TW

stichting
mathematisch
centrum

MC

AFDELING TOEGEPASTE WISKUNDE TC 53/73 AUGUSTUS

N.M. TEMME
GETALSTELSELS EN GETALVOORSTELLINGEN

TW

2e boerhaavestraat 49 amsterdam

Inhoud

Samenvatting

In deze syllabus wordt een overzicht gegeven van getalrepresentaties voor de reële getallen. Zowel binaire representaties als getalstelsels met algemeen grondtal β, $\beta = 2,3,4,\ldots$, komen aan de orde, alsmede de rekenregels voor de conversie van de ene representatie naar de andere. Verdere onderwerpen zijn: floating point notatie, de representatie van de getallen op de computer en afrondfouten.

Getalstelsels en getalvoorstellingen

§1. Opbouw van de reële getallen.

De verzameling van de natuurlijke getallen wordt aangeduid met $\mathbb{N}$ en bevat de getallen 1,2,3,4,... .

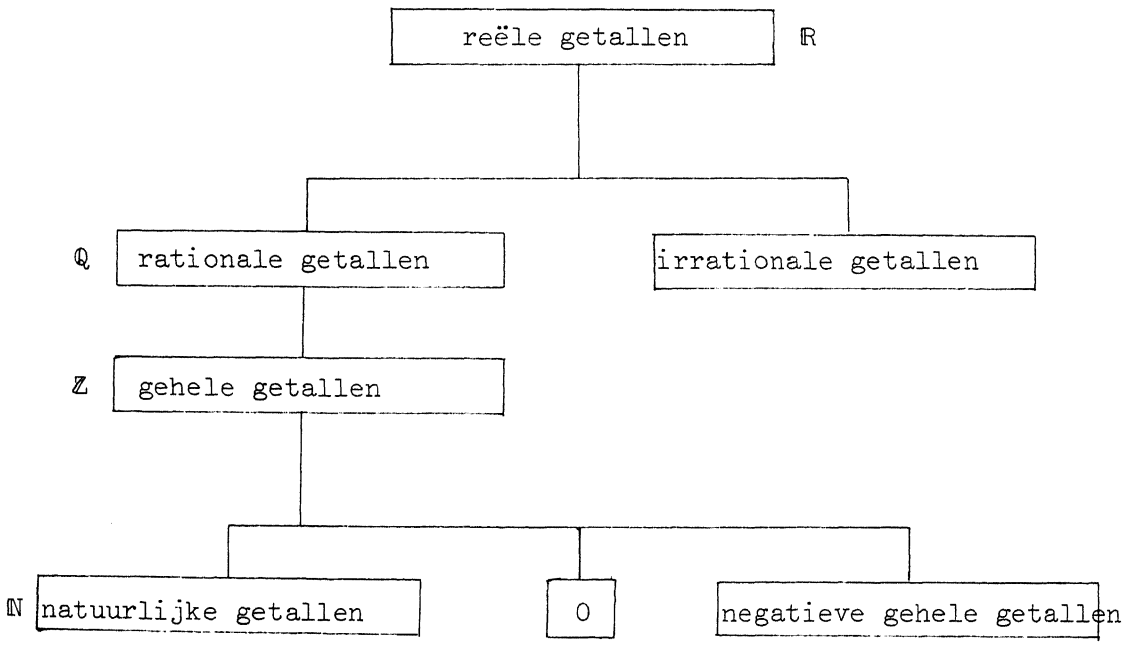
De verzameling der gehele getallen geven we aan met $\mathbb{Z}$. Hierin komen naast de natuurlijke getallen ook de negatieve getallen voor en het getal 0. $\mathbb{Z}$ bevat dus de getallen 0, 1, -1, 2, -2,... .

De verzameling der rationale getallen $\mathbb{Q}$ bevat de getallen p/q, waarbij p en q gehele getallen zijn en $q \neq 0$.

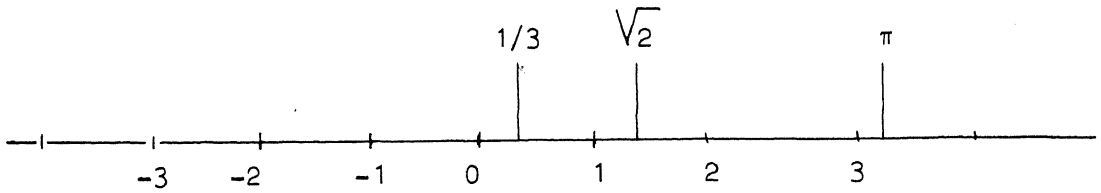
De verzameling der irrationale getallen $\mathbb{Q}'$ bevat getallen als $\sqrt{2}$, ${}^{10}\log 3$, π, die niet als rationaal getal geschreven kunnen worden.

De verzameling $\mathbb{R}$ der reële getallen bevat alle getallen uit $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ en $\mathbb{Q}'$.

Voor de opbouw van de verzameling der reële getallen kan het volgende schema gekonstrueerd worden.



Met de verzameling $\mathbb{R}$ der reële getallen kunnen we de getallen rechts associeren. We kiezen op die lijn een punt dat we de oorsprong noemen en geven daarmee de plaats van het getal 0 aan. (zie figuur).



Elk punt op de lijn representeert een reëel getal en omgekeerd kan elk reëel getal worden gerepresenteerd door een punt van de lijn.

§2. Getalstelsels.

Het decimale getalstelsel gebruikt 10 kentekens, of cijfers, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 om de getallen op te bouwen en aan te geven. Als we een decimaal getal neerschrijven gebruiken we een afkorting van een rekenkundige uitdrukking. Zo kunnen we de notatie 365 beschouwen als een afkorting van $3 \times 10^2 + 6 \times 10^1 + 5 \times 10^0$. Bij niet gehele getallen komen ook negatieve machten van 10 voor. Bijvoorbeeld $7.88 = 7 \times 10^0 + 8 \times 10^{-1} + 8 \times 10^{-2}$. (In Nederland is men gewoon de decimale breuk met een komma aan te geven. Wij zullen hiervoor een punt gebruiken).

Het getal 10 is waarschijnlijk vanouds als basis voor de getallen gekozen omdat de mens over 10 vingers beschikt. We kunnen echter elk willekeurig natuurlijk getal (groter dan 1) als basis van een getalstelsel nemen.
In het algemeen kunnen we een getalstelsel beschouwen met grondtal β waarbij de getallen geschreven kunnen worden als een rekenkundige uitdrukking die machten van β bevat. Het getal kan dan ook weer afgekort worden, maar om aan te geven welk getalstelsel gebruikt wordt zullen we het getal voorzien van een index. Als geen index gebruikt wordt nemen we aan dat als grondtal het getal 10 is gekozen. Als voorbeeld zullen we het getal 433 in verschillende getalstelsels representeren.

Voorbeeld

$$\begin{aligned} 433 &= 6 \times 8^2 + 6 \times 8^1 + 1 \times 8^0 = (661)_8 \\ &= 3 \times 5^3 + 2 \times 5^2 + 1 \times 5^1 + 3 \times 5^0 = (3213)_5 \\ &= 1 \times 3^5 + 2 \times 3^4 + 1 \times 3^3 + 0 \times 3^2 + 0 \times 3^1 + 1 \times 3^0 = (121001)_3 \\ &= 1 \times 2^8 + 1 \times 2^7 + 0 \times 2^6 + 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 \\ &\quad + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = (110\ 110\ 001)_2. \end{aligned}$$

Elk geheel getal kan geschreven worden in een getalstelsel met grondtal β. In het algemeen ziet de representatie er als volgt uit

$$x = \sum_{k=0}^{n} x_k \beta^k = x_0\beta^0 + x_1\beta^1 + x_2\beta^2 + \ldots\ldots + x_n\beta^n.$$

De gehele getallen x_k worden de gewichten genoemd; alle x_k voldoen aan de ongelijkheid

$$0 \leq x_k \leq \beta-1 \quad .$$

Zoals blijkt uit het bovenstaande voorbeeld wordt de lengte van de getalrepresentatie groter naarmate een kleiner grondtal wordt gebruikt. Er zijn echter voor een klein grondtal minder verschillende gewichten nodig om het getal te representeren dan bij een groter. De ons vertrouwde gewichten 0, 1,....., 9 van het tientallig stelsel zullen we ook, voor zover nodig, kunnen gebruiken voor een getalstelsel met $\beta < 10$. Voor $\beta > 10$ hebben we meer symbolen nodig. Zo is voor het twaalf-tallig stelsel een groep van 12 symbolen nodig om de gewichten aan te geven. Om te voorkomen dat de symbolen uit meer dan een teken bestaan kunnen we bij het twaalf-tallig stelsel de groep 0, 1,...., 9 uitbreiden met de symbolen t en e. Het getal 1711 kan dan in het twaalf-tallig stelsel geschreven worden als

$$1711 = 11 \times 12^2 + 10 \times 12^1 + 7 \times 12^0 = e \times 12^2 + t \times 12^1 + 7 \times 12^0 = (et7)_{12}$$

Vraagstukken.

1. In het decimale stelsel kan men de gehele getallen tot en met 999 met drie symbolen (cijfers) weergeven. Hoeveel getallen kan men in het getalstelsel met grondtal β weergeven als men hoogstens n symbolen gebruikt?
2. Welk getal volgt in het binaire getalstelsel (β=2) op het getal $(110\ 111)_2$.
3. Bepaal de som van de volgende getallenparen a en b, waarbij a en b in de tabel gegeven zijn.

a	b
$(101011)_2$	$(10010)_2$
$(1202)_3$	$(1102)_3$
$(43204)_5$	$(22232)_5$

4. Schrijf het getal 500 in het getalstelsel met grondtal respektievelijk 12, 8, 5, 3 en 2.

5. Konstrueer een algemeen algorithme waarmee een geheel getal in decimale representatie omgezet kan worden in het getalstelsel met grondtal β.

6. Het decimale getal 79 kan in een zeker getalstelsel worden aangegeven met 142. Welk grondtal heeft dit stelsel?

§3. Het representeren van niet-gehele getallen.

Getallen waarin een decimale punt voorkomt kunnen in het decimale stelsel met positieve en negatieve machten van 10 worden voorgesteld. Het getal 56.039 kan geschreven worden als

$$5 \times 10^1 + 6 \times 10^0 + 0 \times 10^{-1} + 3 \times 10^{-2} + 9 \times 10^{-3}.$$

De punt kan ook worden gebruikt in andere getalstelsels, om aan te geven of negatieve machten van het grondtal gebruikt worden.
Zo is

$$(073.24)_8 = 7 \times 8^1 + 3 \times 8^0 + 2 \times 8^{-1} + 4 \times 8^{-2} = 59.3125$$
$$(10.1101)_2 = 2^1 + 2^{-1} + 2^{-2} + 2^{-4} = 2.8125$$
$$(20.12)_3 = 2 \times 3^1 + 3^{-1} + 2 \times 3^{-2} = 6 + \frac{3+2}{9} = 6.555\ldots$$

In het algemeen kan een getal in het stelsel met grondtal β worden weergegeven door

$$x = \sum_{k=-m}^{n} x_k \beta^k = x_n \beta^n + x_{n-1} \beta^{n-1} + \ldots \quad x_{-m} \beta^{-m}.$$

Hierin zijn x_k $(k = -m,\ldots,n)$ gehele getallen met $0 \leq x_i \leq \beta - 1$.

Het getal 1/3 heeft in het decimale stelsel een oneindig voortlopende decimale ontwikkeling. In het 3-tallig stelsel is de ontwikkeling echter eindig (ga na!). Het is gemakkelijk in te zien dat er voor een rationaal getal altijd een getalstelsel gevonden kan worden zodat de representatie eindig is. Voorts kan bewezen worden dat de irrationale getallen in elk getalstelsel een oneindige ontwikkeling bezitten (ga na).

We zullen nu nagaan hoe een reëel getal, in decimale representatie, geschreven kan worden in een ander getalstelsel. Elk decimaal getal bestaat uit een geheel getal (in decimale representatie) links van de decimale punt en een gebroken deel. Zo bestaat 176.78 uit het gehele getal 176 en uit de decimale breuk .78. We kunnen het gehele en gebroken deel afzonderlijk beschouwen. Voor het gehele deel verwijzen we naar de vorige paragraaf. We zullen ons dan ook bezig houden met reële getallen tussen 0 en 1.

Veronderstel dat x een reëel getal is met $0 < x < 1$. We zullen de gewichten $x_1, x_2, \ldots$ bepalen in de ontwikkeling van x in het getalstelsel met grondtal β. Alle x_i zijn gehele getallen die voldoen aan $0 \leq x_i \leq \beta-1$. Aangezien $x < 1$ kan deze ontwikkeling dus niet met positieve machten van β beginnen. We kunnen dus schrijven

$$x = x_1\beta^{-1} + x_2\beta^{-2} + x_3\beta^{-3} + \ldots\ldots .$$

Vermenigvuldiging van beide kanten met β geeft

$$x\beta = x_1 + x_2\beta^{-1} + x_3\beta^{-2} + \ldots .$$

Aangezien $x_2\beta^{-1} + x_3\beta^{-2} + \ldots < 1$ volgt hieruit dat x_1 het gehele gedeelte van $x\beta$ is. Hierdoor is x_1 volkomen bepaald. Evenzo blijkt dat x_2 het gehele gedeelte van $(x\beta-x_1)\beta$ is, waarmee x_2 bepaald is. Voortzetting van dit proces levert alle x_i.

Voorbeeld 1. Schrijf 0.3125 in het 8-tallig stelsel

$$8 \times 0.3125 = 2.500 \text{ , dus } x_1 = 2$$
$$8 \times 0.5000 = 4.000 \text{ , dus } x_2 = 4 .$$

Aangezien het gebroken deel 0 is geworden breekt het proces af (alle x_i met $i \geq 3$ zijn 0). Het resultaat is $0.3125 = (0.24)_8$.

Voorbeeld 2. Schrijf 6.8125 in het 2-tallig stelsel.
Het gehele deel 6 kan in het binaire stelsel geschreven worden als $(110)_2$.
Het gebroken gedeelte levert

$$2 \times 0.8125 = 1.6250 \text{ , dus } x_1 = 1$$
$$2 \times 0.6250 = 1.2500 \text{ , dus } x_2 = 1$$
$$2 \times 0.2500 = 0.5000 \text{ , dus } x_3 = 0$$
$$2 \times 0.5000 = 1.0000 \text{ , dus } x_4 = 1 .$$

Het gebroken gedeelte wordt $(1101)_2$, zodat $6.8125 = (110.1101)_2$.

Voorbeeld 3. Schrijf 1/3 in het stelsel met grondtal 4

$$4 \times 1/3 = 4/3 \quad , \quad \text{dus} \quad x_1 = 1$$

$$4 \times 1/3 = 4/3 \quad , \quad \text{dus} \quad x_2 = 1 \quad , \quad \text{etc.}$$

Alle gewichten zijn dus gelijk aan 1; de representatie loopt oneindig voort

$$\frac{1}{3} = 4^{-1} + 4^{-2} + 4^{-3} + 4^{-4} + \ldots\ldots .$$

Vraagstukken.

1. Schrijf het getal 47.11 in getalstelsels met grondtal respektievelijk 8, 5, 3 en 2. Stop de ontwikkeling zonodig 5 plaatsen achter de punt.
2. Onder welke voorwaarde heeft een getal x een eindige representatie in een getalstelsel met grondtal β?
3. Heeft elk binair getal een eindige representatie in het decimale stelsel?
4. Maak een tabel waarin de getallen $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$, $\frac{1}{8}$, $\frac{1}{9}$, $\frac{1}{10}$ en $\frac{1}{12}$ in het twaalftallig stelsel staan aangegeven.
5. Gegeven is het getal $(276)_8$. Schrijf dit getal in het binaire stelsel.
6. Schrijf $(356.364)_8$ in het binaire stelsel.
7. Schrijf $(11001.1111)_2$ in het 8-tallig stelsel.

§4. De floating point notatie.

De plaats van de decimale punt in een reëel getal, dat in decimale ontwikkeling is geschreven, kan verplaatst worden als we dan tevens het getal met een passende macht van 10 vermenigvuldigen. Zo kan 3.14 ook geschreven worden als 31.4×10^{-1}. Als de decimale punt voor het eerste cijfer (van links af geteld) dat ongelijk nul is wordt geplaatst en een passende macht van 10 wordt toegevoegd, dan zeggen we dat het getal geschreven is in decimale floating point notatie.

Voorbeeld. De getallen 1.4142 en 0.001 worden in decimale floating point notatie geschreven als 0.14142×10^{1} en 0.1×10^{-2}.

De algemene gedaante van getallen in de decimale floating point notatie luidt dus

$$x = m\ 10^{\sigma}$$

waarin $0.1 \leq |m| < 1$ en σ een geheel getal is.

We noemen m de mantisse van het getal en σ de wijzer; σ wijst de plaats van de decimale punt aan in de gewone, fixed point notatie.

In het algemeen zal de mantisse de gedaante hebben

$$|m| = x_1 10^{-1} + x_2 10^{-2} + x_3 10^{-3} + \ldots\ldots + x_n 10^{-n} ,$$

met

$$0 \leq x_i \leq 9$$

terwijl nu $x_1 \neq 0$ aangezien anders de decimale punt verder naar rechts had kunnen worden geschoven.

De floating point notatie kan ook voor getalstelsels met een ander grondtal worden ingevoerd. Het getal $(22.13)_3$ wordt in floating point notatie dan $(.2213)_3 \times 3^2$.

In het β-tallig stelsel hebben we

$$x = m\,\beta^{\sigma}$$

met

$$|m| = x_1\beta^{-1} + x_2\beta^{-2} + \ldots\ldots + x_n\beta^{-n} \ ,$$

waarin $0 \leq x_i \leq \beta-1$, $x_1 \neq 0$. Ook hier noemen we m de mantisse en σ de wijzer. Voor de mantisse geldt nu de relatie $\frac{1}{\beta} \leq |m| < 1$.

Vraagstukken.

1. Bewijs dat $\frac{1}{\beta} \leq |m| < 1$.

2. Bewijs dat de getalvoorstelling eenduidig is.

§5. Getalrepresentatie op de computer.

Teneinde snel en efficient te kunnen rekenen worden op een computer de getallen op een bepaalde manier gerepresenteerd. Een van de eerste aspekten waarmee we geconfronteerd worden bij het opslaan van getallen in het geheugen van een computer is de beperkte ruimte die beschikbaar is. We kunnen slechts getallen representeren, die binnen een zeker getalbereik liggen en ook zullen we ons moeten beperken wat betreft het aantal cijfers achter een decimale punt.

De computer werkt met (eindig veel) standaardgetallen in een zeker getalstelsel (meestal met grondtal 2). De representatie van gehele getallen verschilt wezenlijk met die van niet-gehele getallen. We zullen deze groepen dan ook afzonderlijk behandelen. Als grondtal voor het getalstelsel nemen we β.

A. Gehele getallen.

Elk geheel getal kan gerepresenteerd worden in de vorm $(x_1\ x_2\ x_3 \ldots x_l)_\beta$. Als ons per geheel getal n posities ter beschikking staan, plus 1 voor het teken dan kunnen uitsluitend de getallen worden opgenomen met $l \leq n$. We noemen n de maximale woordlengte. Er kunnen zo alle gehele getallen binnen het getalbereik $\pm\ (\beta^n-1)$ exact worden gerepresenteerd. Op deze wijze bezet een geheel getal binnen dit getalbereik één woord in het geheugen van de computer, waarbij n + 1 posities beschikbaar zijn.

B. Niet gehele getallen.

Deze getallen worden in het geheugen gerepresenteerd in floating point representatie. Ook hiervan kunnen slechts eindig veel getallen voorkomen. De gegeven woordlengte dient nu verdeeld te worden over het mantissegedeelte en de wijzer. De wijzer σ is een geheel getal; als de woordlengte hiervan n is, is het bereik van σ weer $\pm\ (\beta^n-1)$. Als voor de mantisse k posities beschikbaar zijn, dan is de algemene gedaante van de mantisse

$$|m| = x_1\beta^{-1} + x_2\beta^{-2} + \ldots\ldots + x_k\beta^{-k}$$

waarin $x_1 \neq 0$; $0 \leq x_i \leq \beta-1$, $i = 1,2,\ldots,k$. We noemen k de woordlengte van de mantisse. Een willekeurig reëel getal bezet dus in het geheugen twee woorden: de wijzer een woord met n posities, de mantisse een woord met k posities. Aangezien de wijzer en de mantisse beide een teken dienen te hebben zijn voor een reëel getal in totaal n + k + 2 posities nodig.

De getallen die in de computer kunnen worden opgeslagen vormen een eindige verzameling. Verder zijn alle getallen rationaal. Als we een willekeurig reëel getal x aanbieden aan de computer dan is er een kleine kans dat dit getal x in de verzameling voorkomt. Er zal dan een benadering x^* van x opgezocht moeten worden.

Veronderstel nu dat het getal x in het β-tallig stelsel de volgende representatie heeft

$$x = m\,\beta^{\sigma}$$

met $$|m| = x_1\beta^{-1} + x_2\beta^{-2} + x_3\beta^{-3} \ldots\ldots\ldots .$$

In het algemeen zal de ontwikkeling oneindig voortlopen. Als de woordlengte van de mantisse k bedraagt dan kan de mantisse van de benadering x^* slechts de eerste k negatieve machten van β bevatten.
De benaderde waarde x^* kan op verschillende manieren bij x gezocht worden.

I. Een computer kan uit de standaardverzameling dat getal x^* opzoeken dat zo dicht mogelijk bij de werkelijke waarde x ligt. In dat geval zeggen we dat de computer netjes afrondt. Welke fout wordt dan gemaakt?

Als $$x_{k+1} + x_{k+2}\beta^{-1} + \ldots\ldots\ldots < \beta/2$$

wordt bij netjes afronden de mantisse m^* van x^*

$$|m^*| = x_1\beta^{-1} + x_2\beta^{-2} + \ldots\ldots + x_k\beta^{-k}$$

zodat

$$|x-x^*| = (x_{k+1} + x_{k+2}\beta^{-1} + \ldots)\beta^{\sigma-k-1}$$

$$< \tfrac{1}{2}\,\beta^{\sigma-k} \quad .$$

Als echter

$$x_{k+1} + x_{k+2}\beta^{-1} + \ldots\ldots\ldots \geq \beta/2$$

dan wordt de mantisse van x^*

$$|m^*| = x_1\beta^{-1} + x_2\beta^{-2} + \ldots + (x_k+1)\beta^{-k}.$$

(Als $x_k = \beta-1$ dan wordt het k-de gewicht gelijk aan nul gekozen en x_{k-1} met 1 verhoogd, met zonodig herhaling van dit proces).
In dit geval is

$$|x^*-x| = \beta^{-k-1+\sigma}\,(\beta-x_{k+1}-x_{k+2}\beta^{-1}\ldots.)$$

$$\leq \tfrac{1}{2}\,\beta^{\sigma-k}\ .$$

Voor beide gevallen geldt $|x-x^*| \leq \frac{1}{2}\,\beta^{\sigma-k}$.
Het verschil $x - x^*$ wordt de <u>absolute fout</u> genoemd.

De absolute waarde van de absolute fout is dus bij netjes afronden hoogstens $\frac{1}{2}\,\beta^{\sigma-k}$.

II. Een andere methode om de x^* te zoeken is de mantisse van x zonder meer af te breken, zodat de mantisse m^* van x^* wordt

$$|m^*| = x_1\beta^{-1} + \ldots\ldots\ldots + x_k\beta^{-k}\ .$$

De absolute fout is nu in absolute waarde

$$|x-x^*| = (x_{k+1}\beta^{-k-1}+x_{k+2}\beta^{-k-2}\ldots.)\ .$$

Aangezien voor alle gewichten geldt $x_i \leq \beta-1$ krijgen we nu

$$|x-x^*| \leq \beta^{\sigma-k-1}(\beta-1)(1+\beta^{-1}+\beta^{-2}+\ldots.)$$

$$= \beta^{\sigma-k}\ .$$

De representatie van x gaat bij deze methode dus gepaard met een absolute fout, die tweemaal zo groot kan zijn als die bij de eerste methode.

Vraagstukken.

1. Bepaal het waardebereik van de reële getallen bij een computer die werkt met grondtal β, terwijl voor de wijzer n posities beschikbaar zijn en voor de mantisse k.

2. Kan het getal 0 als floating point getal worden voorgesteld?

3. Bewijs dat bij netjes afronden geldt

$$-\tfrac{1}{2}\,\beta^{-k} < \frac{x - x^*}{|x|} \leq \tfrac{1}{2}\,\beta^{-k+1}\,.$$

Hierin hebben β, k, x en x^* dezelfde betekenis als in de tekst van deze paragraaf.
(De grootheid $(x-x^*)/x$ wordt de relatieve fout genoemd bij de benadering van x door x^*. We kunnen dus met de floating point notatie alle reële getallen binnen een bepaald getalbereik met een nagenoeg even grote relatieve precisie voorstellen.).

4. Een klein computertje gebruikt bij de representatie van de reële getallen de floating point notatie met $\beta = 2$. Voor de wijzer staan 2 bits ter beschikking plus 1 bit voor het teken; voor de mantisse ook 2 bits en eveneens 1 voor het teken. Bepaal alle standaardgetallen die als floating point getal kunnen worden gerepresenteerd.

5. Door welk standaardgetal x^* (zie vorig vraagstuk) wordt het getal $x = {}^1/_3$ benaderd
   I. als de machine netjes afrondt;
      bepaal de absolute en de relatieve fout.
   II. als de mantisse wordt afgebroken;
      bepaal de absolute en relatieve fout.

Lijst van errata

p. 1 midden, lees: $^{10}\log 3$
naast blok irrationele getallen: $\mathbb{Q}'$

p. 2 regel 1, lees: getallen - rechte

p. 4 regel 3, lees: $\beta \leq 10$

p. 5 vrst. 5, laat weg: in decimale representatie

p. 7 regel 4, "positieve" wordt "niet-negatieve"
regel 1 v.o., lees: $(.1101)_2$

p. 9 regel 2 en 1 v.o., lees: $(22.12)_3$ resp. $(.2212)_3 \times 3^2$

p. 13 regel 4 v.o., lees: $| x - x^* | = \beta^{\sigma}$ (