Numbers

1

The commonly known rational numbers (fractions) are not sufficient for a rigorous foundation of mathematical analysis. The historical development shows that for issues concerning analysis, the rational numbers have to be extended to the real numbers. For clarity we introduce the real numbers as decimal numbers with an infinite number of decimal places. We illustrate exemplarily how the rules of calculation and the order relation extend from the rational to the real numbers in a natural way.

A further section is dedicated to floating point numbers, which are implemented in most programming languages as approximations to the real numbers. In particular, we will discuss optimal rounding and in connection with this the relative machine accuracy.

1.1 The Real Numbers

In this book we assume the following number systems as known:

$$\mathbb{N} = \{1, 2, 3, 4, \ldots\} \qquad \text{the set of natural numbers;}$$

$$\mathbb{N}_0 = \mathbb{N} \cup \{0\} \qquad \text{the set of natural numbers including zero;}$$

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\} \qquad \text{the set of integers;}$$

$$\mathbb{Q} = \left\{ \tfrac{k}{n}\ ;\ k \in \mathbb{Z} \text{ and } n \in \mathbb{N} \right\} \qquad \text{the set of rational numbers.}$$

Two rational numbers $\frac{k}{n}$ and $\frac{\ell}{m}$ are equal if and only if $km = \ell n$. Further an integer $k \in \mathbb{Z}$ can be identified with the fraction $\frac{k}{1} \in \mathbb{Q}$. Consequently, the inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$ are true.

M. Oberguggenberger and A. Ostermann, *Analysis for Computer Scientists*,
Undergraduate Topics in Computer Science,
https://doi.org/10.1007/978-3-319-91155-7_1

Let M and N be arbitrary sets. A *mapping* from M to N is a rule which assigns to each element in M exactly one element in N.[1] A mapping is called *bijective*, if for *each* element $n \in N$ there exists *exactly one* element in M which is assigned to n.

Definition 1.1 Two sets M and N have *the same cardinality* if there exists a bijective mapping between these sets. A set M is called *countably infinite* if it has the same cardinality as $\mathbb{N}$.

The sets $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ have the same cardinality and in this sense are *equally large*. All three sets have an infinite number of elements which can be enumerated. Each enumeration represents a bijective mapping to $\mathbb{N}$. The countability of $\mathbb{Z}$ can be seen from the representation $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\}$. To prove the countability of $\mathbb{Q}$, Cantor's[2] diagonal method is being used:

$$\begin{array}{ccccccccc}
\frac{1}{1} & \rightarrow & \frac{2}{1} & & \frac{3}{1} & \rightarrow & \frac{4}{1} & \cdots \\
 & \swarrow & & \nearrow & & \swarrow & & \\
\frac{1}{2} & & \frac{2}{2} & & \frac{3}{2} & & \frac{4}{2} & \cdots \\
\downarrow & \nearrow & & \swarrow & & & & \\
\frac{1}{3} & & \frac{2}{3} & & \frac{3}{3} & & \frac{4}{3} & \cdots \\
 & \swarrow & & & & & & \\
\frac{1}{4} & & \frac{2}{4} & & \frac{3}{4} & & \frac{4}{4} & \cdots \\
\vdots & & \vdots & & \vdots & & \vdots &
\end{array}$$

The enumeration is carried out in direction of the arrows, where each rational number is only counted at its *first* appearance. In this way the countability of all positive rational number (and therefore all rational numbers) is proven.

To visualise the rational numbers we use a line, which can be pictured as an infinitely long ruler, on which an arbitrary point is labelled as *zero*. The integers are marked equidistantly starting from zero. Likewise each rational number is allocated a specific place on the real line according to its size, see Fig. 1.1.

However, the real line also contains points which do not correspond to rational numbers. (We say that $\mathbb{Q}$ is *not complete*.) For instance, the length of the diagonal d in the unit square (see Fig. 1.2) can be measured with a ruler. Yet, the Pythagoreans already knew that $d^2 = 2$, but that $d = \sqrt{2}$ is not a rational number.

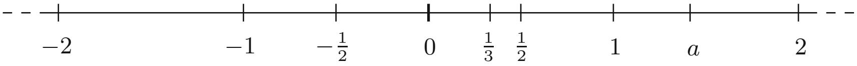

Fig. 1.1 The real line

[1] We will rarely use the term mapping in such generality. The special case of *real-valued functions*, which is important for us, will be discussed thoroughly in Chap. 2.

[2] G. Cantor, 1845–1918.

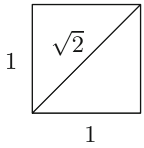

Fig. 1.2 Diagonal in the unit square

Proposition 1.2 $\sqrt{2} \notin \mathbb{Q}$.

Proof This statement is proven indirectly. Assume that $\sqrt{2}$ were rational. Then $\sqrt{2}$ can be represented as a reduced fraction $\sqrt{2} = \frac{k}{n} \in \mathbb{Q}$. Squaring this equation gives $k^2 = 2n^2$ and thus k^2 would be an even number. This is only possible if k itself is an even number, so $k = 2l$. If we substitute this into the above we obtain $4l^2 = 2n^2$ which simplifies to $2l^2 = n^2$. Consequently n would also be even which is in contradiction to the initial assumption that the fraction $\frac{k}{n}$ was reduced. □

As it is generally known, $\sqrt{2}$ is the unique positive root of the polynomial $x^2 - 2$. The naive supposition that all non-rational numbers are roots of polynomials with integer coefficients turns out to be incorrect. There are other non-rational numbers (so-called transcendental numbers) which *cannot* be represented in this way. For example, the ratio of a circle's circumference to its diameter

$$\pi = 3.141592653589793... \notin \mathbb{Q}$$

is transcendental, but can be represented on the real line as half the circumference of the circle with radius 1 (e.g. through unwinding).

In the following we will take up a pragmatic point of view and construct the missing numbers as decimals.

Definition 1.3 A finite decimal number x with l decimal places has the form

$$x = \pm\, d_0.d_1 d_2 d_3 \ldots d_l$$

with $d_0 \in \mathbb{N}_0$ and the single digits $d_i \in \{0, 1, \ldots, 9\}$, $1 \le i \le l$, with $d_l \ne 0$.

Proposition 1.4 (Representing rational numbers as decimals) *Each rational number can be written as a finite or periodic decimal.*

Proof Let $q \in \mathbb{Q}$ and consequently $q = \frac{k}{n}$ with $k \in \mathbb{Z}$ and $n \in \mathbb{N}$. One obtains the representation of q as a decimal by successive division with remainder. Since the remainder $r \in \mathbb{N}$ always fulfils the condition $0 \le r < n$, the remainder will be zero or periodic after a maximum of n iterations. □

Example 1.5 Let us take $q = -\frac{5}{7} \in \mathbb{Q}$ as an example. Successive division with remainder shows that $q = -0.71428571428571...$ with remainders 5, 1, 3, 2, 6, 4, 5, 1, 3, 2, 6, 4, 5, 1, 3, ... The period of this decimal is six.

Each nonzero decimal with a finite number of decimal places can be written as a periodic decimal (with an infinite number of decimal places). To this end one diminishes the last nonzero digit by one and then fills the remaining infinitely many decimal places with the digit 9. For example, the fraction $-\frac{17}{50} = -0.34 = -0.3399999...$ becomes periodic after the third decimal place. In this way $\mathbb{Q}$ can be considered as the set of all decimals which turn periodic from a certain number of decimal places onwards.

Definition 1.6 The set of *real numbers* $\mathbb{R}$ consists of all decimals of the form

$$\pm\, d_0.d_1d_2d_3...$$

with $d_0 \in \mathbb{N}_0$ and digits $d_i \in \{0, ..., 9\}$, i.e. decimals with an infinite number of decimal places. The set $\mathbb{R} \setminus \mathbb{Q}$ is called the set of *irrational* numbers.

Obviously $\mathbb{Q} \subset \mathbb{R}$. According to what was mentioned so far the numbers

$$0.10100100010000 10... \quad \text{and} \quad \sqrt{2}$$

are irrational. There are much more irrational than rational numbers, as is shown by the following proposition.

Proposition 1.7 *The set $\mathbb{R}$ is not countable and has therefore higher cardinality than $\mathbb{Q}$.*

Proof This statement is proven indirectly. Assume the real numbers between 0 and 1 to be countable and tabulate them:

$$\begin{array}{ll}
1 & 0.\,d_{11}\,d_{12}\,d_{13}\,d_{14}... \\
2 & 0.\,d_{21}\,d_{22}\,d_{23}\,d_{24}... \\
3 & 0.\,d_{31}\,d_{32}\,d_{33}\,d_{34}... \\
4 & 0.\,d_{41}\,d_{42}\,d_{43}\,d_{44}... \\
. & ... \\
. & ...
\end{array}$$

With the help of this list, we define

$$d_i = \begin{cases} 1 & \text{if } d_{ii} = 2, \\ 2 & \text{else.} \end{cases}$$

Then $x = 0.d_1d_2d_3d_4...$ is not included in the above list which is a contradiction to the initial assumption of countability. □

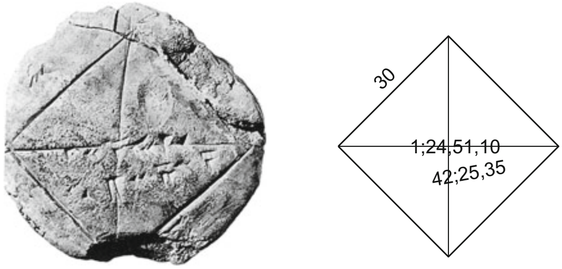

Fig. 1.3 Babylonian cuneiform inscription YBC 7289 (Yale Babylonian Collection, with authorisation) from 1900 before our time with a translation of the inscription according to [1]. It represents a square with side length 30 and diagonals 42; 25, 35. The ratio is $\sqrt{2} \approx 1; 24, 51, 10$

However, although $\mathbb{R}$ contains considerably more numbers than $\mathbb{Q}$, every real number can be approximated by rational numbers to any degree of accuracy, e.g. π to nine digits

$$\pi \approx \frac{314159265}{100000000} \in \mathbb{Q}.$$

Good approximations to the real numbers are sufficient for practical applications. For $\sqrt{2}$, already the Babylonians were aware of such approximations:

$$\sqrt{2} \approx 1; 24, 51, 10 = 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} = 1.41421296\ldots,$$

see Fig. 1.3. The somewhat unfamiliar notation is due to the fact that the Babylonians worked in the sexagesimal system with base 60.

1.2 Order Relation and Arithmetic on $\mathbb{R}$

In the following we write real numbers (uniquely) as decimals with an infinite number of decimal places, for example, we write 0.2999... instead of 0.3.

Definition 1.8 (Order relation) Let $a = a_0.a_1a_2\ldots$ and $b = b_0.b_1b_2\ldots$ be non-negative real numbers in decimal form, i.e. $a_0, b_0 \in \mathbb{N}_0$.

(a) One says that a is *less than or equal to* b (and writes $a \le b$), if $a = b$ or if there is an index $j \in \mathbb{N}_0$ such that $a_j < b_j$ and $a_i = b_i$ for $i = 0, \ldots, j-1$.

(b) Furthermore one stipulates that always $-a \le b$ and sets $-a \le -b$ whenever $b \le a$.

This definition extends the known orders of $\mathbb{N}$ and $\mathbb{Q}$ to $\mathbb{R}$. The interpretation of the order relation $\le$ on the real line is as follows: $a \le b$ holds true, if a is to the left of b on the real line, or $a = b$.

The relation $\leq$ obviously has the following properties. For all $a, b, c \in \mathbb{R}$ it holds that

$$\begin{aligned} &a \leq a \quad \text{(reflexivity)},\\ &a \leq b \ \text{ and } \ b \leq c \quad \Rightarrow \quad a \leq c \quad \text{(transitivity)},\\ &a \leq b \ \text{ and } \ b \leq a \quad \Rightarrow \quad a = b \quad \text{(antisymmetry)}.\end{aligned}$$

In case of $a \leq b$ and $a \neq b$ one writes $a < b$ and calls a *less than* b. Furthermore one defines $a \geq b$, if $b \leq a$ (in words: a *greater than or equal to* b), and $a > b$, if $b < a$ (in words: a *greater than* b).

Addition and multiplication can be carried over from $\mathbb{Q}$ to $\mathbb{R}$ in a similar way. Graphically one uses the fact that each real number corresponds to a segment on the real line. One thus defines the addition of real numbers as the addition of the respective segments.

A rigorous and at the same time *algorithmic* definition of the addition starts from the observation that real numbers can be approximated by rational numbers to any degree of accuracy. Let $a = a_0.a_1a_2...$ and $b = b_0.b_1b_2...$ be two non-negative real numbers. By cutting them off after k decimal places we obtain two rational approximations $a^{(k)} = a_0.a_1a_2...a_k \approx a$ and $b^{(k)} = b_0.b_1b_2...b_k \approx b$. Then $a^{(k)} + b^{(k)}$ is a monotonically increasing sequence of approximations to the yet to be defined number $a + b$. This allows one to *define* $a + b$ as *supremum* of these approximations. To justify this approach rigorously we refer to Chap. 5. The multiplication of real numbers is defined in the same way. It turns out that the real numbers with addition and multiplication $(\mathbb{R}, +, \cdot)$ are a *field*. Therefore the usual rules of calculation apply, e.g., the distributive law

$$(a + b)c = ac + bc.$$

The following proposition recapitulates some of the important rules for $\leq$. The statements can easily be verified with the help of the real line.

Proposition 1.9 *For all $a, b, c \in \mathbb{R}$ the following holds:*

$$\begin{aligned} a \leq b \quad &\Rightarrow \quad a + c \leq b + c,\\ a \leq b \ \text{ and } \ c \geq 0 \quad &\Rightarrow \quad ac \leq bc,\\ a \leq b \ \text{ and } \ c \leq 0 \quad &\Rightarrow \quad ac \geq bc.\end{aligned}$$

Note that $a < b$ does *not* imply $a^2 < b^2$. For example $-2 < 1$, but nonetheless $4 > 1$. However, for $a, b \geq 0$ it always holds that $a < b \ \Leftrightarrow \ a^2 < b^2$.

Definition 1.10 (Intervals) The following subsets of $\mathbb{R}$ are called intervals:

$$\begin{aligned} [a, b] &= \{x \in \mathbb{R}\,;\ a \leq x \leq b\} \quad \text{closed interval};\\ (a, b] &= \{x \in \mathbb{R}\,;\ a < x \leq b\} \quad \text{left half-open interval};\\ [a, b) &= \{x \in \mathbb{R}\,;\ a \leq x < b\} \quad \text{right half-open interval};\\ (a, b) &= \{x \in \mathbb{R}\,;\ a < x < b\} \quad \text{open interval}.\end{aligned}$$

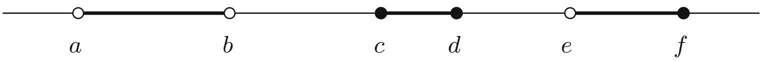

Fig. 1.4 The intervals (a, b), $[c, d]$ and $(e, f]$ on the real line

Intervals can be visualised on the real line, as illustrated in Fig. 1.4.

It proves to be useful to introduce the symbols $-\infty$ (minus infinity) and ∞ (infinity), by means of the property

$$\forall a \in \mathbb{R} : -\infty < a < \infty.$$

One may then define, e.g., the *improper* intervals

$$\begin{aligned}[a, \infty) &= \{x \in \mathbb{R}\ ;\ x \geq a\}\\ (-\infty, b) &= \{x \in \mathbb{R}\ ;\ x < b\}\end{aligned}$$

and furthermore $(-\infty, \infty) = \mathbb{R}$. Note that $-\infty$ and ∞ are only *symbols* and *not* real numbers.

Definition 1.11 The *absolute value* of a real number a is defined as

$$|a| = \begin{cases} a, & \text{if } a \geq 0,\\ -a, & \text{if } a < 0.\end{cases}$$

As an application of the properties of the order relation given in Proposition 1.9 we exemplarily solve some inequalities.

Example 1.12 Find all $x \in \mathbb{R}$ satisfying $-3x - 2 \leq 5 < -3x + 4$. In this example we have the following two inequalities

$$-3x - 2 \leq 5 \quad \text{and} \quad 5 < -3x + 4.$$

The first inequality can be rearranged to

$$-3x \leq 7 \quad \Leftrightarrow \quad x \geq -\frac{7}{3}.$$

This is the first constraint for x. The second inequality states

$$3x < -1 \quad \Leftrightarrow \quad x < -\frac{1}{3}$$

and poses a second constraint for x. The solution to the original problem must fulfil both constraints. Therefore the solution set is

$$S = \left\{x \in \mathbb{R};\ -\frac{7}{3} \leq x < -\frac{1}{3}\right\} = \left[-\frac{7}{3}, -\frac{1}{3}\right).$$

Example 1.13 Find all $x \in \mathbb{R}$ satisfying $x^2 - 2x \geq 3$. By completing the square the inequality is rewritten as

$$(x-1)^2 = x^2 - 2x + 1 \geq 4.$$

Taking the square root we obtain two possibilities

$$x - 1 \geq 2 \qquad \text{or} \qquad x - 1 \leq -2.$$

The combination of those gives the solution set

$$S = \{x \in \mathbb{R}\,;\ x \geq 3 \text{ or } x \leq -1\} = (-\infty, -1] \cup [3, \infty).$$

1.3 Machine Numbers

The real numbers can be realised only partially on a computer. In exact arithmetic, like for example in maple, real numbers are treated as symbolic expressions, e.g. $\sqrt{2} =$ `RootOf(_Z^2-2)`. With the help of the command `evalf` they can be evaluated, exact to many decimal places.

The floating point numbers that are usually employed in programming languages as substitutes for the real numbers have a fixed relative accuracy, e.g. *double precision* with 52 bit mantissa. The arithmetic rules of $\mathbb{R}$ are *not* valid for these machine numbers, e.g.

$$1 + 10^{-20} = 1$$

in double precision. Floating point numbers are standardised by the *Institute of Electrical and Electronics Engineers* IEEE 754-1985 and by the *International Electrotechnical Commission* IEC 559:1989. In the following we give a short outline of these machine numbers. Further information can be found in [20].

One distinguishes between single and double format. The single format (*single precision*) requires 32-bit storage space

V	e	M
1	8	23

The double format (*double precision*) requires 64-bit storage space

V	e	M
1	11	52

Here, $V \in \{0, 1\}$ denotes the sign, $e_{\min} \leq e \leq e_{\max}$ is the exponent (a signed integer) and M is the mantissa of length p

$$M = d_1 2^{-1} + d_2 2^{-2} + \ldots + d_p 2^{-p} \;\cong\; d_1 d_2 \ldots d_p, \qquad d_j \in \{0, 1\}.$$

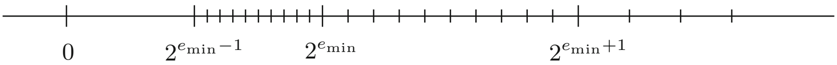

Fig. 1.5 Floating point numbers on the real line

This representation corresponds to the following number x:

$$x = (-1)^V 2^e \sum_{j=1}^{p} d_j 2^{-j}.$$

Normalised floating point numbers in base 2 always have $d_1 = 1$. Therefore, one does not need to store d_1 and obtains for the mantissa

single precision	$p = 24$;
double precision	$p = 53$.

To simplify matters we will only describe the key features of floating point numbers. For the subtleties of the IEEE-IEC standard, we refer to [20].

In our representation the following range applies for the exponents:

	$e_{\min}$	$e_{\max}$
single precision	-125	128
double precision	-1021	1024

With $M = M_{\max}$ and $e = e_{\max}$ one obtains the largest floating point number

$$x_{\max} = \left(1 - 2^{-p}\right) 2^{e_{\max}},$$

whereas $M = M_{\min}$ and $e = e_{\min}$ gives the smallest positive (normalised) floating point number

$$x_{\min} = 2^{e_{\min}-1}.$$

The floating point numbers are *not* evenly distributed on the real line, but their *relative* density is nearly constant, see Fig. 1.5.

In the IEEE standard the following approximate values apply:

	$x_{\min}$	$x_{\max}$
single precision	$1.18 \cdot 10^{-38}$	$3.40 \cdot 10^{38}$
double precision	$2.23 \cdot 10^{-308}$	$1.80 \cdot 10^{308}$

Furthermore, there are special *symbols* like

±`INF`	...	$\pm\infty$
`NaN`	...	not a number, e.g. for *zero divided by zero*.

In general, one can continue calculating with these symbols without program termination.

1.4 Rounding

Let $x = a \cdot 2^e \in \mathbb{R}$ with $1/2 \le a < 1$ and $x_{\min} \le x \le x_{\max}$. Furthermore, let u, v be two adjacent machine numbers with $u \le x \le v$. Then

$$u = \begin{array}{|c|c|c|}\hline 0 & e & b_1 \dots b_p \\ \hline \end{array}$$

and

$$v = u + \begin{array}{|c|c|c|}\hline 0 & e & 00\dots01 \\ \hline \end{array} = u + \begin{array}{|c|c|c|}\hline 0 & e-(p-1) & 10\dots00 \\ \hline \end{array}$$

Thus $v - u = 2^{e-p}$ and the inequality

$$|\mathrm{rd}(x) - x| \le \frac{1}{2}(v-u) = 2^{e-p-1}$$

holds for the optimal *rounding* $\mathrm{rd}(x)$ of x. With this estimate one can determine the *relative error* of the rounding. Due to $\frac{1}{a} \le 2$ it holds that

$$\frac{|\mathrm{rd}(x) - x|}{x} \le \frac{2^{e-p-1}}{a \cdot 2^e} \le 2 \cdot 2^{-p-1} = 2^{-p}.$$

The same calculation is valid for negative x (by using the absolute value).

Definition 1.14 The number $\texttt{eps} = 2^{-p}$ is called *relative machine accuracy*.

The following proposition is an important application of this concept.

Proposition 1.15 *Let $x \in \mathbb{R}$ with $x_{\min} \le |x| \le x_{\max}$. Then there exists $\varepsilon \in \mathbb{R}$ with*

$$rd(x) = x(1+\varepsilon) \quad \textit{and} \quad |\varepsilon| \le \texttt{eps}.$$

Proof We define

$$\varepsilon = \frac{\mathrm{rd}(x) - x}{x}.$$

According to the calculation above, we have $|\varepsilon| \le \texttt{eps}$. □

Experiment 1.16 (Experimental determination of eps) Let z be the smallest positive machine number for which $1 + z > 1$.

$$1 = \begin{array}{|c|c|c|}\hline 0 & 1 & 100\dots00 \\ \hline \end{array}, \qquad z = \begin{array}{|c|c|c|}\hline 0 & 1 & 000\dots01 \\ \hline \end{array} = 2 \cdot 2^{-p}.$$

Thus $z = 2\,\texttt{eps}$. The number z can be determined experimentally and therefore eps as well. (Note that the number z is called eps in MATLAB.)

In IEC/IEEE standard the following applies:

$$\begin{array}{ll}\text{single precision:} & \texttt{eps} = 2^{-24} \approx 5.96 \cdot 10^{-8},\\ \text{double precision:} & \texttt{eps} = 2^{-53} \approx 1.11 \cdot 10^{-16}.\end{array}$$

In double precision arithmetic an accuracy of approximately 16 places is available.

1.5 Exercises

1. Show that $\sqrt{3}$ is irrational.

2. Prove the triangle inequality

$$|a+b| \le |a| + |b|$$

for all $a, b \in \mathbb{R}$.
Hint. Distinguish the cases where a and b have either the same or different signs.

3. Sketch the following subsets of the real line:

$$A = \{x : |x| \le 1\}, \qquad B = \{x : |x-1| \le 2\}, \qquad C = \{x : |x| \ge 3\}.$$

More generally, sketch the set $U_r(a) = \{x : |x-a| < r\}$ (for $a \in \mathbb{R}$, $r > 0$). Convince yourself that $U_r(a)$ is the set of points of distance less than r to the point a.

4. Solve the following inequalities by hand as well as with maple (using `solve`). State the solution set in interval notation.

(a) $4x^2 \le 8x + 1,$ (b) $\dfrac{1}{3-x} > 3 + x,$

(c) $\left|2 - x^2\right| \ge x^2,$ (d) $\dfrac{1+x}{1-x} > 1,$

(e) $x^2 < 6 + x,$ (f) $\big||x| - x\big| \ge 1,$

(g) $|1 - x^2| \le 2x + 2,$ (h) $4x^2 - 13x + 4 < 1.$

5. Determine the solution set of the inequality

$$8(x-2) \ge \frac{20}{x+1} + 3(x-7).$$

6. Sketch the regions in the (x, y)-plane which are given by

(a) $x = y$; (b) $y < x$; (c) $y > x$; (d) $y > |x|$; (e) $|y| > |x|$.

Hint. Consult Sects. A.1 and A.6 for basic plane geometry.

7. Compute the binary representation of the floating point number $x = 0.1$ in single precision IEEE arithmetic.

8. Experimentally determine the relative machine accuracy `eps`.
Hint. Write a computer program in your programming language of choice which calculates the smallest machine number z such that $1 + z > 1$.