1)u_1^n &= (\partial_{n_1} + p_1)u_2^{n-1} & \text{on } \Gamma_1, & & (\partial_{n_2} + p_2)u_2^n &= (\partial_{n_2} + p_2)u_1^n & \text{on } \Gamma_2, \end{aligned}$$

where following Lions p_j can be constants, or functions along the interface, or even operators. The choice of the p_j greatly influences the convergence speed of the method, and optimizing their value in a class of operators leads to the so called optimized Schwarz methods, a term coined in [68]. These methods can be used with and without overlap, and there exist operators which lead to convergence in a finite number of steps also for general decompositions into many subdomains arranged into a sequence, in contrast to the Dirichlet-Neumann and Neumann-Neumann methods. We show

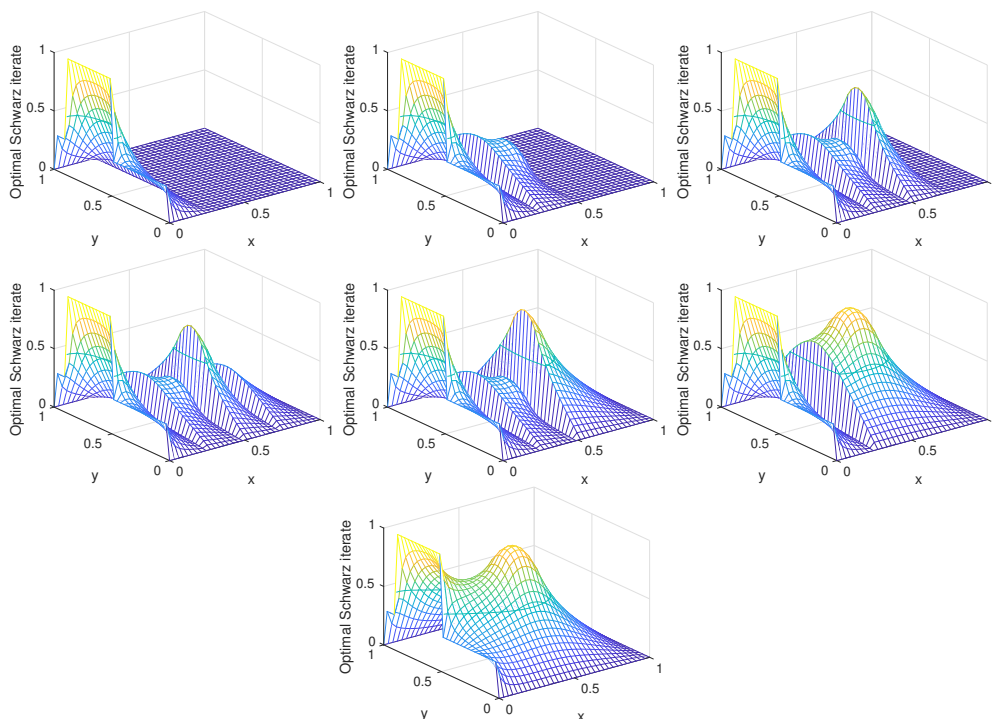


FIG. 9.5. *Sweeping optimal Schwarz method.*

in Fig. 9.5 the convergence of such an alternating optimal Schwarz method⁵⁷ which sweeps over four subdomains forward and backward for the Montreal problem, and converges after one such sweep. The operators used for p_j here are the Dirichlet to Neumann maps of the underlying partial differential equation solved, and at the algebraic level one can interpret the process as an exact block-LU factorization with a forward and backward solve of the linear system, see also the Analytic Incomplete LU preconditioner [72, 73]. With local approximations to these Dirichlet-Neumann operators one can obtain very powerful optimized Schwarz methods for very difficult equations, see for example [75, 76] and references therein.

10. Multigrid Methods.

“(...) serve mainly the direct contact and exchange of experience of research groups which work geographically far away (...)”

(The organizers of the Oberwolfach meeting in July 1976, R. Bulirsch, R.D. Grigorieff, J. Schröder)

In July 1976 W. Hackbusch presented in an Oberwolfach Meeting his new method on “iterative improvement through approximation and smoothing”. Olof Widlund (see Fig. 10.1) in the audience then remarked that a similar idea had independently been discovered by A. Brandt in Israel. Widlund had also told A. Brandt that another

⁵⁷The term optimal Schwarz method for Schwarz methods with transparent boundary conditions appeared already in [67] for time dependent problems, and this use of optimal means really faster is not possible, in contrast to the other common use of optimal meaning just scalable in the domain decomposition literature.

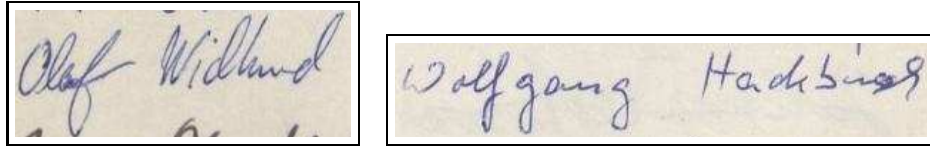
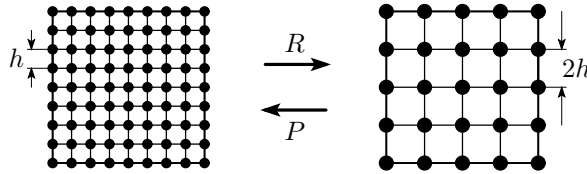


FIG. 10.1. Signatures of O. Widlund and W. Hackbusch in the “Gästebuch IV” of MFO Oberwolfach 1976

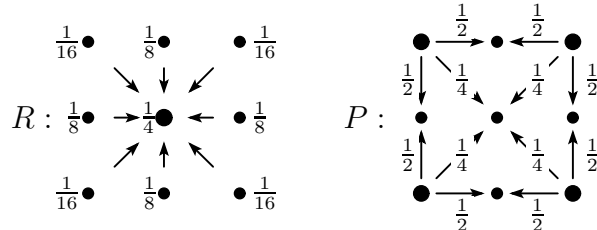
publication on this method was [52] in Russian by R.P. Fedorenko in 1961; see also Fedorenko’s note on the discovery [53]. Still another path to the method was the early work by Nicolaides [132], inspired by ideas from grids in composition methods. In our presentation below we follow the seminal contributions [13] and [83] and private discussions with Achi Brandt.

For the Laplace and Poisson problems, multigrid methods are so effective that they are difficult to beat by other methods, they require 23 floating point operations per grid point and V-cycle, independently of the problem size, and the number of iterations also does not depend on the problem size [82].

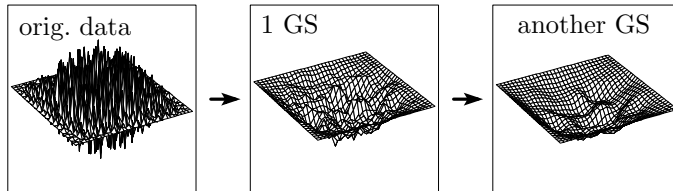
10.1. Two Grid Method. While domain decomposition methods operate on *several domains* for a PDE, multigrid methods operate on *several grids*, by taking advantage of the qualities of each of the grids for the error reduction process. The simplest version operates on two grids, typically with the second grid using twice the mesh size as the first:



In order to move data from one grid to the other, one uses a *restriction* map $R : u_h \mapsto u_{2h}$ as well as a *prolongation* map $P : u_{2h} \mapsto u_h$; in our example as follows:



On the fine grid, Gauss-Seidel converges very slowly for low frequencies (see (6.7)), but has excellent convergence properties for high frequencies (thus called a smoother):



On the coarse grid, however, the solution of the problem is much faster. Multigrid uses this to reduce the errors of the low frequencies and one obtains the following

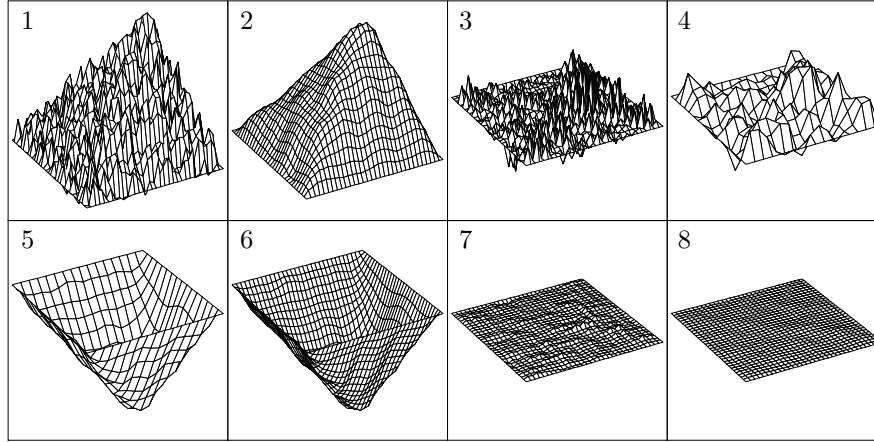


FIG. 10.2. *Convergence of the two-grid method.*

two-grid algorithm (see Figure 10.2, where we demonstrate convergence by solving $-\Delta u = 0$ with zero boundary conditions, starting from a very ugly initial guess u_h in Picture 1):

- 1 \mapsto 2 : Perform 2 Gauss-Seidel iterations on the fine grid;
- 2 \mapsto 3 : Compute the residual $d_h = -Au_h$;
- 3 \mapsto 4 : Restrict this residual d_h to the coarse grid $d_{2h} = Rd_h$;
- 4 \mapsto 5 : Solve on the coarse grid $A_c v_{2h} = d_{2h}$;
- 5 \mapsto 6 : Prolongate v_{2h} to the fine grid as $v_n = Pv_{2h}$;
- 6 \mapsto 7 : Add $u_h + v_h \mapsto u_h$;
- 7 \mapsto 8 : Perform 1 Gauss-Seidel iteration.

We see that all error components are very effectively removed by one such two grid cycle.

10.2. Multi Grid Method. For a general linear system of equations

$$Au = f$$

which represents a discretized PDE, multigrid starts with an initial guess u^0 on the fine grid, and then computes

$$\begin{aligned} u^n &= S(A, u^n, f, \nu_1); \\ u^n &= u^n + PA_c^{-1}R(f - Au^n); \\ u^{n+1} &= S(A, u^n, f, \nu_2); \end{aligned} \tag{10.1}$$

Here $S(A, u^n, f, \nu)$ denotes ν iterations of a smoother, for example damped Jacobi or Gauss-Seidel, P is the prolongation operator we have seen already, often performed by interpolation of missing values from the coarse to the fine grid, R is a restriction operator, which can either just select the necessary values from the fine grid to be used on the coarse grid (called injection). One often also uses $R = CP^T$, where C is an appropriate scaling constant (called full weighting). The coarse problem represented by the matrix A_c can either be a coarse discretization of the underlying problem on a coarse grid, or a coarse matrix obtained from the fine one using the prolongation and restriction operators, $A_c = RAP$ (called Galerkin approach).

The key step in the algorithm (10.1) is not to compute the coarse solution represented symbolically by A_c^{-1} , but to apply the algorithm recursively, using a further

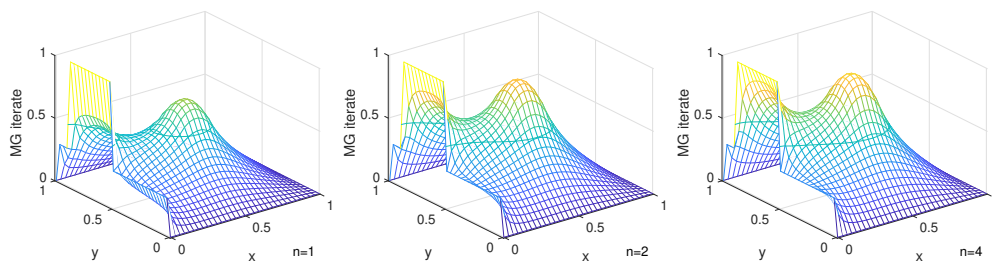


FIG. 10.3. Multigrid for the Montreal problem using V-cycles.

even coarser grid. Doing so leads to the so called V-cycle multigrid method, because one goes from the finest grid down to the coarsest and then back, like a V shape. It is however also possible to visit the different grids differently, which leads to the popular W-cycle, or also to the full multigrid cycle, which starts on the coarse grid and then just does one sequence of inverted V-cycles to visit finer and finer grids.

An example of a multigrid V-cycle applied to the Montreal problem using a damped Jacobi smoother for iteration 1, 2 and 4 is shown in Figure 10.3. We see that convergence is extremely fast, the treatment of different frequency components of the error on appropriate grids is highly effective.

It is important however to note that such excellent convergence depends strongly on the properties of the PDE from which the linear system is obtained. For more general diffusive, Laplace like problems, the performance is always outstanding. For non-symmetric problems one needs to use specially adapted multigrid components, and for time harmonic wave propagation problems like the Helmholtz equation, it is hardly possible to make the multigrid idea an effective, simple solver. This applies however to most of the other iterative methods we have described as well; time harmonic wave propagation problems are simply hard to solve by iteration [48], and the best current methods are based on domain decomposition, see [75, 76] and references therein.

11. Current research and outlook.

“For guidance to the future we should study not Gaussian elimination and its beguiling stability properties, but the diabolically fast conjugate gradient iteration (...) or the convergence in $O(1)$ iteration achieved by multigrid methods for many kinds of problems – or even Borwein and Borwein’s magical AGM iteration for determining 1,000,000 digits of π in an eyeblink. *That is the heart of numerical analysis.*” (Nick Trefethen, *The Definition of Numerical Analysis*, Cornell University, 1992).

The field of iterative methods is more active than ever in numerical analysis, and it is at its core for research and discovery in science and engineering. State of the art books on Newton’s method and its variants like [134, 100, 42] are available, and also for iterative methods for large scale linear systems [169, 91, 149, 114, 36]. There are also works treating both fields [99], and their combination leads to the very powerful Newton-Krylov solvers, which are methods based on inner and outer iterations⁵⁸: the underlying non-linear problem is treated with a Newton method, the outer iteration, and in each iteration the arising linear system in the Newton step is then treated by a second, linear, inner iteration. Furthermore, if the inner iterations need to be

⁵⁸Inner and outer iterations were a key research interest of the late Gene Golub.

preconditioned for performance, methods like the Newton-Krylov-Schwarz methods [19] arise, which are currently among the most powerful parallel solvers for nonlinear partial differential equations. More recently, the preconditioning idea we explained for large scale linear systems has also been introduced directly for non-linear iterations [20]. This is best understood in analogy with the linear case [65]: like one can accelerate convergence of a stationary iterative method for a linear problem using a Krylov method, see Subsection 8.4, one can also accelerate the convergence of a non-linear fixed point iteration using Newton’s method [43]. This way one could also accelerate the multigrid full approximation scheme, which is a nonlinear multigrid fixed point iteration, using Newton’s method. Newton’s method, however, does not have the global acceleration properties that Krylov methods have in the linear case, and non-linear preconditioning is currently not yet well understood: it can even lead to chaotic behavior, while the underlying non-linear fixed point method converges very well [119, 120]. Globalization strategies for Newton’s method, coming from optimization, could then alleviate the situation. A further, not yet well enough explored research avenue is when the extrapolation methods we have seen for linear problems are applied to non-linear vector sequences [121]. In our opinion, as Volker Mehrmann once said for numerical analysis, the golden years of iterative methods are still ahead.

Acknowledgments: The authors are grateful for the insightful remarks of the referees that helped to improve the manuscript.

REFERENCES

- [1] A. Aaboe. Al-Kashi’s iteration method for the determination of $\sin 1^\circ$. *Scripta mathematica*, 20:24–29, 1954.
Cited on page 10.
- [2] A. Aaboe. *Episodes from the Early History of Mathematics*, volume 13. New Mathematical Library 13. The Mathematical Association of America, 1998.
Cited on page 9.
- [3] J. C. Adams. On Newton’s solution of Kepler’s problem. *Monthly Notices of the Royal Astronomical Society*, 43(2):43–49, 1882.
Cited on page 21.
- [4] A. C. Aitken. On Bernoulli’s numerical solution of algebraic equations. *Proc. Roy. Soc. of Edinburgh*, Ser. A(46):289–305, 1926.
Cited on page 50.
- [5] S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta mathematicae*, 3(1):133–181, 1922.
Cited on page 38.
- [6] S. Banach. *Théorie des opérations linéaires*. Warszawa, 1932.
Cited on page 38.
- [7] M. Benzi, M. J. Gander, and G. H. Golub. Optimization of the Hermitian and skew-Hermitian splitting iteration for saddle-point problems. *BIT Numerical Mathematics*, 43(5):881–900, 2003.
Cited on page 69.
- [8] N. Bićanić and K. Johnson. Who was ‘-Raphson’? *International Journal for Numerical Methods in Engineering*, 14(1):148–152, 1979.
Cited on page 22.
- [9] P. E. Bjørstad and O. B. Widlund. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM Journal on Numerical Analysis*, 23(6):1097–1120, 1986.
Cited on page 72.
- [10] A. Blair, N. Metropolis, A. Taub, and M. Tsingou. A study of a numerical solution to a two-dimensional hydrodynamical problem. Technical report, Los Alamos Scientific Laboratory

- of the University of California, 1958.
[Cited on pages 53 and 55.](#)
- [11] R. Bombelli. *L'Algebra parte maggiore dell'arimetica divisa in tre libri*. Rossi, Bologna, 1572.
[Cited on page 8.](#)
 - [12] C. L. Bouton. Discussion of a method for finding numerical square roots. *Annals of Mathematics*, 10(2):167–172, 1909.
[Cited on page 6.](#)
 - [13] A. Brandt. Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems. In *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics 1972*, pages 82–89. Springer, 1973.
[Cited on page 76.](#)
 - [14] C. Brezinski. *Accélération de la convergence en analyse numérique*, volume 584. Springer Verlag, 1977.
[Cited on pages 50, 58 and 59.](#)
 - [15] C. Brezinski. *History of Continued Fractions and Padé Approximants*. Springer Verlag, 1991.
[Cited on page 9.](#)
 - [16] C. Brezinski, G. Meurant, and M. Redivo-Zaglia. *A Journey through the History of Numerical Linear Algebra*. SIAM, 2022.
[Cited on page 41.](#)
 - [17] E. M. Bruins (ed.). Codex Constantinopolitanus. Palatii Veteris no. 1. 3 volumes, E. J. Brill, Leiden, 1964.
[Cited on page 5.](#)
 - [18] S. Cabay and L. Jackson. A polynomial extrapolation method for finding limits and antilimits of vector sequences. *SIAM Journal on Numerical Analysis*, 13(5):734–752, 1976.
[Cited on pages 50, 56 and 57.](#)
 - [19] X.-C. Cai, W. D. Gropp, D. E. Keyes, and M. D. Tidriri. Newton-Krylov-Schwarz methods in cfd. In *Numerical methods for the Navier-Stokes equations: Proceedings of the International Workshop Held at Heidelberg, October 25–28, 1993*, pages 17–30. Springer, 1994.
[Cited on page 79.](#)
 - [20] X.-C. Cai and D. E. Keyes. Nonlinearly preconditioned inexact newton algorithms. *SIAM Journal on Scientific Computing*, 24(1):183–200, 2002.
[Cited on page 79.](#)
 - [21] F. Cajori. Fourier's improvement of the Newton-Raphson method of approximation anticipated by Mourraille. *Bibliotheca mathematica*, 11(3):132–137, 1910.
[Cited on page 26.](#)
 - [22] A. L. Cauchy. *Cours d'analyse de l'École Royale Polytechnique*. Imprimerie Royale. Chez Debure frères, Libraires du Roi et de la Bibliothèque du Roi, 1821.
[Cited on page 37.](#)
 - [23] A. L. Cauchy. Note sur la détermination approximative des racines d'une équation algébrique ou transcendante. In *Leçons sur le Calcul Différentiel*, pages 259–289. Chez de Bure frères, Paris, 1829. Reprinted in *Œuvres* série 2, vol. IV, pp. 573–609.
[Cited on pages 35, 36 and 37.](#)
 - [24] A. L. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus Hebd. Séances Acad. Sci.*, 25:536–538, 1847.
[Cited on page 60.](#)
 - [25] A. Cayley. Application of the Newton-Fourier method to an imaginary root of an equation. *Quart. J. Pure Appl. Math*, 16:179–185, 1879.
[Cited on pages 27, 28, 29 and 30.](#)
 - [26] A. Cayley. The Newton-Fourier imaginary problem. *American Journal of Mathematics*, 2(1):97, 1879.
[Cited on pages 27, 28 and 31.](#)
 - [27] A. Cayley. On the Newton-Fourier imaginary problem. *Proceedings of the Cambridge Philosophical Society*, 3:231–232, 1880.
[Cited on pages 27, 28 and 31.](#)
 - [28] A. Cayley. Sur les racines d'une équation algébrique. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 110:215–218, 1890.
[Cited on pages 27, 28 and 31.](#)
 - [29] M. Channabasappa. On the square root formula in the Bakhshālī manuscript. *Indian Journal of History of Science Calcutta*, 11(2):112–124, 1976.
[Cited on page 6.](#)
 - [30] F. Chaouqui, M. J. Gander, and K. Santugini-Repique. On nilpotent subdomain iterations. In *Domain Decomposition Methods in Science and Engineering XXIII*, pages 125–133.

- Springer, 2017.
[Cited on page 72.](#)
- [31] F. Chaouqui, M. J. Gander, and K. Santugini-Repique. A continuous analysis of Neumann–Neumann methods: Scalability and new coarse spaces. *SIAM Journal on Scientific Computing*, 42(6):A3785–A3811, 2020.
[Cited on page 74.](#)
 - [32] F. Chaouqui, M. J. Gander, and K. Santugini-Repique. A local coarse space correction leading to a well-posed continuous Neumann–Neumann method in the presence of cross points. In *Domain Decomposition Methods in Science and Engineering XXV*, pages 83–91. Springer, 2020.
[Cited on page 74.](#)
 - [33] B. Chaudet-Dumas and M. J. Gander. Cross-points in the Dirichlet-Neumann method I: well-posedness and convergence issues. *Numerical Algorithms*, pages 1–34, 2022.
[Cited on page 73.](#)
 - [34] P. L. Chebyshev. *Théorie des mécanismes connus sous le nom de parallélogrammes*. Imprimerie de l’Académie impériale des sciences, 1853.
[Cited on page 53.](#)
 - [35] G. Ciaramella and M. J. Gander. *Iterative Methods and Preconditioners for Systems of Linear Equations*. SIAM, 2022.
[Cited on pages 50, 62, 67, 68 and 69.](#)
 - [36] G. Ciaramella and M. J. Gander. *Iterative methods and preconditioners for systems of linear equations*. SIAM, 2022.
[Cited on page 78.](#)
 - [37] P. G. Ciarlet and M. Christine. On the Newton-Kantorovich Theorem. *Analysis and Applications*, 10:249–269, 2012.
[Cited on page 40.](#)
 - [38] J. H. Curry, L. Garnett, and D. Sullivan. On the iteration of a rational function: computer experiments with Newton’s method. *Comm. in Math. Physics*, 91(2):267–277, 1983.
[Cited on page 33.](#)
 - [39] G. Dahlquist. Fehlerabschätzungen bei Differenzenmethoden zur numerischen Integration gewöhnlicher Differentialgleichungen. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 31(8-9):239–240, 1951.
[Cited on page 52.](#)
 - [40] J.-L. de Lagrange. Recherches sur la nature et la propagation du son. *Miscellanea Taurinensia t.I, Œuvres t.I*, 1:39–148, 1759.
[Cited on page 43.](#)
 - [41] T. de Smyrne. *Exposition des connaissances mathématiques utiles pour la lecture de Platon*. Hachette, Paris, 1892. Traduite pour la première fois du grec en français par J. Dupuis.
[Cited on page 11.](#)
 - [42] P. Deuffhard. *Newton methods for nonlinear problems. Affine invariance and adaptive algorithms*, volume 35. Springer-Verlag, 2004.
[Cited on pages 40 and 78.](#)
 - [43] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. Nonlinear preconditioning: How to use a nonlinear Schwarz method to precondition Newton’s method. *SIAM Journal on Scientific Computing*, 38(6):A3357–A3380, 2016.
[Cited on page 79.](#)
 - [44] M. Dryja and O. Widlund. An additive variant of the Schwarz alternating method for the case of many subregions. Technical Report 339, Department of Computer Science, Courant Institute, 1987. Also Ultracomputer Note 131.
[Cited on page 72.](#)
 - [45] J.-P. Eckmann. Savez-vous résoudre $z^3 - 1$? *La Recherche*, 14:260–262, 1983.
[Cited on pages 33 and 34.](#)
 - [46] R. P. Eddy. Extrapolating to the limit of a vector sequence. In *Information linkage between applied mathematics and industry*, pages 387–396. Elsevier, 1979.
[Cited on page 58.](#)
 - [47] R. P. Eddy. Acceleration of convergence of a vector sequence by reduced rank extrapolation. Technical report, David W. Taylor Naval Ship Research and Development Center Bethesda MD, 1981.
[Cited on page 50.](#)
 - [48] O. G. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In I. Graham, T. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical analysis of multiscale problems*, pages 325–363. Springer, 2012.

- Cited on page 78.
- [49] L. Euler. De formulis exponentialibus replicatis. In *Acta Academiae Scientiarum Imperialis Petropolitanae*, pages 38–60. Academiae Scientiarum, 1777 (1778). Also in *Opera Omnia* 15(1), pp. 268–297.
- Cited on pages 17 and 18.
- [50] H.-r. Fang and Y. Saad. Two classes of multiseccant methods for nonlinear acceleration. *Numerical linear algebra with applications*, 16(3):197–221, 2009.
- Cited on page 50.
- [51] C. Farhat and F.-X. Roux. A method of Finite Element Tearing and Interconnecting and its parallel solution algorithm. *Int. J. Numer. Meth. Engrg.*, 32(6):1205–1227, 1991.
- Cited on page 74.
- [52] R. P. Fedorenko. A relaxation method for solving elliptic difference equations. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 1(5):922–927, 1961.
- Cited on page 76.
- [53] R. P. Fedorenko. On the history of the multigrid method creation. <https://team.kiam.ru/botchev/fedorenko/>, 2001. Translated by M.A. Botchev.
- Cited on page 76.
- [54] I. Fenyő. Über die lösung der im Banachschen Raume definierten nichtlinearen Gleichungen. *Acta Mathematica Academiae Scientiarum Hungaricae*, 5(1):85–93, 1954. German version of “Banach-terekben értelmzett nemlineáris egyenletekről”, Magyar Tud. Akad. III. Osztályának Közleményei, 3, 1953, pp. 71–83.
- Cited on page 39.
- [55] M. Folkerts, D. Launert, and A. Thom. Jost Bürgi’s method for calculating sines. *Historia mathematica*, 43(2):133–147, 2016.
- Cited on page 13.
- [56] G. Forsythe, M. Hestenes, and J. Rosser. Iterative methods for solving linear equations. In *Bulletin of the American Mathematical Society*, volume 57(6), pages 480–480. Amer. Mathematical Soc. 201 Charles St, Providence, RI 02940-2213, 1951.
- Cited on page 59.
- [57] G. E. Forsythe. Notes. *Mathematical Tables and Other Aids to Computation*, 5(36):255–258, 1951.
- Cited on pages 41 and 47.
- [58] J. Fourier. Question d’analyse algébrique. *Bulletin des sciences par la Société Philomatique de Paris*, pages 61–67, 1818. Also in *Œuvres* II, pp. 243–253.
- Cited on pages 35 and 36.
- [59] J. Fourier. Sur l’usage du théorème de Descartes dans la recherche de la limite des racines. *Bulletin des sciences par la Société Philomatique de Paris*, pages 156–165, 181–187, 1820. Also in *Œuvres* II, pp. 291–309.
- Cited on page 35.
- [60] J. Fourier. *Théorie analytique de la chaleur*. Firmin Didot, Paris, 1822.
- Cited on pages 16, 17 and 71.
- [61] J. Fourier. *Analyse des équations déterminées*. Firmin Didot, Paris, 1831.
- Cited on pages 35 and 36.
- [62] R. W. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60(1):315–339, 1991.
- Cited on page 68.
- [63] M. J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.
- Cited on page 72.
- [64] M. J. Gander. Schwarz methods over the course of time. *Electronic transactions on numerical analysis*, 31:228–255, 2008.
- Cited on pages 71 and 72.
- [65] M. J. Gander. On the origins of linear and non-linear preconditioning. In *Domain decomposition methods in science and engineering XXIII*, pages 153–161. Springer, 2017.
- Cited on page 79.
- [66] M. J. Gander and L. Halpern. Piece-wise constant, linear and oscillatory: a historical introduction to spectral coarse spaces with focus on Schwarz methods. In *Domain Decomposition Methods in Science and Engineering XXVII*. Springer, 2023.
- Cited on page 63.
- [67] M. J. Gander, L. Halpern, and F. Nataf. Optimal convergence for overlapping and non-overlapping Schwarz waveform relaxation. In *11th international conference on domain decomposition methods*, pages 27–36, 1999.

- Cited on page 75.
- [68] M. J. Gander, L. Halpern, and F. Nataf. Optimized Schwarz methods. In *12th international conference on domain decomposition methods*, pages 15–27, 2000.
- Cited on page 74.
- [69] M. J. Gander and F. Kwok. *Numerical analysis of partial differential equations using Maple and Matlab*. SIAM, 2018.
- Cited on page 42.
- [70] M. J. Gander and A. Loneland. SHEM: An optimal coarse space for RAS and its multiscale approximation. In *Domain Decomposition Methods in Science and Engineering XXIII*, pages 313–321. Springer, 2017.
- Cited on page 63.
- [71] M. J. Gander, A. Loneland, and T. Rahman. Analysis of a new harmonically enriched multi-scale coarse space for domain decomposition methods. *arXiv preprint arXiv:1512.05285*, 2015.
- Cited on page 63.
- [72] M. J. Gander and F. Nataf. AILU: a preconditioner based on the analytic factorization of the elliptic operator. *Numer. Linear Algebra Appl.*, 7:505–526, 2000.
- Cited on page 75.
- [73] M. J. Gander and F. Nataf. An incomplete LU preconditioner for problems in acoustics. *J. Comput. Acoust.*, 13(3):455–476, 2005.
- Cited on page 75.
- [74] M. J. Gander and G. Wanner. From Euler, Ritz, and Galerkin to modern computing. *SIAM Review*, 54(4):627–666, 2012.
- Cited on pages 60 and 70.
- [75] M. J. Gander and H. Zhang. A class of iterative solvers for the Helmholtz equation: Factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods. *SIAM Review*, 61(1):3–76, 2019.
- Cited on pages 75 and 78.
- [76] M. J. Gander and H. Zhang. Schwarz methods by domain truncation. *Acta Numerica*, 31:1–134, 2022.
- Cited on pages 75 and 78.
- [77] W. Gander, M. J. Gander, and F. Kwok. *Scientific computing-An introduction using Maple and MATLAB*, volume 11. Springer Science & Business, 2014.
- Cited on page 59.
- [78] C. F. Gauss. Letter to Gerling, December 26, 1823. In *Werke*, volume 9, pages 278–281. Göttingen, 1903.
- Cited on pages 41, 43 and 47.
- [79] S. Glushkov. On approximation method of Leonardo Fibonacci. *Historia Mathematica*, 3:291–296, 1976.
- Cited on page 8.
- [80] G. H. Golub. *The use of Chebyshev matrix polynomials in the iterative solution of linear equations compared to the method of successive overrelaxation*. PhD thesis, University of Illinois at Urbana-Champaign, 1959.
- Cited on pages 50, 55 and 56.
- [81] G. H. Golub and R. S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numerische Mathematik*, 3(1):157–168, 1961.
- Cited on page 55.
- [82] M. Griebel. *Zur Lösung von Finite-Differenzen- und Finite-Element-Gleichungen mittels der Hierarchischen Transformations-Mehrgitter-Methode*. PhD thesis, Institut für Informatik, TU München, 1990. SFB Bericht 342/4/90 A.
- Cited on page 76.
- [83] W. Hackbusch. Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertprobleme. Technical Report 76-12, Institute for Applied Mathematics, University of Cologne, West Germany, Cologne, 1976.
- Cited on page 76.
- [84] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations I: Nonstiff problems*. Springer-Verlag Berlin Heidelberg, 1987. Second edition 1993.
- Cited on page 38.
- [85] Sir T. L. Heath. *The thirteen books of Euclid’s Elements*. University Press, Cambridge, second edition, 3 volumes edition, 1926. Translated from the text of Heiberg with introduction and commentary.

- [Cited on page 11.](#)
- [86] P. Henry. Un mémoire inédit de Lagrange sur le développement successif des courbes. *L'Enseignement Mathématique*, 67(1):95–122, 2021.
[Cited on page 13.](#)
 - [87] P. Henry and G. Wanner. Johann Bernoulli and the cycloid: A theorem for posterity. *Elemente der Mathematik*, 72(4):137–163, 2017.
[Cited on page 53.](#)
 - [88] P. Henry and G. Wanner. Zigzags with Bügi, Bernoulli, Euler and the Seidel–Entringer–Arnol’d triangle. *Elemente der Mathematik*, 74(4):141–168, 2019.
[Cited on page 13.](#)
 - [89] M. R. Hestenes. Conjugacy and gradients. In S. Nash, editor, *A history of scientific computing*, pages 167–179. ACM Press, 1990. New York.
[Cited on pages 63 and 64.](#)
 - [90] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.
[Cited on page 63.](#)
 - [91] A. S. Householder. *The theory of matrices in numerical analysis*. Blaisdell Publishing Co., 1964.
[Cited on pages 12 and 78.](#)
 - [92] J. H. Hubbard and B. B. Hubbard. *Vector Calculus, Linear Algebra, and Differential Forms*. Matrix Editions, 4th edition edition, 2009.
[Cited on page 31.](#)
 - [93] C. G. J. Jacobi. Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden lineären Gleichungen. *Astronomische Nachrichten*, 22(20):297–306, 1845.
[Cited on pages 43, 44 and 45.](#)
 - [94] C. G. J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen*. *Journal für die reine und angewandte Mathematik*, 30:51–94, 1846.
[Cited on page 44.](#)
 - [95] G. Julia. Mémoire sur l’itération des fonctions rationnelles. *Journal de mathématiques pures et appliquées*, 1(8):47–245, 1918.
[Cited on pages 31 and 33.](#)
 - [96] L. V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.
[Cited on pages 38, 39 and 41.](#)
 - [97] L. V. Kantorovich and G. P. Akilov. *Functional analysis in normed spaces*. Fizmatgis, Moscow, 1959. Second edition (with shorter title), translated by H.L. Silcock, Pergamon Press, 1982.
[Cited on page 40.](#)
 - [98] G. R. Kaye. The bakhshālī manuscript, a study in mediæval mathematics. *Archæological Survey of India XLIII*, 1927.
[Cited on page 7.](#)
 - [99] C. T. Kelley. *Iterative methods for linear and nonlinear equations*. SIAM, 1995.
[Cited on page 78.](#)
 - [100] C. T. Kelley. *Solving nonlinear equations with Newton’s method*. SIAM, 2003.
[Cited on page 78.](#)
 - [101] E. S. Kennedy. Parallax theory in Islamic astronomy. *Isis*, 47:33–53, 1956.
[Cited on page 13.](#)
 - [102] E. S. Kennedy. An early method of successive approximations. *Centaurus*, 13(3):248–250, 1969.
[Cited on pages 13 and 14.](#)
 - [103] E. S. Kennedy and W. R. Transue. A medieval iterative algorism. *The American Mathematical Monthly*, 63(2):80–83, 1956.
[Cited on pages 13 and 14.](#)
 - [104] J. Kepler. *Astronomia nova AITIOΛOΓHTOΣ seu physica coelestis, tradita commentariis de motibus stellæ Martis ex observationibus G.V. Tyconis Brahe*, 1609.
[Cited on page 15.](#)
 - [105] J. Kepler. *Epitome AstronomiæCopernicanæUsitatâ Formâ Quæstionum Et Responsionum Conscripta, Inq: VII. Libros Digesta Quorum TRES Hi Priores Sunt De Doctrina Sphærica*. Johannus Plancus, Lentijs ad Danubium, 1618.
[Cited on page 16.](#)
 - [106] D. E. Knuth. Ancient babylonian algorithms. *Communications of the Association for Com-*

- puting Machinery, 15(7):671–677, 1972.
[Cited on page 1.](#)
- [107] N. Kollerstrom. Thomas Simpson and ‘Newton’s method of approximation’: an enduring myth. *The British journal for the history of science*, 25(3):347–354, 1992.
[Cited on page 23.](#)
 - [108] A. N. Krylov. On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined. *Izvestija AN SSSR (News of Academy of Sciences of the USSR), Otdel. mat. i estest. nauk*, 7(4):491–539, 1931.
[Cited on page 59.](#)
 - [109] J. L. Lagrange. *Théorie des fonctions analytiques: contenant les principes du calcul différentiel, dégagés de toute considération d’infiniment petits, d’évanouissans, de limites et de fluxions, et réduits à l’analyse algébrique des quantités finies*. A Paris, de l’imprimerie de la république, Prairial an V (1797); deuxième édition: Gauthier-Villars, 1813.
[Cited on page 35.](#)
 - [110] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282, 1950.
[Cited on page 59.](#)
 - [111] C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.
[Cited on page 59.](#)
 - [112] P. Le Tallec, J. Bourgat, R. Glowinski, and M. Vidrascu. Variational formulation and conjugate gradient algorithm for trace operator in domain decomposition methods. In *SIAM Proceedings of the Second international Symposium on Domain Decomposition Methods for Partial Differential Equations, Los Angeles*, pages 3–16. SIAM, 1989.
[Cited on page 73.](#)
 - [113] T. Lei. The Mandelbrot set, theme and variations. *London Math. Soc. Lecture Note Ser.* 274, 2000.
[Cited on page 31.](#)
 - [114] J. Liesen and Z. Strakos. *Krylov subspace methods: principles and analysis*. Numerical Mathematics and Scie, 2013.
[Cited on page 78.](#)
 - [115] P.-L. Lions. On the Schwarz alternating method. I. In *First international symposium on domain decomposition methods for partial differential equations*, pages 1–42. Paris, France, 1988.
[Cited on page 72.](#)
 - [116] P.-L. Lions. On the Schwarz alternating method III: a variant for nonoverlapping subdomains. In *Third international symposium on domain decomposition methods for partial differential equations*, volume 6, pages 202–223. SIAM Philadelphia, 1990.
[Cited on page 74.](#)
 - [117] B. Mandelbrot. On the quadratic mapping $z \rightarrow z^2 - \mu$ for complex μ and z : the fractal structure of its \mathcal{M} set, and scaling. *Physica D: Nonlinear Phenomena*, 7(1–3):224–239, 1983.
[Cited on page 26.](#)
 - [118] B. B. Mandelbrot, C. J. Evertsz, and M. C. Gutzwiller. *Fractals and Chaos: the Mandelbrot Set and Beyond*, volume 3. Springer, 2004.
[Cited on page 35.](#)
 - [119] C. McCoid and M. J. Gander. Cyclic and chaotic examples in Schwarz-preconditioned Newton methods. In *International Conference on Domain Decomposition Methods*, pages 367–374. Springer, 2022.
[Cited on page 79.](#)
 - [120] C. McCoid and M. J. Gander. Cycles in Newton–Raphson preconditioned by Schwarz (ASPIN and its cousins). In *Domain Decomposition Methods in Science and Engineering XXVI*, pages 265–272. Springer, 2023.
[Cited on page 79.](#)
 - [121] C. McCoid and M. J. Gander. Extrapolation methods as nonlinear Krylov subspace methods. *Linear Algebra and its Applications*, 2023.
[Cited on pages 50 and 79.](#)
 - [122] M. Mešina. Convergence acceleration for the iterative solution of the equations $x = ax + f$. *Computer Methods in Applied Mechanics and Engineering*, 10(2):165–173, 1977.

- Cited on page 58.
- [123] R. Mises and H. Pollaczek-Geiringer. Praktische Verfahren der Gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77, 1929.
- Cited on page 12.
- [124] C. Montelle and K. Ramasubramanian. Determining the sine of one degree in the Sarvasiddhāntarāja of Nityānanda. *SCIAMVS: sources and commentaries in exact sciences*, 19:1–52, 2018.
- Cited on page 10.
- [125] J. Murraille. *Traité de la résolution des équations en général. Première Partie. Des équations invariables*. Jean Mossy, Londres & Marseille, 1768.
- Cited on pages 26 and 27.
- [126] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21(6):1969–1972, 2000.
- Cited on page 69.
- [127] F. Nataf, H. Xiang, V. Dolean, and N. Spillane. A coarse space construction based on local Dirichlet-to-Neumann maps. *SIAM Journal on Scientific Computing*, 33(4):1623–1642, 2011.
- Cited on page 63.
- [128] O. Neugebauer and A. Sachs. Mathematical cuneiform texts. *American Oriental Series*, 29, 1945.
- Cited on pages 2 and 3.
- [129] I. Newton. *Philosophiæ Naturalis Principia Mathematica*. Jussu Societatis Regiæ ac Typis Josephi Streater, Londini, 1687.
- Cited on page 21.
- [130] I. Newton. *La méthode des fluxions, et des suites infinies*. Trans Buffon, Debure, Paris, 1740.
- Cited on page 19.
- [131] I. Newton. *The mathematical principles of natural philosophy*. Daniel Adee, New York, 1846. translated into English by A. Motte.
- Cited on page 21.
- [132] R. A. Nicolaides. On multiple grid and related techniques for solving discrete elliptic systems. *Journal of Computational Physics*, 19(4):418–431, 1975.
- Cited on page 76.
- [133] J. M. Ortega. The Newton-Kantorovich theorem. *Amer. Math. Monthly*, 75:658–660, 1968.
- Cited on page 40.
- [134] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, 1970. SIAM Classics 2000.
- Cited on page 78.
- [135] D. P. O’Leary. Bibliographical memoirs: Gene h. golub 1932–2007. *National Academy of Sciences*, 2018.
- Cited on page 55.
- [136] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.
- Cited on page 68.
- [137] R. A. Parker. *Demotic Mathematical Papyri*. Brown University Press, 1972.
- Cited on pages 3 and 4.
- [138] H.-O. Peitgen and P. H. Richter. *The Beauty of Fractals: Images of Complex Dynamical Systems*. Springer Science & Business Media, 1986.
- Cited on page 31.
- [139] L. Pisano. *Scritti di Leonardo Pisano matematico del secolo decimoterzo pubblicati da Baldassarre Boncompagni*. Tipografia delle scienze matematiche e fisiche, Roma, 1857 & 1862. 2 volumes.
- Cited on pages 2, 4 and 7.
- [140] K. Plofker. *Mathematics in India*. Princeton University Press, 2009.
- Cited on page 9.
- [141] K. Plofker, A. Keller, T. Hayashi, C. Montelle, and D. Wujastyk. The bakhshālī manuscript: A response to the bodleian library’s radiocarbon dating. *History of Science in South Asia*, 5(1):134–150, 2017.
- Cited on page 6.
- [142] J. Raphson. *Analysis æquationum universalis seu ad æquationes algebraicas resolvendas methodus generalis, & expedita, ex nova infinitarum serierum methodo, deducta ac demonstrata: cui annexum est de spatio reali, seu ente infinito conamen mathematico-*

- metaphysicum*, volume 1. edition secunda, Londini, 1697.
Cited on pages 22 and 23.
- [143] L. F. Richardson. IX. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210(459-470):307–357, 1911.
Cited on pages 50, 51 and 52.
 - [144] B. Riemann. *Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse*. PhD thesis, Göttingen, 1851.
Cited on page 70.
 - [145] B. Riemann. Bestimmung einer Function einer veränderlichen complexen Grösse durch Grenz- und Unstetigkeitsbedingungen. *Journal für die reine und angewandte Mathematik*, 54:111–114, 1857. (Gesammelte Werke, p. 96–100).
Cited on page 60.
 - [146] S. J. Rigaud. *Correspondence of Scientific Men of the Seventeenth Century, Including Letters of Barrow, Flamsteed, Wallis, and Newton, printed from the Originals in the Collection of the Right Hounourable the Earl of Macclesfield*, volume two volumes. University Press, Oxford, 1841.
Cited on page 21.
 - [147] B. A. Rosenfeld and J. P. Hogendijk. A mathematical treatise written in the Samarqand observatory of Ulugh Beg. *Zeitschrift für Geschichte der Arabisch-Islamischen Wissenschaften*, 15:25–66, 2003.
Cited on page 10.
 - [148] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of computation*, 37(155):105–126, 1981.
Cited on page 68.
 - [149] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.
Cited on page 78.
 - [150] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3):856–869, 1986.
Cited on page 68.
 - [151] Y. Saad and H. van der Vorst. Iterative solution of linear systems in the XX century. *Numerical Analysis 2000. Vol. III, Linear Algebra. J. Comput. Appl. Math.*, 123:1–33, 2000.
Cited on page 48.
 - [152] H. A. Schwarz. Über einen Grenzübergang durch alternierendes Verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 15:272–286, May 1870.
Cited on pages 70 and 71.
 - [153] L. Sédillot. *Prolégomènes des tables astronomiques d’Oloug-Beg*. Paris, 1853.
Cited on page 10.
 - [154] P. L. Seidel. Über ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherung aufzulösen. In *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich Bayerischen Akademie der Wissenschaften, Band 11, III. Abtheilung*, pages 81–108. Verlag d. Akad., 1874.
Cited on pages 44, 46 and 47.
 - [155] D. Shanks. Non-linear transformations of divergent and slowly convergent sequences. *Journal of Mathematics and Physics*, 34(1-4):1–42, 1955.
Cited on page 50.
 - [156] A. Sidi. Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms. *SIAM journal on numerical analysis*, 23(1):197–209, 1986.
Cited on page 59.
 - [157] A. Sidi. Extrapolation vs. projection methods for linear systems of equations. *Journal of Computational and Applied Mathematics*, 22(1):71–88, 1988.
Cited on page 59.
 - [158] A. Sidi and J. Bridger. Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices. *J. of Comp. and Appl. Math.*, 22:35–61, 1988.
Cited on page 50.
 - [159] L. E. Sigler. *Fibonacci’s Liber Abaci: a translation into modern English of Leonardo Pisano’s book of calculation*. Springer, 2002.
Cited on page 8.

- [160] T. Simpson. *Essays on Several Curious and Useful Subjects: In Speculative and Mix'd Mathematicks illustrated by a Variety of Examples*. H. Woodfall, London, 1740.
Cited on pages 23 and 24.
- [161] J. Smyly. Square roots in Heron of Alexandria. *Hermathena*, 63:18–26, 1944.
Cited on page 4.
- [162] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numerische Mathematik*, 126(4):741–770, 2014.
Cited on page 63.
- [163] E. Stiefel. Über einige Methoden der Relaxationsrechnung. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 3(1):1–33, 1952.
Cited on pages 59, 60, 61, 62, 63, 64, 65 and 66.
- [164] C. Sturm. Analyse d'un mémoire sur la résolution des équations numériques. *Bull. des scien. math. astr. phys. et chim.*, 11:419–425, 1829. Lu à l'Acad. roy. des Scien., le 23 mai 1829; see also *Collected Works of Charles François Sturm*, Birkhäuser, 2009, pp. 345–390.
Cited on page 35.
- [165] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory*, volume 34. Springer, 2006.
Cited on page 72.
- [166] P. Ullrich. Forerunners of the power iteration method in the 16th and 18th centuries. *PAMM*, 22(1):e202200270, 2023.
Cited on page 12.
- [167] H. A. Van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on scientific and Statistical Computing*, 13(2):631–644, 1992.
Cited on page 68.
- [168] B. L. van der Waerden. On pre-babylonian mathematics i. *Archive for History of Exact Sciences*, 23(1):1–25, 1980.
Cited on page 41.
- [169] R. S. Varga. Iterative analysis. *New Jersey*, 322, 1962.
Cited on page 78.
- [170] G. Vedova. Notes on Theon of Smyrna. *The American Mathematical Monthly*, 58(10):675–683, 1951.
Cited on page 11.
- [171] H. Von Koch. Une méthode géométrique élémentaire pour l'étude de certaines questions de la théorie des courbes planes. *Acta mathematica*, 30(1):145–174, 1906.
Cited on page 33.
- [172] J. von Neumann. Appendix II in *A study of a numerical solution to a two-dimensional hydrodynamical problem*. Technical report, Los Alamos Scientific Laboratory of the University of California, 1958.
Cited on pages 50, 53 and 55.
- [173] J. von Neumann. *Collected Works: Vol. 5: Design of Computers, Theory of Automata and Numerical Analysis*. General Editor: A. H. Taub. The Macmillan Company, New York, 1963.
Cited on pages 50 and 53.
- [174] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.
Cited on page 50.
- [175] J. Wallis. *Tractatus Duo: Prior, de cycloide et corporibus inde gentis. Posterior, epistolaris; in qua agitur, de cissoide, et corporibus inde gentis: et de curvarum (...)*. Oxoniæ, 1659.
Cited on page 20.
- [176] J. Wallis. *A treatise of algebra, both historical and practical. Shewing, the original, progress, and advancement thereof, from time to time ; and by what steps it hath attained to the heighth at which now it is*. John Playford, London, 1685.
Cited on page 19.
- [177] G. Warnecke. Ein Brief von C.F Gauß an C.L Gerling–Kleinste Fehlerquadrate und das Gauß-Seidel-Verfahren. *Mathematische Semesterberichte*, 67(1):57–84, 2020.
Cited on page 47.
- [178] K. T. W. Weierstrass. Über das sogenannte Dirichletsche Prinzip. *Werke*, 2:49–54, 1895. Read in the Academy Berlin 14. July 1870.
Cited on page 70.
- [179] C. A. Wilson. From Kepler's laws, so-called, to universal gravitation: empirical factors.

- Archive for History of Exact Sciences*, 6(2):89–170, 1970.
 Cited on page 20.
- [180] P. Wynn. On a device for computing the $e_m(s_n)$ transformation. *Mathematical Tables and Other Aids to Computation*, 10(54):91–96, 1956.
 Cited on page 50.
- [181] P. Wynn. General purpose vector epsilon algorithm Algol procedures. *Numerische Mathematik*, 6(1):22–36, 1964.
 Cited on pages 50 and 59.
- [182] D. M. Young. Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society*, 76(1):92–111, 1954.
 Cited on page 48.
- [183] D. M. Young Jr. *Iterative Methods for Solving Partial Difference Equations of Elliptic Type*. PhD thesis, Harvard University, 1950.
 Cited on page 48.
- [184] T. J. Ypma. Historical development of the Newton–Raphson method. *SIAM review*, 37(4):531–551, 1995.
 Cited on pages 18, 21, 22 and 23.