Wiskunde 2 voor kunstmatige intelligentie (BKI 316)

Bernd Souvignier

najaar 2005

Inhoud

I Voortgezette Analyse **3**

Les 1 Functies van meerdere variabelen 4
- 1.1 Continuïteit 5
- 1.2 Partiële afgeleide en richtingsafgeleide 10
- 1.3 De gradiënt 15
- 1.4 De algemene afgeleide 19

Les 2 Taylor reeksen 24
- 2.1 Interpolatie 24
- 2.2 Taylor veeltermen 26
- 2.3 Taylor reeksen 30
- 2.4 Taylor reeksen voor functies van meerdere variabelen . . . 34

Les 3 Extrema van functies van meerdere variabelen 45
- 3.1 Classificatie van kritieke punten 45
- 3.2 Kritieke punten van functies van meerdere variabelen . . 46
- 3.3 Criterium voor lokale extrema 49
- 3.4 Extrema onder randvoorwaarden 57
- 3.5 De methode van Lagrange multiplicatoren 59

Les 4 Integratie van functies van meerdere variabelen 66
- 4.1 Integratie op (veralgemeende) rechthoeken 66
- 4.2 Integratie over normaalgebieden 69
- 4.3 Substitutie 71
- 4.4 Poolcoördinaten, cilindercoördinaten, sferische coördinaten 76
- 4.5 Toepassingen 81

Les 5 Complexe getallen 85
- 5.1 Constructie van de complexe getallen 85
- 5.2 Oplossen van vergelijkingen 87
- 5.3 Meetkunde van de complexe getallen 88
- 5.4 Complexe conjugatie 93
- 5.5 Machtsverheffen 94
- 5.6 Toepassingen van de complexe getallen 95

Les 6 Complexe functies 98
- 6.1 Complexe exponentiële functie 100
- 6.2 Complexe sinus en cosinus functies 104
- 6.3 Complexe logaritme 107
- 6.4 Differentiëren via Taylor reeksen 109
- 6.5 Appendix: Complexe differentieerbaarheid 111

II Fourier theorie **119**
Les 7 Fourier analyse . . . 120
7.1 Periodieke functies . . . 121
7.2 Trigonometrische benadering . . . 122
7.3 Eigenschappen van de Fourier reeks . . . 127
7.4 Fase verschuivingen . . . 130
7.5 Complexe schrijfwijze . . . 130
7.6 Belangrijke voorbeelden . . . 132
Les 8 Fourier transformatie . . . 139
8.1 Periodieke functies met perioden verschillend van 2π . . . 139
8.2 Van Fourier reeks naar Fourier integraal . . . 140
8.3 Schrijfwijzen van de Fourier transformatie . . . 144
8.4 Eigenschappen van de Fourier transformatie . . . 146
8.5 Het convolutieproduct . . . 149
Les 9 Voorbeelden en toepassingen van de Fourier transformatie . . . 152
9.1 Belangrijke voorbeelden . . . 152
9.2 De Dirac δ-functie . . . 159
9.3 Toepassing: Filters . . . 162
Les 10 Discrete Fourier transformatie . . . 170
10.1 Discretisering . . . 170
10.2 De discrete Fourier transformatie . . . 171
10.3 Voorbeeld van een discrete Fourier transformatie . . . 176
10.4 Eigenschappen van de discrete Fourier transformatie . . . 178
10.5 Snelle (discrete) Fourier transformatie (FFT) . . . 180
10.6 Shannon's aftast-theorema . . . 183

III Probabilistische Modellen **191**
Les 11 Onzekerheid, entropie en informatie . . . 192
11.1 Onzekerheid . . . 192
11.2 Entropie van continue kansverdelingen . . . 198
11.3 Voorwaardelijke entropie . . . 202
11.4 Informatie . . . 205
11.5 Toepassing: Automatische Taalherkenning . . . 207
Les 12 Markov processen en Markov modellen . . . 215
12.1 Markov processen . . . 215
12.2 Stochastische automaten . . . 218
12.3 Markov modellen . . . 219
12.4 Toepassingen van Markov modellen . . . 221
12.5 Markov modellen met verborgen states . . . 224
Les 13 Hidden Markov modellen . . . 229
13.1 Evalueren met behulp van een HMM . . . 230
13.2 States onthullen . . . 233
13.3 Training van een HMM . . . 238
13.4 Toegift: Levenshtein afstand . . . 240

Deel I

Voortgezette Analyse

Les 1 Functies van meerdere variabelen

In het Calculus gedeelte van Wiskunde 1 hebben we ons bijna altijd beperkt tot functies van één variabel, dus functies van de vorm $y = f(x)$ die we makkelijk door hun grafiek in het $x-y$-vlak konden representeren. Helaas is de wereld niet zo eenvoudig dat zich alles door dit soort functies makkelijk laat beschrijven, denk bijvoorbeeld aan het volgende:

(1) Een steentje die je in een meer gooit zal een cirkelvormige golf veroorzaken, waarvan de hoogte van de afstand r van het centrum van de cirkel en van het tijdstip t waarop je kijkt afhangt. De hoogte h is dus een functie van r en t, bijvoorbeeld $h(r,t) = \sin(x+t)\,e^{-t}$.

(2) Voor een ideaal gas geldt (volgens het algemene gaswet van Gay-Lussac en Boyle) de relatie $V = \frac{nRT}{p}$ tussen het volume V, de temperatuur T, de hoeveelheid n van de stof (in mol) en de druk p, waarbij R de universele gasconstante is. Het volume is dus een functie $V = f(n,T,p) = \frac{nRT}{p}$ van de variabelen n, T en p, maar net zo goed is de druk een functie $p = g(n,T,V) = \frac{nRT}{V}$ van n, T en V.

Er zijn verschillende voor de hand liggende gevallen van functies van meerdere veranderlijken die we moeten bekijken:

(i) $f : \mathbb{R}^n \to \mathbb{R}$: Dit zijn functies die van meerdere (n) parameters afhangen, maar slechts één waarde als resultaat opleveren. Een voorbeeld is de functie die de afstand van een punt (x,y,z) in de 3-dimensionale ruimte van de oorsprong aangeeft, namelijk $f(x,y,z) := \sqrt{x^2+y^2+z^2}$.

(ii) $f : \mathbb{R} \to \mathbb{R}^m$: Dit zijn functies die maar van één variabel afhangen, maar meerdere waarden opleveren. Een voorbeeld is de functie $f(t) := (\cos(t), \sin(t))$ die het interval $[0, 2\pi]$ op de eenheidscirkel in het 2-dimensionale vlak afbeeldt.

(iii) $f : \mathbb{R}^n \to \mathbb{R}^m$: Dit is het algemeen geval, waarbij de functie van meerdere parameters afhangt en ook meerdere waarden oplevert. Een voorbeeld van dit soort functies zijn de lineaire functies van de n-dimensionale vectorruimte naar de m-dimensionale vectorruimte, maar ook de functie $f(x,y,z) := (\frac{x}{\sqrt{x^2+y^2+z^2}}, \frac{y}{\sqrt{x^2+y^2+z^2}}, \frac{z}{\sqrt{x^2+y^2+z^2}})$ die een kubus rond de oorsprong op de eenheidskogel afbeeldt.

Als we naar de algemene functies van type (iii) kijken, zien we dat het resultaat uit m componenten opgebouwd is, die zelf functies van de eenvoudigere type (ii) zijn. We kunnen namelijk elke functie $f : \mathbb{R}^n \to \mathbb{R}^m$ beschrijven door

$$f(x_1,\ldots,x_n) = (f_1(x_1,\ldots,x_n), f_2(x_1,\ldots,x_n), \ldots, f_m(x_1,\ldots,x_n)).$$

De functies $f_i(x_1,\ldots,x_m)$ noemen we de *componenten van* f en deze componenten zijn functies $f_i : \mathbb{R}^n \to \mathbb{R}$.

Voorbeeld: De functie $f(x,y,z) : \mathbb{R}^3 \to \mathbb{R}^3$ gegeven door

$$f(x,y,z) := (x\sin(y)\sin(z), x\sin(y)\cos(z), x\cos(y))$$

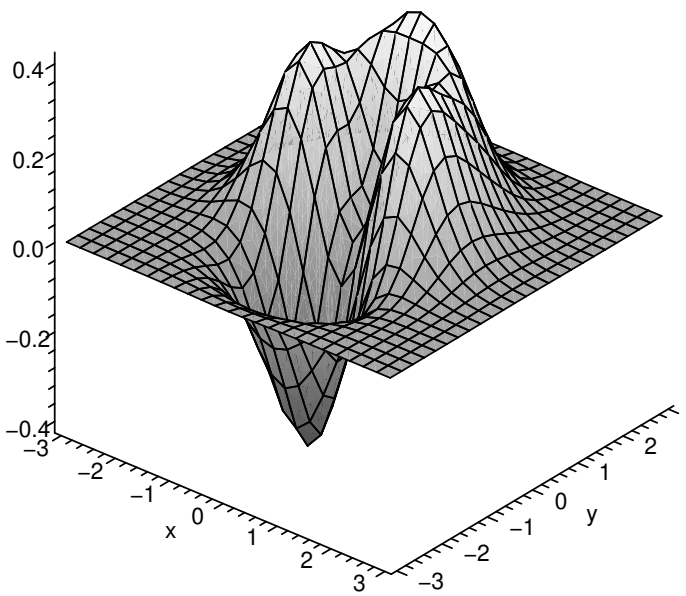
heeft de componenten

$$f_1(x,y,z) = x\sin(y)\sin(z),\ f_2(x,y,z) = x\sin(y)\cos(z),\ f_3(x,y,z) = x\cos(y).$$

Door naar de m componenten $f_i(x_1,\ldots,x_n)$ te kijken kunnen we ons dus meestal beperken tot het geval (ii) van functies die van meerdere variabelen afhangen, maar slechts één waarde opleveren.

Om het schrijfwerk te beperken zullen we vaak voorbeelden van functies van twee variabelen behandelen, dus $f(x,y) : \mathbb{R}^2 \to \mathbb{R}$. Hieraan worden de meeste ideeën wel duidelijk, de veralgemening op grotere aantallen van variabelen vergt meestal weinig nieuwe inzicht.

Een bijzonder voordeel van functies $f(x,y) : \mathbb{R}^2 \to \mathbb{R}$ is dat we hiervan nog net grafieken kunnen tekenen, namelijk door de punten $(x, y, f(x,y))$ in de 3-dimensionale ruimte te bekijken. De functiewaarden vormen een soort *gebergte* boven het $x-y$-vlak waarin we het domein van de functie vinden. In Figuur I.1 is bijvoorbeeld de grafiek van de functie $f(x,y) := (x^2+y^3)\,e^{-x^2-y^2}$ te zien.



Figuur I.1: Grafiek van een functie van twee variabelen.

1.1 Continuïteit

We zullen zien dat de behandeling van functies van meerdere veranderlijken vaak analoog met gewone functies van één veranderlijke loopt, maar er zijn ook belangrijke verschillen waarvan we ons bewust moeten zijn.

Een eerste vraag die we ons kunnen stellen, is, wanneer we een functie continu noemen. Bij een gewone functie hadden we dit intuïtief zo gedefinieerd,

dat een continue functie geen sprongen mag hebben. Bij functies van meerdere variabelen zouden we nu eerst moeten zeggen, wat een sprong eigenlijk is. Maar we hadden ook een meer formele definitie gegeven en deze kunnen we heel makkelijk naar functies van meerdere variabelen vertalen, in feite hoeven we alleen maar de absolute waarde op $\mathbb{R}$ te vervangen door de Euclidische afstand in de n-dimensionale ruimte.

Voor een gewone functie f van één variabele hadden we de volgende definitie gehanteerd:

Definitie: De functie $f : \mathbb{R} \to \mathbb{R}$ heet *continu in het punt* x als er voor iedere $\varepsilon > 0$ een $\delta > 0$ bestaat, zo dat $|f(x) - f(y)| < \varepsilon$ wanneer $|x - y| < \delta$.

In woorden betekent dit dat we voor een gekozen (klein) interval $[f(x) - \varepsilon, f(x) + \varepsilon]$ rond $f(x)$ een interval $[x - \delta, x + \delta]$ rond x kunnen vinden, waarvan de functiewaarden alle in het interval rond $f(x)$ liggen.

Precies hetzelfde idee passen we nu ook bij een functie $f : \mathbb{R}^n \to R$ van n variabelen toe. Het enige verschil is, dat het domein waarop f gedefinieerd is nu vectoren $\mathbf{x} \in \mathbb{R}^n$ in plaats van reële getallen bevat.

> We zullen in deze cursus vectoren altijd met vet gedrukt letters aanduiden, zoals $\mathbf{x}$, $\mathbf{y}$ of $\mathbf{v}$, in tegenstellingen tot gewone reële variabelen zo als x en y.

Van het Lineaire Algebra gedeelte van Wiskunde 1 weten we nog, hoe we de afstand van twee vectoren $\mathbf{x}$ en $\mathbf{y}$ in de n-dimensionale vectorruimte $\mathbb{R}^n$ bepalen, namelijk door de *Euclidische lengte* $\|\mathbf{x} - \mathbf{y}\|$ van de verschilvector $\mathbf{x} - \mathbf{y}$. We zeggen dus, dat de vector $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ dicht bij de vector $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ ligt, als de afstand

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

klein is. We zullen de functie f dus continu in het punt $\mathbf{x}$ noemen als voor alle vectoren $\mathbf{y}$ die dicht bij $\mathbf{x}$ liggen ook de functiewaarden $f(\mathbf{y})$ dicht bij $f(\mathbf{x})$ liggen. De precieze definitie luidt:

Definitie: Een functie $f : \mathbb{R}^n \to \mathbb{R}$ heet *continu in het punt* $\mathbf{x}$ als er voor iedere $\varepsilon > 0$ een $\delta > 0$ bestaat, zo dat $|f(\mathbf{x}) - f(\mathbf{y})| < \varepsilon$ wanneer $\|\mathbf{x} - \mathbf{y}\| < \delta$.

Als we nu echt willen aantonen dat een zekere functie continu is, zien we dat er wel verschillen zijn tussen functies van één en van meerdere variabelen. Voor een gewone functie van één variabel testen we continuïteit in principe zo:

We bepalen de limiet $\lim_{y \to x+} f(y)$ van $f(y)$ als we met y van rechts tegen x aanlopen, en de limiet $\lim_{y \to x-} f(y)$ van $f(y)$ als we met y van links tegen x aanlopen. Als deze twee limieten bestaan en dezelfde waarde hebben is de functie continu.

Bij een functie van meerdere variabelen is dit niet meer zo makkelijk. We kunnen namelijk uit elke willekeurige richting tegen $\mathbf{x}$ aanlopen. En dat hoeft

niet eens op een rechte lijn te gebeuren, we kunnen ook in een spiraal rond $\mathbf{x}$ lopen, in een zigzag of langs elke willekeurige kromme die uiteindelijk steeds dichter bij $\mathbf{x}$ komt. Omdat er zo veel mogelijkheden zijn, komt het er op neer dat men de continuïteit rechtstreeks met de definitie bewijst, die onafhankelijk van een gekozen pad naar $\mathbf{x}$ toe is.

Maar meestal is er een slimmere methode om de continuïteit van een functie te bewijzen: Men heeft ergens een lijst van heel eenvoudige functies waarvan de continuïteit bekend is, bijvoorbeeld veeltermfuncties zo als $f(x,y) = x^2 - 3xy + 2y^3$ en standaardfuncties zo als de *exponentiële functie*, de *sinus* en de *cosinus*. De grap is nu dat sommen producten en samenstellingen van continue functies ook weer continu zijn. Voor de meeste functies volgt dus de continuïteit heel makkelijk, omdat ze met deze operaties uit eenvoudige continue functies opgebouwd kunnen worden. De volgende stelling geeft de precieze voorwaarden aan:

Stelling: Laten $f(\mathbf{x})$, $g(\mathbf{x})$ functies zijn die op een gemeenschappelijk domein $D \subseteq \mathbb{R}^n$ continu zijn.

(i) De som $f(\mathbf{x}) + g(\mathbf{x})$, het product $f(\mathbf{x}) \cdot g(\mathbf{x})$ en de schaling $c \cdot f(\mathbf{x})$ (met $c \in \mathbb{R}$) zijn continu op D.

(ii) Als $f(\mathbf{x}) \neq 0$ op D, dan is ook $\frac{1}{f(\mathbf{x})}$ continu op D.

(iii) Zij $h(x) : \mathbb{R} \to \mathbb{R}$ een functie die continu op $I \subseteq \mathbb{R}$ is en stel dat $f(\mathbf{x}) \in I$ voor alle $\mathbf{x} \in D$. Dan is de samenstelling $h(f(\mathbf{x}))$ continu op D.

Uit punt (iii) volgt in het bijzonder dat functies zo als $e^{x^2+y^2}$ of $\sin(xy - z)$ continu zijn. Problemen leveren meestal alleen maar breuken op, waar de noemer niet 0 mag worden, en de *logaritme* en wortels, die een positief argument moeten hebben. Daarom is bijvoorbeeld de functie $\log(1 - x^2 - y^2)$ alleen maar continu op het gebied waar $f(x,y) = 1 - x^2 - y^2 > 0$ is, d.w.z. voor $x^2 + y^2 < 1$, met andere woorden binnen een cirkel met straal 1 rond $(0,0)$.

Soms kan een functie met nulpunten in de noemer wel door een geschikte definitie van functiewaarden continu naar de nulpunten van de noemer voortgezet worden. Voor gewone functies kennen we dit al: De functie $f(x) := \frac{x^2-1}{x-1}$ is voor $x = 1$ niet gedefinieerd, maar omdat voor $x \neq 1$ geldt dat $f(x) = \frac{(x-1)(x+1)}{x-1} = x + 1$, laat zich $f(x)$ door $f(1) := 2$ tot een continue functie voortzetten.

Om een functie continu naar een nulpunt van de noemer voort te zetten, is het noodzakelijk dat de nulpunten van de noemer ook nulpunten van de teller zijn (want anders gaat de functie naar oneindig). In zo'n geval moet onderzocht worden, hoe de functie zich in de buurt van de nulpunt precies gedraagt.

Voorbeelden:

(1) De functie $f(x,y) := \frac{x^2}{\sqrt{x^2+y^2}}$ laat zich door $f(0,0) := 0$ tot een continue functie voortzetten. Er geldt namelijk voor $(x,y) \neq (0,0)$ dat $\frac{x^2}{\sqrt{x^2+y^2}} \leq \frac{x^2+y^2}{\sqrt{x^2+y^2}} = \sqrt{x^2+y^2}$ en voor $(x,y) \to (0,0)$ gaat $\sqrt{x^2+y^2} \to 0$.

(2) De functie $f(x,y) := \frac{x^2-y^2}{x-y}$ is op de lijn met $x = y$ niet gedefinieerd. Maar voor $x \neq y$ geldt dat $f(x,y) = \frac{(x-y)(x+y)}{x-y} = x + y$. Daarom laat zich $f(x,y)$ met $f(x,x) := 2x$ tot de lijn met $x = y$ voortzetten.

(3) De functie $f(x,y) := \frac{\sin(xy)}{x^2+y^2}$ laat zich niet continu naar $(x,y) = (0,0)$ voortzetten. In Wiskunde 1 hadden we namelijk gezien dat uit $\sin(x)' = \cos(x)$ volgt dat $\lim_{x\to 0} \frac{\sin(x)-\sin(0)}{x} = \cos(0) = 1$, en hieruit volgt dat op de lijn $x = y$ geldt dat $f(x,x) = \frac{\sin(x^2)}{2x^2} \to \frac{1}{2}$ voor $x \to 0$. Maar als we op de x-as tegen $(0,0)$ aan lopen, is $f(x,0) = \frac{\sin(0)}{x^2} = 0$. We zouden de functie dus tegelijkertijd met $f(0,0) = \frac{1}{2}$ en met $f(0,0) = 0$ moeten voortzetten en dit is natuurlijk onmogelijk.

Opdracht 1

(i) Laat zien dat de functie $f(x,y) := \frac{x^3+2x^2+xy^2+2y^2}{x^2+y^2}$ een continue voortzetting naar het punt $(0,0)$ heeft. (Hint: Probeer de noemer als factor van de teller te vinden.)

(ii) Hoe moet de functie $f(x,y) := \frac{x^3-y^3}{x-y}$ op de lijn $x = y$ gedefinieerd worden zo dat de functie op het hele $x-y$-vlak continu is?

(iii) Ga na dat de functie $f(x,y,z) := \frac{\sin(xyz)}{xyz}$ voor $(x,y,z) \to (0,0,0)$ tegen de waarde 1 gaat. (Hint: Denk aan de limiet $\frac{\sin(x)}{x}$ voor $x \to 0$.)

Een afschrikkend voorbeeld: De grotere vrijheid van paden die men in $\mathbb{R}^n$ tegenover $\mathbb{R}$ heeft, heeft wel soms verrassende effecten. We zullen hier een voorbeeld van bekijken, namelijk de functie

$$f(x,y) := \frac{(y^2-x)^2}{y^4+x^2}$$

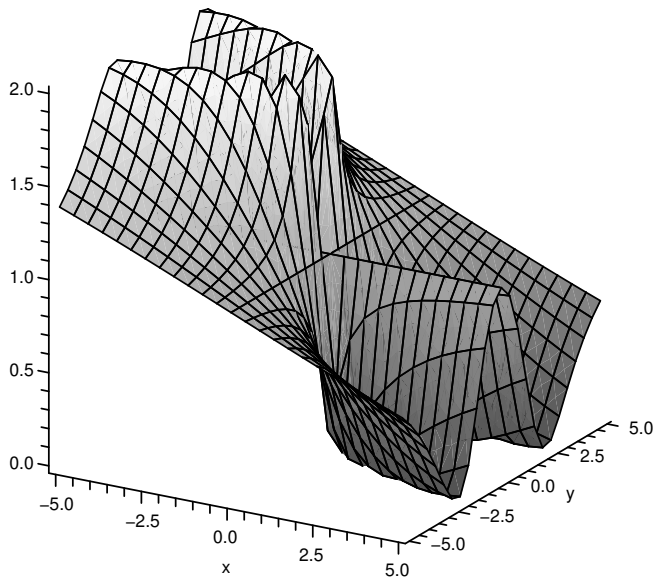
waarvan de grafiek in Figuur I.2 te zien valt.

Voor $(x,y) = (0,0)$ is $f(x,y)$ natuurlijk niet gedefinieerd, maar het zou kunnen dat zich $f(x,y)$ door een geschikte definitie van $f(0,0)$ tot een continue functie laat voortzetten.

Als we langs de x-as naar $(0,0)$ lopen is $y = 0$, dus $f(x,y) = \frac{x^2}{x^2} = 1$. Lopen we langs de y-as, is $x = 0$ en $f(x,y) = \frac{y^4}{y^4} = 1$. Als de functie überhaupt een continue voortzetting heeft, dan moet deze dus noodzakelijk de waarde $f(0,0) = 1$ hebben.

Als we nu langs een willekeurige lijn lopen (behalve de x- en y-as) kunnen we dit door de punten $(x,y) = (t,ct)$ beschrijven, die op de lijn met $y = cx$ liggen en voor $t \to 0$ lopen we op zo'n lijn tegen $(0,0)$ aan. Als we (t,ct) voor (x,y) invullen, krijgen we

$$\begin{aligned} f(x,y) &= \frac{(c^2t^2-t)^2}{c^4t^4+t^2} = \frac{c^4t^4+t^2-2c^2t^3}{c^4t^4+t^2} = 1 - \frac{2c^2t^3}{c^4t^4+t^2} \\ &= 1 - \frac{2c^2t}{c^4t^2+1} \to 1 - \frac{0}{0+1} = 1 \text{ voor } t \to 0. \end{aligned}$$



Figuur I.2: Grafiek van de functie $f(x,y) := \frac{(y^2-x)^2}{y^4+x^2}$.

We hebben dus aangetoond dat $f(x,y)$ op elke lijn naar $(0,0)$ de limiet 1 heeft. Ook al zouden we nu misschien denken, dat de functie zich inderdaad met $f(0,0) = 1$ tot een continue functie laat voortzetten, laten we nu zien dat $f(x,y)$ met deze voortzetting *niet continu* in $(0,0)$ is.

Als we namelijk nog eens goed naar de teller $(y^2 - x)^2$ van $f(x,y)$ kijken, zien we dat die voor $x = y^2$ gelijk aan 0 is. Dit betekent dat voor alle punten op de parabool $(x,y) = (t^2,t)$ de functiewaarde 0 is. Maar voor $t \to 0$ lopen we op deze parabool ook tegen het punt $(0,0)$ aan, dus zijn er punten die willekeurig dicht bij $(0,0)$ liggen waarvoor $f(x,y)$ de waarde 0 heeft en die ligt helaas niet dicht bij 1 (neem bijvoorbeeld $\varepsilon = \frac{1}{2}$).

Nu dat we het weten, kunnen we ook in Figuur I.2 zien, dat er onderaan een soort rand is, waar de functie op de functiewaarde 0 blijft, terwijl alle lijnen door $(x,y) = (0,0)$ inderdaad door de functiewaarde 1 gaan.

Het voorbeeld laat zien dat de vraag of een functie continu is bij functies van meerdere variabelen soms enigszins subtiel kan zijn. Maar omdat we in de praktijk nauwelijks functies tegen komen die niet continu zijn, zullen we hier verder geen aandacht meer aan besteden.

Opdracht 2 *Laat zien dat de functie* $f(x,y) := \frac{x^2-y^2}{x^2+y^2}$ *geen continue voortzetting naar* $(x,y) = (0,0)$ *heeft. (Hint: Bekijk verschillende lijnen door het nulpunt.)*

1.2 Partiële afgeleide en richtingsafgeleide

Bij functies van één veranderlijke hebben we gezien dat de afgeleide van een functie de snelheid aangeeft waarmee de functie in een punt verandert. Dit was heel nuttig om minima en maxima van functies te vinden, maar ook om te zien of de functie stijgt of daalt en waar de punten zijn waar ze het snelste stijgt. Een van de interpretaties van de afgeleide was, dat hij de richtingscoëfficiënt van de raaklijn aan de grafiek van de functie aangeeft. Dit kunnen we helaas niet rechtstreeks op functies van meerdere veranderlijken veralgemenen. Zo is bijvoorbeeld de grafiek van een functie van twee veranderlijken (met één waarde) een soort gebergte boven het $x-y$-vlak waar men in een punt voor elke richting van het $x-y$-vlak een raaklijn kan aanleggen, en voor verschillende richtingen zullen deze raaklijnen zeker verschillende stijgingen hebben. We moeten daarom iets beter kijken, wat we als afgeleide willen definiëren.

Een eerste idee (die we in Wiskunde 1 al kort hebben bekeken) is, dat we de richtingen van de raaklijnen op richtingen langs de coördinaatassen beperken. Als we bij een functie $f(x,y)$ de raaklijn in het punt (x_0,y_0) in de richting van de x-as willen bepalen, leggen we een lijn door de punten (x_0,y_0) en (x_0+h,y_0) en laten h tegen 0 lopen. Bij dit proces blijft y_0 altijd ongedeerd, want omdat we de raaklijn in de richting van de x-as bepalen, moet de y-waarde constant blijven. Maar als de y-waarde de vaste waarde y_0 heeft, is de functie $f(x,y)$ eigenlijk een functie $g(x)=f(x,y_0)$ van slechts één variabel.

Voor een functie van twee variabelen kunnen we dit idee ook grafisch illustreren: De grafiek van zo'n functie kunnen we zien als de verzameling van punten $(x,y,z=f(x,y))$ in de 3-dimensionale ruimte, net zo als we de grafiek van een gewone functie als de verzameling van punten $(x,y=f(x))$ in het 2-dimensionale vlak bekijken. Als we nu y tot een constante y_0 verklaren, dan kijken we naar de doorsnede van de grafiek $(x,y,f(x,y))$ met het vlak dat bepaald is door de vergelijking $y=y_0$, dus we bekijken de punten $(x,y_0,f(x,y_0))$. Maar de tweede coördinaat hierbij is natuurlijk volstrekt overbodig, en als we deze schrappen, houden we de punten $(x,f(x,y_0))$ over. Dit zijn gewoon de punten van de grafiek van de functie $g(x):=f(x,y_0)$ van één variabel (want y_0 is een constante).

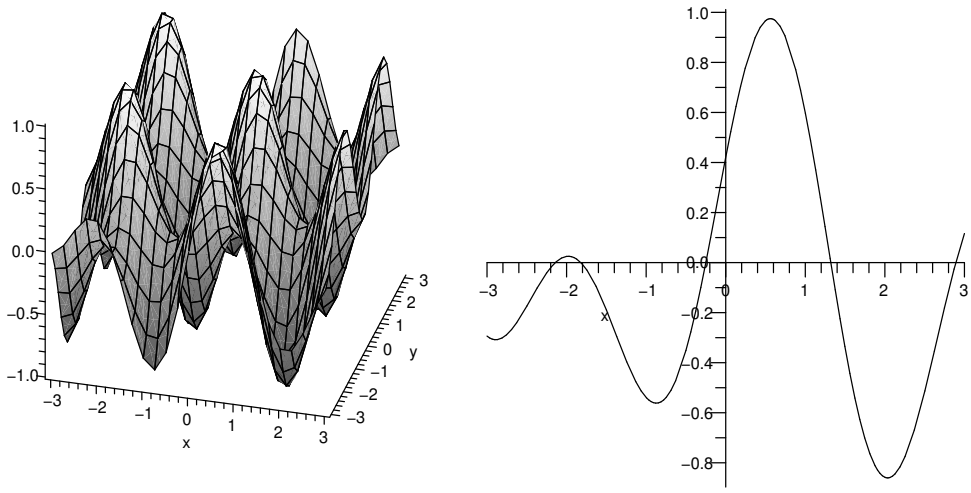
De plaatjes in Figuur I.3 laten links de functie $f(x,y):=\sin(2x+y)\cos(\frac{x}{2}-y)$ en rechts de doorsnede door deze grafiek voor $y=\frac{1}{2}$ zien, dus de functie $g(x):=f(x,\frac{1}{2})=\sin(2x+\frac{1}{2})\cos(\frac{x}{2}-\frac{1}{2})$.

Het is nu voor de hand liggend hoe we de afgeleide in het punt (x_0,y_0) in de richting van de x-as definiëren, namelijk gewoon als de richtingscoëfficiënt van de raaklijn aan de gewone functie $g(x)=f(x,y_0)$ in het punt x_0. De afgeleide in de richting van de x-as noemen we de *partiële afgeleide* van f naar x.

Als we ons nu aan de definitie

$$g'(x)=\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}$$

van de gewone afgeleide van een functie van één variabel herinneren, kunnen we rechtstreeks de definitie voor de partiële afgeleide in dit geval aangeven, dit



Figuur I.3: Functie $f(x, y)$ en doorsnede door deze functie voor $y = 0.5$

is namelijk de limiet

$$\lim_{h\to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}.$$

Natuurlijk is er geen reden, waarom de x-as beter dan de y-as zou zijn, we kunnen net zo goed ook de raaklijn in de richting van de y-as bekijken en krijgen dan de partiële afgeleide van f naar y. Maar nu dat we het idee hebben gezien dat tot op één na alle variabelen tot constanten verklaard worden, kunnen we ook meteen de algemene definitie van de partiële afgeleide geven.

Definitie: Voor een functie $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ definiëren we de *partiële afgeleide* van f naar de variabel x_i als de limiet

$$\lim_{h\to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}$$

als deze limiet bestaat. De partiële afgeleide wordt meestal met $\frac{\partial f}{\partial x_i}$ genoteerd, maar vaak ook kort als f_{x_i} geschreven.

Merk op: Om een partiële afgeleide uit te rekenen, gebruiken we nooit deze limiet-definitie. Omdat we alleen maar de variabel x_i gaan veranderen, kunnen we de andere variabelen $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ als constanten (net zo als de constanten 5 of π in de functie $g(x) = e^{5x} \sin(\pi x)$) behandelen. Op deze manier interpreteren we de functie $f(x_1, \ldots, x_n)$ als een nieuwe functie $g(x_i)$ van één variabel (waarin de andere variabelen x_j als constanten voorkomen) en deze functie leiden we nu met de bekende rekenregels naar zijn variabel x_i af.

Voorbeelden:

(1) Zij $f(x, y)$ gegeven door $f(x, y) := x^3 y + e^{xy^2}$. Dan geldt (let op de kettingregel):

$$\frac{\partial f}{\partial x} = 3x^2 y + y^2\, e^{xy^2} \quad \text{en} \quad \frac{\partial f}{\partial y} = x^3 + 2xy\, e^{xy^2}.$$

(2) Zij $f(x,y,z)$ de functie van drie variabelen, gegeven door $f(x,y,z) := x\log(yz)$, dan geldt:

$$\frac{\partial f}{\partial x} = \log(yz), \quad \frac{\partial f}{\partial y} = x(yz)^{-1}z = \frac{x}{y}, \quad \frac{\partial f}{\partial z} = x(yz)^{-1}y = \frac{x}{z}.$$

OPDRACHT 3 *Bepaal de partiële afgeleiden* $\frac{\partial f}{\partial x}$ *en* $\frac{\partial f}{\partial y}$ *van de functies* $f(x,y) := 2x^2 - xy + y^2$ *en* $g(x,y) := \cos(xy) + x\cos(y)$.

Merk op: De partiële afgeleiden $\frac{\partial f}{\partial x_i}$ zijn zelfs ook weer functies van $\mathbb{R}^n$ naar $\mathbb{R}$, want ook al behandelen we $x_2, x_3, \ldots$ bij het afleiden naar x_1 als constanten, heeft de partiële afgeleide afhankelijk van deze constanten toch verschillende waarden.

Als we voor een functie $f(x,y)$ van twee variabelen de waarde van $\frac{\partial f}{\partial x}$ in een bepaald punt (x_0, y_0) willen aanduiden, noteren we dit met $\frac{\partial f}{\partial x}(x_0, y_0)$ of $\frac{\partial f}{\partial x}|_{x_0,y_0}$. Net zo schrijven we bij een algemene functie $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ van n variabelen $\frac{\partial f}{\partial x_i}(\mathbf{x}_0)$ of $\frac{\partial f}{\partial x_i}|_{\mathbf{x}_0}$ voor de waarde van de i-de partiële afgeleide in het punt $\mathbf{x}_0$.

Voorbeeld: Voor $f(x,y) := x^2y + y^3$ geldt $\frac{\partial f}{\partial x} = 2xy$ en $\frac{\partial f}{\partial y} = x^2 + 3y^2$. De waarden van de twee partiële afgeleiden in het punt $(x,y) = (1,2)$ zijn $\frac{\partial f}{\partial x}(1,2) = 2 \cdot 1 \cdot 2 = 4$ en $\frac{\partial f}{\partial y}(1,2) = 1^2 + 3 \cdot 2^2 = 13$.

Hogere partiële afgeleiden

Net zo als we bij gewone functies de afgeleide van de afgeleide kunnen bepalen en zo tot de hogere afgeleiden $f''(x), f'''(x), f^{(i)}(x)$ komen, kunnen we ook partiële afgeleide itereren. Hierbij hebben we echter veel meer keuze want we kunnen iedere keer een andere variabele kiezen waar we naar afleiden.

In het voorbeeld met $f(x,y,z) = x\log(yz)$ hadden we de partiële afgeleide $\frac{\partial f}{\partial x} = \log(yz)$ gevonden. Als we dit bijvoorbeeld partieel naar y afleiden, schrijven we dit als

$$\frac{\partial}{\partial y}(\frac{\partial f}{\partial x}) = \frac{\partial^2 f}{\partial y\,\partial x} = \frac{1}{yz}z = \frac{1}{y}.$$

De alternatieve notatie f_x in plaats van $\frac{\partial f}{\partial x}$ voor de partiële afgeleide geeft aanleiding tot de notatie $f_{xy} = (f_x)_y = \frac{\partial^2 f}{\partial y\,\partial x}$ voor de geïtereerde partiële afgeleide. Het lijkt erg onhandig dat de volgorde van de variabelen in de twee notaties verruild is, maar we zullen straks zien dat dit geen probleem voorstelt.

Als we meerdere keer naar dezelfde variabel afleiden, is er een verdere notatie gebruikelijk: $\frac{\partial^2 f}{\partial x\,\partial x} = f_{xx}$ wordt kort geschreven als $\frac{\partial^2 f}{\partial x^2}$. Merk op dat hier niet naar een nieuwe variabel x^2 wordt afgeleid, maar twee keer naar de variabel x (de ∂^2 in de teller geeft aan dat het om een tweede afgeleide gaat).

Als we voor een functie van twee variabelen de tweede partiële afgeleiden bepalen, hebben we hiervoor $2^2 = 4$ mogelijkheden, want voor de eerste en voor de tweede afgeleide kunnen we telkens een van de twee variabelen kiezen. Bij een functie van drie variabelen krijgen we op deze manier zelfs $3^2 = 9$ tweede afgeleiden en bij een functie van n variabelen zijn het er n^2.

Voorbeeld:

(1) Voor de functie $f(x,y) := x^3y + e^{xy^2}$ (waarvan we boven al de eerste partiële afgeleiden hebben bepaald) krijgen we de volgende tweede partiële afgeleiden:

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}(3x^2y + y^2\, e^{xy^2}) = 6xy + y^2\; y^2\, e^{xy^2} = 6xy + y^4\, e^{xy^2} \\
\frac{\partial^2 f}{\partial y\, \partial x} &= \frac{\partial}{\partial y}(3x^2y + y^2\, e^{xy^2}) = 3x^2 + 2y\, e^{xy^2} + y^2\; 2xy\, e^{xy^2} \\
&= 3x^2 + (2y + 2xy^3)\, e^{xy^2} \\
\frac{\partial^2 f}{\partial x\, \partial y} &= \frac{\partial}{\partial x}(x^3 + 2xy\, e^{xy^2}) = 3x^2 + 2y\, e^{xy^2} + 2xy\; y^2\, e^{xy^2} \\
&= 3x^2 + (2y + 2xy^3)\, e^{xy^2} \\
\frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}(x^3 + 2xy\, e^{xy^2}) = 2x\, e^{xy^2} + 2xy\; 2xy\, e^{xy^2} = (2x + 4x^2y^2)\, e^{xy^2}
\end{aligned}$$

(2) Voor de functie $f(x,y,z) := x\log(yz)$ schrijven we de tweede partiële afgeleiden in een 3×3 schema, waarbij de rijen met de variabel voor de eerste afgeleide en de kolommen met de variabel voor de tweede afgeleide corresponderen. Het schema ziet er als volgt uit:

$$\begin{array}{ccc}
\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\log(yz) = 0 & \frac{\partial^2 f}{\partial y\, \partial x} = \frac{\partial}{\partial y}\log(yz) = \frac{1}{y} & \frac{\partial^2 f}{\partial z\, \partial x} = \frac{\partial}{\partial z}\log(yz) = \frac{1}{z} \\
\frac{\partial^2 f}{\partial x\, \partial y} = \frac{\partial}{\partial x}\frac{x}{y} = \frac{1}{y} & \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\frac{x}{y} = -\frac{x}{y^2} & \frac{\partial^2 f}{\partial z\, \partial y} = \frac{\partial}{\partial z}\frac{x}{y} = 0 \\
\frac{\partial^2 f}{\partial x\, \partial z} = \frac{\partial}{\partial x}\frac{x}{z} = \frac{1}{z} & \frac{\partial^2 f}{\partial y\, \partial z} = \frac{\partial}{\partial y}\frac{x}{z} = 0 & \frac{\partial^2 f}{\partial z^2} = \frac{\partial}{\partial z}\frac{x}{z} = -\frac{x}{z^2}
\end{array}$$

Merk op: In beide voorbeelden kunnen we constateren, dat het geen verschil maakt of we eerst partieel naar x en dan naar y afleiden, of andersom, en dit geldt ook voor alle andere paren van variabelen. Na twee voorbeelden te controleren geloven we natuurlijk niet meer aan een toeval, maar het is ook niet vanzelfsprekend dat dit daadwerkelijk altijd zou gelden. Hier zit inderdaad een serieuze stelling achter, de *Stelling van Schwarz*, die een niet helemaal triviaal bewijs heeft (die door Leonhard Euler werd gevonden). Gelukkig is de uitspraak van de stelling wel zo eenvoudig als we maar zouden kunnen hopen, namelijk dat we (onder zwakke voorwaarden) de volgorde van partiële afgeleiden mogen verruilen.

Stelling van Schwarz: Als voor een functie $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ de tweede partiële afgeleiden $\frac{\partial^2 f}{\partial x_i\, \partial x_j}$ en $\frac{\partial^2 f}{\partial x_j\, \partial x_i}$ bestaan en continu zijn, dan zijn ze gelijk. De volgorde van de partiële afgeleiden speelt in dit geval dus geen rol.

Als we eens ervan uit gaan dat we het altijd met voldoende goedachtige functies te maken hebben (met continue hogere partiële afgeleiden, dus), hoeven we niet op de volgorde van de partiële afgeleiden te letten. We hoeven ons dus ook van de verwarring over de notaties $\frac{\partial^2 f}{\partial y\,\partial x} = f_{xy}$ niets aan te trekken.

We merken nog op dat we door herhaalde toepassing van de Stelling van Schwarz ook voor derde, vierde en hogere partiële afgeleiden kunnen laten zien dat de volgorde van de afgeleiden geen rol speelt. Er geldt dus bijvoorbeeld dat

$$\frac{\partial^3 f}{\partial x\,\partial y\,\partial z} = \frac{\partial^2}{\partial x\,\partial y}(\frac{\partial f}{\partial z}) = \frac{\partial^2}{\partial y\,\partial x}(\frac{\partial f}{\partial z}) = \frac{\partial}{\partial y}(\frac{\partial^2 f}{\partial x\,\partial z}) = \frac{\partial}{\partial y}(\frac{\partial^2 f}{\partial z\,\partial x}) = \frac{\partial^3 f}{\partial y\,\partial z\,\partial x}.$$

Op eenzelfde manier laat zich aantonen dat

$$\frac{\partial^4 g}{\partial y\,\partial x^3} = \frac{\partial^4 g}{\partial x\,\partial y\,\partial x^2} = \frac{\partial^4 g}{\partial x^2\,\partial y\,\partial x} = \frac{\partial^4 g}{\partial x^3\,\partial y}.$$

Opdracht 4 *Ga voor $f(x,y,z) := z\,e^{xy} + yz^3x^2$ door expliciet partieel af te leiden na dat $\frac{\partial^3 f}{\partial x\,\partial y\,\partial z} = \frac{\partial^3 f}{\partial z\,\partial y\,\partial x}$.*

De richtingsafgeleide

Bij de partiële afgeleide hebben we de raaklijnen in de richting van de coördinaatassen bekeken. Het is natuurlijk net zo goed toegestaan de raaklijn in een andere richting te bekijken, bijvoorbeeld als men de verandering van de functie in de richting van de lijn met $x = y$ wil weten. Algemeen kunnen we een richting steeds met een *richtingsvector* $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ in het domein $D \subseteq \mathbb{R}^n$ van de functie $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ aangeven. Hierbij nemen we altijd aan dat $\mathbf{v}$ een vector van lengte 1 is, dus dat $\|\mathbf{v}\| = \sqrt{v_1^2 + \ldots + v_n^2} = 1$.

Als we nu een raaklijn in de richting van zo'n richtingsvector $\mathbf{v}$ willen bepalen, bekijken we net als bij de partiële afgeleide de lijn door de punten $\mathbf{x}$ en $\mathbf{x} + h \cdot \mathbf{v}$ en laten h tegen 0 gaan. Hierbij is het belangrijk dat de richtingsvector $\mathbf{v}$ op lengte 1 genormeerd is. We noemen dan de limiet

$$\lim_{h \to 0} \frac{f(\mathbf{x} + h \cdot \mathbf{v}) - f(\mathbf{x})}{h}$$

de *richtingsafgeleide* van f in de richting $\mathbf{v}$ en noteren dit met $\frac{\partial f}{\partial \mathbf{v}}$.

De gewone partiële afgeleiden vinden we nu als speciale richtingsafgeleiden terug, namelijk als richtingsafgeleiden met betrekking tot de basisvectoren $\mathbf{e}_i$ van de standaardbasis $B = \left\{ \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\}$.

Als we de partiële afgeleiden van een functie al kennen, kunnen we hieruit de richtingsafgeleide voor een willekeurige richting makkelijk berekenen, er geldt namelijk voor $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$:

$$\frac{\partial f}{\partial \mathbf{v}} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot v_i = \frac{\partial f}{\partial x_1} \cdot v_1 + \ldots + \frac{\partial f}{\partial x_n} \cdot v_n.$$

Helaas zullen we de reden voor deze samenhang pas in de volgende les nader toelichten.

Voorbeeld: We bepalen de richtingsafgeleide van de functie $f(x, y, z) := xyz$ in de richting van de vector $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$. Omdat deze vector lengte $\sqrt{2}$ heeft, hebben we de richtingsvector $\mathbf{v} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ nodig. Voor de partiële afgeleiden geldt $\frac{\partial f}{\partial x} = yz$, $\frac{\partial f}{\partial y} = xz$ en $\frac{\partial f}{\partial z} = xy$, dus krijgen we de richtingsafgeleide $\frac{\partial f}{\partial \mathbf{v}} = yz \cdot \frac{1}{\sqrt{2}} + yz \cdot 0 + xy \cdot \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}(yz + xy)$.

Opdracht 5 *Bepaal de richtingsafgeleide van $f(x, y, z) := e^x + yz$ in de richting van de vector $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ (let op dat de vector niet lengte 1 heeft).*

1.3 De gradiënt

We zijn nu klaar om een van de meest belangrijke begrippen voor functies van meerdere variabelen te behandelen, namelijk de *gradiënt* van een functie. Dit is eigenlijk helemaal niets nieuws, we schrijven gewoon de verschillende partiële afgeleiden $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$ van een functie $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ in een vector:

$$\nabla f(\mathbf{x}) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

heet de *gradiënt van f in* $\mathbf{x}$, waarbij het symbool ∇ als *nabla* te lezen is. Soms wordt ook de notatie $\operatorname{grad} f$ in plaats van ∇f gehanteerd.

De gradiënt van een functie geeft voor ieder punt $\mathbf{x}$ in het domein van de functie een vector aan, dus is ∇f zelf een functie van $\mathbb{R}^n$ naar $\mathbb{R}^n$.

Voorbeeld: Zij $f(x, y) := e^{xy} + \sin(xy)$, dan is $\frac{\partial f}{\partial x} = y\, e^{xy} + y \cos(xy)$ en $\frac{\partial f}{\partial y} = x\, e^{xy} + x \cos(xy)$, dus

$$\nabla f(x, y) = \begin{pmatrix} y\, e^{xy} + y \cos(xy) \\ x\, e^{xy} + x \cos(xy) \end{pmatrix} = (e^{xy} + \cos(xy)) \begin{pmatrix} y \\ x \end{pmatrix}.$$

Opdracht 6 *Bepaal de gradiënten van de functies* $f(x,y) := x\,e^{x^2+y^2}$ *en* $g(x,y,z) := xy^2 + yz^2 + zx^2$.

Als we nu nog eens naar de richtingsafgeleide in de richting van de vector $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ kijken, kunnen we dit met behulp van de gradiënt herschrijven als

$$\frac{\partial f}{\partial \mathbf{v}} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot v_i = \frac{\partial f}{\partial x_1} \cdot v_1 + \ldots + \frac{\partial f}{\partial x_n} \cdot v_n = \nabla f \cdot \mathbf{v}$$

waarbij we met het product $\mathbf{a} \cdot \mathbf{b}$ van twee vectoren $\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ en $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ het *(standaard)inproduct*

$$\sum_{i=1}^{n} a_i b_i = a_1 b_1 + \ldots + a_n b_n$$

van deze twee vectoren bedoelen. Dus:

Merk op: De richtingsafgeleide $\frac{\partial f}{\partial \mathbf{v}}$ van een functie $f(\mathbf{x})$ in de richting van de vector $\mathbf{v}$ (van lengte 1) is het inproduct $\nabla f \cdot \mathbf{v}$ van $\mathbf{v}$ met de gradiënt ∇f.

Als we deze inzicht combineren met het feit dat de richtingsafgeleide de verandering van de functie in de richting van $\mathbf{v}$ aangeeft, komen we al een heel stuk verder met de interpretatie van de gradiënt. Als we namelijk het inproduct van de gradiënt ∇f met alle vectoren $\mathbf{v}$ met lengte 1 bekijken, kunnen we meteen zeggen voor welke $\mathbf{v}$ het inproduct de maximale waarde aanneemt. In Wiskunde 1 hadden we namelijk gezien dat voor het inproduct $\nabla f \cdot \mathbf{v}$ geldt dat

$$\nabla f \cdot \mathbf{v} = \|\nabla f\| \cdot \|\mathbf{v}\| \cdot \cos(\varphi)$$

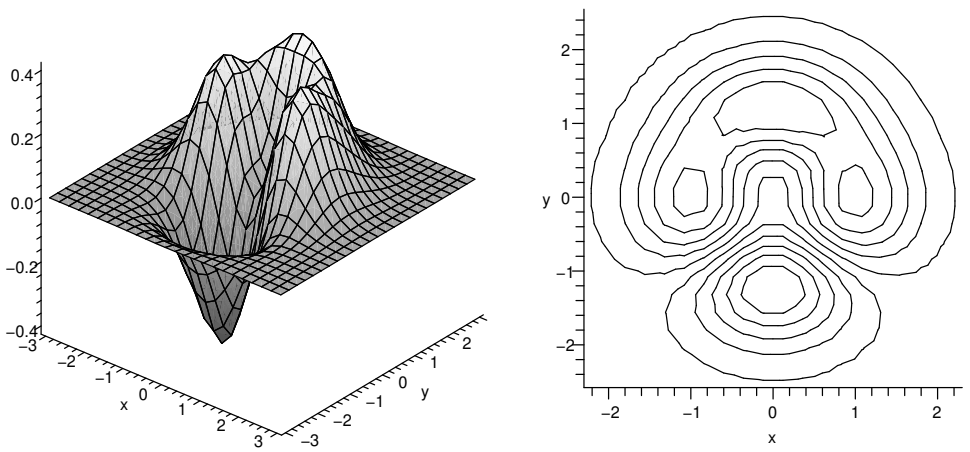
waarbij φ de hoek tussen de vectoren ∇f en $\mathbf{v}$ is. Maar omdat de richtingsvectoren $\mathbf{v}$ alle dezelfde lengte $\|\mathbf{v}\| = 1$ hebben, geldt $\nabla f \cdot \mathbf{v} = \|\nabla f\| \cdot \cos(\varphi)$ en dit is maximaal als $\cos(\varphi) = 1$, dus als $\varphi = 0$. Het inproduct is dus maximaal als $\mathbf{v}$ precies in de richting van ∇f wijst. Omgekeerd is het inproduct minimaal als $\cos(\varphi) = -1$, dus als $\varphi = 180°$ en dit betekent dat $\mathbf{v}$ in de tegengestelde richting van ∇f wijst. We hebben dus de volgende stelling ingezien:

Stelling: De gradiënt $\nabla f(\mathbf{x})$ wijst in de richting van de maximale toename van de functie f in het punt $\mathbf{x}$. De tegengestelde richting $-\nabla f(\mathbf{x})$ is de richting van de snelste afname van de functie.

Deze stelling laat zich goed onthouden door te zeggen dat een knikker in de richting $-\nabla f(\mathbf{x})$ loopt als hij in het punt $\mathbf{x}$ op de grafiek van de functie $f(\mathbf{x})$ neergezet wordt.

In plaats van de maximale waarde van het inproduct kunnen we ook eens naar het geval kijken, dat het inproduct $\nabla f \cdot \mathbf{v} = 0$ is. Dit betekent aan de ene kant dat de gradiënt loodrecht op $\mathbf{v}$ staat want $\cos(\varphi) = 0$ voor een rechte hoek φ. Maar aan de andere kant geeft $\nabla f \cdot \mathbf{v}$ de verandering van f in de richting van $\mathbf{v}$ aan en $\nabla f \cdot \mathbf{v} = 0$ betekent dus dat de functie in de richting van $\mathbf{v}$ niet van waarde verandert.

We komen zo naar het begrip van *niveaukrommen*: Dit zijn de krommen in het $x - y$-vlak waarop de functie $f(x, y)$ dezelfde waarde heeft. Iedereen heeft wel eens niveaukrommen gezien, dit zijn namelijk juist de hoogtelijnen op een topografische kaart. In Figuur I.4 zijn naast de grafiek ook niveaukrommen voor de functie $f(x, y) = (x^2 + y^3)\, e^{-x^2-y^2}$ te zien.



Figuur I.4: Grafiek en niveaukrommen voor de functie $f(x, y) = (x^2 + y^3)\, e^{-x^2-y^2}$.

De richting in die een functie in een punt $\mathbf{x}$ niet verandert is juist de raaklijn aan de niveaukromme door $\mathbf{x}$. En deze richting staat loodrecht op de gradiënt. Natuurlijk zijn er twee mogelijke vectoren, maar de raaklijn heeft ook twee richtingen. Onze tweede nieuwe inzicht over de gradiënt luidt dus:

Stelling: De gradiënt $\nabla f(\mathbf{x})$ staat loodrecht op de raaklijn aan de niveaukromme door het punt $\mathbf{x}$.

Voorbeeld: Voor de functie $f(x, y) := (x^2 + y^3)\, e^{-x^2-y^2}$ geldt:

$$\frac{\partial f}{\partial x} = 2x\, e^{-x^2-y^2} - 2x(x^2+y^3)\, e^{-x^2-y^2}, \quad \frac{\partial f}{\partial y} = 3y^2\, e^{-x^2-y^2} - 2y(x^2+y^3)\, e^{-x^2-y^2},$$

en dus

$$\nabla f(x, y) = e^{-x^2-y^2} \begin{pmatrix} 2x - 2x^3 - 2xy^3 \\ 3y^2 - 2x^2y - 2y^4 \end{pmatrix}.$$

Voor $(x, y) = (1, 2)$ hebben we bijvoorbeeld $\nabla f(1, 2) = e^{-5} \begin{pmatrix} -16 \\ -24 \end{pmatrix}$ en we kunnen in het rechterplaatje van Figuur I.4 controleren dat deze vector inderdaad loodrecht op de niveaukromme staat.

Voor de punten met $x = 0$, dus de punten op de y-as, hebben we $\nabla f(0, y) = e^{-y^2} \begin{pmatrix} 0 \\ 3y^2 - 2y^4 \end{pmatrix}$ en dit betekent dat de niveaukrommen loodrecht op de y-as staan. Op een soortgelijke manier volgt uit $\nabla f(x, 0) = e^{-x^2} \begin{pmatrix} 2x - 2x^3 \\ 0 \end{pmatrix}$ dat de niveaukrommen ook loodrecht op de x-as staan. Ook dit kunnen we in het plaatje terugvinden.

In de discussie hierboven hebben we functies van twee veranderlijken bekeken. Als we nu naar het algemene geval van functies van n variabelen kijken, verandert niet zo vreselijk veel. De gradiënt geeft nog steeds de richting in $\mathbb{R}^n$ aan waar de functie het snelste toeneemt.

Om te zien wat er met de niveaukrommen gebeurt, kijken we eerst naar het geval van een functie $f(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$. Als voorbeeld nemen we de functie $f(x, y, z) := \sqrt{x^2 + y^2 + z^2}$ die de afstand van het nulpunt aangeeft. Voor een vaste waarde r zijn de punten met $f(x, y, z) = r$ juist het oppervlak van een kogel met straal 0. Maar het woord opper*vlak* geeft het al aan: De punten met een vaste waarde voor $f(x, y, z)$ liggen op een krom vlak in de 3-dimensionale ruimte. Net zo als een kromme in het 2-dimensionale vlak bij uitvergroten steeds meer op een rechte lijn lijkt, wordt ook een oppervlak in de 3-dimensionale ruimte bij uitvergroten bij benadering een plat vlak.

Als we ons tot kleine omgevingen van een punt beperken, kunnen we ook in het algemeen zeggen dat de punten in $\mathbb{R}^3$ waarop een functie $f(x, y, z)$ dezelfde waarde heeft op een oppervlak in de 3-dimensionale ruimte liggen en dit noemen we een *niveauvlak*. Aan het niveauvlak door een gegeven punt $\mathbf{x} \in \mathbb{R}^3$ kunnen we ook weer een raakvlak aanleggen, en het feit dat de functie $f(x, y, z)$ in de richtingen van dit raakvlak niet verandert laat zien, dat de gradiënt $\nabla f(\mathbf{x})$ loodrecht op dit raakvlak staat.

Voor een algemene functie $f(x_1, \ldots, x_n)$ van n variabelen vormen de punten met constante functiewaarden veralgemeende oppervlakken, die bij uitvergroten op $n-1$-dimensionale hypervlakken lijken (het woord *hypervlak* wordt algemeen voor een $n-1$-dimensionale deelruimte van een n-dimensionale vectorruimte gebruikt). We noemen de oppervlakken waarop de functie dezelfde waarde heeft *niveauhypervlakken* en we kunnen ook hier aan het niveauhypervlak door een punt $\mathbf{x} \in \mathbb{R}^n$ een raakhypervlak leggen. Dit is een $n-1$-dimensionale deelruimte van $\mathbb{R}^n$ met de richtingen waarin de functie niet van waarde verandert. Dit betekent dat het juist het hypervlak is dat loodrecht op de gradiënt $\nabla f(\mathbf{x})$ staat.

De gradiëntenmethode

Een belangrijke toepassing van de gradiënt zijn benaderingsmethoden voor het vinden van maxima (of minima) van functies, te weten de *gradiëntenmethoden*. Vaak is het namelijk (zelfs bij gewone functies van één veranderlijke) zo, dat maxima niet expliciet bepaald kunnen worden en daarom numeriek benaderd worden. Voor gewone functies kan men dit heel naïf doen:

Kies een stapsgrote Δx en loop zo lang in stappen van Δx naar rechts tot dat de functie niet meer toeneemt. (Als de functie bij de eerste stap al kleiner wordt loop je naar links in plaats van rechts.) Op deze manier kom je (als de functie enigszins glad en Δx voldoende klein is) in de buurt van een lokaal maximum van de functie. Als de benadering nog niet goed genoeg is, kan je natuurlijk vanaf het gevonden punt met een kleinere Δx nog verder door gaan.

Als we hetzelfde idee voor functies van meerdere variabelen willen toepassen, hebben we het probleem een richting te kiezen, want er zijn nu oneindig veel richtingen in plaats van slechts twee (rechts en links). Maar omdat men zo snel mogelijk een maximum wil bereiken, is het voor de hand liggend de richting van de snelste toename van de functie te kiezen, en dat is juist de richting van de gradiënt.

Men kiest daarom ook hier een stapsgrote $\Delta\mathbf{x}$ en loopt bij iedere stap om $\Delta\mathbf{x}$ in de richting van de gradiënt in het laatste punt. Omdat het bepalen van de gradiënt een dure operatie is, wordt in de praktijk meestal pas een nieuwe richting gekozen, als op de lijn in de richting van de gradiënt een maximum van de functie is gevonden (zo als bij gewone functies van één veranderlijke).

Natuurlijk zijn bij ingewikkelde functies vaak ook de partiële afgeleiden niet analytisch te berekenen, maar deze laten zich benaderen door

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}$$

voor een voldoende kleine waarde van h.

1.4 De algemene afgeleide

Met de begrippen die we tot nu toe hebben behandeld, kunnen we nu ook het begrip van de algemene afgeleide van een functie van n veranderlijken formuleren.

Een van de ideeën achter de afgeleide van een gewone functies is, dat de functie in een kleine omgeving van een punt door een lijn benaderd kan worden. Als we namelijk de grafiek van een functie $f(x)$ rond een gekozen punt x_0 steeds meer uitvergroten, lijkt de grafiek (behalve voor exotische functies, die we hier buiten beschouwing laten) steeds meer op de raaklijn in het punt $(x_0, f(x_0))$ met stijging $f'(x_0)$.

Als we nu over een functie van twee veranderlijken nadenken, hebben we ons de grafiek van de functie nu al een paar keer als een soort gebergte voorgesteld. Als we dit nu uitvergroten, wordt het gebergte gewoon een vlakte in de ruimte die de grafiek in het gegeven punt raakt, dus het raakvlak. De vraag is nu hoe we de functiewaarden op het raakvlak kunnen bepalen, maar in principe hebben we al gezien dat dit juist door de richtingsafgeleide aangegeven wordt:

We hadden gezien dat de functie in de richting van een richtingsvector $\mathbf{v}$ van lengte 1 om het inproduct $\nabla f \cdot \mathbf{v}$ toeneemt (of afneemt). Maar als we nu op een vlak lopen, is de toename langs een veelvoud $c\mathbf{v}$ van $\mathbf{v}$ juist c keer de toename langs $\mathbf{v}$ en omdat het inproduct lineair is, geldt $c(\nabla f \cdot \mathbf{v}) = \nabla f \cdot (c\mathbf{v})$. Maar dit

betekent dat de toename van f langs een willekeurige vector $\mathbf{v}$ gegeven is door het inproduct $\nabla f \cdot \mathbf{v}$. Er geldt dus:

Stelling: Op het raakvlak aan de grafiek van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ is de toename van de functie $f(\mathbf{x})$ langs een willekeurige vector $\mathbf{v}$ gegeven door het inproduct $\nabla f(\mathbf{x}_0) \cdot \mathbf{v}$.

Analoog met de benadering $f(x) - f(x_0) \approx f'(x_0)(x - x_0)$ voor een gewone functie van één veranderlijke krijgen we zo de uitspraak dat

$$f(\mathbf{x}) - f(\mathbf{x}_0) \approx \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

voor vectoren $\mathbf{x}$ die dicht bij $\mathbf{x}_0$ liggen.

De limiet definitie voor de afgeleide van een gewone functie was

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \text{ of te wel } \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) = 0$$

als deze limiet bestaat. We kunnen dit makkelijk herschrijven tot de volgende definitie: De functie $f(x) : \mathbb{R} \to \mathbb{R}$ heeft in het punt x_0 de afgeleide $f'(x_0)$ als geldt dat

$$\lim_{x \to x_0} \frac{f(x) - f(x_0) - f'(x_0) \cdot (x - x_0)}{x - x_0} = 0.$$

Merk op dat vermenigvuldiging met $f'(x_0)$ een lineaire afbeelding $h \mapsto f'(x_0) \cdot h$ op de 1-dimensionale vectorruimte $\mathbb{R} = \mathbb{R}^1$ geeft. Deze definitie laat zich nu als volgt op functies $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ van n veranderlijken en met m componenten veralgemenen:

Definitie: De functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ heeft in het punt $\mathbf{x}_0$ de *afgeleide* $T := T_{\mathbf{x}_0}$ als $T : \mathbb{R}^n \to \mathbb{R}^m$ een lineaire afbeelding is waarvoor geldt dat

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{f(\mathbf{x}) - f(\mathbf{x}_0) - T(\mathbf{x} - \mathbf{x}_0)}{\|\mathbf{x} - \mathbf{x}_0\|} = 0.$$

Hierbij is de toepassing $T(\mathbf{x} - \mathbf{x}_0)$ van T op de vector $\mathbf{x} - \mathbf{x}_0$ gegeven door het matrix product van de matrix van T met $\mathbf{x} - \mathbf{x}_0$.

Uit de discussie van boven weten we dat voor functies $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ met een enkele component geldt dat $f(\mathbf{x}) - f(\mathbf{x}_0) \approx \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$. Hieruit volgt dat er helemaal geen keuze voor de lineaire afbeelding $T_{\mathbf{x}_0}$ is, er geldt namelijk in dit geval noodzakelijk dat

$$T_{\mathbf{x}_0} = (\nabla f(\mathbf{x}_0))^{tr} = \left(\tfrac{\partial f}{\partial x_1}(\mathbf{x}_0), \ldots, \tfrac{\partial f}{\partial x_n}(\mathbf{x}_0) \right),$$

de afgeleide is dus juist de getransponeerde (gespiegelde) van de gradiënt. Merk op dat voor een vector $\mathbf{v} \in \mathbb{R}^n$ het matrix product $\left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)(\mathbf{v}) = (\nabla f)^{tr}(\mathbf{v})$ hetzelfde is als het inproduct $\nabla f \cdot \mathbf{v}$.

De Jacobi matrix

Ten slotte komen we terug op een algemene functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$, die door de componenten $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ beschreven is. Voor iedere van de componenten $f_i(\mathbf{x})$ krijgen we de afgeleide $T_i = (\nabla f_i)^{tr}$ als een rijvector. Als we deze rijvectoren nu als rijen in een matrix schrijven, krijgen we

$$J := \begin{pmatrix} T_1 \\ \vdots \\ T_m \end{pmatrix} = \begin{pmatrix} (\nabla f_1)^{tr} \\ \vdots \\ (\nabla f_m)^{tr} \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Deze $m \times n$-matrix $J = (\frac{\partial f_i}{\partial x_j})$ met de partiële afgeleiden van de componenten van $f(\mathbf{x})$ heet de *Jacobi matrix* of kort *Jacobiaan* van $f(\mathbf{x})$. (Let op, soms wordt de term *Jacobiaan* ook voor de determinant van deze matrix gebruikt.)

We weten nu dat voor iedere component f_i van f geldt dat

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{f_i(\mathbf{x}) - f_i(\mathbf{x}_0) - T_i(\mathbf{x} - \mathbf{x}_0)}{\|\mathbf{x} - \mathbf{x}_0\|} = 0.$$

Maar als we nu de componenten tot een m-dimensionale vector samenvoegen, krijgen we hieruit meteen dat

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{f(\mathbf{x}) - f(\mathbf{x}_0) - J(\mathbf{x} - \mathbf{x}_0)}{\|\mathbf{x} - \mathbf{x}_0\|} = 0.$$

We hebben dus de volgende belangrijke stelling ingezien:

Stelling: De afgeleide van een functie $f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) : \mathbb{R}^n \to \mathbb{R}^m$ is gegeven door de Jacobi matrix

$$J := \begin{pmatrix} (\nabla f_1)^{tr} \\ \vdots \\ (\nabla f_m)^{tr} \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Voorbeeld: De functie $f(x, y, z) := (z\, e^x, -y\, e^z)$ heeft de Jacobi matrix

$$\begin{pmatrix} z\, e^x & 0 & e^x \\ 0 & -e^z & -y\, e^z \end{pmatrix}.$$

Opdracht 7 *Bepaal de Jacobi matrices van de volgende functies:*

(i) $f : \mathbb{R}^3 \to \mathbb{R}^2$ *gegeven door* $f(x, y, z) := (x - y, y + z)$;

(ii) $f : \mathbb{R}^2 \to \mathbb{R}^3$ *gegeven door* $f(x, y) := (x + y, x - y, xy)$.

Belangrijke begrippen in deze les

- functie van meerdere variabelen

- continuïteit
- partiële afgeleide
- verruilen van de volgorde van partiële afgeleiden
- richtingsafgeleide
- gradiënt
- niveaukrommen
- Jacobi matrix

Opgaven

1. Geef voor de volgende functies de maximale domeinen aan, waarop ze continu gedefinieerd kunnen worden en maak een schets van deze domeinen:

   (i) $f(x,y) := \log((16 - x^2 - y^2)(x^2 + y^2 - 4))$,

   (ii) $f(x,y) := \sqrt{6 - (2x + 3y)}$.

2. Ga na of de volgende functies continu naar de aangegeven punten voortgezet kunnen worden.

   (i) $f(x,y) = \frac{x}{x^2+y^2}$ naar $(x,y) = (0,0)$,

   (ii) $f(x,y) = \frac{x^2+y^2}{y}$ naar $(x,y) = (0,0)$,

   (iii) $f(x,y) = \frac{y^3}{x^2+y^2}$ naar $(x,y) = (0,0)$,

   (iv) $f(x,y) = \frac{\sin(x-y)}{\cos(x+y)}$ naar $(x,y) = (0,0)$,

   (v) $f(x,y) = \frac{\cos(xy)}{1-x-\cos(y)}$ naar $(x,y) = (1,\pi)$,

   (vi) $f(x,y) = \frac{\sin(xy)}{x^2+y^2}$ naar $(x,y) = (0,0)$.

3. Bepaal voor de volgende functies de partiële afgeleiden $\frac{\partial f}{\partial x}$ en $\frac{\partial f}{\partial x}$:

   (i) $f(x,y) := xy$;

   (ii) $f(x,y) := e^{xy}$;

   (iii) $f(x,y) := x\cos(x)\cos(y)$;

   (iv) $f(x,y) := (x^2 + y^2)\log(x^2 + y^2)$.

4. Bepaal voor de volgende functies de partiële afgeleiden $\frac{\partial f}{\partial x}$ en $\frac{\partial f}{\partial x}$:

   (i) $f(x,y) := x\, e^{x^2+y^2}$;

   (ii) $f(x,y) := \frac{x^2+y^2}{x^2-y^2}$;

   (iii) $f(x,y) := e^{xy}\log(x^2 + y^2)$;

   (iv) $f(x,y) := \frac{x}{y}$;

   (v) $f(x,y) := \cos(y\, e^{xy})\sin(x)$.

5. Bepaal voor de volgende functies de tweede partiële afgeleiden $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial x\,\partial y}$, $\frac{\partial^2 f}{\partial y\,\partial x}$ en $\frac{\partial^2 f}{\partial y^2}$ en laat hiermee in het bijzonder zien dat $\frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial x}$:
   (i) $f(x,y) := \cos(xy^2)$;
   (ii) $f(x,y) := e^{-xy^2} + y^3x^4$;
   (iii) $f(x,y) := \log(x-y)$;
   (iv) $f(x,y) := \sin(x^2 - 3xy)$.
6. Laat voor $f(x,y,z,w) := e^{xyz}\sin(xw)$ expliciet zien dat $\frac{\partial^3 f}{\partial w\,\partial z\,\partial x} = \frac{\partial^3 f}{\partial x\,\partial w\,\partial z}$.
7. Bepaal voor de volgende functies de richtingsafgeleiden in de richtingen van de aangegeven vectoren (die niet op lengte 1 genormeerd zijn):
   (i) $f(x,y) := x + 2xy - 3y^2$ in de richting van $\begin{pmatrix}3\\4\end{pmatrix}$;
   (ii) $f(x,y) := \log(x^2+y^2)$ in de richting van $\begin{pmatrix}2\\1\end{pmatrix}$;
   (iii) $f(x,y) := e^x\cos(\pi y)$ in de richting van $\begin{pmatrix}-1\\2\end{pmatrix}$.
8. Bepaal de gradiënten van de functies:
   (i) $f(x,y,z) := x\,e^{-x^2-y^2-z^2}$;
   (ii) $f(x,y,z) := \frac{xyz}{x^2+y^2+z^2}$;
   (iii) $f(x,y,z) := z^2\,e^x\cos(y)$.
9. Bepaal de gradiënten $\nabla f(\mathbf{x})$ van de volgende functies in de aangegeven punten $\mathbf{x}$:
   (i) $f(x,y) := x^2 + 2y^3$ in $(x,y) = (1,1)$;
   (ii) $f(x,y,z) := (x+z)\,e^{x-y}$ in $(x,y,z) = (1,1,1)$;
   (iii) $f(x,y,z) := x^2 + y^2 - z^2$ in $(x,y,z) = (0,0,1)$;
   (iv) $f(x,y,z) := \log(x^2+y^2+z^2)$ in $(x,y,z) = (1,0,1)$.
10. Bepaal voor de volgende functies de richting waarin de functie in het punt $(x,y) = (1,1)$ het snelste toeneemt:
   (i) $f(x,y) := x^2 + 2y^2$;
   (ii) $f(x,y) := x^2 - 2y^2$;
   (iii) $f(x,y) := e^x\sin(y)$;
   (iv) $f(x,y) := e^x\sin(x) - e^{-x}\cos(y)$.
11. Bepaal de Jacobi matrices van de volgende functies:
   (i) $f : \mathbb{R}^2 \to \mathbb{R}^2$ gegeven door $f(x,y) := (e^x, \sin(xy))$;
   (ii) $f : \mathbb{R}^3 \to \mathbb{R}^3$ gegeven door $f(x,y,z) := (x+z, y-5z, x-y)$;
   (iii) $f : \mathbb{R}^2 \to \mathbb{R}^3$ gegeven door $f(x,y) := (x\,e^y + \cos(y), x, x+e^y)$;
   (iv) $f : \mathbb{R}^3 \to \mathbb{R}^2$ gegeven door $f(x,y,z) := (x + e^z + y, yx^2)$;
   (v) $f : \mathbb{R}^2 \to \mathbb{R}^3$ gegeven door $f(x,y) := (xy\,e^{xy}, x\sin(y), 5xy^2)$.

Les 2 Taylor reeksen

We hebben in Wiskunde 1 een aantal belangrijke reële functies gezien, bijvoorbeeld de exponentiële functie $\exp(x)$ of de trigonometrische functies $\sin(x)$ en $\cos(x)$. Toen hebben we wel eigenschappen van deze functies aangegeven, bijvoorbeeld dat $\exp(x)$ gekarakteriseerd is door $\exp(x)' = \exp(x)$ en $\exp(0) = 1$. Maar hoe we de waarde van zo'n functie echt kunnen berekenen, of hoe een zakrekenmachine, GRM of computer de waarden van zo'n functie berekent, hebben we toen niet gezien.

De methode die hiervoor (met zekere variaties) wordt gebruikt, is een functie te benaderen door een veelterm. Dit lijkt op het eerste gezicht te eenvoudig om efficiënt te kunnen werken, maar men kan bewijzen dat op een begrensd gebied een continue functie door veeltermen willekeurig goed benaderd kan worden. Naarmate de gewenste nauwkeurigheid toeneemt zijn hiervoor echter veeltermen van hogere graad nodig.

2.1 Interpolatie

Eén mogelijkheid om een benaderende veelterm te vinden, gebruikt de methode van *interpolatie*. We weten dat door 2 punten (x_1, y_1) en (x_2, y_2) met verschillende x-coördinaten $x_1 \neq x_2$ een eenduidige lineaire functie vastgelegd is, met als grafiek de lijn door deze twee punten, te weten

$$l(x) = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1.$$

Net zo leggen 3 punten (weer met verschillende x-waarden) in het $x - y$-vlak eenduidig een parabool, dus een kwadratische functie vast. Op een analoge manier laat zich aantonen dat door $n + 1$ punten met verschillende x-waarden een eenduidige veelterm van graad n vastgelegd is. De punten x_i waarop de waarden gegeven zijn noemt men ook *basispunten* of *roosterpunten*. De veelterm door de punten laat zich met behulp van de *Lagrange interpolatie* zelfs expliciet aangeven:

De veelterm $p(x)$ van graad n door de punten $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ is gegeven door

$$p(x) := y_0 \cdot L_0(x) + y_1 \cdot L_1(x) + \ldots + y_n \cdot L_n(x), \text{ waarbij}$$

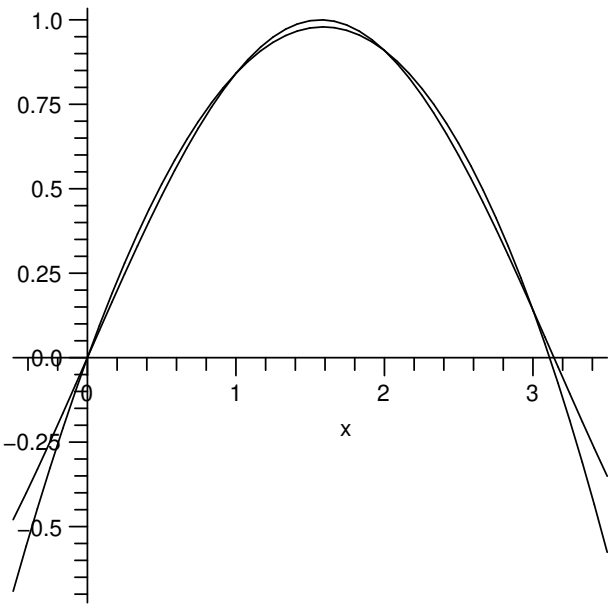
$$L_k(x) := \frac{x - x_0}{x_k - x_0} \cdot \frac{x - x_1}{x_k - x_1} \cdots \frac{x - x_{k-1}}{x_k - x_{k-1}} \cdot \frac{x - x_{k+1}}{x_k - x_{k+1}} \cdots \frac{x - x_n}{x_k - x_n}.$$

De grap is dat de hulpfuncties $L_k(x)$ zo gemaakt zijn dat $L_k(x_k) = 1$ en $L_k(x_i) = 0$ voor $i \neq k$.

In Figuur I.5 zien we dat de sinus functie goed door een derdegraads interpolatie $p(x)$ benaderd wordt, die door de vier basispunten $0, 1, 2, 3$ loopt en

gegeven is door:

$$\begin{aligned}
p(x) &= \frac{\sin(1)}{2}x(x-2)(x-3) - \frac{\sin(2)}{2}x(x-1)(x-3) + \frac{\sin(3)}{6}x(x-1)(x-2) \\
&= (\frac{\sin(1)}{2} - \frac{\sin(2)}{2} + \frac{\sin(3)}{6})\,x^3 + (-\frac{5\sin(1)}{2} + 2\sin(2) - \frac{\sin(3)}{2})\,x^2 \\
&+ (3\sin(1) - \frac{3\sin(2)}{2} + \frac{\sin(3)}{3})\,x \\
&\approx -0.0104\,x^3 - 0.3556\,x^2 + 1.2075\,x.
\end{aligned}$$



Figuur I.5: Benadering van $\sin(x)$ door de Lagrange interpolatie door de punten met x-waarden $0, 1, 2, 3$.

De interpolatie laat zich op een enigszins voor de hand liggende manier ook op functies van meerdere variabelen veralgemenen. Dit heeft in het bijzonder voor punten in de 3-dimensionale ruimte veel toepassingen. Zo worden bijvoorbeeld voor oppervlakken (zo als gezichten) alleen maar een aantal roosterpunten opgeslagen en de punten ertussen door interpolatie berekend om een glad oppervlak te krijgen. Op deze manier laat zich de beweging van een *computer animated* figuur in de *virtual reality* reconstrueren uit gemeten bewegingen van een echte figuur, waarbij alleen maar op enkele punten sensoren zitten.

Er zijn echter ook een paar problemen als we de interpolatie willen gebruiken om een functie te benaderen:

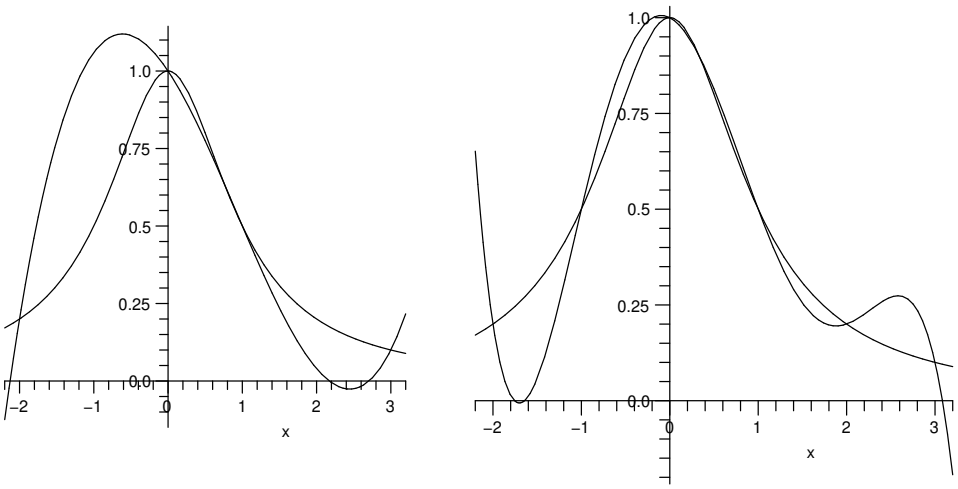
(i) Omdat we een functie $f(x)$ door een veelterm willen benaderen, moeten we de y_k in principe al als functiewaarden $y_k = f(x_k)$ kennen, terwijl we eigenlijk de veelterm juist bepalen om de functiewaarden te kunnen benaderen.

(ii) Afhankelijk van hoe we de x_k kiezen, kan de veelterm $p(x)$ sterk van de functie $f(x)$ afwijkingen, bijvoorbeeld als de x_k te grote afstanden hebben.

Het laatste punt is in Figuur I.6 geïllustreerd: We benaderen de functie

$$f(x) := \frac{1}{1+x^2}$$

op het interval $[-2, 3]$ door interpolaties van graad 3 en van graad 5. In het linkerplaatje is de interpolatie met de vier basispunten $-2, 0, 1, 3$ te zien, in het rechterplaatje de interpolatie met de zes basispunten $-2, -1, 0, 1, 2, 3$. In beide gevallen wijkt $f(x)$ sterk van de interpolatie af, alleen maar op het interval $[0, 1.5]$ is de benadering enigszins redelijk.



Figuur I.6: Interpolatie van $\frac{1}{1+x^2}$ met 4 en 6 basispunten.

In principe zouden we op intervallen waar de functie $f(x)$ heel sterk verandert veel basispunten x_k willen hebben, terwijl op stukken waar de functie bijna lineair is niet zo veel basispunten nodig zijn. Om hierover te kunnen beslissen moeten we naar de hogere afgeleiden van $f(x)$ kijken en dit idee geeft aanleiding tot een andere manier om een benaderende veelterm te bepalen. Deze gebruikt inderdaad de hogere afgeleiden van $f(x)$, maar slechts in een enkele punt.

2.2 Taylor veeltermen

Het idee bij de *Taylor veelterm* is, een functie $f(x)$ in de omgeving van één punt x_0 te benaderen. Hiervoor wordt een veelterm geconstrueerd die in het punt x_0 niet alleen maar dezelfde functiewaarde als $f(x)$ heeft, maar ook dezelfde eerste afgeleide, dezelfde tweede afgeleide, enzovoorts.

Definitie: De *Taylor veelterm* $p(x) := p_{f,n,x_0}(x)$ van graad n voor een continue functie $f(x)$ in het punt x_0 is gedefinieerd door de eigenschappen:

$$p(x_0) = f(x_0), \quad p'(x_0) = f'(x_0), \quad p''(x_0) = f''(x_0), \ \ldots, \ p^{(n)}(x_0) = f^{(n)}(x_0).$$

Dit betekent dat de Taylor veelterm in het punt x_0 dezelfde afleidingen (tot orde n) heeft als $f(x)$.

Men gaat na dat deze eigenschappen inderdaad een eenduidige veelterm van graad n bepalen, namelijk de veelterm

$$p(x) = c_n(x - x_0)^n + c_{n-1}(x - x_0)^{n-1} + \ldots + c_1(x - x_0) + c_0$$

met coëfficiënten

$$c_0 = f(x_0), \quad c_1 = f'(x_0), \quad c_2 = \frac{f''(x_0)}{2!}, \quad \ldots \quad c_n = \frac{f^{(n)}(x_0)}{n!}.$$

Dat dit inderdaad de coëfficiënten zijn, ziet men door het uitschrijven van de afgeleiden van $p(x)$, invullen van x_0 en vergelijken met $f^{(n)}(x_0)$:

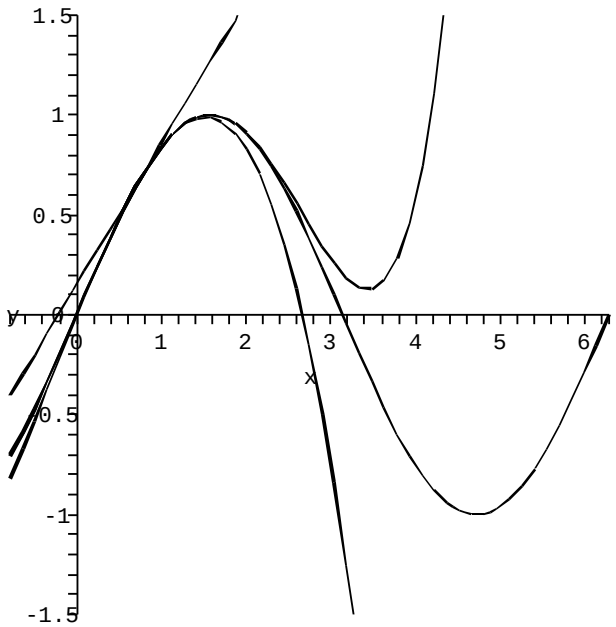
$$\begin{aligned}
p(x) &= c_n(x - x_0)^n + \ldots + c_3(x - x_0)^3 + c_2(x - x_0)^2 + c_1(x - x_0) + c_0 \\
&\Rightarrow p(x_0) = c_0, \text{ uit } f(x_0) = p(x_0) = c_0 \text{ volgt dan } c_0 = f(x_0) \\
p'(x) &= n \cdot c_n(x - x_0)^{n-1} + \ldots + 3 \cdot c_3(x - x_0)^2 + 2 \cdot c_2(x - x_0) + c_1 \\
&\Rightarrow p'(x_0) = c_1, \text{ uit } f'(x_0) = p'(x_0) = c_1 \text{ volgt dan } c_1 = f'(x_0) \\
p''(x) &= n(n-1) \cdot c_n(x - x_0)^{n-2} + \ldots + 3 \cdot 2 \cdot c_3(x - x_0) + 2 \cdot c_2 \\
&\Rightarrow p''(x_0) = 2 \cdot c_2, \text{ uit } f''(x_0) = p''(x_0) = 2 \cdot c_2 \text{ volgt dan } c_2 = \frac{f''(x_0)}{2} \\
p'''(x) &= n(n-1)(n-2) \cdot c_n(x - x_0)^{n-3} + \ldots + 3 \cdot 2 \cdot c_3 \\
&\Rightarrow p'''(x_0) = 3 \cdot 2 \cdot c_3 \Rightarrow c_3 = \frac{f'''(x_0)}{3!} \quad \text{enzovoorts.}
\end{aligned}$$

Het idee achter de geëiste eigenschappen is dat het overeenstemmen van de hogere afgeleiden ervoor zorgt dat de grafiek van de Taylor veelterm zich rond het punt x_0 steeds beter aan de grafiek van $f(x)$ vlijt.

In Figuur I.7 zien we het effect van Taylor veeltermen van verschillende graad. Terwijl de veelterm van graad 1 de grafiek van $\sin(x)$ alleen maar raakt (omdat $\sin(x)$ geen rechte stukken heeft), is de veelterm van graad 3 al een redelijke benadering tussen ongeveer 0.5 en 1.5, voordat hij naar beneden wegduikt. Met de veelterm van graad 5 gaat het iets langer goed, maar dan gaat deze naar boven weg.

Meestal wordt de Taylor veelterm niet in de boven aangegeven volgorde van de termen geschreven (dus beginnend met de hoogste), maar andersom. Verder wordt het feit dat de Taylor veelterm de functie slechts in een kleine omgeving van x_0 goed benadert weergegeven door x in de vorm $x = x_0 + h$ te schrijven (waarbij men stiekem verondersteld dat h klein is). Dit geeft de volgende vorm van de Taylor veelterm:

$$p(x_0 + h) = f(x_0) + f'(x_0) \cdot h + \frac{f''(x_0)}{2} \cdot h^2 + \ldots + \frac{f^{(n)}(x_0)}{n!} \cdot h^n.$$



Figuur I.7: Benadering van $\sin(x)$ door de Taylor veeltermen van graad 1, 3 en 5 in het punt $x_0 = \frac{\pi}{4}$

Opdracht 8 *Bepaal voor $f(x) := \sqrt{x}$ de Taylor veelterm van graad 4 in het punt $x_0 = 1$.*

Over het verband tussen de benaderde functie $f(x)$ en de Taylor veelterm $p(x)$ valt echter nog veel meer te zeggen, in het bijzonder laat zich expliciet een afschatting voor de fout $|f(x_0 + h) - p(x_0 + h)|$ aangeven (die natuurlijk van h afhangt). We definiëren hiervoor de n-de *foutterm* $R_n(h)$ door

$$f(x_0 + h) = p(x_0 + h) + R_n(h) \quad \text{of} \quad R_n(h) = f(x_0 + h) - p(x_0 + h).$$

Er zijn verschillende manieren om de foutterm $R_n(h)$ aan te geven, de drie meest belangrijke zijn de volgende:

$$R_n(h) = \frac{1}{(n+1)!} \cdot f^{(n+1)}(t) \cdot h^{n+1} \text{ voor een } t \in [0, h] \text{ (Lagrange vorm)}$$

$$R_n(h) = \frac{1}{n!} \cdot f^{(n+1)}(t) \cdot (h - t)^n h \text{ voor een } t \in [0, h] \text{ (Cauchy vorm)}$$

$$R_n(h) = \frac{1}{n!} \int_0^h f^{(n+1)}(t) \cdot (h - t)^n \, dt \text{ (integraal vorm)}$$

Het bewijs dat de foutterm op deze manieren geschreven kan worden is niet eens zo moeilijk. Bijvoorbeeld berust de Lagrange vorm op de *middelwaardestelling*, die zegt, dat er voor een (differentieerbare)

functie $f(x)$ op een interval $[a, b]$ een punt $c \in [a, b]$ in het interval ligt waar de raaklijn aan $f(x)$ dezelfde stijging heeft als de gemiddelde stijging $\frac{f(b)-f(a)}{b-a}$ van $f(x)$ op het interval $[a, b]$, d.w.z. waar geldt dat $f(b) - f(a) = f'(c)(b - a)$. Dit klopt omdat de raaklijn niet overal een grotere of overal een kleinere stijging dan de gemiddelde stijging kan hebben.

Vooral de Lagrange vorm van de foutterm is vaak geschikt om de fout bij de benadering door de Taylor veelterm af te schatten. Dit lukt in het bijzonder als zich makkelijk een grens voor de $(n + 1)$-de afgeleide van $f(x)$ op het interval $[0, h]$ laat aangeven.

Bijvoorbeeld geldt voor $f(x) = \sin(x)$ dat $|f^{(n+1)}(t)| \leq 1$ voor alle t, omdat de afgeleiden $\pm\sin(t)$ of $\pm\cos(t)$ zijn en deze functies alleen maar waarden tussen -1 en 1 hebben.

Net zo goed ziet men in dat voor $f(x) = \exp(x)$ voor $t \in [0, h]$ (met $h > 0$) geldt dat $|f^{(n+1)}(t)| \leq e^h$ omdat de afgeleiden de exponentiële functie reproduceren en de exponentiële functie monotoon stijgend is.

Voorbeelden:

(1) Voor $f(x) := \sin(x)$ en $x_0 = 0$ hebben we $f(x_0) = \sin(0) = 0$, $f'(x_0) = \cos(0) = 1$, $f''(x_0) = -\sin(0) = 0$, $f'''(x_0) = -\cos(0) = -1$ en $f^{(4)}(x_0) = \sin(0) = 0$. De Taylor veelterm van graad 4 van $\sin(x)$ rond $x_0 = 0$ is dus

$$p(x) = 0 + x - 0 - \frac{1}{6}x^3 + 0 = x - \frac{1}{6}x^3.$$

Als we nu bijvoorbeeld de waarde van $\sin(\frac{\pi}{10})$ willen bepalen, krijgen we de benadering $\sin(\frac{\pi}{10}) \approx \frac{\pi}{10} - \frac{1}{6}(\frac{\pi}{10})^3 = \frac{\pi}{10} - \frac{\pi^3}{6000} \approx 0.30899$.

De fout kunnen we afschatten met $|R_4(\frac{\pi}{10})| < \frac{1}{5!} \cdot 1 \cdot (\frac{\pi}{10})^5 < 2.6 \cdot 10^{-5} = 0.000026$, dus hebben we aangetoond dat $0.30896 < \sin(\frac{\pi}{10}) < 0.30902$.

(2) We kunnen het Euler getal e benaderen met behulp van de Taylor ontwikkeling van $f(x) := \exp(x)$ in $x_0 = 0$. Omdat $f^{(n)}(x) = \exp(x)$ geldt $f^{(n)}(0) = 1$ voor alle n. We krijgen daarom de n de Taylor veelterm

$$p(h) = 1 + h + \frac{h^2}{2} + \frac{h^3}{3!} + \ldots + \frac{h^n}{n!}.$$

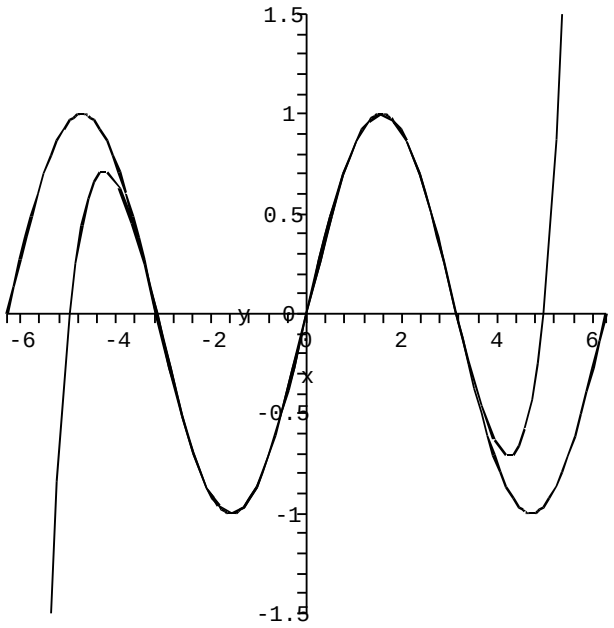
Als benadering van $e = f(1)$ krijgen we met $n = 6$ bijvoorbeeld $p(1) = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{1}{720} \approx 2.718$.

De fout kunnen we afschatten met $\frac{1}{(n+1)!}\, e^t \cdot 1^{n+1}$ voor een $t \in [0, 1]$, maar omdat we e juist willen berekenen, moeten we hier een grove afschatting voor e nemen, bijvoorbeeld $e^t \leq e < 3$. Hieruit volgt dat $|R_n(1)| < \frac{3}{7!} = \frac{3}{5040} \approx 0.0006$, dus hebben we aangetoond dat $2.7174 < e < 2.7186$.

Opdracht 9 *Bepaal met behulp van een Taylor veelterm de waarde van* $\sin(1)$ *op* 8 *decimalen, d.w.z. met een fout van hoogstens* 10^{-8}.

2.3 Taylor reeksen

Door de graad van de Taylor veelterm te laten groeien, krijgen we (in goedaardige gevallen) een steeds betere benadering van een functie, en meestal ook op een groter interval. In Figuur I.8 zien we bijvoorbeeld, dat de Taylor veelterm van graad 10 (deze keer in het punt $x_0 = 0$ berekend) tussen $-\pi$ en π bijna niet van de functie $\sin(x)$ te onderscheiden is.



Figuur I.8: Benadering van $\sin(x)$ door de Taylor veelterm van graad 10 in het punt $x_0 = 0$

Men krijgt dus het idee dat de benadering steeds beter wordt als we de graad verhogen, en het beste zou zijn, helemaal niet te stoppen maar oneindig door te gaan. Dit doen we dus!

Hier wordt nu duidelijk waarom het nuttig is de Taylor veelterm niet met de hoogste term $\frac{f^{(n)}(x_0)}{n!}h^n$ te beginnen, maar met de laagste term $f(x_0)$.

Machtsreeksen

Als men een algemene veeltermen met opstijgende termen schrijft, krijgt men iets als

$$p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n = \sum_{i=0}^{n} a_ix^i.$$

Als we nu niet met een term a_nx^n stoppen, maar ook voor een willekeurig hoge graad i steeds nog een coëfficiënt a_i van x^i definiëren (die weliswaar 0 mag zijn),

noemen we dit geen veelterm meer, maar een *reeks*, soms voor alle duidelijkheid zelfs een *oneindige reeks*. Een reeks is dus een uitdrukking van de vorm

$$r(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots = \sum_{n=0}^{\infty} a_n x^n.$$

Veeltermen zijn nu een speciale vorm van reeksen, een veelterm van graad n is namelijk een reeks waarbij vanaf $i = n + 1$ geldt dat $a_i = 0$ is.

Er zijn twee verschillende manieren hoe men tegen een oneindige reeks aan kan kijken, van een meer algebraïsch standpunt of van een meer analytisch standpunt.

- Algebraïsch kan men een rij zien als een abstracte uitdrukking die gerepresenteerd is door de oneindige rij $(a_0, a_1, a_2, \ldots)$ van coëfficiënten. Twee reeksen laten zich net zo optellen als men veeltermen optelt, namelijk door de coëfficiënten van iedere term x^n bij elkaar op te tellen.

  Desnoods laten zich twee reeksen ook met elkaar vermenigvuldigen, dit geeft het zogeheten *convolutieproduct* of *Cauchy product* waarbij men in het product van de twee reeksen weer de coëfficiënten van de termen x^n bepaald. Dit gebeurt in principe ook hetzelfde als bij veeltermen, omdat alleen maar de termen tot graad n in de twee reeksen een bijdrage aan de term x^n in het product kunnen leveren. Er geldt

  $$(\sum_{n=0}^{\infty} a_n x^n) \cdot (\sum_{n=0}^{\infty} b_n x^n) = \sum_{n=0}^{\infty} c_n x^n \text{ met } c_n = \sum_{i=0}^{n} a_i b_{n-i},$$

  bijvoorbeeld is $c_3 = a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0$.
- Analytisch zien we $f(x) = \sum_{n=0}^{\infty} a_n x^n$ als functie, waarbij we voor de functiewaarde in het punt x een limiet moeten bekijken, namelijk de limiet

  $$f(x) = \lim_{n \to \infty} \sum_{i=0}^{n} a_i x^i.$$

  Deze limiet hoeft niet voor iedere x te bestaan, dus hebben we het alleen maar voor degene waarden x echt met een functie te maken, waar deze limiet inderdaad bestaat.

Voorbeeld: Als we een voorbeeld van een reeks willen bekijken dat geen veelterm is, schiet misschien als eerste de reeks door het hoofd waarbij alle coëfficiënten $a_n = 1$ zijn, dus de reeks

$$r(x) = 1 + x + x^2 + \ldots = \sum_{n=0}^{\infty} x^n.$$

Deze reeks heet de *meetkundige reeks* en heeft de mooie eigenschap dat voor $|x| < 1$ geldt dat

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Dit gaat men na, door het product $(1+x+\ldots+x^n)(1-x)$ uit te schrijven als $(1+x+\ldots+x^n)(1-x) = (1+x+\ldots+x^n)-(x+x^2+\ldots+x^{n+1}) = 1-x^{n+1}$. Voor $|x|<1$ is $\lim_{n\to\infty} x^{n+1} = 0$, dus gaat $(1+x+\ldots+x^n)(1-x) \to 1$ en we moeten alleen nog door $(1-x)$ delen om de boven aangegeven formule te krijgen.

De Taylor reeks van een functie

We gaan nu van de Taylor veeltermen een reeks maken door de graad van de veeltermen naar oneindig te laten gaan. De reeks die we zo krijgen noemen we natuurlijk Taylor reeks.

Definitie: De oneindige reeks $T(x) := T_{f,x_0}(x)$ gedefinieerd door

$$\begin{aligned} T(x) &:= \lim_{n\to\infty} p_{f,n,x_0}(x) \\ &= f(x_0) + f'(x_0)(x-x_0) + \ldots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \ldots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n. \end{aligned}$$

heet de *Taylor reeks* van $f(x)$ in het punt x_0.

Over de Taylor reeks als limiet van de Taylor veeltermen moeten we een paar belangrijke opmerkingen kwijt:

(1) Natuurlijk geldt $T(x_0) = f(x_0)$, omdat alle termen in de som voor $n > 0$ wegvallen. Maar het kan zijn dat de reeks $T(x)$ voor geen elke $x \neq x_0$ convergeert, dus een limiet heeft. Voor goedaardige functies (zo als we die in de praktijk kunnen verwachten) geldt dit wel, tenminste in een zeker interval rond x_0. De vraag of en waar de Taylor reeks van een functie convergeert, geeft aanleiding tot belangrijke en diepe stellingen in de wiskunde, maar is in deze cursus een minder belangrijke vraagstelling, omdat we veronderstellen dat we het met voldoende gladde functies te maken hebben.

(2) Zelfs als de Taylor reeks $T(x)$ voor een zekere waarde x convergeert, hoeft de limiet niet de juiste functiewaarde $f(x)$ te zijn. Een afschrikkend voorbeeld hiervoor is de functie

$$f(x) := e^{-\frac{1}{x^2}}$$

die door $f(0) := 0$ continu naar 0 voortgezet wordt. Deze functie is zelfs willekeurig vaak differentieerbaar, en er geldt $f^{(n)}(0) = 0$ voor alle n. Dit betekent, dat de coëfficiënten van de Taylor reeks van $f(x)$ in het punt 0 alle 0 zijn en dus de Taylor reeks $T(x) = 0$ is. Aan de andere kant is $f(x) \neq 0$ voor $x \neq 0$, dus geeft de Taylor reeks de functie alleen maar in het nulpunt $x_0 = 0$ weer.

(3) De grootste afstand r zo dat $T(x)$ voor alle x met $|x-x_0|$ naar de goede waarde $f(x)$ convergeert noemt men de *convergentiestraal* van $T(x)$.

Als we bij een functie $f(x)$ ervan uit gaan dat de Taylor reeks $T(x)$ in de punten waar hij convergeert ook naar de goede waarde $f(x)$ convergeert, schrijven we gewoon $f(x)$ in plaats van $T(x)$. Net als bij de Taylor veeltermen is het ook hier gebruikelijk, duidelijk te maken dat x dicht bij het punt x_0 ligt door $x = x_0 + h$ te schrijven. Op deze manier krijgt men de volgende vorm van de Taylor reeks:

$$f(x_0 + h) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} h^n.$$

Voorbeelden:

(1) $\exp(x)$:
We bekijken de Taylor reeks in het punt $x_0 = 0$, daar geldt $\exp^{(n)}(x_0) = \exp(0) = 1$ voor alle $n \in \mathbb{N}$. De Taylor veelterm van graad n voor $\exp(x)$ in het punt 0 is dus

$$1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \ldots + \frac{1}{n!}x^n = \sum_{i=0}^{n} \frac{x^i}{i!}$$

en de Taylor reeks in het punt 0 is de limiet $n \to \infty$ van deze veeltermen, dus

$$T(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots.$$

Er laat zich aantonen dat deze Taylor reeks convergentiestraal ∞ heeft, d.w.z. dat $T(x)$ voor elke $x \in \mathbb{R}$ naar $\exp(x)$ convergeert. Omdat de noemers met $n!$ heel snel groeien, is de convergentie erg goed en kunnen we de exponentiële functie al met weinig termen goed benaderen.

(2) $\sin(x)$:
Ook voor de sinus functie bepalen we de Taylor reeks in $x_0 = 0$. Merk op dat $\sin''(x) = -\sin(x)$, $\sin'''(x) = -\sin'(x)$ en $\sin^{(4)}(x) = \sin(x)$, hieruit volgt dat de afgeleiden (beginnend met de 0-de afgeleide $\sin(0) = 0$) van $\sin(x)$ in het punt $x = 0$ periodiek $0, 1, 0, -1, 0, 1, 0, -1, 0$ enz. zijn. Dit betekent voor algemene n dat

$$\sin^{(2n)}(0) = 0 \quad \text{en} \quad \sin^{(2n+1)}(0) = (-1)^n.$$

De Taylor reeks van $\sin(x)$ in het punt 0 is dus

$$T(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \ldots.$$

(3) $\cos(x)$:
Analoog met de sinus geldt voor de cosinus dat $\cos''(x) = -\cos(x)$, $\cos'''(x) = -\cos'(x)$ en $\cos^{(4)}(x) = \cos(x)$, dus zijn hier de afgeleiden in het punt $x_0 = 0$ periodiek $1, 0, -1, 0, 1, 0, -1, 0, 1$ enz. We hebben dus

$$\cos^{(2n)}(0) = (-1)^n \quad \text{en} \quad \cos^{(2n+1)}(0) = 0.$$

De Taylor reeks van $\cos(x)$ in het punt 0 is dus

$$T(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \dots.$$

Net als bij de exponentiële functie is de convergentiestraal ook bij de Taylor reeks van de sinus en cosinus functies oneindig. Verder is ook hier de convergentie van de Taylor reeks erg goed, zo dat we snel een goede benadering van $\sin(x)$ of $\cos(x)$ vinden.

(4) $\log(x)$:
Omdat de logaritme voor $x = 0$ niet gedefinieerd is, moeten we hier de Taylor reeks in een andere punt bepalen, en we kiezen hiervoor $x = 1$. Maar omdat het uiteindelijk toch prettiger is om een reeks met termen x^n en niet $(x-1)^n$ te hebben, gebruiken we een klein trucje: In plaats van $\log(x)$ kijken we naar de functie $f(x) := \log(x+1)$ en bepalen hiervoor de Taylor reeks in het punt $x_0 = 0$.

Er geldt $f'(x) = \frac{1}{x+1} = (x+1)^{-1}$, $f''(x) = (-1) \cdot (x+1)^{-2}$, $f'''(x) = (-1)(-2) \cdot (x+1)^{-3}$, $f^{(4)}(x) = (-1)(-2)(-3) \cdot (x+1)^{-4}$ en algemeen

$$f^{(n)}(x) = (-1)(-2)\dots(-(n-1)) \cdot (x+1)^{-n} = (-1)^{n-1}(n-1)! \frac{1}{(x+1)^n}.$$

In het bijzonder is $f^{(n)}(0) = (-1)^{n-1}(n-1)!$ en dus $\frac{f^{(n)}(0)}{n!} = \frac{(-1)^{n-1}}{n}$. De Taylor reeks van $\log(x+1)$ in het punt 0 is dus

$$T(x) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots.$$

Merk op dat deze Taylor reeks hoogstens convergentiestraal 1 kan hebben, omdat de logaritme in 0 niet gedefinieerd is. Dit is echter ook de convergentiestraal, we kunnen met deze reeks dus alleen maar waarden in het interval $(0, 2)$ bepalen. Maar dit stelt geen probleem voor het berekenen van $\log(x)$ voor, want voor $x \geq 2$ is $\frac{1}{x} \leq \frac{1}{2}$ en dit ligt wel in het interval $[0, 2]$ en we berekenen $\log(x)$ met behulp van de relatie $\log(x) = -\log(\frac{1}{x})$.

Er valt wel nog op te merken, dat de convergentie van de reeks voor de logaritme veel slechter is dan die voor $\exp(x)$, $\sin(x)$ of $\cos(x)$, omdat de noemers met n en niet met $n!$ groeien.

Opdracht 10 *Bepaal de Taylor reeks van $f(x) := \sqrt{x+1}$ in het punt $x_0 = 0$.*

2.4 Taylor reeksen voor functies van meerdere variabelen

We hebben in het eerste deel van deze les gekeken hoe we een gewone functie van één variabel door een oneindige reeks kunnen beschrijven of door veeltermen kunnen benaderen die we door afbreken van de Taylor reeks krijgen. De benadering van een functie $f(x)$ door de Taylor veelterm van graad n wordt vaak ook kort de n-de benadering van $f(x)$ genoemd. De meest belangrijke van

deze benaderingen zijn de 0de (ook al lijkt het flauw), de 1ste of *lineaire* en de 2de of *kwadratische* benadering.

We zullen nu kijken, hoe we Taylor veeltermen op functies van meerdere veranderlijken kunnen veralgemenen. Het idee dat we hierbij hanteren is hetzelfde: We benaderen een functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ door een veelterm $p(\mathbf{x})$ in de n variabelen die in een vaste punt $\mathbf{x}_0$ dezelfde functiewaarde heeft als $f(\mathbf{x})$ en ook dezelfde eerste, tweede enz. partiële afgeleiden.

Om de notatie overzichtelijk te houden, zullen we weer vooral naar functies van twee variabelen kijken, de resultaten laten zich dan makkelijk ook voor n variabelen formuleren.

Om te beginnen moeten we afspreken wat een veelterm van meerdere variabelen is, en wat de graad daarvan is.

Een *veelterm* in de twee variabelen x en y is een (eindige) som van termen van de vorm cx^iy^j, waarbij we c de *coëfficiënt* van x^iy^j noemen. Een zuiver product x^iy^j van machten van de variabelen noemt men ook en *monoom*. De *graad* van een term cx^iy^j is de som $i + j$ van de machten van de variabelen en de graad van een veelterm is het maximum van de graden van de termen. Een veelterm van graad 0 heet een *constante functie*, een veelterm van graad 1 een *lineaire veelterm* of *lineaire functie* en een veelterm van graad 2 een *kwadratische veelterm* of *kwadratische functie*.

Voorbeeld: De algemene kwadratische veelterm van twee variabelen is van de vorm

$$p(x, y) = a + b_1x + b_2y + c_1x^2 + c_2xy + c_3y^2.$$

We zullen nu de algemene definitie van de begrippen *veelterm* en *graad* voor n variabelen geven, dit vergt enigszins veel indices en ingewikkelde notaties, maar met het geval van twee variabelen voor ogen zal het wel begrijpelijk zijn.

Definitie: Een *veelterm* in de n variabelen $x_1, \ldots, x_n$ is een eindige som van termen van de vorm

$$cx_1^{i_1} \cdot \ldots \cdot x_n^{i_n}.$$

De *graad* van een term $cx_1^{i_1} \cdot \ldots \cdot x_n^{i_n}$ is de som $i_1 + \ldots + i_n$ van de machten van de variabelen. De *graad van een veelterm* is het maximum van de graden van de termen in de veelterm.

Voorbeeld:

(i) De algemene lineaire veelterm in $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$p(\mathbf{x}) = a + b_1x_1 + \ldots + b_nx_n = a + \sum_{i=1}^{n} b_ix_i.$$

(ii) De algemene kwadratische veelterm is van de vorm

$$p(\mathbf{x}) = a + \sum_{i=1}^{n} b_ix_i + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_ix_j.$$

Als we nu een functie $f(x, y)$ door een lineaire veelterm $p(x, y)$ willen benaderen, kunnen we eisen dat $p(x_0, y_0) = f(x_0, y_0)$ voor een vast gekozen punt (x_0, y_0) en dat de eerste partiële afgeleiden van $p(x, y)$ overeenkomen met de eerste partiële afgeleiden van $f(x, y)$ in het punt (x_0, y_0).

Voor het gemak nemen we nu nog aan, dat het vast gekozen punt (x_0, y_0) het nulpunt $(0, 0)$ is, we komen naar een algemeen punt terug door x door $h_1 := x - x_0$ en y door $h_2 := y - y_0$ te vervangen.

Voor de algemene kwadratische veelterm

$$p(x, y) = a + b_1 x + b_2 y + c_1 x^2 + c_2 xy + c_3 y^2$$

geldt

$$\frac{\partial p}{\partial x} = b_1 + 2c_1 x + c_2 y \quad \text{en} \quad \frac{\partial p}{\partial y} = b_2 + c_2 x + 2c_3 y.$$

Hieruit volgt

$$p(0, 0) = a, \quad \frac{\partial p}{\partial x}(0, 0) = b_1, \quad \frac{\partial p}{\partial y}(0, 0) = b_2$$

en om te bereiken dat $p(0, 0) = f(0, 0)$, $\frac{\partial p}{\partial x}(0, 0) = \frac{\partial f}{\partial x}(0, 0)$ en $\frac{\partial p}{\partial y}(0, 0) = \frac{\partial f}{\partial y}(0, 0)$ moet gelden dat

$$a = f(0, 0), \quad b_1 = \frac{\partial f}{\partial x}(0, 0), \quad b_2 = \frac{\partial f}{\partial y}(0, 0).$$

Voor de *lineaire benadering* van $f(x, y)$ in het punt $(x, y) = (0, 0)$ krijgen we dus de lineaire veelterm

$$p(x, y) = f(0, 0) + \frac{\partial f}{\partial x}(0, 0)\, x + \frac{\partial f}{\partial y}(0, 0)\, y.$$

Omdat $p(x, y)$ juist zo gekozen is dat de functiewaarde en de eerste partiële afgeleiden met $f(x, y)$ overeenkomen, geeft $p(x, y)$ juist het raakvlak aan de grafiek van $f(x, y)$ in het punt (x_0, y_0) aan.

We kunnen de lineaire benadering met behulp van de gradiënt nog iets eenvoudiger schrijven: Met de notaties $\mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ en $\mathbf{h} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt:

$$p(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h}.$$

Met deze notatie hebben we inderdaad de algemene vorm van de Taylor veelterm van graad 1 gevonden:

Stelling: Voor een functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is de Taylor veelterm van graad 1 voor het punt $\mathbf{x}_0$ gegeven door

$$p(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h} = f(\mathbf{x}_0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, h_i$$

waarbij h_i de i-de component van de vector $\mathbf{h}$ is, dus $\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$.

Om naar de Taylor veelterm van graad 2 te komen, moeten we nu ervoor zorgen dat ook de tweede partiële afgeleiden van $p(x,y)$ in het punt $(0,0)$ met de tweede partiële afgeleiden van $f(x,y)$ in dit punt overeenkomen. moeten we nog de tweede partiële van $p(x,y)$ bepalen. Hiervoor krijgen we:

$$\frac{\partial^2 p}{\partial x^2} = 2c_1, \quad \frac{\partial^2 p}{\partial x\,\partial y} = \frac{\partial^2 p}{\partial y\,\partial x} = c_2, \quad \frac{\partial^2 p}{\partial y^2} = 2c_3$$

en gelijk zetten met de tweede partiële afgeleiden van $f(x,y)$ in het punt $(0,0)$ geeft:

$$c_1 = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(0,0), \quad c_2 = \frac{\partial^2 f}{\partial x\,\partial y}(0,0), \quad c_3 = \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(0,0).$$

Voor de *kwadratische benadering* van $f(x,y)$ in het punt $(x,y) = (0,0)$ krijgen we dus de kwadratische veelterm

$$\begin{aligned} p(x,y) &= f(0,0) + \frac{\partial f}{\partial x}(0,0)\,x + \frac{\partial f}{\partial y}(0,0)\,y \\ &\quad + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(0,0)\,x^2 + \frac{\partial^2 f}{\partial x\,\partial y}(0,0)\,xy + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(0,0)\,y^2 \\ &= f(0,0) + \frac{\partial f}{\partial x}(0,0)\,x + \frac{\partial f}{\partial y}(0,0)\,y \\ &\quad + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}(0,0)\,x^2 + \frac{\partial^2 f}{\partial x\,\partial y}(0,0)\,xy + \frac{\partial^2 f}{\partial y\,\partial x}(0,0)\,yx + \frac{\partial^2 f}{\partial y^2}(0,0)\,y^2\right). \end{aligned}$$

Het opsplitsen van $\frac{\partial^2 f}{\partial x\,\partial y}(0,0)\,xy$ in de som $\frac{1}{2}(\frac{\partial^2 f}{\partial x\,\partial y}(0,0)\,xy + \frac{\partial^2 f}{\partial y\,\partial x}(0,0)\,yx)$ lijkt niet echt een vereenvoudiging, maar we zullen nu zien dat we hier wel iets aan hebben.

Net zo als we de eerste partiële afgeleiden in een vector, de gradiënt samengevat hebben, kunnen we de tweede partiële afgeleiden in een matrix samenvatten, voor een functie $f(x,y)$ van twee variabelen geeft dit de 2×2-matrix

$$H := H_f := \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y\,\partial x} \\ \frac{\partial^2 f}{\partial x\,\partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

Deze matrix heet de *Hesse matrix* van f. Voor een algemene functie $f(\mathbf{x})$ van n variabelen is de Hesse matrix een $n \times n$-matrix met in de i-de rij de partiële afgeleiden naar de i-de variabel x_i en in de j-de kolom de partiële afgeleiden naar x_j, dus op positie (i,j) staat de tweede partiële afgeleide naar x_i en x_j.

Definitie: Voor een functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ heet de matrix

$$H := H_f := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n\,\partial x_1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1\,\partial x_n} & \frac{\partial^2 f}{\partial x_2\,\partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

met $H_{ij} = \frac{\partial^2 f}{\partial x_j\,\partial x_i}$ de *Hesse matrix* van $f(\mathbf{x})$.

Omdat volgens de Stelling van Schwarz geldt dat $\frac{\partial^2 f}{\partial x_j\,\partial x_i} = \frac{\partial^2 f}{\partial x_i\,\partial x_j}$, is de Hesse matrix een symmetrische matrix, d.w.z. er geldt $H = H^{tr}$.

We noteren met $H(\mathbf{x}_0)$ (of $H(x,y)$ in het geval van twee variabelen) de Hesse matrix waarvoor de partiële afgeleiden in het punt $\mathbf{x}_0$ (of het punt (x,y)) geëvalueerd zijn.

Voor $\mathbf{h} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt dan:

$$\begin{aligned}\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h} &= \frac{\partial^2 f}{\partial x^2}(\mathbf{x}_0)\,x^2 + \frac{\partial^2 f}{\partial y\,\partial x}(\mathbf{x}_0)\,xy + \frac{\partial^2 f}{\partial x\,\partial y}(\mathbf{x}_0)\,yx + \frac{\partial^2 f}{\partial y^2}(\mathbf{x}_0)\,y^2 \\ &= \frac{\partial^2 f}{\partial x^2}(\mathbf{x}_0)\,x^2 + 2\frac{\partial^2 f}{\partial y\,\partial x}(\mathbf{x}_0)\,xy + \frac{\partial^2 f}{\partial y^2}(\mathbf{x}_0)\,y^2.\end{aligned}$$

Met behulp van de Hesse matrix kunnen we Taylor veelterm van graad 2 daarom als volgt schrijven:

$$p(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h} + \frac{1}{2}\,\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h}.$$

Ook hier hebben we met deze notatie de algemene vorm van de Taylor veelterm van graad 2 gevonden, er geldt:

Stelling: Voor een functie $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is de Taylor veelterm van graad 1 voor het punt $\mathbf{x}_0$ gegeven door

$$p(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h} + \frac{1}{2}\,\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h}.$$

waarbij h_i de i-de component van de vector $\mathbf{h}$ is.

Voorbeelden:

(1) We bekijken de functie

$$f(x,y) := e^x \cos(y)$$

in het punt $(x_0, y_0) = (0,0)$. Er geldt

$$\frac{\partial f}{\partial x} = e^x \cos(y), \quad \frac{\partial f}{\partial y} = -e^x \sin(y),$$

en dus geldt voor de gradiënt

$$\nabla f = \begin{pmatrix} e^x \cos(y) \\ -e^x \sin(y) \end{pmatrix} \quad \text{en} \quad \nabla f(0,0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Voor de tweede partiële afgeleiden geldt

$$\frac{\partial^2 f}{\partial x^2} = e^x \cos(y), \quad \frac{\partial^2 f}{\partial y\,\partial x} = -e^x \sin(y), \quad \frac{\partial^2 f}{\partial y^2} = -e^x \cos(y),$$

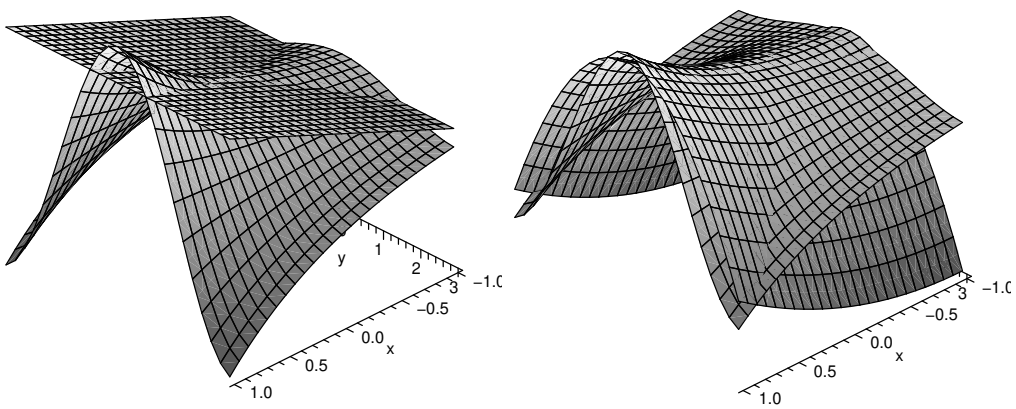
en dus krijgen we de Hesse matrix

$$H = \begin{pmatrix} e^x \cos(y) & -e^x \sin(y) \\ -e^x \sin(y) & -e^x \cos(y) \end{pmatrix} \quad \text{met} \quad H(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Omdat $f(0,0) = 1$, zijn de Taylor veeltermen van graad 1 en 2 van $f(x,y)$ gegeven door

$$p_1(x,y) = 1 + x \quad \text{en} \quad p_2(x,y) = 1 + x + \frac{1}{2}x^2 - \frac{1}{2}y^2.$$

In Figuur I.9 zijn de grafiek van de functie en de grafieken van de benadering door de Taylor veeltermen van graad 1 en 2 te zien. Het is duidelijk dat de lineaire benadering het raakvlak aan de grafiek geeft en men ziet goed dat de kwadratische benadering in een omgeving van $(0,0)$ al redelijk goed is.



Figuur I.9: Benadering van $e^x \cos(y)$ door Taylor veeltermen van graad 1 en 2.

(2) We bekijken de functie

$$f(x,y) := \sin(xy)$$

in het punt $(x_0, y_0) = (1, \frac{\pi}{2})$. Er geldt

$$\frac{\partial f}{\partial x} = y\,\cos(xy), \quad \frac{\partial f}{\partial y} = x\,\cos(xy),$$

en dus geldt voor de gradiënt

$$\nabla f = \begin{pmatrix} y\,\cos(xy) \\ x\,\cos(xy) \end{pmatrix} \quad \text{en} \quad \nabla f(1, \frac{\pi}{2}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

De lineaire benadering van $f(x,y)$ is dus de constante $f(0,0) = 1$.

Voor de tweede partiële afgeleiden geldt

$$\frac{\partial^2 f}{\partial x^2} = -y^2 \sin(xy), \ \frac{\partial^2 f}{\partial y\,\partial x} = \cos(xy) - xy\,\sin(xy), \ \frac{\partial^2 f}{\partial y^2} = -x^2 \sin(xy),$$

en dus krijgen we de Hesse matrix

$$H = \begin{pmatrix} -y^2 \sin(xy) & \cos(xy) - xy\,\sin(xy) \\ \cos(xy) - xy\,\sin(xy) & -x^2 \sin(xy) \end{pmatrix}$$

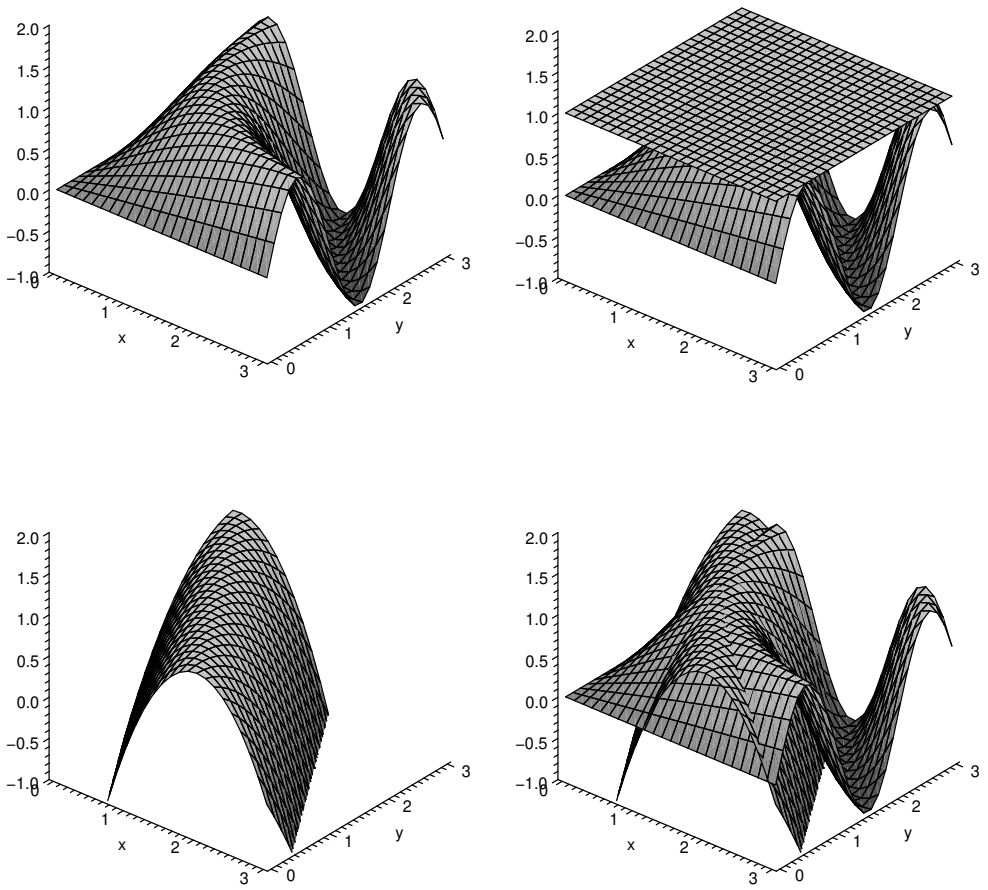
en dus

$$H(0,0) = \begin{pmatrix} -\frac{\pi^2}{4} & -\frac{\pi}{2} \\ -\frac{\pi}{2} & -1 \end{pmatrix}.$$

De Taylor veeltermen van graad 2 van $f(x,y)$ is dus gegeven door

$$\begin{aligned} p(x,y) &= 1 + \frac{1}{2}\left(-\frac{\pi^2}{4}(x-1)^2 - 2\frac{\pi}{2}(x-1)(y-\frac{\pi}{2}) - (y-\frac{\pi}{2})^2\right) \\ &= 1 - \frac{\pi^2}{8}(x-1)^2 - \frac{\pi}{2}(x-1)(y-\frac{\pi}{2}) - \frac{1}{2}(y-\frac{\pi}{2})^2). \end{aligned}$$

In Figuur I.10 zijn boven de grafiek van de functie en de (constante) lineaire benadering te zien, beneden de grafiek van de kwadratische benadering apart en de kwadratische benadering samen met de grafiek van de functie.



Figuur I.10: Benadering van $\sin(xy)$ door Taylor veeltermen van graad 1 en 2.

Opdracht 11 *Bepaal voor* $f(x,y) := e^{x^2+y^2}$ *de Taylor veelterm van graad* 2 *in het punt* $(0,0)$.

Taylor reeksen voor functies van meerdere veranderlijken

Als we tot Taylor veeltermen van hogere graden dan 2 en misschien zelfs tot Taylor reeksen voor functies van meerdere variabelen willen komen, wordt het vergelijken van de partiële afgeleiden van een veelterm met algemene coëfficiënten met de partiële afgeleiden van $f(\mathbf{x})$ snel erg onhandig. Maar gelukkig kunnen we de Taylor veeltermen en Taylor reeks voor functies van meerdere variabelen met een klein trucje ook afleiden uit de Taylor reeks voor een functie van één variabel.

Als we een functie $f(\mathbf{x})$ in een kleine omgeving van een vast gekozen punt $\mathbf{x}_0$ bekijken, kunnen we dit schrijven als $f(\mathbf{x}_0 + \mathbf{h})$ voor een (korte) vector $\mathbf{h}$. Hieruit maken we nu (kunstmatig) een nieuwe functie $g(t)$ van één variabel, namelijk

$$g(t) := f(\mathbf{x}_0 + t \cdot \mathbf{h}).$$

We krijgen de Taylor reeks van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ door $g(t)$ rond $t = 0$ in een Taylor reeks te ontwikkelen en voor $f(\mathbf{x}_0 + \mathbf{h})$ geldt $f(\mathbf{x}_0 + \mathbf{h}) = g(1)$. Er geldt dus:

$$\begin{aligned} f(\mathbf{x}_0 + h) = g(1) &= g(0) + g'(0) \cdot 1 + \frac{1}{2} g''(0) \cdot 1^2 + \ldots + \frac{1}{n!} g^{(n)}(0) \cdot 1^n + \ldots \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} g^{(n)}(0). \end{aligned}$$

Voor de Taylor reeks van $f(\mathbf{x})$ moeten we dus alleen maar de afgeleiden van $g(t)$ in het punt $t = 0$ bepalen. De grap is nu dat de afgeleide $g'(t)$ juist de richtingsafgeleide van $f(\mathbf{x})$ in de richting van $\mathbf{h}$ is, voor een vector $\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$ met componenten h_i geldt dus

$$g'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + t\mathbf{h})\, h_i \quad \text{en dus} \quad g'(0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, h_i.$$

Voor de tweede afgeleide $g''(t)$ moeten we weer de richtingsafgeleide in de richting van $\mathbf{h}$ nemen, dit geeft

$$g''(t) = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_j\, \partial x_i}(\mathbf{x}_0 + t\mathbf{h})\, h_i h_j \quad \text{en} \quad g''(0) = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_j\, \partial x_i}(\mathbf{x}_0)\, h_i h_j.$$

Op een soortgelijke manier krijgen we voor de derde afgeleide

$$\begin{aligned} g'''(t) &= \sum_{k=1}^{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^3 f}{\partial x_k\, \partial x_j\, \partial x_i}(\mathbf{x}_0 + t\mathbf{h})\, h_i h_j h_k, \\ g'''(0) &= \sum_{k=1}^{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^3 f}{\partial x_k\, \partial x_j\, \partial x_i}(\mathbf{x}_0)\, h_i h_j h_k \end{aligned}$$

en we kunnen in principe zo doorgaan tot hogere afgeleiden, de n-de term is dan een n-voudige som van de n-de partiële afgeleiden van $f(\mathbf{x})$. Maar ook deze sommen zijn nog niet echt prettig om op te schrijven, daarom zullen we nu nog naar een veel gebruikte notatie kijken, waarmee de Taylor reeks voor meerdere variabelen bijna net zo eenvoudig wordt als de Taylor reeks voor functies van één variabel.

Het idee is, de partiële afgeleide als een soort bewerkingsvoorschrift te beschouwen, die we op een functie toepassen en die we een *operator* noemen. In principe is dit niets nieuws, want ook de gewone afgeleide kunnen we zien als een operator $'$ die uit een functie $f(x)$ een andere functie $f'(x)$ maakt. Net zo interpreteren we nu de partiële afgeleide als een operator $\frac{\partial}{\partial x}$ die uit de functie $f(\mathbf{x})$ de nieuwe functie $\frac{\partial}{\partial x} f = \frac{\partial f}{\partial x}$ maakt.

Met deze operatoren kunnen we nu op een voor de hand liggende manier rekenen, voor de som $\frac{\partial}{\partial x} + \frac{\partial}{\partial y}$ van twee operatoren moeten we hiervoor aangeven, welke nieuwe functie de toepassing van $\frac{\partial}{\partial x} + \frac{\partial}{\partial y}$ op een functie $f(\mathbf{x})$ geeft. Als verdere bewerkingen definiëren we ook de vermenigvuldiging van twee operatoren en het vermenigvuldigen van een operator met een constante:

$$
\begin{aligned}
(c \cdot \frac{\partial}{\partial x_i}) f &:= c \cdot \frac{\partial f}{\partial x_i} \\
(\frac{\partial}{\partial x_i} + \frac{\partial}{\partial x_j}) f &:= \frac{\partial f}{\partial x_i} + \frac{\partial f}{\partial x_j} \\
(\frac{\partial}{\partial x_i} \cdot \frac{\partial}{\partial x_j}) f &:= \frac{\partial^2 f}{\partial x_i \, \partial x_j}.
\end{aligned}
$$

Hiermee krijgen we bijvoorbeeld

$$
(\frac{\partial}{\partial x} + \frac{\partial}{\partial y})^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y \, \partial x} + \frac{\partial^2}{\partial x \, \partial y} + \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial x^2} + 2\frac{\partial^2}{\partial y \, \partial x} + \frac{\partial^2}{\partial y^2}.
$$

Merk op: Dit is in principe dezelfde formule als de binomische formule

$$
(a+b)^2 = a^2 + 2ab + b^2
$$

en dit betekent dat we met de partiële afgeleiden als operatoren juist zo rekenen als met gewone variabelen.

Met behulp van de operatoren kunnen we het inproduct van een vector $\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}$ met de gradiënt ∇f herschrijven als

$$
\nabla f(\mathbf{x}_0) \cdot \mathbf{h} = (h_1 \cdot \frac{\partial}{\partial x_1} + \ldots + h_n \cdot \frac{\partial}{\partial x_n}) f(\mathbf{x}_0)
$$

waarbij het invullen van $\mathbf{x}_0$ aan de rechterkant natuurlijk *na* het toepassen van de operatoren gebeurt.

Net zo krijgen we voor het product met de Hesse matrix

$$\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h} = (h_1 \cdot \frac{\partial}{\partial x_1} + \ldots + h_n \cdot \frac{\partial}{\partial x_n})^2 f(\mathbf{x}_0).$$

De Taylor reeks voor een functie van n variabelen laat zich nu schrijven als:

$$f(\mathbf{x}_0 + \mathbf{h}) = \sum_{n=0}^{\infty} \frac{1}{n!}(h_1 \cdot \frac{\partial}{\partial x_1} + \ldots + h_n \cdot \frac{\partial}{\partial x_n})^n f(\mathbf{x}_0).$$

Voorbeeld: We berekenen de derdegraads term van de Taylor reeks voor een functie $f(x, y)$ van twee variabelen. Er geldt $(a+b)^3 = a^3+3a^2b+3ab^2+b^3$, hieruit volgt

$$(h_1\frac{\partial}{\partial x} + h_2\frac{\partial}{\partial y})^3 = h_1^3\frac{\partial^3}{\partial x^3} + 3h_1^2h_2\frac{\partial^3}{\partial y\,\partial x^2} + 3h_1h_2^2\frac{\partial^3}{\partial y^2\,\partial x} + h_2^3\frac{\partial^3}{\partial y^3}$$

dus is de derdegraads term van de Taylor reeks

$$\frac{1}{3!}\left(\frac{\partial^3 f}{\partial x^3}(\mathbf{x}_0)\,h_1^3 + 3\frac{\partial^3 f}{\partial y\,\partial x^2}(\mathbf{x}_0)\,h_1^2h_2 + 3\frac{\partial^3 f}{\partial y^2\,\partial x}(\mathbf{x}_0)\,h_1h_2^2 + \frac{\partial^3 f}{\partial y^3}(\mathbf{x}_0)\,h_2^3\right).$$

Belangrijke begrippen in deze les

- Lagrange interpolatie
- Taylor veelterm
- foutterm
- oneindige reeksen
- Taylor reeks
- lineaire/kwadratische benadering
- Hesse matrix
- partiële afgeleide als operator

Opgaven

12. Bepaal voor $f(x) := (1 + x)^a$ met $a \in \mathbb{R}$ de Taylor veelterm van graad 3 in $x_0 = 0$.
13. Bepaal voor de volgende functies de Taylor veelterm van graad 5 in $x_0 = 0$:
    (i) $f(x) := \sqrt[3]{1 + x} = (1 + x)^{\frac{1}{3}}$;
    (ii) $f(x) := \frac{1}{1+x^2}$;
    (iii) $f(x) := \frac{1}{\sqrt{1-x^2}}$;

(iv) $f(x) := \frac{1}{3+x}$;
(v) $f(x) := \sin(2x^2)$;
(vi) $f(x) := e^{-3x}$.

14. Zij $f(x) := \sqrt[3]{8+x} = (8+x)^{\frac{1}{3}}$.
    (i) Bepaal voor $f(x)$ de Taylor veelterm van graad 4 in $x_0 = 0$.
    (ii) Vind met behulp van de Taylor veelterm in (i) een benadering voor $\sqrt[3]{9}$.
    (iii) Geef met behulp van de Lagrange vorm van de foutterm een afschatting voor de mogelijke fout van de benadering in (ii).

15. Bepaal voor de volgende functies de Taylor reeksen in de aangeven punten:
    (i) $f(x) := \frac{1}{x}$ in $x_0 = 1$;
    (ii) $f(x) := e^x$ in $x_0 = 2$;
    (iii) $f(x) := \sin(x+1)$ in $x_0 = -1$;
    (iv) $f(x) := \log(x)$ in $x_0 = 2$.

16. Vind de Taylor reeks voor $f(x) := \arctan(x)$ in $x_0 = 0$. (Hint: Er geldt $\arctan'(x) = \frac{1}{1+x^2}$ en dit laat zich voor $x^2 < 1$ schrijven als meetkundige reeks $\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \ldots = \sum_{n=0}^{\infty}(-1)^n x^{2n}$.)

17. Bepaal voor de functie $f(x,y) := \sqrt{x^2+y^3}$ de Taylor veelterm van graad 2 in het punt $(1,2)$. Gebruik de Taylor veelterm om de waarde van $\sqrt{1.02^2 + 1.97^3}$ te benaderen en vergelijk het resultaat met de *juiste* waarde (volgens een rekenmachine).

18. Bepaal voor de volgende functies de Taylor veelterm van graad 2 in het punt $(x_0, y_0) = (0,0)$:
    (i) $f(x,y) := \frac{1}{1+x^2+y^2}$;
    (ii) $f(x,y) := e^{x+y}$;
    (iii) $f(x,y) := e^{-x^2-y^2}\cos(xy)$;
    (iv) $f(x,y) := \sin(xy) + \cos(xy)$.

19. Bepaal voor de volgende functies de Taylor veelterm van graad 2 in de aangegeven punten (x_0, y_0):
    (i) $f(x,y) := \frac{1}{2+xy^2}$ in $(0,0)$;
    (ii) $f(x,y) := \log(1+x+y+xy)$ in $(0,0)$;
    (iii) $f(x,y) := \arctan(x+xy)$ in $(0,-1)$;
    (iv) $f(x,y) := x^2 + xy + y^3$ in $(1,-1)$;
    (v) $f(x,y) := \sin(2x+3y)$ in $(0,0)$;
    (vi) $f(x,y) := \frac{\sin(x)}{y}$ in $(\frac{\pi}{2}, 1)$;
    (vii) $f(x,y) := \frac{1+x}{1+x^2+y^4}$ in $(0,0)$;
    (viii) $f(x,y) := e^{(x-1)^2}\cos(y)$ in $(1,0)$.

20. Bepaal voor de volgende functies de Taylor veelterm van graad 3 in de aangegeven punten (x_0, y_0):
    (i) $f(x,y) := \frac{1}{2+x-2y}$ in $(2,1)$;
    (ii) $f(x,y) := \log(x^2+y^2)$ in $(1,0)$;
    (iii) $f(x,y) := \cos(x+\sin(y))$ in $(0,0)$.

Les 3 Extrema van functies van meerdere variabelen

Bij gewone functies van één variabel hebben we in Wiskunde 1 de vraag behandeld hoe we minima en maxima van een functie kunnen vinden. Het belangrijkste criterium was dat een differentieerbare functie in een lokaal extremum (minimum of maximum) een horizontale raaklijn heeft, de afgeleide van de functie in een extremum is dus noodzakelijk nul. Punten met deze eigenschap hebben we *kritieke punten* genoemd, naast de speciale punten waar de functie niet differentieerbaar is en de randpunten van het interval waarop we de functie bekijken.

3.1 Classificatie van kritieke punten

Het kan zijn dat een functie in een punt een horizontale raaklijn heeft, zonder in dit punt een extremum te hebben. Dit is bijvoorbeeld het geval voor de functie $f(x) = x^3$ in het punt $x = 0$. Zo'n punt noemt men ook een *zadelpunt*. Het verschil tussen een zadelpunt en een echt extremum laat zich aan de hand van de tweede afgeleide $f''(x)$ beschrijven:

Bij een minimum in het punt x_0 is de functie links van x_0 (dus voor $x < x_0$) dalend en rechts van x_0 stijgend, dus is de eerste afgeleide $f'(x)$ links van x_0 negatief en rechts van x_0 positief. Dit betekent dat $f'(x)$ rond x_0 stijgend is en dus geldt $f''(x_0) > 0$. Net zo volgt uit $f'(x_0) = 0$ en $f''(x_0) < 0$, dat de functie $f(x)$ in het punt x_0 een maximum heeft.

Met onze kennis van Taylor reeksen kunnen we dit nu ook van een andere kant bekijken. We ontwikkelen $f(x)$ rond een kritiek punt x_0 met $f'(x_0) = 0$ in een Taylor reeks, dit geeft:

$$f(x_0 + h) = f(x_0) + \frac{1}{2} f''(x_0)\, h^2 + \dots$$

In een kleine omgeving van x_0 kunnen we de hogere termen h^n tegenover h^2 verwaarlozen, en om na te gaan of $f(x)$ in x_0 een lokaal extremum heeft, kunnen we in plaats van $f(x)$ de kwadratische functie

$$g(h) := f(x_0) + \frac{1}{2} f''(x_0)\, h^2$$

bekijken. Maar dit is juist de vergelijking van een parabool met toppunt $(0, f(x_0))$ en deze parabool is naar boven geopend als $f''(x_0) > 0$ en naar beneden geopend als $f''(x_0) < 0$. Voor $f''(x_0) > 0$ heeft $f(x)$ dus een minimum en voor $f''(x_0) < 0$ een maximum.

Als in een kritiek punt x_0 ook de tweede afgeleide $f''(x_0) = 0$ is, kunnen we nog steeds niet beslissen of de functie een minimum, maximum of een zadelpunt heeft. In dit geval moeten we de hogere afgeleiden bepalen tot dat we naar een n komen met $f^{(n)}(x_0) \neq 0$. Dan kunnen we weer de Taylor reeks van $f(x)$ in x_0 bepalen, deze is

$$f(x_0 + h) = f(x_0) + \frac{1}{n!} f^{(n)}(x_0)\, h^n + \dots$$

en in een kleine omgeving van x_0 kunnen we in plaats van $f(x)$ naar de functie

$$g(h) := f(x_0) + \frac{1}{n!} f^{(n)}(x_0)\, h^n$$

kijken. Nu zijn er drie mogelijke gevallen:

(1) Als n oneven is, heeft $f(x)$ in x_0 een zadelpunt, omdat $g(h)$ (tot op een verschuiving en een schaling na) een functie van de vorm h^3, h^5, enz. is.

(2) Als n even is en $f^{(n)}(x_0) > 0$, heeft $f(x)$ in x_0 een minimum, want in dit geval is $g(h)$ naar boven geopend (net zo als de functies $2x^4$ of πx^6).

(3) Als n even is en $f^{(n)}(x_0) < 0$, heeft $f(x)$ in x_0 een maximum, want in dit geval is $g(h)$ naar beneden geopend (net zo als $-3x^4$ of $-\sqrt{2}x^6$).

We krijgen dus de volgende classificatie voor kritieke punten van differentieerbare functies:

Stelling: Zij $f(x)$ een in het punt x_0 differentieerbare functie met $f'(x_0) = 0$ en zij $n \geq 2$ de kleinste n met $f^{(n)}(x_0) \neq 0$. Dan geldt:

(i) $f(x)$ heeft in x_0 een minimum als n even is en $f^{(n)}(x_0) > 0$;

(ii) $f(x)$ heeft in x_0 een maximum als n even is en $f^{(n)}(x_0) < 0$;

(iii) $f(x)$ heeft in x_0 een zadelpunt als n oneven is.

3.2 Kritieke punten van functies van meerdere variabelen

We kijken nu naar de vraag hoe we lokale extrema van functies van meerdere veranderlijken kunnen vinden. De ideeën die we hierbij hanteren zijn in principe hetzelfde als bij de gewone functies, we moeten alleen maar de afgeleide door de partiële afgeleiden vervangen.

We starten weer met de beschrijving van de consequenties van een extremum. Als een functie $f(\mathbf{x})$ in een punt $\mathbf{x}_0$ een lokaal extremum heeft, kunnen we naar de partiële afgeleiden in dit punt kijken. Maar bij de partiële afgeleide $\frac{\partial f}{\partial x_i}$ bekijken we de verandering van een functie $g(x_i)$ die we uit $f(\mathbf{x})$ krijgen, door de andere variabelen $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ als constanten te beschouwen. Hieruit volgt dat de functie $f(\mathbf{x})$ alleen maar een extremum kan hebben, als de functie $g(x_i)$ van één variabel een extremum heeft, en dit betekent dat $\frac{\partial f}{\partial x_i}(\mathbf{x}_0) = 0$ moet zijn. Omdat dit argument voor iedere variabel x_i geldt, krijgen we als noodzakelijke voorwaarde voor een extremum in $\mathbf{x}_0$, dat de gradiënt in $\mathbf{x}_0$ nul moet zijn, dus:

Stelling: Als een functie $f(\mathbf{x})$ in een punt $\mathbf{x}_0$ een lokaal minimum of maximum heeft, dan geldt

$$\nabla f(\mathbf{x}_0) = 0.$$

Deze stelling kunnen we ook uit de interpretatie van de gradiënt afleiden, want de gradiënt $\nabla f(\mathbf{x}_0)$ wijst in de richting van de snelste toename van $f(\mathbf{x})$. Maar in een maximum mag de functie in geen enkele richting toenemen, dus

moet de gradiënt nul zijn. Net zo wijst $-\nabla f(\mathbf{x}_0)$ in de richting van de snelste afname van de functie, en in een minimum mag de functie in geen richting afnemen, dus moet ook hier de gradiënt nul zijn.

We kunnen ook vanuit het perspectief van de Taylor reeks argumenteren. De lineaire benadering

$$f(\mathbf{x}_0 + \mathbf{h}) \approx T(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h}$$

geeft het raakvlak aan de grafiek van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ aan. Maar in een lokaal extremum moet het raakvlak horizontaal zijn, dus moet de lineaire benadering in een minimum of maximum een constante zijn, en dit betekent ook weer dat $\nabla f(\mathbf{x}_0) = 0$ moet zijn.

Net zo als bij gewone functies noemen we de punten $\mathbf{x}_0$ die aan de noodzakelijke voorwaarde $\nabla f(\mathbf{x}_0) = 0$ voldoen, de *kritieke punten* van $f(\mathbf{x})$.

Definitie: De punten $\mathbf{x}_0$ waar voor een differentieerbare functie $f(\mathbf{x})$ de gradiënt $\nabla f(\mathbf{x}_0) = 0$ is, heten *kritieke punten* van $f(\mathbf{x})$. De kritieke punten zijn juist de kandidaten voor lokale minima of maxima van $f(\mathbf{x})$.

Voorbeelden:

(1) Een open doos met rechthoekig grondvlak moet een bepaald volume V bevatten. Wat zijn de optimale afmetingen van de doos zo dat we zo weinig materiaal als mogelijk nodig hebben?

Als de zijden van het grondvlak afmetingen x en y hebben, moet de hoogte z van de doos $z = \frac{V}{xy}$ zijn. De oppervlakte van de doos is dus een functie $A(x, y)$ van de afmetingen van het grondvlak en er geldt

$$A(x, y) = xy + 2x\frac{V}{xy} + 2y\frac{V}{xy} = xy + 2\frac{V}{y} + 2\frac{V}{x}.$$

Voor de partiële afgeleiden geldt

$$\frac{\partial A}{\partial x} = y - \frac{2V}{x^2} \quad \text{en} \quad \frac{\partial A}{\partial y} = x - \frac{2V}{y^2}$$

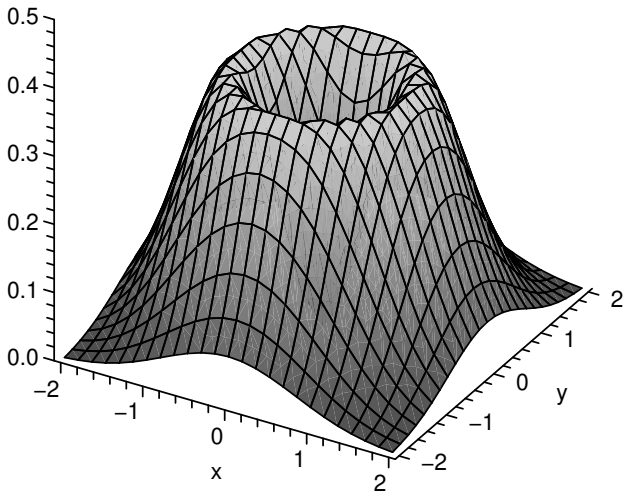
en uit $\frac{\partial A}{\partial x} = 0$ volgt $y = \frac{2V}{x^2}$. Dit ingevuld in $\frac{\partial A}{\partial y} = 0$ geeft $x = \frac{2V}{y^2} = \frac{x^4}{2V}$. Hieruit volgt $x = 0$ of $x^3 = 2V$, waarbij de eerste oplossing wegvalt, omdat in dit geval $y = \frac{2V}{x^2} = \infty$ zou moeten zijn.

Uit $y = \frac{2V}{x^2}$ volgt nu $y^3 = \frac{8V^3}{x^6} = \frac{8V^3}{4V^2} = 2V$, dus is $x = y = \sqrt[3]{2V}$ en we krijgen het niet erg verrassende resultaat dat het grondvlak een vierkant is.

Voor de hoogte z van de doos geldt dat $z = \frac{V}{xy} = \sqrt[3]{\frac{V^3}{4V^2}} = \sqrt[3]{\frac{V}{4}} = \frac{1}{2}\sqrt[3]{2V} = \frac{1}{2}x$, dus is de doos half zo hoog als lang en breed.

(2) We bepalen de kritieke punten van de functie

$$f(x, y) := (x^2 + y^2)\, e^{-x^2-y^2}$$



Figuur I.11: Grafiek van de functie $f(x,y) = (x^2+y^2)e^{-x^2-y^2}$.

waarvan de grafiek (in de vorm van een vulkaan) in Figuur I.11 te zien is. Voor de partiële afgeleiden geldt

$$\frac{\partial f}{\partial x} = 2x\,e^{-x^2-y^2} - 2x(x^2+y^2)\,e^{-x^2-y^2} = 2x\,e^{-x^2-y^2}(1-x^2-y^2)$$
$$\frac{\partial f}{\partial y} = 2y\,e^{-x^2-y^2} - 2y(x^2+y^2)\,e^{-x^2-y^2} = 2y\,e^{-x^2-y^2}(1-x^2-y^2)$$

en hieruit volgt dat $\nabla f(x,y) = 0$ voor $(x,y) = (0,0)$ en voor (x,y) met $x^2+y^2 = 1$, dus voor punten op een cirkel met straal 1 rond $(0,0)$. Het eerste geval geeft het minimum in het centrum van de vulkaan, het tweede geval geeft de lokale maxima op de rand van de vulkaan.

Merk op dat we de extreme alleen maar met behulp van de grafiek van de functie als minima of maxima hebben geïdentificeerd. Hoe we dit zonder grafiek kunnen herkennen, gaan we straks behandelen.

(3) We bekijken de functie $f(x,y) := x^2y + xy^2$. Er geldt

$$\frac{\partial f}{\partial x} = 2xy + y^2 = y(2x+y) \quad \text{en} \quad \frac{\partial f}{\partial y} = x^2 + 2xy = x(x+2y).$$

Uit $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ volgt $x = y = 0$, want voor $x \neq 0$ volgt $y = -2x$ en $y = -\frac{1}{2}x$ en dit is onmogelijk. Dus is $(x,y) = (0,0)$ het enige kritieke punt. Maar op de lijn $x = y$ is de functie $f(x,y)$ gelijk aan $2x^3$, en is dus < 0 voor $x < 0$ en > 0 voor $x > 0$, dus heeft de functie in $(0,0)$ geen maximum of minimum.

Opdracht 12 *Vind de kritieke punten voor* $f(x, y) := x^3 + y^3 - 3x - 12y + 20$.

3.3 Criterium voor lokale extrema

Het voorbeeld (3) van de functie $f(x, y) = x^2y + xy^2$ laat zien dat (net als bij gewone functies van één variabel) een functie van meerdere veranderlijken in een kritiek punt niet noodzakelijk een maximum of minimum hoeft te hebben.

Definitie: Een kritiek punt $\mathbf{x}_0$ met $\nabla f(\mathbf{x}_0) = 0$ die geen extremum van de functie $f(\mathbf{x})$ is noemt men een *zadelpunt* van $f(\mathbf{x})$. In een zadelpunt vindt men in iedere (willekeurig kleine) omgeving van $\mathbf{x}_0$ punten $\mathbf{x}$ met $f(\mathbf{x}) > f(\mathbf{x}_0)$ en punten met $f(\mathbf{x}) < f(\mathbf{x}_0)$.

De vraag is nu, hoe we erover kunnen beslissen of een kritiek punt een minimum, maximum of een zadelpunt is. Hiervoor zullen we analoog met het geval van gewone functies de Taylor reeks in het kritieke punt $\mathbf{x}_0$ gebruiken, beter gezegd bekijken we de kwadratische benadering van $f(\mathbf{x})$ door de Taylor veelterm van graad 2.

We veronderstellen vanaf nu dat $\mathbf{x}_0$ een kritiek punt van de functie $f(\mathbf{x})$ is, dus dat $\nabla f(\mathbf{x}_0) = 0$ en we noteren met $H(\mathbf{x}_0)$ de Hesse matrix geëvalueerd in het punt $\mathbf{x}_0$. Dan is de kwadratische benadering van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ gegeven door

$$f(\mathbf{x}_0 + \mathbf{h}) \approx T(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \frac{1}{2}\,\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h}.$$

Als we de functie $f(\mathbf{x})$ alleen maar in een kleine omgeving van $\mathbf{x}_0$ bekijken, kunnen we de hogere termen van de Taylor reeks verwaarlozen, het gedrag van de functie wordt dan door de kwadratische benadering weergegeven. We krijgen nu rechtstreeks het volgende criterium voor minima en maxima:

Criterium:

(1) De functie $T(\mathbf{x})$ en dus ook de functie $f(\mathbf{x})$ heeft een *minimum* in $\mathbf{x}_0$ als $T(\mathbf{x})$ vanuit $\mathbf{x}_0$ in alle richtingen $\mathbf{h}$ toeneemt, d.w.z. als $\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h} > 0$ voor alle richtingen $\mathbf{h} \neq 0$.

(2) De functie $T(\mathbf{x})$ en dus ook de functie $f(\mathbf{x})$ heeft een *maximum* in $\mathbf{x}_0$ als $T(\mathbf{x})$ vanuit $\mathbf{x}_0$ in alle richtingen $\mathbf{h}$ afneemt, d.w.z. als $\mathbf{h}^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h} < 0$ voor alle richtingen $\mathbf{h} \neq 0$.

(3) Als er een richting $\mathbf{h}_1$ bestaat met $\mathbf{h}_1^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h}_1 > 0$ en een andere richting $\mathbf{h}_2$ met $\mathbf{h}_2^{tr} \cdot H(\mathbf{x}_0) \cdot \mathbf{h}_2 < 0$, dan heeft $T(\mathbf{x})$ en dus ook $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ een zadelpunt.

Om dit criterium toe te passen, moeten we dus voor de symmetrische matrix $H := H(\mathbf{x}_0)$ beslissen of de producten $\mathbf{h}^{tr} \cdot H \cdot \mathbf{h}$ altijd positief, altijd negatief of geen van de twee zijn. Dit is eigenlijk een vraagstelling uit de Lineaire Algebra, die in het verband met inproducten ter sprake komt.

Positief definiete matrices

Definitie: Zij A een symmetrische $n \times n$-matrix, d.w.z. $A_{ij} = A_{ji}$ voor alle i, j (kort: $A^{tr} = A$).

(i) A heet *positief definiet* als $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} > 0$ voor alle $\mathbf{v} \neq 0$.

(ii) A heet *positief semidefiniet* als $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} \geq 0$ voor alle $\mathbf{v}$.

(iii) A heet *negatief definiet* als $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} < 0$ voor alle $\mathbf{v} \neq 0$.

(iv) A heet *negatief semidefiniet* als $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} \leq 0$ voor alle $\mathbf{v}$.

(v) Als er vectoren $\mathbf{v}_1$ en $\mathbf{v}_2$ bestaan met $\mathbf{v}_1^{tr} \cdot A \cdot \mathbf{v}_1 > 0$ en $\mathbf{v}_2^{tr} \cdot A \cdot \mathbf{v}_2 < 0$, heet A *indefiniet*.

Het idee achter deze definitie is, dat men algemeen met behulp van een symmetrische matrix A een bilineaire afbeelding op paren van vectoren kan definiëren door $(v, w) \mapsto v^{tr} \cdot A \cdot w$. Deze afbeelding is lineair in beide argumenten en is symmetrisch, d.w.z. verruilen van de argumenten verandert de waarde niet. Als A positief definiet is, laat zich met $\|v\| := \sqrt{v^{tr} \cdot A \cdot v}$ een lengte voor de vectoren definiëren.

Merk op: Uit de definite en uit $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} > 0 \Leftrightarrow \mathbf{v}^{tr} \cdot (-A) \cdot \mathbf{v} < 0$ volgt rechtstreeks, dat een matrix A positief definiet is dan en slechts dan als de tegengestelde matrix $-A$ negatief definiet is. Evenzo volgt dat A negatief definiet is dan en slechts dan als $-A$ positief definiet is.

Voorbeelden:

(1) De matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is positief definiet, want voor $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} = x^2 + y^2 > 0$ voor $\mathbf{v} \neq 0$.

(2) Voor de matrix $A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ en $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} = ax^2 + by^2$. Voor $a, b > 0$ is A positief definiet, voor $a, b < 0$ is A negatief definiet. In het geval $a \cdot b < 0$ is A indefiniet, met $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ en $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ volgt bijvoorbeeld voor $a > 0$ en $b < 0$ dat $\mathbf{v}_1^{tr} \cdot A \cdot \mathbf{v}_1 = a > 0$ en $\mathbf{v}_2^{tr} \cdot A \cdot \mathbf{v}_2 = b < 0$.

(3) De matrix $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is positief semidefiniet, want voor $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} = x^2 + 2xy + y^2 = (x + y)^2 \geq 0$. De matrix is niet positief definiet, want voor $y = -x$ is $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} = 0$ zonder dat $\mathbf{v} = 0$ is.

(4) De matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is indefiniet, want voor $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ geldt $\mathbf{v}^{tr} \cdot A \cdot \mathbf{v} = 2xy$, voor $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is dus $\mathbf{v}_1^{tr} \cdot A \cdot \mathbf{v}_1 = 2 > 0$ en voor $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ is $\mathbf{v}_2^{tr} \cdot A \cdot \mathbf{v}_2 = -2 < 0$.

Algemeen Voorbeeld: We bekijken de $n \times n$-diagonaalmatrix

$$A := \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix} \text{ met } A_{ii} = d_i \text{ en } A_{ij} = 0 \text{ voor } i \neq j.$$

Er geldt:

(a) A is positief definiet als $d_i > 0$ voor alle i;

(b) A is positief semidefiniet als $d_i \geq 0$ voor alle i;

(c) A is negatief definiet als $d_i < 0$ voor alle i;

(d) A is negatief semidefiniet als $d_i \leq 0$ voor alle i;

(e) A is indefiniet als er i en j bestaan met $d_i > 0$ en $d_j < 0$.

In het bijzonder geldt voor een diagonaalmatrix A die positief (of negatief) semidefiniet maar niet positief (of negatief) definiet is, dat $\det(A) = 0$, omdat in dit geval minstens een $d_i = 0$ is en $\det(A) = d_1 \cdot d_2 \cdot \ldots \cdot d_n$ geldt.

Het algemene voorbeeld geeft het cruciale idee, hoe we kunnen testen of een matrix positief of negatief definiet is. De volgende stelling geeft hiervoor een criterium aan. Hierbij bedoelen met de *linksboven $k \times k$-deelmatrix* van een matrix A de deelmatrix van A waarvoor de indices slechts tussen 1 en k lopen (in plaats van tussen 1 en n):

$$\left(\begin{array}{ccc|ccc} A_{11} & \ldots & A_{1k} & \ldots & A_{1n} \\ \vdots & & \vdots & & \vdots \\ A_{k1} & \ldots & A_{kk} & \ldots & A_{kn} \\ \hline \vdots & & \vdots & & \vdots \\ A_{n1} & \ldots & A_{nk} & \ldots & A_{nn} \end{array}\right).$$

Stelling: Zij A een symmetrische $n \times n$-matrix, dan geldt:

(i) A is positief definiet dan en slechts dan als alle linksboven $k \times k$-deelmatrices van A positieve determinant hebben.

(ii) A is negatief definiet dan en slechts dan als de linksboven $k \times k$-deelmatrices van A alternerend negatieve en positieve determinant hebben, dus als de 1×1-deelmatrix negatieve determinant heeft, de 2×2-deelmatrix positieve determinant, de 3×3-deelmatrix negatieve determinant enz.

Equivalent (en eenvoudiger) geldt: A is negatief definiet dan en slechts dan als de matrix $-A$ positief definiet is, dus als alle linksboven $k \times k$-deelmatrices van $-A$ positieve determinant hebben.

(iii) Als $\det(A) \neq 0$ is, is A indefiniet dan en slechts dan als nog A nog $-A$ positief definiet zijn.

We zien rechtstreeks in dat deze stelling voor diagonaalmatrices geldt. De grap is nu, dat we door een basistransformatie iedere symmetrische matrix op diagonaalvorm kunnen brengen, en een basistransformatie bewaart de eigenschap van een matrix positief of negatief definiet te zijn.

> De attente lezer is natuurlijk gewaar geworden dat de stelling in het geval $\det(A) = 0$ geen uitspraak erover maakt of de matrix positief of negatief semidefiniet is of indefiniet. Hiervoor zou men de matrix A inderdaad door een basistransformatie op diagonaalvorm moeten brengen, dan laat het zich weer makkelijk aan de diagonaalelementen aflezen.
>
> We gaan dit probleem hier echter niet verdiepen, omdat het geval dat de Hesse matrix determinant 0 heeft in de praktijk nauwelijks een rol speelt. Om in zo'n geval te beslissen of een kritiek punt een minimum, maximum of zadelpunt is, zou men net zo als in het geval $f''(x_0) = 0$ voor gewone functies naar hogere partiële afgeleiden dan de tweede moeten kijken.

Voorbeeld: Een 2×2-matrix $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is positief definiet als $a > 0$ en $\det(A) = ac - b^2 > 0$. De matrix A is negatief definiet als $a < 0$ en $\det(A) = ac - b^2 > 0$. Als $\det(A) < 0$, is A indefiniet.

Toepassing op functies van twee variabelen

Als we de uitspraak van het vorige voorbeeld voor de Hesse matrix

$$H := H(x_0, y_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(x_0, y_0) & \frac{\partial^2 f}{\partial x\, \partial y}(x_0, y_0) \\ \frac{\partial^2 f}{\partial x\, \partial y}(x_0, y_0) & \frac{\partial^2 f}{\partial y^2}(x_0, y_0) \end{pmatrix}$$

van een functie van twee veranderlijken herformuleren, krijgen we een handige stelling over de kritieke punten van een functie van twee veranderlijken.

Stelling: Zij $f(x, y)$ een functie van twee variabelen en zij (x_0, y_0) een kritiek punt, d.w.z. een punt met $\nabla f(x_0, y_0) = (0, 0)$. Verder zij $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}$ de Hesse matrix van $f(x, y)$ in (x_0, y_0), d.w.z.

$$H_{11} = \frac{\partial^2 f}{\partial x^2}(x_0, y_0), \quad H_{12} = \frac{\partial^2 f}{\partial x\, \partial y}(x_0, y_0), \quad H_{22} = \frac{\partial^2 f}{\partial y^2}(x_0, y_0).$$

(i) $f(x, y)$ heeft in het punt (x_0, y_0) een *lokaal minimum* als

$$H_{11} > 0 \text{ en } H_{11}H_{22} - H_{12}^2 > 0, \text{ dus als}$$

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0 \quad \text{en} \quad \frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - (\frac{\partial^2 f}{\partial x\, \partial y}(x_0, y_0))^2 > 0.$$

(ii) $f(x,y)$ heeft in het punt (x_0,y_0) een *lokaal maximum* als

$$H_{11} < 0 \text{ en } H_{11}H_{22} - H_{12}^2 > 0, \text{ dus als}$$

$$\frac{\partial^2 f}{\partial x^2}(x_0,y_0) < 0 \quad \text{en} \quad \frac{\partial^2 f}{\partial x^2}(x_0,y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0,y_0) - (\frac{\partial^2 f}{\partial x\,\partial y}(x_0,y_0))^2 > 0.$$

(iii) $f(x,y)$ heeft in het punt (x_0,y_0) een *zadelpunt* als

$$H_{11}H_{22} - H_{12}^2 < 0, \text{ dus als}$$

$$\frac{\partial^2 f}{\partial x^2}(x_0,y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0,y_0) - (\frac{\partial^2 f}{\partial x\,\partial y}(x_0,y_0))^2 < 0.$$

Zo als eerder opgemerkt is deze stelling niet van toepassing als $\det(H) = 0$. In dit geval is de Hesse matrix positief of negatief semidefiniet en is er een vector $\mathbf{h} \neq 0$ met $\mathbf{h}^{tr} \cdot H \cdot \mathbf{h} = 0$. Deze situatie is analoog met het geval $f''(x_0) = 0$ voor gewone functies, waar pas de hogere termen van de Taylor reeks aangeven of het punt een extremum of een zadelpunt is. Dit geldt ook voor de functies van meerdere variabelen, men moet de hogere termen van de Taylor reeks raadplegen om te beslissen hoe zich de functie in een richting $\mathbf{h}$ met $\mathbf{h}^{tr} \cdot H \cdot \mathbf{h} = 0$ gedraagt. In de praktijk speelt dit probleem echter een minder belangrijke rol, dus zullen we genoegen nemen met het geval $\det(H) \neq 0$.

Voorbeeld 1: We bekijken de functie

$$f(x,y) := x^3 + 6xy^2 - 2y^3 - 12x.$$

Er geldt

$$\frac{\partial f}{\partial x} = 3x^2 + 6y^2 - 12 \quad \text{en} \quad \frac{\partial f}{\partial y} = 12xy - 6y^2 = 6y(2x - y).$$

Uit $\frac{\partial f}{\partial y} = 0$ volgt $y = 0$ of $y = 2x$. In het eerste geval volgt uit $\frac{\partial f}{\partial x} = 0$ dat $3x^2 = 12$, dus $x = \pm 2$. In het geval $y = 2x$ moet gelden dat $3x^2 + 24x^2 = 12$, dus $x^2 = \frac{12}{27} = \frac{4}{9}$, dus $x = \pm\frac{2}{3}$. Er zijn dus vier kritieke punten:

$$(2,0), \quad (-2,0), \quad (\frac{2}{3},\frac{4}{3}), \quad (-\frac{2}{3},-\frac{4}{3}).$$

Voor de analyse van de kritieke punten hebben we de tweede partiële afgeleiden nodig, er geldt

$$\frac{\partial^2 f}{\partial x^2} = 6x, \quad \frac{\partial^2 f}{\partial x\,\partial y} = 12y, \quad \frac{\partial^2 f}{\partial y^2} = 12x - 12y.$$

De Hesse matrix is dus

$$H = \begin{pmatrix} 6x & 12y \\ 12y & 12x - 12y \end{pmatrix}$$

en $\det(H) = 6x(12x - 12y) - (12y)^2 = 72x^2 - 72xy - 144y^2$. Voor de kritieke punten geeft dit de volgende tabel:

kritiek punt	H_{11}	$\det(H)$	type
$(2,0)$	12	288	minimum
$(-2,0)$	-12	288	maximum
$(\frac{2}{3},\frac{4}{3})$	4	-288	zadelpunt
$(-\frac{2}{3},-\frac{4}{3})$	-4	-288	zadelpunt

Voorbeeld 2: We onderzoeken de kritieke punten van de functie

$$f(x,y) := (x^2-y^2)\,e^{\frac{-x^2-y^2}{2}}.$$

Er geldt

$$\frac{\partial f}{\partial x} = (2x - x(x^2-y^2))\,e^{\frac{-x^2-y^2}{2}} \quad \text{en} \quad \frac{\partial f}{\partial y} = (-2y - y(x^2-y^2))\,e^{\frac{-x^2-y^2}{2}}.$$

De exponentiële functie wordt nooit 0, dus vinden we de kritieke punten als oplossingen van de vergelijkingen

$$x(2-(x^2-y^2)) = 0 \quad \text{en} \quad y(-2-(x^2-y^2)) = 0.$$

Omdat x^2-y^2 niet tegelijkertijd de waarden 2 en -2 kan hebben, moet $x=0$ of $y=0$ zijn, dit geeft de kritieke punten $(0,0)$, $(\pm\sqrt{2},0)$ en $(0,\pm\sqrt{2})$.

Voor de tweede partiële afgeleiden, dus de elementen H_{ij} van de Hesse matrix, krijgen we:

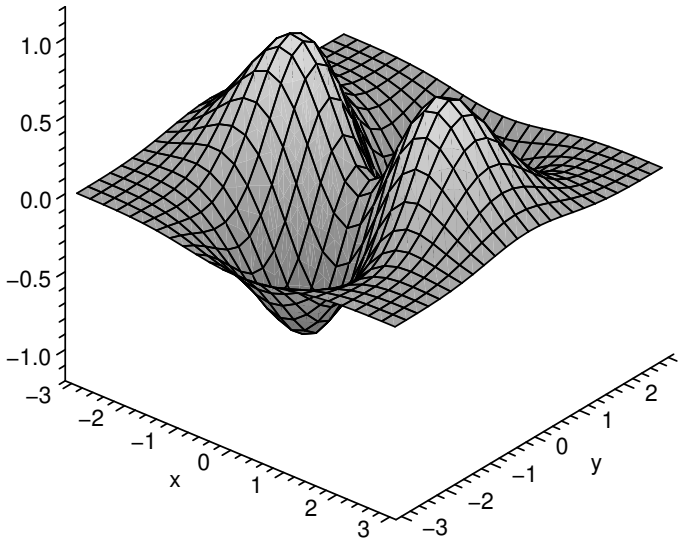
$$\begin{aligned}
H_{11} &= \frac{\partial^2 f}{\partial x^2} = (2-5x^2+x^2(x^2-y^2)+y^2)\,e^{\frac{-x^2-y^2}{2}} \\
H_{12} = H_{21} &= \frac{\partial^2 f}{\partial x\,\partial y} = xy(x^2-y^2)\,e^{\frac{-x^2-y^2}{2}} \\
H_{22} &= \frac{\partial^2 f}{\partial y^2} = (5y^2-2+y^2(x^2-y^2)-x^2)\,e^{\frac{-x^2-y^2}{2}}.
\end{aligned}$$

Hiermee vinden we de volgende tabel voor de kritieke punten:

kritiek punt	H_{11}	H_{12}	H_{22}	$\det(H) = H_{11}H_{22} - H_{12}^2$	type
$(0,0)$	2	0	-2	-4	zadelpunt
$(\sqrt{2},0)$	$-4e^{-1}$	0	$-4e^{-1}$	$16e^{-2}$	maximum
$(-\sqrt{2},0)$	$-4e^{-1}$	0	$-4e^{-1}$	$16e^{-2}$	maximum
$(0,\sqrt{2})$	$4e^{-1}$	0	$4e^{-1}$	$16e^{-2}$	minimum
$(0,-\sqrt{2})$	$4e^{-1}$	0	$4e^{-1}$	$16e^{-2}$	minimum

In de grafiek van de functie in Figuur I.12 kunnen we controleren dat onze analyse van de kritieke punten inderdaad klopt.

Opdracht 13 *Bepaal de kritieke punten van de functie $f(x,y) := 4x^3-3x^2y+y^3-9y$ en ga na of de punten minima, maxima of zadelpunten zijn.*



Figuur I.12: Grafiek van de functie $f(x,y) = (x^2 - y^2)e^{\frac{-x^2-y^2}{2}}$.

Functies van meer dan twee variabelen

Als we bij een functie $f(\mathbf{x})$ van meer dan twee variabelen willen testen of een kritiek punt een minimum, maximum of een zadelpunt is, is het meestal het eenvoudigste het criterium van de linksboven deelmatrices op het concrete voorbeeld toe te passen. Algemene formules zijn al voor 3 variabelen behoorlijk afschrikkend, we beperken ons daarom tot de algemene stelling en een voorbeeld.

Stelling: Zij $f(\mathbf{x})$ een functie van n variabelen, zij $\mathbf{x}_0$ een kritiek punt van $f(\mathbf{x})$, d.w.z. $\nabla f(\mathbf{x}_0) = 0$ en zij $H := H(\mathbf{x}_0)$ de Hesse matrix van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$, d.w.z. $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0)$. Dan geldt:

(i) $f(\mathbf{x})$ heeft in het punt $\mathbf{x}_0$ een lokaal minimum als H positief definiet is;

(ii) $f(\mathbf{x})$ heeft in het punt $\mathbf{x}_0$ een lokaal maximum als H negatief definiet is;

(iii) $f(\mathbf{x})$ heeft in het punt $\mathbf{x}_0$ een zadelpunt als H indefiniet is.

Voorbeeld: Zij de functie $f(x,y,z)$ gegeven door

$$f(x,y,z) := x^2y + y^2z + z^2 - 2x.$$

Om de kritieke punten te vinden moeten we de eerste partiële afgeleiden bepalen. Er geldt

$$\frac{\partial f}{\partial x} = 2xy - 2, \quad \frac{\partial f}{\partial y} = x^2 + 2yz, \quad \frac{\partial f}{\partial z} = y^2 + 2z.$$

Uit $\frac{\partial f}{\partial z} = 0$ volgt $z = -\frac{y^2}{2}$. Dit ingevuld in $\frac{\partial f}{\partial y} = 0$ geeft $x^2 = y^3$ en hiermee volgt uit $\frac{\partial f}{\partial x} = 0$ dat $y^{\frac{5}{2}} = 1$. Dit geeft $y = 1$, $x = 1$ en $z = -\frac{1}{2}$, het enige kritieke punt is dus $\mathbf{x}_0 = (1, 1, -\frac{1}{2})$.

Om na te gaan of dit een minimum, maximum of een zadelpunt is, moeten we nu de tweede partiële afgeleiden bepalen. We krijgen

$$\frac{\partial^2 f}{\partial x^2} = 2y, \quad \frac{\partial^2 f}{\partial x\,\partial y} = 2x, \quad \frac{\partial^2 f}{\partial x\,\partial z} = 0, \quad \frac{\partial^2 f}{\partial y^2} = 2z, \quad \frac{\partial^2 f}{\partial y\,\partial z} = 2y, \quad \frac{\partial^2 f}{\partial z^2} = 2$$

dit geeft in het punt $\mathbf{x}_0 = (1, 1, -\frac{1}{2})$ de Hesse matrix

$$H(\mathbf{x}_0) = \begin{pmatrix} 2 & 2 & 0 \\ 2 & -1 & 2 \\ 0 & 2 & 2 \end{pmatrix}$$

Voor de linksboven deelmatrices van $H(\mathbf{x}_0)$ geldt

$$\det(\begin{pmatrix} 2 \end{pmatrix}) = 2 > 0, \quad \det(\begin{pmatrix} 2 & 2 \\ 2 & -1 \end{pmatrix}) = -6 < 0, \quad \det(H) = -20.$$

De matrix is dus indefiniet en het kritieke punt $\mathbf{x}_0$ is een zadelpunt.

Toepassing: Regressielijn

Als belangrijke toepassing voor extrema van functies van meerdere veranderlijken bekijken we het bepalen van de regressielijn

$$L(x) := ax + b$$

door een verzameling $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ van punten. Het idee is hierbij, dat er een lineaire samenhang $y = Ax + B$ tussen de x- en y-coördinaten van de punten verondersteld wordt, maar dat de parameters A en B onbekend zijn. Hiervoor wordt er uit een steekproef van punten een schatting a voor A en b voor B gemaakt door een lijn zo door de punten te leggen, dat de som van de kwadratische afwijkingen $(y_i - (ax_i + b))^2$ tussen de *gemeten* y-waarden y_i en de *berekende* waarden $\hat{y}_i := ax_i + b$ minimaal wordt. We kijken dus naar de functie

$$f(a, b) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

en moeten a en b zo bepalen dat $f(a, b)$ minimaal wordt.

Voor de partiële afgeleiden geldt

$$\frac{\partial f}{\partial a} = -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b) \quad \text{en} \quad \frac{\partial f}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

en uit $\nabla f = 0$ volgt dat de gezochte parameters a en b oplossingen zijn van het lineaire stelsel vergelijkingen

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}.$$

Uit de tweede vergelijking volgt in het bijzonder dat de regressielijn $ax+b$ door het zwaartepunt $(\overline{x},\overline{y})$ met $\overline{x}=\frac{1}{n}\sum_{i=1}^n x_i$ en $\overline{y}=\frac{1}{n}\sum_{i=1}^n y_i$ loopt, met deze notatie luidt de tweede vergelijking namelijk $n\overline{x}a+nb=n\overline{y}$, dus geldt

$$b=\overline{y}-a\overline{x}.$$

Ook de eerste vergelijking kunnen we met een geschikte notatie voor gemiddelden iets eenvoudiger schrijven, met $\overline{x^2}=\frac{1}{n}\sum_{i=1}^n x_i^2$ en $\overline{xy}=\frac{1}{n}\sum_{i=1}^n x_iy_i$ krijgen we de vergelijking $\overline{x^2}a+\overline{x}b=\overline{xy}$. Als we in deze vergelijking b door $\overline{y}-a\overline{x}$ vervangen, krijgen we $\overline{x^2}a+\overline{x}(\overline{y}-a\overline{x})=\overline{xy}$ en opgelost naar a geeft dit de bekende vergelijking

$$a=\frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x^2}-\overline{x}^2}$$

voor de richtingscoëfficiënt van de regressielijn.

We moeten nu nog nagaan dat we inderdaad een minimum van de functie hebben gevonden. Hiervoor bepalen we de tweede partiële afgeleiden, er geldt

$$\frac{\partial^2 f}{\partial a^2}=2\sum_{i=1}^n x_i^2,\quad \frac{\partial^2 f}{\partial a\,\partial b}=2\sum_{i=1}^n x_i,\quad \frac{\partial^2 f}{\partial b^2}=2n,$$

we krijgen dus de Hesse matrix

$$H=2\begin{pmatrix}\sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & n\end{pmatrix}.$$

Het is duidelijk dat $H_{11}=2\sum_{i=1}^n x_i^2>0$, omdat dit een som van kwadraten is. Met de notatie $\overline{x}=\frac{1}{n}\sum_{i=1}^n x_i$ krijgen we verder

$$\begin{aligned}n(\sum_{i=1}^n(x_i-\overline{x})^2)&=n\sum_{i=1}^n x_i^2-n\sum_{i=1}^n 2x_i\overline{x}+n\cdot n\overline{x}=n\sum_{i=1}^n x_i^2-n^2\overline{x}\\&=n\sum_{i=1}^n x_i^2-(\sum_{i=1}^n x_i)^2=\det(\frac{1}{2}H).\end{aligned}$$

Aan de linkerkant staat een som van kwadraten, dus is $\det(\frac{1}{2}H)>0$ en dus ook $\det(H)>0$. De Hesse matrix is dus altijd positief definiet, in het bijzonder ook in het kritieke punt (a,b) en dus hebben we een minimum gevonden.

3.4 Extrema onder randvoorwaarden

Vaak zijn problemen waarbij we minima of maxima van functies moeten bepalen zo geformuleerd, dat de oplossingen aan zekere randvoorwaarden moeten voldoen. In principe hebben we zo iets al gezien. In het voorbeeld van de open doos hadden we gezegd, dat de doos een bepaald volume V moet hebben. In principe moeten we dus voor een doos met de afmetingen x, y en z de oppervlakte $f(x,y,z)=xy+2xz+2yz$ onder de randvoorwaarde minimaliseren dat $xyz=V$ geldt. Deze randvoorwaarde hadden we gebruikt om de variabel z te verwijderen, door z door $\frac{V}{xy}$ te vervangen.

Meestal is het helaas zo, dat we een randvoorwaarde niet zo makkelijk naar één van de variabelen kunnen oplossen, of dat de functie die we dan krijgen erg ingewikkeld wordt. We zullen daarom nu naar een methode kijken, hoe we bij gegeven randvoorwaarden een minimum of maximum van een functie kunnen bepalen.

Het probleem voor functies van twee variabelen luidt als volgt: Voor een functie $f(x, y)$ is een extremum (maximum of minimum) gezocht onder de randvoorwaarde dat $g(x, y) = 0$.

> Een typische randvoorwaarde is, dat we een extremum op de rand van een begrensd oppervlak zo als een cirkelschijf willen bepalen. Als we bijvoorbeeld alleen maar de punten op een cirkel van straal 3 rond het punt $(1, 1)$ willen bekijken, hebben we de punten nodig waarvoor geldt dat $(x-1)^2 + (y-1)^2 = 3^2$, de randvoorwaarde is dan
>
> $$g(x, y) = (x-1)^2 + (y-1)^2 - 3^2 = 0.$$

We gaan nu na dat in een extremum (x_0, y_0) noodzakelijk geldt dat de gradiënten van $f(x, y)$ en $g(x, y)$ in dit punt parallel zijn, dus dat $\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0)$ voor een zekere λ.

De punten (x, y) met $g(x, y) = 0$ vormen een niveaukromme van $g(x, y)$, dus staat ∇g in punten die aan de randvoorwaarde voldoen steeds loodrecht op de raaklijn aan $g(x, y)$. Stel nu dat in een punt (x_0, y_0) de gradiënt ∇f niet loodrecht op de raaklijn aan $g(x, y)$ staat, dan is de projectie $\nabla f_{||}$ van ∇f op de raaklijn aan $g(x, y)$ niet 0. Aan de ene kant kunnen we nu op de niveaukromme $g(x, y) = 0$ in de richting van $\nabla f_{||}$ lopen, omdat dit de richting van de raaklijn aan $g(x, y)$ is. Aan de andere kant neemt $f(x, y)$ in de richting van $\nabla f_{||}$ toe, omdat $\nabla f_{||}$ (als projectie van ∇f) niet loodrecht of ∇f staat. Als we vanuit het punt (x_0, y_0) in de richting van $\nabla f_{||}$ lopen, neemt $f(x, y)$ dus toe, als we in de tegengestelde richting $-\nabla f_{||}$ lopen, neemt $f(x, y)$ af, dus is (x_0, y_0) nog een maximum nog een minimum.

We kunnen hetzelfde argument met behulp van Taylor reeksen ook iets algemener formuleren, namelijk zo, dat het ook voor een functie $f(\mathbf{x})$ van n variabelen geldt. In dit geval is de randvoorwaarde gegeven door $g(\mathbf{x}) = 0$, waarbij ook $g(\mathbf{x})$ een functie van n variabelen is.

Omdat $g(\mathbf{x}) = 0$ een niveauoppervlak is, staat de gradiënt $\nabla g(\mathbf{x}_0)$ loodrecht op het raakvlak aan $g(\mathbf{x})$ in het punt $\mathbf{x}_0$. De richtingen $\mathbf{h}$ met $\nabla g(\mathbf{x}_0) \cdot h = 0$ zijn dus juist de richtingen in die we vanuit $\mathbf{x}_0$ mogen lopen, om verder aan de randvoorwaarde $g(\mathbf{x}) = 0$ te voldoen. Dit volgt ook uit de Taylor reeks voor $g(\mathbf{x})$: Voor kleine $\mathbf{h}$ geldt $g(\mathbf{x}_0 + \mathbf{h}) = g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot \mathbf{h}$ en als ook $\mathbf{x}_0 + \mathbf{h}$ aan de randvoorwaarde voldoet is $g(\mathbf{x}_0 + \mathbf{h}) = g(\mathbf{x}_0)$, dus moet $\nabla g(\mathbf{x}_0) \cdot \mathbf{h} = 0$ gelden.

Nu bekijken we de Taylor reeks van $f(\mathbf{x})$ in het punt $\mathbf{x}_0$. De lineaire benadering is

$$f(\mathbf{x}_0 + \mathbf{h}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot \mathbf{h}$$

en in een extremum onder de randvoorwaarde $g(\mathbf{x}) = 0$ moet gelden, dat $\nabla f(\mathbf{x}_0) \cdot \mathbf{h} = 0$ voor alle richtingen $\mathbf{h}$ in die we onder de randvoorwaarde $g(\mathbf{x}) = 0$ mogen lopen. Maar we hebben net gezegd dat de mogelijke richtingen $\mathbf{h}$ juist de richtingen zijn die loodrecht op $\nabla g(\mathbf{x}_0)$ staan, dus moet in een extremum gelden dat

$$\nabla f(\mathbf{x}_0) \cdot \mathbf{h} = 0 \text{ voor alle } \mathbf{h} \text{ met } \nabla g(\mathbf{x}_0) \cdot \mathbf{h} = 0.$$

Dit betekent dat $\nabla f(\mathbf{x}_0)$ loodrecht op alle $\mathbf{h}$ staat, die zelf loodrecht op $\nabla g(\mathbf{x}_0)$ staan, dus moet $\nabla f(\mathbf{x}_0)$ loodrecht op het raakvlak aan $g(\mathbf{x})$ in het punt $\mathbf{x}_0$ staan. Maar de enige vectoren die loodrecht op dit raakvlak staan, zijn de (positieve en negatieve) veelvouden van $\nabla g(\mathbf{x}_0)$ en hieruit volgt dat

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0).$$

We hebben dus de volgende stelling ingezien:

Stelling: In een lokaal extremum $\mathbf{x}_0$ onder de randvoorwaarde $g(\mathbf{x}) = 0$ geldt dat $\nabla f(\mathbf{x}_0)$ en $\nabla g(\mathbf{x}_0)$ lineair afhankelijk zijn, dus dat

$$\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0).$$

3.5 De methode van Lagrange multiplicatoren

Deze stelling hierboven geeft aanleiding tot de volgende methode: De functie

$$L(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda g(\mathbf{x})$$

van de $n + 1$ variabelen $x_1, \ldots, x_n$ en λ heet de *Lagrange functie* van $f(\mathbf{x})$ onder de randvoorwaarde $g(\mathbf{x}) = 0$. De variabel λ heet hierbij de *Lagrange multiplicator*. Voor de Lagrange functie geldt:

Stelling: Als $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ een extremum onder de randvoorwaarde $g(\mathbf{x}) = 0$ heeft, dan is $\mathbf{x}_0$ een kritiek punt van de Lagrange functie $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$.

Stel namelijk dat $\mathbf{x}_0$ een kritiek punt van $L(\mathbf{x}, \lambda)$ is. Uit $\frac{\partial L}{\partial x_i} = 0$ voor $1 \leq i \leq n$ volgt dat $\frac{\partial f}{\partial x_i} + \lambda \frac{\partial g}{\partial x_i} = 0$ voor alle i en dus $\nabla f(\mathbf{x}_0) = -\lambda \nabla g(\mathbf{x}_0)$. (Merk op dat we tegenover de stelling van boven λ door $-\lambda$ hebben vervangen.)

Omdat $\frac{\partial L}{\partial \lambda} = g(\mathbf{x})$, volgt uit $\frac{\partial L}{\partial \lambda} = 0$ dat het punt $\mathbf{x}_0$ inderdaad aan de randvoorwaarde voldoet.

Voorbeeld 1: We bepalen de extrema van de functie

$$f(x, y) = x^2 + 2y^2$$

op de rand van de cirkel met straal 1 rond $(0, 0)$. De punten op de cirkel kunnen we beschrijven door de randvoorwaarde $x^2 + y^2 = 1$, dus is de functie $g(x, y)$ gegeven door $g(x, y) := x^2 + y^2 - 1$. De Lagrange functie is dus

$$L(x, y, \lambda) = x^2 + 2y^2 + \lambda(x^2 + y^2 - 1).$$

Er geldt

$$\frac{\partial L}{\partial x} = 2x + 2\lambda x, \quad \frac{\partial L}{\partial y} = 4y + 2\lambda y, \quad \frac{\partial L}{\partial \lambda} = x^2 + y^2 - 1.$$

Uit $\frac{\partial L}{\partial x} = 0$ volgt $2x(1+\lambda) = 0$, dus $x = 0$ of $\lambda = -1$. Uit $\frac{\partial L}{\partial y} = 0$ volgt $2y(2+\lambda) = 0$, dus $y = 0$ of $\lambda = -2$. Omdat λ niet tegelijkertijd -1 en -2 kan zijn, is noodzakelijk $x = 0$ of $y = 0$.

We krijgen dus de kritieke punten $(1,0)$, $(-1,0)$, $(0,1)$ en $(0,-1)$ van de Lagrange functie.

Het is natuurlijk in dit voorbeeld niet moeilijk om in te zien dat onder de randvoorwaarde $x^2 + y^2 = 1$ geldt, dat $f(x,y) = (x^2+y^2) + y^2 = 1 + y^2$, dus vinden we minima in $(\pm 1, 0)$ en maxima in $(0, \pm 1)$.

Voorbeeld 2: We bepalen de minima en maxima van de functie

$$f(x,y) := xy$$

op de cirkel met $x^2 + y^2 = 1$. De Lagrange functie is

$$L(x,y,\lambda) = xy + \lambda(x^2 + y^2 - 1)$$

en voor de partiële afgeleiden krijgen we

$$\frac{\partial L}{\partial x} = y + 2\lambda x, \quad \frac{\partial L}{\partial y} = x + 2\lambda y, \quad \frac{\partial L}{\partial \lambda} = x^2 + y^2 - 1.$$

Voor $x = 0$ volgt rechtstreeks $y = 0$ en andersom, dus moet volgens de randvoorwaarde noodzakelijk gelden dat $x \neq 0$ en $y \neq 0$. Uit de eerste twee vergelijkingen volgt dat $\frac{x}{y} = \frac{y}{x} = -2\lambda$, dit geeft $x^2 = y^2$ en hieruit volgt met de derde vergelijking dat $2x^2 = 1$, dus $x = \pm\frac{1}{\sqrt{2}}$ en evenzo $y = \pm\frac{1}{\sqrt{2}}$.

Men gaat na dat de vier punten $(\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}})$ inderdaad voldoen aan $\nabla L = 0$, dus kritieke punten van de Lagrange functie zijn, en er geldt $f(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = f(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) = \frac{1}{2}$ en $f(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) = f(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = -\frac{1}{2}$, dus zijn de eerste twee punten maxima en de laatste twee punten minima van de functie.

Voorbeeld 3: We willen het punt (x,y,z) op het oppervlak gegeven door de vergelijking $z = x^2 + y^2$ bepalen, dat het dichtst bij het punt $P := (1, 1, \frac{1}{2})$ ligt. De functie $f(x,y,z)$ die we moeten bekijken is in dit geval de afstand van (x,y,z) van het punt P. Maar omdat wortels vaak onhandig zijn kijken we liever naar het kwadraat van de afstand, dus naar de functie

$$f(x,y,z) := (x-1)^2 + (y-1)^2 + (z - \frac{1}{2})^2 = x^2 - 2x + y^2 - 2y + z^2 - z + \frac{9}{4}.$$

De randvoorwaarde is in dit geval gegeven door $g(x,y,z) = 0$ met $g(x,y,z) = x^2 + y^2 - z$ en we krijgen de Lagrange functie

$$L(x,y,z,\lambda) = x^2 - 2x + y^2 - 2y + z^2 - z + \frac{9}{4} + \lambda(x^2 + y^2 - z).$$

Voor de partiële afgeleiden geldt

$$\frac{\partial L}{\partial x} = 2x - 2 + 2\lambda x,\ \frac{\partial L}{\partial y} = 2y - 2 + 2\lambda y,\ \frac{\partial L}{\partial z} = 2z - 1 - \lambda,\ \frac{\partial L}{\partial \lambda} = x^2 + y^2 - z$$

en op 0 zetten van de partiële afgeleiden naar x, y en z geeft

$$x = \frac{1}{1+\lambda}, \quad y = \frac{1}{1+\lambda}, \quad z = \frac{1+\lambda}{2}.$$

Dit ingevuld in $x^2+y^2 = z$ geeft $\frac{2}{(1+\lambda)^2} = \frac{1+\lambda}{2}$ en dus $(1+\lambda)^3 = 4$ of $\lambda = \sqrt[3]{4}-1$. Hieruit krijgen we voor x, y en z:

$$x = \frac{1}{\sqrt[3]{4}}, \quad y = \frac{1}{\sqrt[3]{4}}, \quad z = \frac{\sqrt[3]{4}}{2} = \frac{1}{\sqrt[3]{2}}.$$

Het gezochte punt op het oppervlak $z = x^2 + y^2$ met de kleinste afstand van P is dus $(\frac{1}{\sqrt[3]{4}}, \frac{1}{\sqrt[3]{4}}, \frac{1}{\sqrt[3]{2}})$.

Natuurlijk hadden we de vergelijking $x^2 + y^2 - z = 0$ ook naar z kunnen oplossen en $x^2 + y^2$ in plaats van z in de functie $f(x, y, z)$ invullen. Op deze manier krijgen we de nieuwe functie

$$h(x,y) = x^2-2x+y^2-2y+(x^2+y^2)^2-(x^2+y^2)+\frac{9}{4} = x^4+2x^2y^2+y^4-2x-2y+\frac{9}{4}$$

waarvan we het minimum zonder randvoorwaarden mogen bepalen. Er geldt

$$\frac{\partial h}{\partial x} = 4x^3+4xy^2-2 = 4x(x^2+y^2)-2 \quad \text{en} \quad \frac{\partial h}{\partial y} = 4y^3+4x^2y-2 = 4y(x^2+y^2)-2$$

en uit $\frac{\partial h}{\partial x} = 0$ volgt $x^2 + y^2 = \frac{1}{2x}$. Dit ingevuld in $\frac{\partial h}{\partial y} = 0$ geeft $\frac{4y}{2x} = 2$, dus moet $x = y$ gelden. Als we dit weer in $\frac{\partial h}{\partial x} = 0$ invullen, krijgen we $8x^3 = 2$ en dus $x = y = \frac{1}{\sqrt[3]{4}}$. Uit $z = x^2 + y^2$ volgt nu weer dat $z = \frac{1}{\sqrt[3]{2}}$.

> Gelukkig geven beide methodes dezelfde oplossing. In het algemeen is het echter niet mogelijk, een variabel (zo als hier z) te elimineren, maar zelfs als dit lukt zijn de vergelijkingen $\nabla f = 0$ vaak moeilijk op te lossen. Zo als in dit voorbeeld geeft de methode met Lagrange multiplicatoren meestal eenvoudigere vergelijkingen, maar het zijn er natuurlijk meer vergelijkingen en meer onbekenden. Alles heeft zijn prijs!

Opdracht 14 *Bepaal het maximum van de functie $f(x, y, z) := x + z$ onder de randvoorwaarde $x^2 + y^2 + z^2 = 1$.*

Meerdere randvoorwaarden

Tot nu toe zijn we er altijd van uitgegaan dat de randvoorwaarde door één functie $g(\mathbf{x}) = 0$ gegeven is. Maar soms moet men ook naar meerdere randvoorwaarden kijken, bijvoorbeeld als het maximum van een functie op een kromme gezocht is, die gegeven is als doorsnede van twee oppervlakken in de ruimte.

In feite verandert bij meerdere randvoorwaarden niet zo erg veel: Als r randvoorwaarden gegeven zijn door $g_k(\mathbf{x}) = 0$ voor $1 \leq k \leq r$ moet in een extremum $\mathbf{x}_0$ gelden dat de gradiënt $\nabla f(\mathbf{x}_0)$ een lineaire combinatie van de gradiënten $\nabla g_k(\mathbf{x}_0)$ van de randvoorwaarden is. Er moeten dus coëfficiënten $\lambda_1, \ldots, \lambda_r$ bestaan met

$$\nabla f(\mathbf{x}_0) = -\lambda_1 \nabla g_1(\mathbf{x}_0) - \ldots - \lambda_r \nabla g_r(\mathbf{x}_0).$$

Ook dit is niet erg moeilijk in te zien: De toegelaten richtingen $\mathbf{h}$ in de we onder de randvoorwaarden vanuit een punt $\mathbf{x}_0$ mogen lopen, liggen in de doorsnede van de raakvlakken een de $g_k(\mathbf{x})$ in het punt $\mathbf{x}_0$. Maar de vectoren die loodrecht op deze doorsnede van raakvlakken staan, zijn juist de lineaire combinaties van de gradiënten $\nabla g_k(\mathbf{x}_0)$.

Men definieert nu de Lagrange functie

$$L(\mathbf{x}, \lambda_1, \ldots, \lambda_r) := f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \ldots + \lambda_r g_r(\mathbf{x})$$

met Lagrange multiplicatoren $\lambda_1, \ldots, \lambda_r$ en vindt de extrema van $f(\mathbf{x})$ onder de randvoorwaarden $g_k(\mathbf{x})$ met behulp van de volgende stelling:

Stelling: Als $f(\mathbf{x})$ in het punt $\mathbf{x}_0$ een extremum onder de randvoorwaarden $g_1(\mathbf{x}) = 0, \ldots, g_r(\mathbf{x}) = 0$ heeft, dan is $\mathbf{x}_0$ een kritiek punt van de Lagrange functie $L(\mathbf{x}, \lambda_1, \ldots, \lambda_r) = f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \ldots + \lambda_r g_r(\mathbf{x})$.

Voorbeeld: We bepalen de extrema van de functie $f(x, y, z) := 5x + y - 3z$ op de doorsnede van het vlak met de vergelijking $x + y + z = 0$ en de kogel van straal 1 rond $(0, 0, 0)$, dus onder de randvoorwaarden $g_1(x, y, z) := x + y + z = 0$ en $g_2(x, y, z) := x^2 + y^2 + z^2 - 1 = 0$. De Lagrange functie is

$$L(x, y, z, \lambda_1, \lambda_2) := 5x + y - 3z + \lambda_1(x + y + z) + \lambda_2(x^2 + y^2 + z^2 - 1)$$

en we krijgen de partiële afgeleiden

$$\frac{\partial L}{\partial x} = 5 + \lambda_1 + 2\lambda_2 x, \quad \frac{\partial L}{\partial y} = 1 + \lambda_1 + 2\lambda_2 y, \quad \frac{\partial L}{\partial z} = -3 + \lambda_1 + 2\lambda_2 z$$

$$\frac{\partial L}{\partial \lambda_1} = x + y + z, \quad \frac{\partial L}{\partial \lambda_2} = x^2 + y^2 + z^2 - 1.$$

Optellen van de drie eerste vergelijkingen geeft in verband met de vierde, dat $3 + 3\lambda_1 = 0$, dus $\lambda_1 = -1$. Hieruit volgt met de tweede dat $y = 0$ en dus $z = -x$. De laatste vergelijking geeft nu $2x^2 = 1$, dus $x = \pm\frac{1}{\sqrt{2}}$ en we krijgen als kritieke punten

$$\mathbf{x}_1 = (\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}) \quad \text{en} \quad \mathbf{x}_2 = (-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}).$$

Men gaat na dat $f(\mathbf{x}_1) = 4\sqrt{2}$ en $f(\mathbf{x}_2) = -4\sqrt{2}$, het eerste punt is dus een maximum, het tweede een minimum.

De kritieke punten van de Lagrange functie geven net als de kritieke punten van functies zonder randvoorwaarden alleen maar kandidaten voor minima of maxima. Om erover te beslissen of een punt inderdaad een minimum of maximum is, moet men op een iets slimmere manier dan zonder randvoorwaarden naar de tweede partiële afgeleiden kijken.

Maar bij dit soort vraagstukken is het bepalen van de kritieke punten meestal het grotere probleem, vaak volgt uit de samenhang dat een kritiek punt alleen maar een minimum of maximum kan zijn. We zullen deze vraag dus buiten beschouwing laten.

Toepassing: Entropie

Voor een discrete kansverdeling met kansen $p_1, \ldots, p_n$ voor n mogelijke uitkomsten definieert men de *entropie* H door

$$H := -\sum_{i=1}^{n} p_i \; {}^2\log(p_i),$$

waarbij we met ${}^2\log(x)$ de logaritme met basis 2 noteren (dus ${}^2\log(2^x) = x$). De entropie is een maat voor de onzekerheid die we over de uitkomsten volgens de gegeven kansverdeling hebben. Bijvoorbeeld zijn we bij een experiment met 2 mogelijke uitkomsten met $p_1 = 0.9$ en $p_2 = 0.1$ veel minder onzeker over de uitkomst dan bij een experiment met $p_1 = p_2 = 0.5$.

We zullen nu aantonen, dat de uniforme verdeling met $p_i = \frac{1}{n}$ voor alle kansverdelingen voor n uitkomsten de hoogste entropie heeft. De randvoorwaarde die we hanteren is natuurlijk $p_1 + \ldots + p_n = 1$, omdat we het over kansverdelingen hebben. Als Lagrange functie krijgen we

$$L(p_1, \ldots, p_n, \lambda) = -\sum_{i=1}^{n} p_i \; {}^2\log(p_i) + \lambda(\sum_{i=1}^{n} p_i).$$

Merk op dat ${}^2\log(x) = \frac{\log(x)}{\log(2)}$, dus is ${}^2\log'(x) = \frac{1}{\log(2)x}$. Er geldt

$$\frac{\partial L}{\partial p_i} = {}^2\log(p_i) + p_i \frac{1}{\log(2)p_i} + \lambda = {}^2\log(p_i) + \frac{1}{\log(2)} + \lambda$$

en uit $\frac{\partial L}{\partial p_i} = 0$ volgt dus ${}^2\log(p_i) = -\frac{1}{\log(2)} - \lambda$. In het bijzonder moeten dus alle p_i gelijk zijn en uit de randvoorwaarde volgt dan natuurlijk $p_i = \frac{1}{n}$.

Voor de entropie van de uniforme verdeling geldt dan $H = -\,{}^2\log(\frac{1}{n}) = {}^2\log(n)$, dus heeft een verdeling met 2^m mogelijke uitkomsten de entropie m. Algemeen geeft de entropie H aan hoeveel bits gemiddeld nodig zijn om de uitkomsten van een experiment te coderen.

Belangrijke begrippen in deze les

- kritieke punten
- lokale maxima/minima

- positief/negatief definiet
- extrema onder randvoorwaarden
- Lagrange functie, Lagrange multiplicatoren

Opgaven

21. De uitwerking van een hoeveelheid van x μg (microgram) van een medicijn is op een tijdstip t na de inneming gegeven door

$$f(x,t) = x^2(a-x)t^2e^{-t},$$

waarbij a de maximaal mogelijke hoeveelheid van de medicijn is. Wat is de maximale uitwerking van de medicijn die bereikt kan worden, en voor welke hoeveelheid wordt deze op welk tijdstip bereikt?

22. Vind de kritieke punten, lokale maxima, minima en zadelpunten van de volgende functies:

(i) $f(x,y) := x^3 + 6xy^2 - 2y^3 - 12x$;
(ii) $f(x,y) := xy\, e^{-(x^2+y^2)}$;
(iii) $f(x,y) := \frac{x}{1+x^2+y^2}$;
(iv) $f(x,y,z) := x^2y + y^2z + z^2 - 2x$.

23. Vind de kritieke punten van de volgende functies en beslis waar minima, maxima of zadelpunten liggen:

(i) $f(x,y) := x^2 - y^2 + xy$;
(ii) $f(x,y) := x^2 + y^2 + 3xy$;
(iii) $f(x,y) := e^{1+x^2-y^2}$;
(iv) $f(x,y) := x^2 - 3xy + 5x - 2y + 6y^2 + 8$;
(v) $f(x,y) := \sin(x^2+y^2)$;
(vi) $f(x,y) := \cos(x^2+y^2)$;
(vii) $f(x,y) := y + x\sin(y)$;
(viii) $f(x,y) := e^x\cos(y)$.

24. Vind de kritieke punten van de volgende functies en beslis waar minima, maxima of zadelpunten liggen:

(i) $f(x,y) := xy + \frac{1}{x} + \frac{1}{y}$;
(ii) $f(x,y) := \log(2+\sin(xy))$;
(iii) $f(x,y) := x\sin(y)$;
(iv) $f(x,y) := (x+y)(xy+1)$.

25. Zij $f(x,y) := x^2 + y^2 + kxy$ waarbij k een constante is. Bepaal de kritieke punten van $f(x,y)$. Voor welke waarden van k heeft $f(x,y)$ een extremum, voor welke een zadelpunt?

26. Zij $f(x,y) := \frac{1}{xy}$. Vind het punt op de grafiek van $f(x,y)$ dat het dicht bij de oorsprong $(0,0,0)$ in $\mathbb{R}^3$ ligt, d.w.z. vind het punt $(x, y, \frac{1}{xy})$ met de kleinste afstand van $(0,0,0)$.

27. Vind de extrema van de volgende functies onder de aangegeven randvoorwaarden:

    (i) $f(x,y,z) := x - y + z$ onder de randvoorwaarde $x^2 + y^2 + z^2 = 2$;

    (ii) $f(x,y) := x - y$ onder de randvoorwaarde $x^2 - y^2 = 2$;

    (iii) $f(x,y) := x$ onder de randvoorwaarde $x^2 + 2y^2 = 3$;

    (iv) $f(x,y) := 3x + 2y$ onder de randvoorwaarde $2x^2 + 3y^2 = 3$;

    (v) $f(x,y,z) := x + y + z$ onder de randvoorwaarden $x^2 - y^2 = 1$ en $2x + z = 1$.

28. De temperatuur $T = T(x,y,z)$ op het oppervlak van een kogel van straal 1 rond de oorsprong $(0,0,0)$ is gegeven door $T(x,y,z) = xz + yz$. Vind de *hot spots* op de kogel, dus de punten met maxima van de temperatuur.

29. Een open doos moet volume V hebben. De onderkant van de doos moet stabiel zijn en de voorkant van de doos moet er mooi uitzien, daarom is het materiaal voor deze twee zijden van de doos vijf keer zo duur als het materiaal voor de andere drie zijden. Wat zijn de afmetingen van de goedkoopste doos met deze eigenschappen?

30. Bepaal het punt op de kromme gegeven door $17x^2 + 12xy + 8y^2 = 100$ die het dichtst bij de oorsprong $(0,0)$ ligt. (De kromme is in feite een ellips die schuin in het vlak ligt.)

31. Bepaal (bij benadering) het punt op de grafiek van $y = \log(x)$ die het dichtst bij het punt $(1,1)$ ligt.

    (Je moet hierbij uiteindelijk een nulpunt van een gewone functie van x bepalen, die alleen maar numeriek maar niet analytisch te vinden is.)

32. Bepaal het maximum van de functie $f(x_1, \ldots, x_n) = x_1 \cdot \ldots \cdot x_n$ onder de randvoorwaarde $x_1 + \ldots + x_n = a$ voor een zekere $a > 0$. Veronderstel hierbij dat $x_i \geq 0$ voor alle i.

    Concludeer uit het resultaat dat het meetkundig gemiddelde $\sqrt[n]{x_1 \cdot \ldots \cdot x_n}$ uit positieve getallen x_i steeds kleiner of gelijk aan het rekenkundig gemiddelde $\frac{1}{n}(x_1 + \ldots + x_n)$ is.

Les 4 Integratie van functies van meerdere variabelen

In deze les gaan we het omgekeerde van de afgeleide, de integratie bekijken, en zien hoe we deze voor functies van meerdere variabelen definiëren en uitrekenen. De functies waar we het hierbij over hebben zijn weer functies van n variabelen die waarden in $\mathbb{R}$ hebben, dus functies $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$.

Net zo als we met de integraal voor een gewone functie van één variabel de oppervlakte onder een grafiek berekenen, geeft de integraal voor een functie van twee variabelen het volume onder de grafiek van de functie aan. Analoog geeft voor een algemene functie van n variabelen de integraal een (veralgemeend) volume in de $n+1$-dimensionale ruimte aan.

Meerdimensionale integralen hebben veel toepassingen in de patroonverwerking, bijvoorbeeld is de intensiteit van een plaatje een functie van de $x-y$-coördinaten en is de totale intensiteit op een gebied de integraal van de intensiteit over dit gebied. Maar ook voor kansverdelingen van gecombineerde stochasten die door een dichtheidsfunctie gegeven zijn, moeten we meerdimensionale integralen berekenen om de kans op uitkomsten in een zeker interval te vinden of de verwachtingswaarde te bepalen.

We zullen in deze les vooral functies van twee of drie variabelen behandelen, omdat deze belangrijke toepassingen hebben en het schrijfwerk hierbij nog beperkt is. Het algemene geval werkt echter op een analoge manier en bevat geen verdere complicaties.

4.1 Integratie op (veralgemeende) rechthoeken

Voor een functie $f(x)$ van één variabel hebben we de integraal $\int_a^b f(x)\,dx$ gedefinieerd als limiet van de som $\sum_{i=0}^{N-1} f(a+i\Delta x)\Delta x$, waarbij $N \cdot \Delta x = b-a$, dus

$$\int_a^b f(x)\,dx := \lim_{N\to\infty} \sum_{i=0}^{N-1} f(a+i\Delta x)\cdot \Delta x, \text{ waarbij } \Delta x = \frac{b-a}{N}.$$

Het idee hierbij is, de oppervlakte onder de grafiek van $f(x)$ te benaderen door een rij van rechthoeken van breedte Δx en hoogte $f(a+i\Delta x)$.

> In de wiskunde is een iets algemenere definitie gebruikelijk, waarbij men punten $a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b$ kiest en de som $\sum_{i=0}^{N-1} f(x_i)(x_{i+1}-x_i)$ bekijkt. Dit betekent gewoon dat de rechthoeken niet alle even breed hoeven te zijn. Voor de limiet is het dan wel noodzakelijk dat het maximum van de intervallen $(x_{i+1}-x_i)$ tegen 0 gaat.
>
> Voor de functies waar we het hier over hebben is onze eenvoudigere definitie echter voldoende, de gevallen waar de definities tot verschillende resultaten leiden zijn erg pathologisch.

Dit idee kunnen we nu als volgt op functies van meerdere veranderlijke veralgemenen: Uit een interval $[a, b]$ voor de variabel x wordt bij twee variabelen x en y de combinatie van twee intervallen, één voor x en één voor y, dit geeft de rechthoek

$$R_{a,b,c,d} := \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b,\ c \leq y \leq d\}.$$

Analoog krijgen we voor een functie van drie variabelen een blok

$$B_{a,b,c,d,e,f} := \{(x, y, z) \in \mathbb{R}^3 \mid a \leq x \leq b,\ c \leq y \leq d,\ e \leq z \leq f\}.$$

Algemeen geeft bij n variabelen $x_1, \ldots, x_n$ de combinatie van n intervallen $[a_1, b_1], \ldots, [a_n, b_n]$ de n-dimensionale rechthoek

$$R_{a_1,b_1,\ldots,a_n,b_n} := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid a_i \leq x_i \leq b_i \text{ voor } i = 1, \ldots, n\}.$$

De integratie over een rechthoek wordt nu analoog met het geval van één variabel gedefinieerd als limiet van de som over pilaren met als grondvlak een rechthoek met zijden $\Delta x, \Delta y$ en hoogte $f(a + i\Delta x, c + j\Delta y)$. Het volume van zo'n pilaar is natuurlijk $f(a + i\Delta x, c + j\Delta y) \cdot \Delta x \Delta y$. Men definieert dus:

$$\int_{R_{a,b,c,d}} f(x, y)\, dA = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} f(a + i\Delta x, c + j\Delta y) \cdot \Delta x \Delta y,$$

waarbij $N = \frac{b-a}{\Delta x}$ en $M = \frac{d-c}{\Delta y}$. Hierbij schrijven we het symbool dA voor het *differentiaal* van een oppervlakte element, dus voor de limiet van de rechthoeken met zijden $\Delta x, \Delta y$.

> In een algemenere definitie wordt de rechthoek $R_{a,b,c,d}$ in kleine stukken ΔA_i gesplitst, die niet noodzakelijk rechthoekig hoeven te zijn. Men kiest nu in elk stuk A_i een punt (x_i, y_i) en benadert de integraal door de som $\sum_{i=1}^{N} f(x_i, y_i)\Delta A_i$. Voor de limiet moet de diameter van de ΔA_i tegen 0 gaan.
>
> Ook hier geldt, dat dit voor redelijke functies geen verschil met onze eenvoudigere definitie geeft. Als redelijk beschouwen we hierbij functies, die stuksgewijs continu zijn.

In de definitie hebben we het met twee limieten tegelijkertijd te maken, met de limiet $\Delta x \to 0$ en de limiet $\Delta y \to 0$. Deze kunnen we op verschillende manieren berekenen, we kunnen of eerst de limiet over Δx en dan die over Δy uitvoeren, of andersom, of we kunnen de twee tegelijkertijd tegen 0 laten gaan. Het is niet vanzelfsprekend dat de verschillende manieren in elk geval hetzelfde resultaat geven, en bij zekere functies is dit helaas ook niet het geval. Maar we mogen hier weer ervan uitgaan, dat het voor de functies die we in de praktijk tegenkomen wel goed gaat en dat we altijd in de aangename situatie zijn die door de stelling van Fubini weergegeven wordt:

Stelling van Fubini: Als de integraal $\int_{R_{a,b,c,d}} f(x,y)\ dA$ bestaat, dan bestaan ook de functies $g(y) := \int_a^b f(x,y)\ dx$ en $h(x) := \int_c^d f(x,y)\ dy$ en er geldt

$$\begin{aligned}\int_{R_{a,b,c,d}} f(x,y)\ dA &= \int_c^d g(y)\ dy = \int_c^d \left(\int_a^b f(x,y)\ dx\right) dy \\ &= \int_a^b h(x)\ dx = \int_a^b \left(\int_c^d f(x,y)\ dy\right) dx.\end{aligned}$$

Merk op dat we in $g(y) := \int_a^b f(x,y)\ dx$ bij de integratie de variabel y als constante beschouwen, dit is dus een gewone integratie van één veranderlijke. Hetzelfde geldt voor de functie $h(x)$. We kunnen dus een integraal over een functie van twee variabelen uitwerken door eerst over een van de variabelen te integreren, en vervolgens over de andere, dus door *geïtereerde integraties* van een veranderlijke.

Voorbeeld 1: Zij $f(x,y) := 2x + 3y$ en $R := [0,2] \times [3,4]$. Dan is

$$\begin{aligned}\int_R f(x,y)\ dA &= \int_0^2 \left(\int_3^4 (2x+3y)\ dy\right) dx = \int_0^2 \left((2xy + \frac{3}{2}y^2)\Big|_3^4\right) dx \\ &= \int_0^2 (8x + 24 - 6x - \frac{27}{2})\ dx = \int_0^2 (2x + \frac{21}{2})\ dx \\ &= (x^2 + \frac{21}{2}x)\Big|_0^2 = 4 + 21 = 25.\end{aligned}$$

We kunnen ook eerst over x en dan over y integreren:

$$\begin{aligned}\int_R f(x,y)\ dA &= \int_3^4 \left(\int_0^2 (2x+3y)\ dx\right) dy = \int_3^4 \left((x^2 + 3xy)\Big|_0^2\right) dy \\ &= \int_3^4 (4+6y)\ dy = (4y + 3y^2)\Big|_3^4 = 16 + 48 - 12 - 27 = 25.\end{aligned}$$

We zien dat in dit voorbeeld de tweede manier iets makkelijker is dan de eerste, maar de resultaten zijn natuurlijk hetzelfde.

Voorbeeld 2: Zij $f(x,y) := e^{x+y}$ en $R := [1,2] \times [1,2]$. Dan is

$$\begin{aligned}\int_R f(x,y)\ dA &= \int_1^2 \left(\int_1^2 e^{x+y}\ dy\right) dx = \int_1^2 \left(\int_1^2 e^x \cdot e^y\ dy\right) dx \\ &= \int_1^2 \left(\int_1^2 e^y\ dy\right) e^x\ dx = \int_1^2 \left(e^y\Big|_1^2\right) e^x\ dx \\ &= \int_1^2 (e^2 - e)e^x\ dx = (e^2 - e)\int_1^2 e^x\ dx = (e^2 - e)\cdot e^x\Big|_1^2 \\ &= (e^2 - e)(e^2 - e) = (e^2 - e)^2.\end{aligned}$$

Voor functies van drie variabelen geldt iets soortgelijks als voor functies van twee variabelen, we moeten nu over kleine volume elementen (blokken)

$\Delta x \Delta y \Delta z$ integreren, die in de limiet tot een differentiaal dV van een volume element wordt. Ook de integratie over de kleine volume elementen kunnen we weer opsplitsen in drie gewone integraties, er geldt:

$$\int_{B_{a,b,c,d,e,f}} f(x,y,z)\ dV = \int_a^b \left(\int_c^d \left(\int_e^f f(x,y,z)\ dx \right) dy \right) dz.$$

Ook hier kunnen we een andere volgorde voor de integraties kiezen, het maakt niets uit of we eerst over x, y of z integreren. Soms scheelt een geschikte keuze van de volgorde zelfs een hoop rekenwerk.

Merk op: In principe is het natuurlijk logisch, dat de eerste integraal $\int_a^b$ met grenzen a en b bij de laatste differentiaal dz hoort, de tweede integraal $\int_c^d$ bij de voorlaatste differentiaal dy enzovoorts. Maar men is vaak iets slordig met de haakjes en ook met de volgorde, en bij ingewikkelde functies wordt de notatie alsnog onoverzichtelijk. Daarom is er een vaak gebruikte conventie, de differentiaal meteen achter de bijhorende integraal te plaatsen om zo duidelijk te maken voor welke integratie variabel de grenzen van dit integraalteken gelden. In plaats van de schrijfwijze hierboven vind je dus ook vaak:

$$\int_{B_{a,b,c,d,e,f}} f(x,y,z)\ dV = \int_a^b dx \int_c^d dy \int_e^f dz\ f(x,y,z).$$

Voorbeeld: Zij $f(x,y,z) := \frac{x^2 z^3}{1+y^2}$ en $R := [0,1] \times [0,1] \times [0,1]$. Dan is

$$\begin{aligned}
\int_R f(x,y,z)\ dV &= \int_0^1 \left(\int_0^1 \left(\int_0^1 \frac{x^2 z^3}{1+y^2}\ dx \right) dz \right) dy \\
&= \int_0^1 \left(\int_0^1 \left(\frac{x^3 z^3}{3(1+y^2)} \Big|_0^1 \right) dz \right) dy = \int_0^1 \left(\int_0^1 \frac{z^3}{3(1+y^2)}\ dz \right) dy \\
&= \int_0^1 \left(\frac{z^4}{12(1+y^2)} \Big|_0^1 \right) dy = \int_0^1 \frac{1}{12(1+y^2)}\ dy \\
&= \frac{1}{12} \arctan(y) \Big|_0^1 = \frac{1}{12}(\frac{\pi}{4} - 0) = \frac{\pi}{48}.
\end{aligned}$$

Opdracht 15 *Bepaal voor $f(x,y) := 2xy + 3y^2$ de integraal $\int_R f(x,y)\ dA$ voor de rechthoek $R = [a,b] \times [c,d] = \{(x,y) \mid x \in [a,b], y \in [c,d]\}$ door geïtereerde integratie over x en y. Laat zien dat het resultaat niet van de volgorde van de integraties afhangt.*

4.2 Integratie over normaalgebieden

Het lijkt natuurlijk erg beperkend als we alleen maar over rechthoek gebieden kunnen integreren. In feite is de beperking niet zo groot, want we kunnen een willekeurig gebied benaderen door een combinatie van kleine rechthoeken en als de onderverdeling voldoende fijn is, kunnen we ervan uit gaan dat de fout die we hierbij maken klein (verwaarloosbaar) is. Maar voor gebieden die alleen maar door krommen begrensd zijn (zo als een cirkel), moeten we hiervoor vaak een redelijk groot aantal rechthoeken bekijken om een redelijke benadering te krijgen, en dit is ook weer een beetje vervelend.

Er is echter een algemenere klasse van gebieden dan rechthoek gebieden, waarvoor we de integraal rechtstreeks kunnen uitrekenen, dit zijn de *normaalgebieden*. In het 2-dimensionale geval zijn dit de gebieden van de vorm

$$B = \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, \varphi_1(x) \leq y \leq \varphi_2(x)\},$$

d.w.z. gebieden die door twee rechte lijnen $x = a$ en $x = b$ evenredig met de y-as en de twee krommen $\varphi_1(x)$ en $\varphi_2(x)$ begrensd zijn die elkaar niet snijden. In dit geval geldt

$$\int_B f(x, y)\ dA = \int_a^b \left(\int_{\varphi_1(x)}^{\varphi_2(x)} f(x, y)\ dy \right) dx.$$

Net zo goed kan men natuurlijk ook twee lijnen evenredig met de x-as en twee krommen $\psi_1(y)$ en $\psi_2(y)$ als grenzen hebben, dit geeft ook een normaalgebied, te weten

$$B = \{(x, y) \in \mathbb{R}^2 \mid c \leq y \leq d, \psi_1(y) \leq x \leq \psi_2(y)\}$$

dan wordt de integraal over B berekend als

$$\int_B f(x, y)\ dA = \int_c^d \left(\int_{\psi_1(y)}^{\psi_2(y)} f(x, y)\ dx \right) dy.$$

Voorbeeld 1: We berekenen de integraal van de functie $f(x, y) := x^2 y$ over de halfcirkel B van straal 1 rond $(0, 0)$ die boven de x-as ligt. De halfcirkel B is begrensd door de lijnen $x = -1$ en $x = 1$ en de krommen $\varphi_1(x) = 0$ en $\varphi_2(x) = \sqrt{1 - x^2}$. We hebben dus

$$\begin{aligned}
\int_B f(x, y)\ dA &= \int_{-1}^1 \left(\int_0^{\sqrt{1-x^2}} x^2 y\ dy \right) dx = \int_{-1}^1 \left(\frac{1}{2} x^2 y^2 \Big|_0^{\sqrt{1-x^2}} \right) dx \\
&= \int_{-1}^1 \frac{1}{2} x^2 (1 - x^2)\ dx = \frac{1}{2} \int_{-1}^1 (x^2 - x^4)\ dx = \frac{1}{2} \left(\frac{x^3}{3} - \frac{x^5}{5} \right) \Big|_{-1}^1 \\
&= \frac{1}{2} (\frac{1}{3} - \frac{1}{5} - \frac{-1}{3} + \frac{-1}{5}) = \frac{1}{3} - \frac{1}{5} = \frac{2}{15}.
\end{aligned}$$

Voorbeeld 2: We bepalen de integraal $\int_D x^3 y + \cos(x)\ dA$ op de driehoek D met hoekpunten $(0, 0)$, $(\frac{\pi}{2}, 0)$, $(\frac{\pi}{2}, \frac{\pi}{2})$. De driehoek is begrensd door de lijnen $x = 0$ en $x = \frac{\pi}{2}$ en door de functies $\varphi_1(x) = 0$ en $\varphi_2(x) = x$. Hiermee krijgen we:

$$\begin{aligned}
\int_D x^3 y + \cos(x)\ dA &= \int_0^{\frac{\pi}{2}} \left(\int_0^x x^3 y + \cos(x)\ dy \right) dx \\
&= \int_0^{\frac{\pi}{2}} (\frac{1}{2} x^3 y^2 + \cos(x) y) \Big|_0^x\ dx = \int_0^{\frac{\pi}{2}} (\frac{1}{2} x^5 + x \cos(x))\ dx \\
&= (\frac{1}{12} x^6 + x \sin(x) + \cos(x)) \Big|_0^{\frac{\pi}{2}} = \frac{\pi^6}{768} + \frac{\pi}{2} - 1
\end{aligned}$$

(merk op dat met partiële integratie geldt dat $\int x \cos(x)\ dx = x \sin(x) - \int \sin(x)\ dx = x \sin(x) + \cos(x)$).

Opdracht 16 *Bepaal de integraal* $\int_G f(x,y)\,dA$ *van de functie* $f(x,y) := x+y$ *op het gebied* G *gegeven door* $G := \{(x,y) \mid 0 \le x \le 1, 1 \le y \le e^x\}$.

In drie dimensies zijn normaalgebieden begrensd door een gebied B in het $x-y$-vlak (bijvoorbeeld) en twee functies $\varphi_1(x,y)$ en $\varphi_2(x,y)$, die de variabel z inschakelen. Dan geldt

$$\int_V f(x,y,z)\,dV = \int_B \left(\int_{\varphi_1(x,y)}^{\varphi_2(x,y)} f(x,y,z)\,dz \right) dA.$$

Na het uitwerken van de binnenste integraal over z is dit terug gebracht tot een integratie met twee variabelen op het 2-dimensionale gebied B, en het zou dus handig zijn als B ook weer een normaalgebied is.

Opdracht 17 *Laat zien dat het gebied* B *dat tussen de grafieken van* $y = x^2$ *en* $y = x$ *ligt een normaalgebied is en bepaal de oppervlakte van het gebied* B. *Bereken verder de integraal* $\int_B 1 + 2xy\,dA$.

4.3 Substitutie

Een belangrijke methode in de integratie van gewone functies van één variabel is de substitutie. Het idee hierbij is, de integratievariabel x door een geschikte nieuwe variabel u te vervangen zo dat de integratie makkelijker wordt. Als we in de integraal $\int f(x)\,dx$ de variabel x door een nieuwe variabel u willen vervangen, moeten we de samenhang van x en u kennen, en dit drukken we uit door x te schrijven als een functie $x = x(u)$ van u. Als we nu een functie $g(u)$ definiëren door $g(u) := f(x(u))$ dan zegt de substitutie regel dat

$$\int f(x)\,dx = \int g(u)x'(u)\,du = \int f(x(u))x'(u)\,du.$$

Als we ons nu nog eens herinneren dat de integraal gedefinieerd is als de limiet van de som $\sum f(x_i)(x_i - x_{i-1}) = \sum f(x_i)\Delta x$, kunnen we precies de reden zien, waarom de differentiaal dx door de nieuwe differentiaal $x'(u)du$ vervangen moet worden. In een kleine omgeving van u vervangen we de functie $x(u)$ door de lineaire benadering van de Taylor reeks, dus door de lineaire functie $x(u+\Delta u) \approx x(u) + x'(u)\Delta u$. Maar hieruit volgt dat

$$\Delta x = x(u+\Delta u) - x(u) = x'(u)\Delta u,$$

de afgeleide $x'(u)$ geeft dus juist aan hoe groot de stappen Δx worden waarin we x veranderen als we u in stappen van Δu veranderen.

Als we nu weer naar de limiet $\Delta x \to 0$ kijken, krijgen we de relatie

$$dx = x'(u)\,du$$

voor de differentialen.

De Jacobi matrix

Het idee van de substitutie voor gewone functies gaan we nu veralgemenen op functies van meerdere variabelen. Zij $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ een functie van de n variabelen $x_1, \ldots, x_n$.

Stel we willen nu nieuwe variabelen $u_1, \ldots, u_n$ hanteren, dan hangen de x_i van de nieuwe variabelen u_j af, en we schrijven x_i als functie

$$x_i(\mathbf{u}) = x_i(u_1, \ldots, u_n).$$

Net zo als boven kunnen we nu in een kleine omgeving van $\mathbf{u}$ de functie $x_i(\mathbf{u})$ door de lineaire benadering vervangen, dit geeft

$$x_i(\mathbf{u} + \Delta\mathbf{u}) = x_i(\mathbf{u}) + \nabla x_i(\mathbf{u}) \cdot \Delta\mathbf{u}.$$

Als we de componenten $x_i(\mathbf{u})$ nu in een vector schrijven, krijgen we een functie $\mathbf{x}(\mathbf{u}) : \mathbb{R}^n \to \mathbb{R}^n$ gegeven door

$$\mathbf{x}(\mathbf{u}) := \begin{pmatrix} x_1(\mathbf{u}) \\ \vdots \\ x_n(\mathbf{u}) \end{pmatrix}$$

en als lineaire benadering hiervan krijgen we:

$$\mathbf{x}(\mathbf{u} + \Delta\mathbf{u}) = \begin{pmatrix} x_1(\mathbf{u} + \Delta\mathbf{u}) \\ \vdots \\ x_n(\mathbf{u} + \Delta\mathbf{u}) \end{pmatrix} = \begin{pmatrix} x_1(\mathbf{u}) \\ \vdots \\ x_n(\mathbf{u}) \end{pmatrix} + \begin{pmatrix} \nabla x_1(\mathbf{u})^{tr} \\ \vdots \\ \nabla x_n(\mathbf{u})^{tr} \end{pmatrix} \Delta\mathbf{u}.$$

Maar de matrix $J := \begin{pmatrix} \nabla x_1(\mathbf{u})^{tr} \\ \vdots \\ \nabla x_n(\mathbf{u})^{tr} \end{pmatrix}$ is een oude bekende, in de i-de rij staat namelijk in de j-de kolom de afgeleide van $x_i(\mathbf{u})$ naar de variabel u_j, dus de partiële afgeleide $\frac{\partial x_i}{\partial u_j}$. Dit betekent, dat J juist de *Jacobi matrix* van $\mathbf{x}(\mathbf{u})$ is en we hebben gevonden dat

$$\mathbf{x}(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{x}(\mathbf{u}) + J \cdot \Delta\mathbf{u} \quad \text{en dus} \quad \Delta\mathbf{x} := \mathbf{x}(\mathbf{u} + \Delta\mathbf{u}) - \mathbf{x}(\mathbf{u}) = J \cdot \Delta\mathbf{u}.$$

Dit is volledig analoog met de formule voor gewone functies, de Jacobi matrix J is de veralgemening van de afgeleide $x'(u)$ die we toen hadden.

Betekenis van de Jacobiaan

We zullen nu de rol van de Jacobi matrix voor de substitutie van functies van meerdere variabelen toelichten. Hiervoor kijken we eerst naar een functie van twee variabelen, x en y. We kiezen twee nieuwe variabelen u en v en schrijven x en y als functies van u en v, dus $x = x(u, v)$, $y = y(u, v)$. Met behulp van de Jacobi matrix J kunnen we nu de functie

$$(x(u, v), y(u, v)) : \begin{pmatrix} u \\ v \end{pmatrix} \to \begin{pmatrix} x(u, v) \\ y(u, v) \end{pmatrix}$$

in een omgeving van (u, v) door de Taylor veelterm van graad 1 benaderen, voor het verschil van de functiewaarden geldt dan:

$$\begin{pmatrix} \Delta x(u,v) \\ \Delta y(u,v) \end{pmatrix} := \begin{pmatrix} x(u+\Delta u, v+\Delta v) - x(u,v) \\ y(u+\Delta u, v+\Delta v) - y(u,v) \end{pmatrix} = J \cdot \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}$$
$$= \begin{pmatrix} \frac{\partial x(u,v)}{\partial u} & \frac{\partial x(u,v)}{\partial v} \\ \frac{\partial y(u,v)}{\partial u} & \frac{\partial y(u,v)}{\partial v} \end{pmatrix} \cdot \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}.$$

Bij de substitutie van functies met één variabel hebben we gezien dat we de differentiaal dx door $x'(u)\,du$ moeten vervangen. De vraag is nu, hoe in het geval van twee variabelen de differentiaal $dA = dx\,dy$ met de nieuwe differentiaal $du\,dv$ samenhangt.

Om hier uit te komen, gaan we even een stap terug en interpreteren de integraal weer als som van pilaren over kleine rechthoeken met zijden $\Delta x, \Delta y$. Zo'n rechthoek moeten we nu door de nieuwe variabelen u en v beschrijven, en bij benadering lukt dit in een punt (x, y) met behulp van de Jacobi matrix door de vergelijking

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = J \cdot \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} \quad \text{of} \quad \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = J^{-1} \cdot \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}.$$

Omdat J en dus ook J^{-1} een lineaire afbeelding is, is het beeld van de rechthoek met zijden $\Delta x, \Delta y$ onder J^{-1} een parallellogram. De vraag is nu wat de oppervlakte van dit parallellogram is.

De oplossing van dit vraagstuk is verrassend eenvoudig, we hebben alleen maar de determinant van J nodig.

Stelling: De absolute waarde van de determinant $\det(A)$ van een $n \times n$-matrix A geeft het *volume* van het parallellepipedum aan, dat door de kolommen van de matrix A opgespannen wordt.

Met andere woorden is $|\det(A)|$ het volume van het beeld onder A van de eenheidsvierkant (eenheidskubus, eenheidshyperkubus, enz.) die door de standaardbasis opgespannen wordt.

Deze stelling kunnen we als volgt inzien:

Voor een diagonaalmatrix A is het opspansel van de kolommen van A een rechthoek, blok, enzovoorts, en het volume hiervan is het product van de absolute waarden van de elementen op de diagonaal. Maar dit is ook de absolute waarde van de determinant van de matrix A.

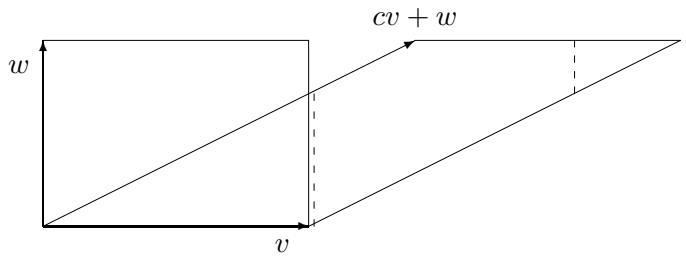
Verder weten we dat we elke matrix door elementaire transformaties op diagonaal vorm kunnen brengen, we moeten dus alleen maar kijken, wat er met het volume gebeurt als we een elementaire transformatie toepassen:

(i) Als we twee kolommen verwisselen, verandert het parallellepipedum niet, het volume blijft dus hetzelfde. De determinant wordt hierbij met -1 vermenigvuldigd, maar de absolute waarde blijft gelijk.

(ii) Als we een kolom met een factor $c \neq 0$ vermenigvuldigen, wordt ook het volume van het parallellepipedum $|c|$ keer zo groot. Maar in dit geval wordt ook de determinant met c vermenigvuldigd.

(iii) Als we een veelvoud van een vector bij een andere optellen, verandert de determinant niet, dus mag ook het volume bij deze transformatie niet veranderen. Omdat hierbij alleen maar twee vectoren een rol spelen, is het voldoende dit in het 2-dimensionale geval te bekijken. De schets hieronder licht dit toe.

De rechthoek opgespannen door de vectoren v en w en het parallellogram opgespannen door v en $cv + w$ hebben dezelfde oppervlakte, omdat de oppervlakte van een parallellogram gelijk is aan het product van de grondzijde en de hoogte.



Dat de rechthoek en het parallellogram dezelfde oppervlakte hebben, laat zich ook door knippen en plakken aantonen, als we het parallellogram langs de twee stippellijnen in stukken snijden, zien we makkelijk in dat de delen de rechthoek precies overdekken.

De stelling hierboven toegepast op de Jacobi matrix betekent dat het parallellogram met zijden $\Delta u, \Delta v$ oppervlakte $|\det(J^{-1})| \cdot \Delta x \Delta y$ heeft en hieruit volgt omgekeerd dat

$$\Delta x \Delta y = |\det(J)| \cdot \Delta u \Delta v.$$

Door nu weer de limieten $\Delta x \to 0$ en $\Delta y \to 0$ te nemen, krijgen we dat voor de differentialen geldt dat

$$dx\, dy = |\det(J)|\, du\, dv.$$

Omdat de determinant van de Jacobi matrix zo'n belangrijke rol speelt, heeft deze ook een eigen naam, ze heet *Jacobiaan*.

Het argument dat we net op twee variabelen hebben toegepast, geldt natuurlijk volledig analoog voor functies van meerdere veranderlijken. We transformeren de variabelen $x_1, \ldots, x_n$ op nieuwe variabelen $u_1, \ldots, u_n$ met $x_i = x_i(u_1, \ldots, u_n)$ en bepalen de Jacobi matrix J met $J_{ij} = \frac{\partial x_i}{\partial u_j}$, dan geldt:

$$\begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{pmatrix} = J \cdot \begin{pmatrix} \Delta u_1 \\ \vdots \\ \Delta u_n \end{pmatrix}$$

en tussen de n-dimensionale volumes van de blok $\Delta x_1 \ldots \Delta x_n$ en het parallellepipedum $\Delta u_1 \ldots \Delta u_n$ bestaat de relatie

$$\Delta x_1 \ldots \Delta x_n = |\det(J)| \cdot \Delta u_1 \ldots \Delta u_n.$$

Voor de differentialen van de volume elementen geldt dus:

$$dx_1\, dx_2 \ldots dx_n = |\det(J)|\, du_1\, du_2 \ldots du_n.$$

De Jacobiaan speelt dus bij functies van meerdere veranderlijken precies de rol van de afgeleide in het geval van functies van één veranderlijke.

Substitutieregel voor functies van meerdere variabelen

We kunnen nu de substitutieregel voor functies van meerdere veranderlijken formuleren. Voor het gemak doen we dit eerst voor functies van twee veranderlijken en geven dan de algemene regel aan.

Substitutieregel voor twee variabelen: Voor een coördinatentransformatie naar nieuwe variabelen u en v met $x = x(u,v)$, $y = y(u,v)$ en Jacobi matrix $J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$ wordt een functie $f(x,y)$ met betrekking tot de nieuwe coördinaten geschreven als $g(u,v)$ met $g(u,v) := f(x(u,v), y(u,v))$.

Voor de integraal van $f(x,y)$ over een gebied $B \subseteq \mathbb{R}^2$ geldt dan in de nieuwe coördinaten:

$$\int_B f(x,y)\, dx\, dy = \int_{B'} g(u,v)\, |\det(J)|\, du\, dv.$$

Hierbij moet het gebied B' in het $u-v$-vlak zo gekozen worden, dat (x,y) over B loopt als (u,v) over B' loopt, waarbij elke punt in B precies een keer voorkomt. Dit betekent dat de afbeelding $\begin{pmatrix} u \\ v \end{pmatrix} \to \begin{pmatrix} x(u,v) \\ y(u,v) \end{pmatrix}$ een bijectieve (omkeerbare) afbeelding van B' naar B is.

> In de praktijk spelen vooral speciale coördinatentransformaties een rol die we hieronder gaan bespreken. Bij deze transformaties laat zich de vraag of de afbeelding omkeerbaar is eenvoudig beantwoorden.
>
> Algemeen is de omkeerbaarheid een lastige vraag. Als de Jacobi matrix in een punt een inverteerbare matrix is, is de functie in een kleine omgeving van dit punt omkeerbaar (men noemt de functie dan lokaal inverteerbaar in dit punt). Maar hieruit volgt helaas niet dat de functie globaal omkeerbaar op een gebied B is, er bestaan zelfs functies die in iedere punt van een gebied B lokaal inverteerbaar zijn, maar niet omkeerbaar op B.

Voorbeeld: We bepalen de integraal van $f(x,y) := e^{\frac{y}{x+y}}$ op de driehoek gegeven door $B = \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 1,\, 0 \le y \le 1-x\}$. De driehoek is een normaalgebied en in principe kunnen we de integratie opsplitsen in twee in elkaar geschakelde gewone integraties, namelijk

$$\int_B f(x,y)\, dA = \int_0^1 \int_0^{1-x} e^{\frac{y}{x+y}}\, dy\, dx.$$

Het probleem is, dat deze integraal niet zo eenvoudig op te lossen is.

Een slimme transformatie van de variabelen is

$$x + y = u, \quad y = uv, \quad \text{dus } x = u - uv, \quad y = uv.$$

Merk op dat de transformatie juist zo gekozen is dat $e^{\frac{y}{x+y}} = e^{\frac{uv}{u}} = e^v$ wordt.

We gaan na dat (x, y) over B loopt als (u, v) over de eenheidsvierkant $B' = \{(u, v) \in \mathbb{R}^2 \mid 0 \leq u \leq 1,\ 0 \leq v \leq 1\}$ loopt: Ten eerste is duidelijk dat $x \geq 0$ en $y \geq 0$ voor $(u, v) \in B'$. Verder is $x = u(1 - v) \leq 1$, omdat $u \leq 1$ en $1 - v \leq 1$ zijn. Net zo is $y = uv \leq 1$. Ten slotte is $y \leq 1 - x \Leftrightarrow uv \leq 1 - u + uv \Leftrightarrow u \leq 1$, dus geldt ook $y \leq 1 - x$.

Omgekeerd moeten we nagaan dat alle punten van B echt doorlopen worden. Maar we kunnen de transformatie expliciet inverteren, er geldt

$$u = x + y \quad \text{en} \quad v = \frac{y}{x + y}$$

en omdat $x + y \leq 1$ en $\frac{y}{x+y} \leq 1$ voor $x, y \geq 0$ kunnen we voor iedere punt (x, y) een punt (u, v) aangeven, die door de transformatie op (x, y) wordt afgebeeld.

De Jacobi matrix en de Jacobiaan van de transformatie zijn

$$J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1 - v & -u \\ v & u \end{pmatrix} \quad \text{en} \quad \det(J) = (1 - v)u - (-uv) = u.$$

Met de transformatie op de nieuwe variabelen u en v krijgen we dus:

$$\int_B f(x, y)\, dA = \int_0^1 \int_0^1 e^v \cdot u\, du\, dv = \int_0^1 e^v \int_0^1 u\, du\, dv = \int_0^1 e^v \frac{1}{2}\, dv = \frac{1}{2}(e - 1).$$

Algemene substitutieregel voor meerdere veranderlijken: We vervangen de coördinaten $x_1, \ldots, x_n$ door nieuwe coördinaten $u_1, \ldots, u_n$ zo dat $x_i = x_i(u_1, \ldots, u_n)$ een functie van de nieuwe coördinaten wordt en noteren met J de Jacobi matrix van de coördinatentransformatie, d.w.z. $J_{ij} = \frac{\partial x_i}{\partial u_j}$.

Herschrijven van een functie $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ in de nieuwe coördinaten geeft een nieuwe functie

$$g(\mathbf{u}) = g(u_1, \ldots, u_n) := f(x_1(u_1, \ldots, u_n), \ldots, x_n(u_1, \ldots, u_n)).$$

Voor de integraal van $f(\mathbf{x})$ over een gebied $B \subseteq \mathbb{R}^n$ geldt dan met betrekking tot de nieuwe coördinaten:

$$\int_B f(\mathbf{x})\, dx_1 \ldots dx_n = \int_{B'} g(\mathbf{u})\, |\det(J)|\, du_1 \ldots du_n.$$

4.4 Poolcoördinaten, cilindercoördinaten, sferische coördinaten

De belangrijkste toepassingen van substitutie bij functies van meerdere variabelen zijn transformaties tussen verschillende standaard stelsels van coördinaten.

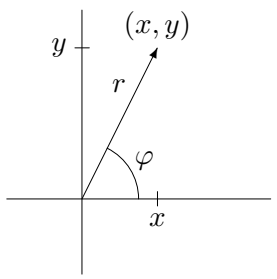
Als functies in het 2-dimensionale vlak alleen maar van de afstand van de oorsprong afhangen, is het vaak handig het probleem op *poolcoördinaten* te transformeren. Hierbij wordt een punt (x, y) door zijn afstand van de oorsprong en door een hoek beschreven.

Ook in de 3-dimensionale ruimte zijn er naast de gewone cartesische coördinaten nog twee andere stelsels coördinaten, die geschikt zijn voor zekere situaties, namelijk de *cilindercoördinaten* en de *sferische coördinaten* (ook *kogelcoördinaten* genoemd).

Poolcoördinaten

Bij functies van twee variabelen zijn vaak *poolcoördinaten* handig, in het bijzonder als het over integratie van functies op ronde gebieden gaat.

Het idee bij de poolcoördinaten is, een punt (x, y) door zijn afstand r van de oorsprong en door de hoek tussen de lijn door de oorsprong en (x, y) en de positieve x-as te beschrijven, zo als in de schets hieronder te zien:



Figuur I.13: Poolcoördinaten

Tussen de gewone coördinaten x, y en de poolcoördinaten r, φ bestaat het volgende verband:

$$\begin{aligned} x &= r\cos(\varphi), & y &= r\sin(\varphi) \\ r &= x^2 + y^2, & \tan(\varphi) &= \frac{y}{x}. \end{aligned}$$

In het bijzonder wordt een cirkelschijf $B(0, R) := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq R^2\}$ in poolcoördinaten een rechthoek, namelijk $[0, R] \times [0, 2\pi]$.

Voor de partiële afgeleiden van de transformatie $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos(\varphi) \\ r\sin(\varphi) \end{pmatrix}$ geldt:

$$\frac{\partial x}{\partial r} = \cos(\varphi), \quad \frac{\partial x}{\partial \varphi} = -r\sin(\varphi), \quad \frac{\partial y}{\partial r} = \sin(\varphi), \quad \frac{\partial y}{\partial \varphi} = r\cos(\varphi).$$

Hieruit volgt dat de Jacobi matrix J gelijk is aan

$$J = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{pmatrix}$$

en de Jacobiaan $\det(J)$ is dus

$$\det(J) = r\cos^2(\varphi) + r\sin^2(\varphi) = r.$$

Voor een functie $f(x, y)$ en $g(r, \varphi) := f(r\cos(\varphi), r\sin(\varphi))$ geldt dus de substitutieregel

$$\int f(x, y)\, dx\, dy = \int g(r, \varphi)\, r\ dr\, d\varphi.$$

Een eerste toepassing van de poolcoördinaten is natuurlijk het berekenen van de oppervlakte van een cirkel met straal R. Dit kunnen we berekenen door de constante functie $f(x,y) = 1$ over het gebied $B := B(0,R) := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq R^2\}$ te integreren. Maar in poolcoördinaten wordt B de rechthoek $B' = [0,R] \times [0,2\pi]$, want als (r,φ) over B' loopt, loopt (x,y) precies een keer over B. We hebben dus

$$\int_B 1\, dx\, dy = \int_{B'} r\, dr\, d\varphi = \int_0^{2\pi} \left(\int_0^R r\, dr \right) d\varphi = \int_0^{2\pi} \left(\frac{1}{2} r^2 \Big|_0^R \right) d\varphi$$
$$= \int_0^{2\pi} \frac{1}{2} R^2\, d\varphi = \frac{1}{2} R^2 \varphi \Big|_0^{2\pi} = \pi R^2.$$

Opdracht 18 *Bepaal de integraal* $\int_B (x^2 + 2xy)\, dA$ *op de halfcirkel* $B = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1, y \geq 0$. *Transformeer de functie (en het gebied) hiervoor op poolcoördinaten. (Herinnering: Met partiële integratie volgt* $\int \cos^2(x) = \sin(x)\cos(x) + \int \sin^2(x) = \sin(x)\cos(x) + \int (1 - \cos^2(x)) = \frac{1}{2}(\sin(x)\cos(x) + x)$.*)*

Toepassing: Normale verdeling

Een iets verrassendere toepassing van poolcoördinaten is dat we nu de integraal over de Gauss-functie e^{-x^2} kunnen berekenen die we in de normale verdeling tegenkomen en waarvan we tot nu toe de integraal niet analytisch konden bepalen.

Hiervoor bekijken we de analoge functie in twee variabelen, namelijk de functie $f(x,y) := e^{-(x^2+y^2)}$ en integreren deze functie één keer over een cirkel van straal R en één keer over een vierkant met lengte $2a$.

Zij eerst $B := B(0,R) := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq R^2\}$, dan is

$$\int_B e^{-(x^2+y^2)}\, dx\, dy = \int_0^R \left(\int_0^{2\pi} d\varphi \right) e^{-r^2} r\, dr = 2\pi \int_0^R e^{-r^2} r\, dr$$
$$= 2\pi(-\frac{1}{2} e^{-r^2})\Big|_0^R = -\pi(e^{-R^2} - 1) = \pi(1 - e^{-R^2})$$

Zij nu $B' := V(-a,a) := \{(x,y) \in \mathbb{R}^2 \mid |x| \leq a, |y| \leq a\}$ de vierkant met lengte $2a$ rond 0, dan is

$$\int_{B'} e^{-(x^2+y^2)}\, dx\, dy = \int_{-a}^a \left(\int_{-a}^a e^{-(x^2+y^2)} dx \right) dy = \int_{-a}^a \left(\int_{-a}^a e^{-x^2} dx \right) e^{-y^2}\, dy$$
$$= (\int_{-a}^a e^{-x^2}\, dx)(\int_{-a}^a e^{-y^2}\, dy) = (\int_{-a}^a e^{-x^2}\, dx)^2$$

Maar de cirkel $B(0,a)$ van straal a ligt volledig in het vierkant $V(-a,a)$ en dit ligt wederom volledig in de cirkel $B(0,\sqrt{2}\,a)$ met straal $\sqrt{2}\,a$. Omdat de functie $e^{-(x^2+y^2)} > 0$ is, volgt hieruit

$$\pi(1 - e^{-a^2}) \leq (\int_{-a}^a e^{-x^2}\, dx)^2 \leq \pi(1 - e^{-2a^2}).$$

Als we nu de limiet $a \to \infty$ laten lopen, wordt de rechter en de linker zijde van deze ongelijkheden gelijk aan π, en dus hebben we bewezen dat

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

Cilindercoördinaten

In de 3-dimensionale ruimte komt het vaak voor dat een probleem symmetrisch ten opzichte van een rotatie as is. Dit is bijvoorbeeld het geval voor het elektrische veld rond een rechte geleider. Bij dit soort problemen zijn *cilindercoördinaten* heel praktisch, die veronderstellen dat de rotatie-as de z-as is. Het idee van de cilindercoördinaten is, een punt (x, y, z) te beschrijven door poolcoördinaten voor het $x - y$-vlak en de gewone z-coördinaat.

Dit geeft:

$$x = r\cos(\varphi), \qquad y = r\sin(\varphi), \qquad z = z,$$

waarbij $r > 0$ en $\varphi \in [0, 2\pi)$. De Jacobi matrix J hiervan is

$$J = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \varphi} & \frac{\partial z}{\partial z} \end{pmatrix} = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) & 0 \\ \sin(\varphi) & r\cos(\varphi) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

en de Jacobiaan is

$$\det(J) = r\cos^2(\varphi) + r\sin^2(\varphi) = r.$$

Hieruit volgt voor een functie $f(x, y, z)$ en $g(r, \varphi, z) := f(r\cos(\varphi), r\sin(\varphi), z)$:

$$\int f(x, y, z)\, dx\, dy\, dz = \int g(r, \varphi, z)\, r\ dr\, d\varphi\, dz.$$

Sferische coördinaten

Bij functies op de 3-dimensionale ruimte die eigenlijk alleen maar van de afstand van een punt afhangen (zo als de gravitatie kracht of de intensiteit van een geïdealiseerde bron van licht) worden vaak *sferische coördinaten* toegepast. Het idee is, een punt door zijn afstand en twee ruimtelijke hoeken aan te geven.

Men splitst de vector van de oorsprong naar het punt (x, y, z) in zijn projecties in het $x - y$-vlak en op de z-as. De projectie in het $x - y$-vlak wordt door poolcoördinaten r en φ aangegeven en de projectie op de z-as met behulp van de hoek θ tussen (x, y, z) en de z-as (zie de schets in Figuur I.14).

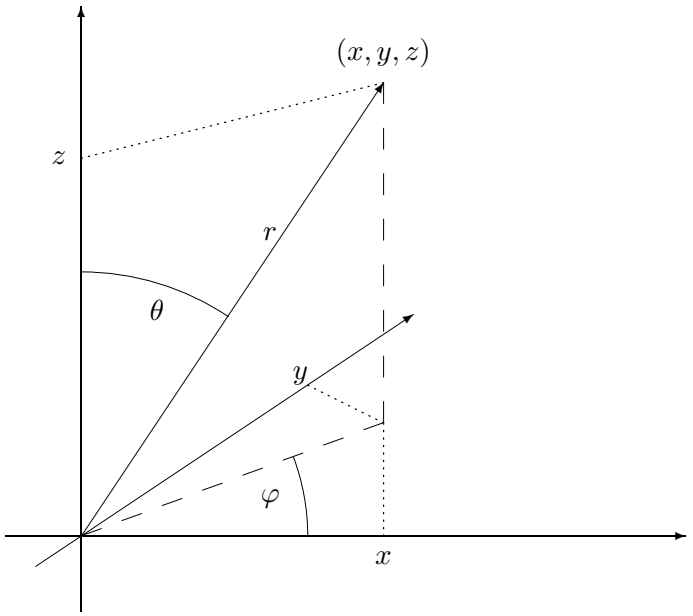
De coördinatentransformatie luidt:

$$x = r\cos(\varphi)\sin(\theta), \qquad y = r\sin(\varphi)\sin(\theta), \qquad z = r\cos(\theta)$$

waarbij $r > 0$, $\varphi \in [0, 2\pi)$, $\theta \in [0, \pi]$.

De Jacobi matrix J hiervan is

$$J = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} & \frac{\partial y}{\partial \theta} \\ \frac{\partial \theta}{\partial r} & \frac{\partial \theta}{\partial \varphi} & \frac{\partial \theta}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos(\varphi)\sin(\theta) & -r\sin(\varphi)\sin(\theta) & r\cos(\varphi)\cos(\theta) \\ \sin(\varphi)\sin(\theta) & r\cos(\varphi)\sin(\theta) & r\sin(\varphi)\cos(\theta) \\ \cos(\theta) & 0 & -r\sin(\theta) \end{pmatrix}$$



Figuur I.14: Sferische coördinaten

en voor de Jacobiaan $\det(J)$ krijgt men in dit geval

$$\begin{aligned}\det(J) &= -\,r^2\cos^2(\varphi)\sin^3(\theta) - r^2\sin^2(\varphi)\cos^2(\theta)\sin(\theta)\\ &\quad -\,r^2\cos^2(\varphi)\cos^2(\theta)\sin(\theta) - r^2\sin^2(\varphi)\sin^3(\theta)\\ &= -\,r^2\sin^3(\theta) - r^2\cos^2(\theta)\sin(\theta)\\ &= -\,r^2\sin(\theta).\end{aligned}$$

Omdat $\sin(\theta) > 0$ is $|\det(J)| = r^2\sin(\theta)$ en dus

$$\int f(x,y,z)\;dx\,dy\,dz = \int g(r,\varphi,\theta)\,r^2\sin(\theta)\;dr\,d\varphi\,d\theta,$$

waarbij $g(r,\varphi,\theta) := f(r\cos(\varphi)\sin(\theta), r\sin(\varphi)\sin(\theta), r\cos(\theta))$.

Een alternatieve versie van de sferische coördinaten gebruikt voor de hoek θ in plaats van de hoek tussen (x,y,z) en de z-as de hoek tussen (x,y,z) en het $x-y$-vlak. Dit geeft

$$x = r\cos(\varphi)\cos(\theta), \quad y = r\sin(\varphi)\cos(\theta), \quad z = r\sin(\theta),$$

waarbij $r > 0$, $\varphi \in [0, 2\pi)$, $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.
In dit geval wordt $dx\,dy\,dz = r^2\cos(\theta)\;dr\,d\varphi\,d\theta$.

De eenvoudigste toepassing van sferische coördinaten is het bepalen van het volume V van een kogel $B := B(0,R)$ van straal R. De functie $f(x,y,z)$ is in

dit geval $f(x,y,z) = 1$, dus hebben we

$$\begin{aligned} V &= \int_B 1\, dx\, dy\, dz = \int_0^{\pi} \int_0^{2\pi} \int_0^{R} r^2 \sin(\theta)\, dr\, d\varphi\, d\theta \\ &= \int_0^{\pi} \int_0^{2\pi} \frac{1}{3} R^3 \sin(\theta)\, d\varphi\, d\theta = \int_0^{\pi} \frac{2\pi}{3} R^3 \sin(\theta)\, d\theta = \frac{2\pi}{3} R^3 (-\cos(\theta))\Big|_0^{\pi} \\ &= \frac{4\pi}{3} R^3. \end{aligned}$$

4.5 Toepassingen

Oppervlaktes, volumes

Een belangrijke toepassing voor integralen over functies van meerdere variabelen is het bepalen van oppervlaktes en volumes. Voorbeelden hiervan hebben we al gezien, namelijk de oppervlakte van een cirkel en het volume van een kogel. De manier van aanpak is steeds dezelfde: Men integreert de constante functie die overal de waarde 1 heeft over het gebied waarvan men de oppervlakte of het volume wil bepalen. De kunst ligt hierbij meestal niet zo zeer in de integratie, maar in het beschrijven van het gebied. Soms is het mogelijk een gebied in meerdere delen te splitsen die als normaalgebieden te beschrijven zijn en vaak helpt een geschikte keuze van nieuwe coördinaten.

Vaak is het ook handig één van de standaard coördinatentransformaties te combineren met een verdere substitutie. Een voorbeeld hiervoor is het berekenen van de oppervlakte van een ellips.

Zij E een ellips rond het nulpunt $(0,0)$ met hoofdassen van lengte a en b in de richtingen van de x-as en de y-as, dan wordt E beschreven door:

$$E = \{(x,y) \in \mathbb{R}^2 \mid \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1\}.$$

Door de transformatie op nieuwe coördinaten u, v met $x = au$ en $y = bv$ wordt E op de eenheidscirkel $B(0,1)$ getransformeerd, want $\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1 \Leftrightarrow \frac{(au)^2}{a^2} + \frac{(bv)^2}{b^2} = u^2 + v^2 \leq 1$. De Jacobi matrix voor deze substitutie is heel eenvoudig, we hebben $J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ en dus $\det(J) = ab$. Hieruit volgt dat

$$\int_E 1\, dx\, dy = \int_{B(0,1)} ab\, du\, dv = ab \int_{B(0,1)} 1\, du\, dv = \pi ab.$$

Zwaartepunten

Een verdere toepassing van meervoudige integralen is het berekenen van zwaartepunten van objecten. Het zwaartepunt is een soort gemiddelde van het object en in drie dimensies kunnen de coördinaten (x_s, y_s, z_s) van het zwaartepunt van een object B met volume V berekenen als $x_s = \frac{1}{V} \int_B x\ dx\, dy\, dz$ $y_s = \frac{1}{V} \int_B y\ dx\, dy\, dz$ $z_s = \frac{1}{V} \int_B z\ dx\, dy\, dz$ waarbij we veronderstellen dat de dichtheid van het object overal hetzelfde is.

Maar we kunnen het zwaartepunt ook berekenen als de dichtheid niet constant is, maar een functie $\rho(x,y,z)$. De massa van B berekenen we door $M = \int_B \rho(x,y,z)\ dx\,dy\,dz$, dus geeft de functie $\frac{1}{M}\int_B \rho(x,y,z)\ dx\,dy\,dz$ de verdeling van de massa over B aan. Deze verdelingsfunctie moeten we nu gewoon in de integralen invullen en krijgen zo $x_s = \frac{1}{M}\int_B x\rho(x,y,z)\ dx\,dy\,dz$ $y_s = \frac{1}{M}\int_B y\rho(x,y,z)\ dx\,dy\,dz$ $z_s = \frac{1}{M}\int_B z\rho(x,y,z)\ dx\,dy\,dz$ Het speciaal geval voor constante dichtheid $\rho(x,y,z) = \rho$ volgt hieruit met $V = \int_B dx\,dy\,dz$, omdat dat $M = \int_B \rho\ dx\,dy\,dz = \rho \cdot V$, dus $\frac{\rho}{M} = \frac{1}{V}$.

> In het kader van de kansrekening is het taalgebruik is anders en het zwaartepunt is een oude bekende. Als $f(x,y)$ of $f(x,y,z)$ de dichtheidsfunctie van een meerdimensionale kansverdeling is, heet het zwaartepunt namelijk de *verwachtingswaarde* van de kansverdeling. De dichtheidsfunctie speelt precies de rol van de functie $\frac{1}{M}\rho(x,y,z)$ voor de verdeling van de massa, want de integraal over het hele gebied is gelijk aan 1.

Als voorbeeld berekenen we het zwaartepunt van een halfkogel H met straal R rond het nulpunt $(0,0,0)$ die boven het $x-y$-vlak ligt. We gaan van constante dichtheid uit. Een halfkogel beschrijven we het makkelijkste met sferische coördinaten, we moeten alleen maar over de hoek θ tussen (x,y,z) en de z-as nadenken. Bij een volle kogel loopt die van 0 tot π en voor punten in het $x-y$-vlak is $\theta = \frac{\pi}{2}$, dus loopt θ nu van 0 tot $\frac{\pi}{2}$. Uit symmetrie redenen is het duidelijk dat het zwaartepunt op de z-as moet liggen, daarom hoeven we alleen maar de z-coördinaat uit te rekenen. Omdat een volle kogel van straal R het volume $\frac{4}{3}\pi R^3$ heeft, heeft H het volume $V = \frac{2}{3}\pi R^3$. Er geldt:

$$\begin{aligned}
z_s &= \frac{1}{V}\int_H z\ dx\,dy\,dz = \frac{1}{V}\int_0^R\int_0^{2\pi}\int_0^{\frac{\pi}{2}} r\cos(\theta)\,r^2\sin(\theta)\ d\theta\,d\varphi\,dr \\
&= \frac{1}{V}\int_0^R\int_0^{2\pi} r^3(\frac{1}{2}\sin^2(\theta)\big|_0^{\frac{\pi}{2}})\ d\varphi\,dr = \frac{1}{V}\int_0^R\int_0^{2\pi}\frac{1}{2}r^3\ d\varphi\,dr \\
&= \frac{1}{V}\int_0^R \pi r^3\ dr = \frac{1}{V}\frac{1}{4}\pi r^4\big|_0^R = \frac{\pi}{4V}R^4 = \frac{\pi}{4\frac{2}{3}\pi R^3}R^4 = \frac{3}{8}R.
\end{aligned}$$

Hetzelfde voorbeeld kunnen we ook in cilindercoördinaten uitwerken. In het $x-y$-vlak loopt de straal r dan van 0 tot R, de hoek φ van 0 tot 2π en de z-variabel loopt van 0 tot $\sqrt{R^2-r^2}$. Hiermee krijgen we:

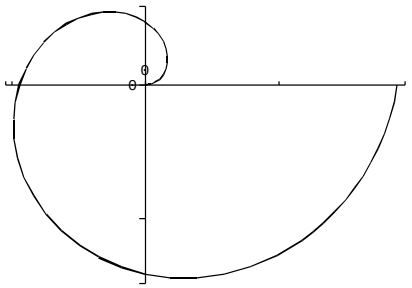
$$\begin{aligned}
z_s &= \frac{1}{V}\int_H z\ dx\,dy\,dz = \frac{1}{V}\int_0^R\int_0^{2\pi}\int_0^{\sqrt{R^2-r^2}} z\,r\ dz\,d\varphi\,dr \\
&= \frac{1}{V}\int_0^R\int_0^{2\pi}(\frac{1}{2}z^2\big|_0^{\sqrt{R^2-r^2}})\,r\ d\varphi\,dr = \frac{1}{V}\int_0^R\int_0^{2\pi}\frac{1}{2}(R^2-r^2)r\ d\varphi\,dr \\
&= \frac{1}{V}\int_0^R \pi(R^2-r^2)r\ dr = \frac{\pi}{V}\int_0^R (rR^2-r^3)\ dr = \frac{\pi}{V}(\frac{1}{2}r^2R^2-\frac{1}{4}r^4)\big|_0^R \\
&= \frac{\pi}{V}\frac{1}{4}R^4 = \frac{3}{8}R.
\end{aligned}$$

Belangrijke begrippen in deze les

- integratie van functies van meerdere variabelen
- geïtereerde integratie
- integratie over rechthoek gebieden
- integratie over normaalgebieden
- Jacobi matrix, Jacobiaan
- substitutie voor functies van meerdere variabelen
- coördinatentransformatie
- poolcoördinaten
- cilindercoördinaten, sferische coördinaten

Opgaven

33. Bereken de volgende 2-dimensionale integralen:
    (i) $\int_B (xy + y^2)\, dA$ met $B = [0,1] \times [0,1]$,
    (ii) $\int_B \sin(x+y)\, dA$ met $B = [0, \frac{\pi}{2}] \times [0, \frac{\pi}{2}]$,
    (iii) $\int_B (x + y^2)\, dA$, waarbij B de driehoek met hoekpunten $(0,0)$, $(1,0)$ en $(0,1)$ is.
    (iv) $\int_B (x^2 + y^2)\, dA$, waarbij B de driehoek met hoekpunten $(0,0)$, $(1,0)$ en $(\frac{1}{2}, \frac{1}{2})$ is.
34. Zij B het gebied tussen de grafieken van $\varphi_1(x) := x^3$ en $\varphi_2(x) := x^2$. Bereken de integralen $\int_B x\, dA$ en $\int_B y\, dA$.
35. Bereken de integraal $\int_B \frac{\sin(x)}{x}\, dA$ voor de driehoek B met hoekpunten $(0,0)$, $(1,0)$ en $(1,1)$. Let op dat hierbij de volgorde van de integraties een rol speelt, want de integraal $\int \frac{\sin(x)}{x}\, dx$ laat zich niet zonder integraal teken schrijven.
36. Beschrijf het gebied $B := \{(x,y) \in \mathbb{R}^2 \mid 3y \leq x \leq 3, 0 \leq y \leq 1\}$. en bereken de integraal $\int_B e^{x^2}\, dA$. Dit lukt helaas alleen maar voor een van de twee mogelijke volgordes van integratie.
37. Bereken de oppervlakte van het gebied tussen de archimedische spiraal (zie Figuur I.15) gegeven door $r = a\varphi$, $0 \leq \varphi \leq 2\pi$ en de x-as. Merk op dat de spiraal in poolcoördinaten aangegeven is.
38. Bereken op het gebied $B := \{(x,y) \in \mathbb{R}^2 \mid 0 \leq x \leq 1, x^2 \leq y \leq \sqrt{x}\}$ de integraal $\int_B (x^2 + y)\, dx\, dy$.
39. Bepaal de integraal $\int_0^\infty \int_0^\infty e^{-(x^2+y^2)} x^2\, dA$ met behulp van een transformatie op poolcoördinaten.



Figuur I.15: Archimedische spiraal

40. Bereken het volume van de (onregelmatige) tetraëder die begrensd is door de drie coördinaatvlakken $x = 0$, $y = 0$ en $z = 0$ en het vlak met $z = 2 - 2x - y$.

41. Een cirkelvormige boor van straal R snijdt uit een kogel van straal $2R$ een cilinder langs de z-as uit. Wat is het volume van de cilinder?

42. Een halfkogel H van straal R die op het $x - y$-vlak ligt heeft een niet constante dichtheidsfunctie, de dichtheid hangt namelijk af van de afstand van het grondvlak: $\rho(x, y, z) = az$ voor een $a > 0$. Bereken het zwaartepunt van de halfkogel.

Les 5 Complexe getallen

Iedereen weet, dat kwadraten van getallen positieve getallen zijn. Dat is vaak erg praktisch, we weten bijvoorbeeld dat de functie $f(x) := x^2+1$ steeds positief is en in het bijzonder geen nulpunten heeft. Daarom is bijvoorbeeld ook de functie $f(x) := \frac{1}{x^2+1}$ voor alle waarden van x gedefinieerd, omdat de noemer nooit 0 wordt.

Aan de andere kant is het feit, dat kwadraten positief zijn, ook een bron van frustratie, we kunnen namelijk vergelijkingen van de vorm $X^2 = a$ voor $a < 0$ niet oplossen.

Nu is het een typische eigenschap van wiskundigen, dat ze in een voor gewone mensen hopeloze situatie (een situatie zonder oplossing) toch vooruit gaan: ze definiëren gewoon iets, waarmee ze verder kunnen.

> Er zijn mensen die beweren dat wiskundigen mensen zijn die geen weg weten met de reële wereld en zich daarom hun eigen wereld definiëren waarin ze zich thuis voelen.

In sommige gevallen zijn de nieuw gedefinieerde objecten misschien niet zo erg nuttig, maar in het geval van de oplossingen van kwadratische vergelijkingen bleek de definitie van de *complexe getallen* een echt succesverhaal te zijn.

5.1 Constructie van de complexe getallen

Het idee achter de *complexe getallen* is dat we een oplossing van de vergelijking $X^2 = -1$ definiëren en kijken wat er gebeurt als we deze oplossing aan de reële getallen $\mathbb{R}$ toevoegen.

Definitie: We noteren de (symbolische) oplossing van de kwadratische vergelijking $X^2 = -1$ met de letter i en noemen i de *imaginaire eenheid*. Voor de imaginaire eenheid i geldt dus dat $i^2 = -1$.

Wat betekent het nu dat we i aan de reële getallen *toevoegen*? We willen zeker dat we i met een willekeurig reëel getal y kunnen vermenigvuldigen, dit geeft getallen van de vorm $i \cdot y$.

Het aardige is dat we hiermee al uit ieder reëel getal de wortel kunnen trekken, want voor $a \geq 0$ konden we dit al eerder en voor $a < 0$ is $-a > 0$, dus bestaat er een $b \in \mathbb{R}$ met $b^2 = -a$ en we hebben $(i \cdot b)^2 = i^2 \cdot b^2 = (-1) \cdot (-a) = a$.

Maar we willen getallen natuurlijk ook optellen, daardoor krijgen we alle getallen van de vorm $x + i \cdot y$ met $x, y \in \mathbb{R}$. Het leuke aan dit verhaal is, dat dit al voldoende is, d.w.z. dat we geen verdere getallen meer nodig hebben om goed met i te kunnen rekenen:

- Het *optellen* van getallen van de vorm $x + i \cdot y$ gebeurt componentsgewijs, waarbij we x en y als de componenten van het getal beschouwen, dus

$$(x_1 + i \cdot y_1) + (x_2 + i \cdot y_2) = (x_1 + x_2) + i \cdot (y_1 + y_2).$$

- Voor het *vermenigvuldigen* moeten we het product van twee getallen gewoon uitwerken:

$$\begin{aligned}(x_1 + i \cdot y_1)(x_2 + i \cdot y_2) &= x_1x_2 + i \cdot (x_1y_2 + y_1x_2) + i^2 \cdot y_1y_2 \\ &= (x_1x_2 - y_1y_2) + i \cdot (x_1y_2 + y_1x_2).\end{aligned}$$

Voorbeeld: We hebben $(1 + i \cdot 2) + (3 - i) = (1 + 3) + i \cdot (2 - 1) = 4 + i$ en $(1 + i \cdot 2)(3 - i) = (3 - (-2)) + i \cdot (-1 + 6) = 5 + i \cdot 5$.

Definitie: De verzameling

$$\mathbb{C} := \{x + i \cdot y \mid x, y \in \mathbb{R}\}$$

met de bewerkingen

$$\begin{aligned}(x_1 + i \cdot y_1) + (x_2 + i \cdot y_2) &= (x_1 + x_2) + i \cdot (y_1 + y_2) \\ (x_1 + i \cdot y_1)(x_2 + i \cdot y_2) &= (x_1x_2 - y_1y_2) + i \cdot (x_1y_2 + y_1x_2).\end{aligned}$$

heet het *lichaam* $\mathbb{C}$ *der complexe getallen.*

Notatie: Net zo goed als $1 + i \cdot 2$ kunnen we natuurlijk ook $1 + 2i$ schrijven en in feite is $x + yi$ in het algemeen de gebruikelijkere schrijfwijze dan $x + i \cdot y$.

We mogen de verzameling $\mathbb{C}$ een *lichaam* noemen, omdat optellen en vermenigvuldigen commutatief ($a + b = b + a$ en $ab = ba$) en associatief ($(a + b) + c = a + (b + c)$ en $(ab)c = a(bc)$) zijn en omdat de vermenigvuldiging distributief over het optellen is ($a(b + c) = ab + ac$ en $(a + b)c = ac + bc$). Deze eigenschappen erven de complexe getallen gewoon van de reële getallen.

> **Let op:** In de wiskunde staat steeds het symbool i voor de imaginaire eenheid. Maar omdat in de natuurkunde en elektrotechniek traditioneel de letter I (en vroeger ook i) voor de *stroomsterkte* gebruikt wordt, wordt in deze disciplines meestal j voor de imaginaire eenheid geschreven.

Omdat we de complexe getallen verkregen hebben door i aan de reële getallen toe te voegen, zijn de reële getallen in de complexe getallen bevat, namelijk als de getallen van de vorm $x + i \cdot 0$ met $x \in \mathbb{R}$. Aan de andere kant noemt men de getallen van de vorm $i \cdot y$ met $y \in \mathbb{R}$ *zuiver imaginair.*

We hebben al gezien dat een complex getal $z \in \mathbb{C}$ eenduidig door twee reële getallen beschreven wordt, namelijk door $x, y \in \mathbb{R}$ zo dat $z = x + i \cdot y$.

De eerste component x van $z = x + i \cdot y$ heet het *reële deel* van z, genoteerd met $x = \Re(z)$ en de tweede component y heet het *imaginaire deel* van z, genoteerd met $y = \Im(z)$. Er geldt dus:

$$z = \Re(z) + i \cdot \Im(z).$$

> Het is gebruikelijk dat complexe getallen of complexe variabelen z heten, terwijl reële getallen x en y heten. Dit is natuurlijk geen garantie, maar als je de letter z in een formule tegen komt, is dit een sterk signaal dat je het misschien met complexe getallen te maken hebt.

Opdracht 19 *Schrijf de volgende complexe getallen in de standaard vorm $z = x + yi$:*

(i) $(1 + 2i) - 3(5 - 2i)$;

(ii) $(4 - 3i)(8 + i) + 5 - i$;

(iii) $(2 + i)^2$.

5.2 Oplossen van vergelijkingen

We hebben boven al gezien, dat we met behulp van de getallen $i \cdot y$ uit elk reëel getal de wortel kunnen trekken. Dit betekent, dat elke kwadratische veelterm $f(x) = ax^2 + bx + c$ een nulpunt heeft, want de *abc*-formule

$$x_{1,2} = -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a^2} - \frac{c}{a}}$$

geeft de nulpunten expliciet aan en we hoeven alleen maar de wortel uit het reële getal $\frac{b^2}{4a^2} - \frac{c}{a}$ te trekken.

Maar de situatie is nog veel beter, we kunnen namelijk zelfs uit een willekeurig *complex getal* de wortel trekken:

Gezocht is een complex getal $z = x + i \cdot y$ met $z^2 = a + i \cdot b$ voor een gegeven complex getal $a + i \cdot b$. Uit $z^2 = (x^2 - y^2) + i \cdot (xy + yx) = (x^2 - y^2) + i \cdot 2xy$ volgt $x^2 - y^2 = a$ en $2xy = b$. Hieruit berekenen we $a^2 + b^2 = x^4 - 2x^2y^2 + y^4 + 4x^2y^2 = x^4 + 2x^2y^2 + y^4 = (x^2 + y^2)^2$, dus hebben we $x^2 + y^2 = \sqrt{a^2 + b^2}$ (merk op dat $a^2 + b^2$ positief is).

Door de vergelijkingen $x^2 + y^2 = \sqrt{a^2 + b^2}$ en $x^2 - y^2 = a$ bij elkaar op te tellen en van elkaar af te trekken krijgen we

$$x^2 = \frac{1}{2}\sqrt{a^2 + b^2} + \frac{1}{2}a \quad \text{en} \quad y^2 = \frac{1}{2}\sqrt{a^2 + b^2} - \frac{1}{2}a$$

en omdat $\sqrt{a^2 + b^2} \geq |a|$ hebben deze vergelijkingen reële oplossingen x en y.

We moeten wel opletten of we voor x en y de positieve of de negatieve wortel kiezen, want er moet gelden dat $2xy = b$. Als $b \geq 0$ moeten we bij x en y hetzelfde teken kiezen (beide positief of beide negatief), als $b < 0$ moeten x en y verschillende tekens hebben. Het zal geen verrassing zijn dat we in ieder geval twee oplossingen vinden, want met $z^2 = a + bi$ geldt ook $(-z)^2 = a + bi$.

Uit de *abc*-formule volgt nu rechtstreeks dat elke kwadratische veelterm met coëfficiënten in $\mathbb{C}$ ook een nulpunt in $\mathbb{C}$ heeft, of anders gezegd, dat elke kwadratische vergelijking een oplossing in $\mathbb{C}$ heeft. Maar er geldt een veel sterker resultaat, namelijk de

Hoofdstelling van de algebra (I): Elke veelterm met coëfficiënten in $\mathbb{C}$ heeft een nulpunt in $\mathbb{C}$.

Als een veelterm $f(z)$ een nulpunt a_1 heeft, dan kunnen we (met behulp van een staartdeling) $f(z)$ schrijven als $f(z) = (z - a_1)g(z)$ waarbij $g(z)$ een veelterm van lagere graad is. Maar ook $g(z)$ heeft volgens de hoofdstelling van de algebra een nulpunt a_2, en dus kunnen we doorgaan en $f(z)$ schrijven als $f(z) = (z - a_1)(z - a_2)h(z)$ waarbij de graad van $h(z)$ al om 2 kleiner is dan die van $f(z)$.

Uiteindelijk kunnen we een veelterm $f(z) = c_n z^n + c_{n-1} z^{n-1} + c_1 z + c_0$ op deze manier schrijven als $f(z) = c_n(z - a_1)(z - a_2) \dots (z - a_n)$, waarbij de a_i de (niet noodzakelijk verschillende) nulpunten van $f(z)$ zijn. Omdat $z - a$ een lineaire functie is hebben we zo de volgende variatie van de hoofdstelling van de algebra ingezien:

Hoofdstelling van de algebra (II): Elke veelterm met coëfficiënten in $\mathbb{C}$ laat zich (over $\mathbb{C}$) schrijven als een product van lineaire factoren.

Merk op: Over de reële getallen geldt slechts de zwakkere uitspraak: Elke veelterm met coëfficiënten in $\mathbb{R}$ laat zich (over $\mathbb{R}$) schrijven als een product van lineaire en kwadratische factoren.

Opdracht 20 *Bepaal de (complexe) nulpunten van de volgende veeltermen:*

(i) $z^2 - 2z + 5$;

(ii) $z^2 + \frac{1}{3}z + \frac{1}{2}$;

(iii) $z^3 + 2z^2 + 2z + 1$ *(een nulpunt is makkelijk te gokken).*

5.3 Meetkunde van de complexe getallen

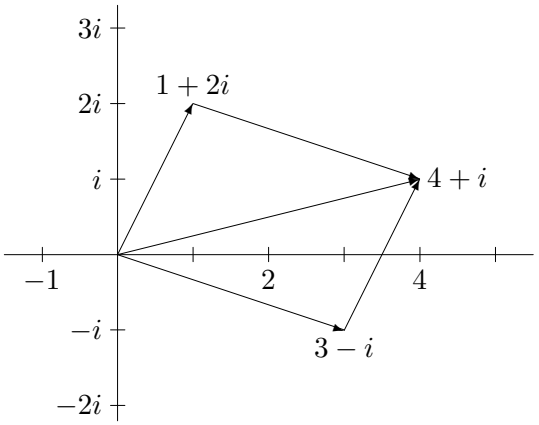
We hebben ons tot nu toe tot algebraïsche eigenschappen van de complexe getallen beperkt, maar een belangrijke rol spelen ook de meetkundige eigenschappen. We hebben gezien, dat een reëel getal via het reële en imaginaire deel met een paar van reële getallen correspondeert. Dit geeft een identificatie van de complexe getallen met het gewone 2-dimensionale vlak $\mathbb{R}^2$, het getal $z = x + i \cdot y$ correspondeert hierbij met het punt (x, y) en op grond van deze correspondentie spreekt men ook vaak van het *complexe vlak* in plaats van de complexe getallen.

We hebben al gezien dat het optellen van complexe getallen componentsgewijs voor het reële en imaginaire deel gebeurt. Maar dat is precies de manier hoe we vectoren optellen en daarom is het redelijk voor de hand liggend het getal $z = x + i \cdot y$ met de 2-dimensionale vector $\begin{pmatrix} x \\ y \end{pmatrix}$ te identificeren. Het voordeel ervan, bij complexe getallen aan vectoren in plaats van punten te denken, is dat we van vectoren weten hoe we ze optellen, terwijl we hiervoor bij punten toch stiekem weer vectoren zouden gebruiken.

In de taal van de lineaire algebra zeggen we, dat de complexe getallen $\mathbb{C}$ een 2-dimensionale $\mathbb{R}$-vectorruimte vormen, en de boven aangegeven correspondentie met $\mathbb{R}^2$ vinden we door de standaardbasis $\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$ van $\mathbb{R}^2$ met de basis $(1, i)$ van $\mathbb{C}$ te identificeren.

Het plaatje in Figuur I.16 geeft het optellen van de complexe getallen $1 + 2i$ en $3 - i$ in het complexe vlak weer.

Een voor de hand liggende vraag is nu natuurlijk, of ook de vermenigvuldiging van complexe getallen een mooie meetkundige interpretatie heeft. Dit is inderdaad het geval, maar het verhaal is iets ingewikkelder dan voor het optellen.



Figuur I.16: Optellen in het complexe vlak

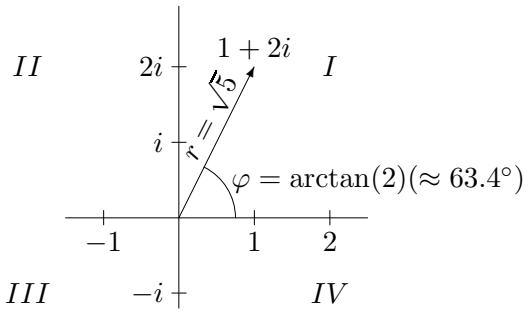
Poolcoördinaten

Om te beginnen, hebben we hiervoor en andere manier van beschrijving van punten in het 2-dimensionale vlak nodig, namelijk de *poolcoördinaten* die we ook bij de integratie van functies van meerdere veranderlijken al zijn tegengekomen.

Elk punt P in het vlak $\mathbb{R}^2$ kan behalve met zijn coördinaten (x, y) ook in de vorm (r, φ) aangegeven worden, waarbij r de afstand van het nulpunt is en φ de hoek tussen de x-as en de verbinding van het nulpunt met P (tegen de klok gemeten). Tussen de gewone *cartesische* coördinaten (x, y) en de poolcoördinaten (r, φ) bestaat het volgende verband:

$$x = r\cos(\varphi) \qquad y = r\sin(\varphi)$$

$$r = \sqrt{x^2 + y^2} \qquad \tan(\varphi) = \frac{y}{x}.$$



Figuur I.17: Poolcoördinaten

Merk op dat de relatie $\tan(\varphi) = \frac{y}{x}$ de hoek φ nog niet eenduidig vast legt, omdat $\tan(x)$ een periode van π en niet van 2π heeft. De omkeerfunctie $\arctan(x)$ heeft waarden tussen $-\frac{\pi}{2}$ en $\frac{\pi}{2}$ en is negatief voor $x < 0$ en positief voor $x > 0$. We moeten daarom voor de vier kwadranten tussen de assen van het complexe vlak aparte definities nemen:

$I:\ x > 0, y \geq 0$: $\varphi = \arctan(\frac{y}{x})$

$II:\ x<0, y\geq 0$: $\varphi = \arctan(\frac{y}{x}) + \pi$

$III:\ x<0, y<0$: $\varphi = \arctan(\frac{y}{x}) + \pi$

$IV:\ x>0, y<0$: $\varphi = \arctan(\frac{y}{x}) + 2\pi$

Voor (x, y) met $x = 0$ hebben we $\varphi = \frac{\pi}{2}$ als $y > 0$ en $\varphi = \frac{3\pi}{2}$ als $y < 0$. Voor het nulpunt zelf is de hoek niet gedefinieerd.

Opdracht 21 *Bepaal de poolcoördinaten van de volgende complexe getallen:*

$$2i, \quad 1-i, \quad \sqrt{3}+i, \quad -2-\sqrt{12}\,i.$$

We hadden eerder al gezegd dat de gewone coördinaten van een punt z in het complexe vlak het reële en imaginaire deel van z heten. Maar ook de poolcoördinaten r en φ van een complex getal z hebben een speciale betekenis:

Definitie: Zij $z = x + yi$ een complex getal met poolcoördinaten (r, φ):

(i) Het getal $r = \sqrt{x^2+y^2}$ heet de *absolute waarde* of *modulus* van z en wordt genoteerd met $|z|$. Dit is de gewone (euclidische) afstand van het nulpunt in $\mathbb{R}^2$ en komt voor $z \in \mathbb{R}$ overeen met de gewone absolute waarde van een reëel getal.

(ii) Het getal φ met $\tan(\varphi) = \frac{y}{x}$ heet het *argument* van z en wordt met $\arg(z)$ genoteerd.

Voor een complex getal z met $|z| = r$ en $\arg(z) = \varphi$ geldt dus

$$z = r \cdot (\cos(\varphi) + i \cdot \sin(\varphi)).$$

We hebben al gezien dat voor twee complexe getallen $z_1 = x_1 + i \cdot y_1$ en $z_2 = x_2 + i \cdot y_2$ het product $z_1 z_2$ gegeven is door $z_1 z_2 = (x_1x_2 - y_1y_2) + i \cdot (x_1y_2 + y_1x_2)$. Als we z_1 en z_2 in poolcoördinaten schrijven, dus $z_1 = (r_1, \varphi_1)$ en $z_2 = (r_2, \varphi_2)$, geeft dit volgens de boven aangegeven transformaties:

$$\begin{aligned}
z_1 \cdot z_2 &= r_1(\cos(\varphi_1) + i \cdot \sin(\varphi_1)) \cdot r_2(\cos(\varphi_2) + i \cdot \sin(\varphi_2)) \\
&= r_1 \cos(\varphi_1) r_2 \cos(\varphi_2) - r_1 \sin(\varphi_1) r_2 \sin(\varphi_2) \\
&\quad + i \cdot (r_1 \cos(\varphi_1) r_2 \sin(\varphi_2) + r_1 \sin(\varphi_1) r_2 \cos(\varphi_2)) \\
&= r_1 r_2 (\cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2)) \\
&\quad + i \cdot r_1 r_2 (\cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2)) \\
&= r_1 r_2 \cos(\varphi_1 + \varphi_2) + i \cdot r_1 r_2 \sin(\varphi_1 + \varphi_2).
\end{aligned}$$

In de laatste stap hebben we een optelheorema toegepast dat we in Wiskunde 1 al een keer zijn tegengekomen. Hier is een korte herinnering: Een rotatie in het 2-dimensionale vlak om een hoek van φ beschrijven we met betrekking tot de standaardbasis $\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right)$ door de matrix $\begin{pmatrix}\cos(\varphi) & -\sin(\varphi)\\ \sin(\varphi) & \cos(\varphi)\end{pmatrix}$. Maar een rotatie om $\varphi_1 + \varphi_2$ kunnen

we ook zien als de samenstelling van eerst een rotatie om φ_1 en vervolgens een rotatie om φ_2. De matrix van de samenstelling van twee rotaties is het product van de matrices van de enkele rotaties. Dit geeft de matrix vergelijking

$$\begin{pmatrix} \cos(\varphi_1+\varphi_2) & -\sin(\varphi_1+\varphi_2) \\ \sin(\varphi_1+\varphi_2) & \cos(\varphi_1+\varphi_2) \end{pmatrix} = \begin{pmatrix} \cos(\varphi_1) & -\sin(\varphi_1) \\ \sin(\varphi_1) & \cos(\varphi_1) \end{pmatrix} \cdot \begin{pmatrix} \cos(\varphi_2) & -\sin(\varphi_2) \\ \sin(\varphi_2) & \cos(\varphi_2) \end{pmatrix}$$

$$= \begin{pmatrix} \cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\cos(\varphi_2) & -\cos(\varphi_1)\sin(\varphi_2) - \sin(\varphi_1)\cos(\varphi_2) \\ \sin(\varphi_1)\cos(\varphi_2) + \cos(\varphi_1)\sin(\varphi_2) & -\sin(\varphi_1)\cos(\varphi_2) + \cos(\varphi_1)\cos(\varphi_2) \end{pmatrix}$$

en een vergelijk van de matrixelementen geeft in het bijzonder:

$$\begin{aligned} \cos(\varphi_1+\varphi_2) &= \cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\cos(\varphi_2) \\ \sin(\varphi_1+\varphi_2) &= \cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2). \end{aligned}$$

De gevonden relatie

$$z_1 z_2 = r_1 r_2 \cos(\varphi_1+\varphi_2) + i\cdot r_1 r_2 \sin(\varphi_1+\varphi_2) = r_1 r_2(\cos(\varphi_1+\varphi_2) + i\cdot\sin(\varphi_1+\varphi_2))$$

betekent nu het volgende:

Voor complexe getallen z_1 met poolcoördinaten (r_1, φ_1) en z_2 met poolcoördinaten (r_2, φ_2) heeft het product $z_1 z_2$ de poolcoördinaten $(r_1 r_2, \varphi_1+\varphi_2)$, d.w.z. de absolute waarden r_1 en r_2 worden vermenigvuldigd en de argumenten φ_1 en φ_2 opgeteld.

Merk op: Twee complexe getallen worden met elkaar vermenigvuldigd door hun absolute waarden te vermenigvuldigen en hun argumenten bij elkaar op te tellen.

Meetkundig uitgedrukt vermenigvuldigen we een complex getal z_1 met een complex getal z_2 door z_1 met de lengte (absolute waarde) van z_2 te schalen en vervolgens om het argument van z_2 (tegen de klok) te draaien.

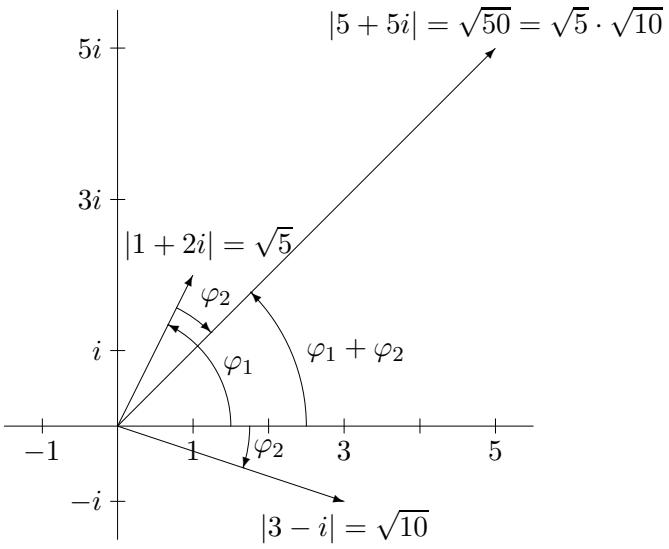
Voor het product $(1+2i)(3-i)$ geeft het plaatje in Figuur I.18 de meetkundige interpretatie van de vermenigvuldiging weer.

Worteltrekken

We komen even terug op het worteltrekken voor complexe getallen. We hadden gezien hoe we voor een gegeven complex getal $x+i\cdot y$ een complex getal $z = a+i\cdot b$ kunnen vinden met $z^2 = x + i\cdot y$. Maar met de meetkundige interpretatie van de vermenigvuldiging is dit eigenlijk veel makkelijker.

Een complex getal w met poolcoördinaten (r, φ) heeft de wortel z met poolcoördinaten $(\sqrt{r}, \frac{\varphi}{2})$. Merk op dat ook het getal met poolcoördinaten $(\sqrt{r}, \frac{\varphi}{2} + \pi)$ een wortel is, want omdat het argument steeds tussen 0 en 2π ligt, is $2\cdot(\varphi/2+\pi) = \varphi + 2\pi = \varphi$. Dit is geen verrassing, want voor een complex getal z met argument $\arg(z) = \varphi$ is $\arg(-z) = \varphi + \pi$ en natuurlijk weten we dat $z^2 = (-z)^2$, dus met z is ook $-z$ een wortel uit w.

Op dezelfde manier kunnen we ook n-de machtswortels trekken. Een complex getal w met $|w| = r$ en $\arg(w) = \varphi$ heeft als n-de machtswortel het getal



Figuur I.18: Vermenigvuldigen in het complexe vlak

$z = \sqrt[n]{r}(\cos(\frac{\varphi}{n}) + i\cdot\sin(\frac{\varphi}{n}))$, dus moeten we uit de absolute waarde de n-de wortel trekken en het argument door n delen. Ook hier zijn behalve van het getal z met $\arg(z) = \frac{\varphi}{n}$ de getallen met absolute waarde $\sqrt[n]{r}$ en argumenten $\frac{\varphi}{n} + \frac{2\pi k}{n}$ voor $k = 1, \ldots, n-1$ n-de machtswortels uit w, want bij het vermenigvuldigen met n worden al deze hoeken gelijk aan φ.

Een belangrijke toepassing van de meetkundige interpretatie van het vermenigvuldigen van complexe getallen is de *Regel van de Moivre*. Een complex getal z met absolute waarde 1 kunnen we schrijven als $z = \cos(\varphi) + i\cdot\sin(\varphi)$, waarbij $\varphi = \arg(z)$. Maar de n-de macht z^n kunnen we nu makkelijk berekenen, de absolute waarde is nog steeds 1 en het argument is het n-voud van het argument van z, dus $\arg(z^n) = n\arg(z) = n\varphi$. Dit betekent dat $z^n = \cos(n\varphi) + i\cdot\sin(n\varphi)$ en dit geeft de

Regel van de Moivre:

$$(\cos(\varphi) + i\cdot\sin(\varphi))^n = \cos(n\varphi) + i\cdot\sin(n\varphi).$$

Als toepassing hiervan kunnen we eenvoudig formules voor de *sinus* of *cosinus* van het dubbele of drievoud van een hoek afleiden, bijvoorbeeld:

$$\begin{aligned}\cos(2x) &= \Re(\cos(2x) + i\cdot\sin(2x)) = \Re((\cos(x) + i\cdot\sin(x))^2)\\ &= \Re(\cos^2(x) - \sin^2(x) + 2i\cdot\cos(x)\sin(x))\\ &= \cos^2(x) - \sin^2(x)\end{aligned}$$

$$\begin{aligned}\sin(3x) &= \Im(\cos(3x) + i\cdot\sin(3x)) = \Im((\cos(x) + i\cdot\sin(x))^3)\\ &= \Im(\cos^3(x) + 3i\cdot\cos^2(x)\sin(x) - 3\cos(x)\sin^2(x) - i\cdot\sin^3(x))\\ &= 3\cos^2(x)\sin(x) - \sin^3(x).\end{aligned}$$

We hebben gezien dat we met complexe getallen net zo goed als met reële getallen kunnen reken (ook al is de vermenigvuldiging iets ingewikkelder) en dat we vergelijkingen veel beter kunnen oplossen dan in $\mathbb{R}$. Maar er is ook een belangrijk nadeel van de complexe getallen tegenover de reële getallen: We kunnen van twee reële getallen steeds zeggen dat één van de twee groter is dan de andere (als ze niet gelijk zijn). We zeggen namelijk dat $x > y$ als $x - y > 0$ en voor elk getal $x \in \mathbb{R}$ geldt $x > 0$, $x = 0$ of $-x > 0$. Verder is voor twee positieve getallen $x, y > 0$ ook de som $x + y$ en het product xy positief.

Een ordening met deze eigenschappen kunnen we op $\mathbb{C}$ niet construeren, want als er een $z \in \mathbb{C}$ is met $z > 0$, dan is $z^2 > 0$. Maar voor $z \neq 0$ is of $z > 0$ of $-z > 0$ en dus is in elk geval $z^2 > 0$. Omdat we elk complex getal a in de vorm $a = z^2$ kunnen schrijven, zijn dus alle getallen $z \in \mathbb{C}$ positief. In het bijzonder is $1 > 0$ en $-1 > 0$ en dus $0 = -1 + 1 > 0$ (omdat -1 en 1 positief zijn) en dit is onmogelijk. De enige manier om de complexe getallen zo te ordenen dat sommen en producten van positieve getallen weer positief zijn, is de triviale ordeningen, waar alle getallen even groot zijn als 0, maar daar hebben niets aan.

5.4 Complexe conjugatie

Een belangrijke operatie op de complexe getallen is de *complexe conjugatie* die van een getal $z = x + i \cdot y$ het nieuwe getal

$$\overline{z} := x - i \cdot y$$

maakt, dat de *complex geconjugeerde van* z heet. Soms wordt de complex geconjugeerde ook met z^* in plaats van $\overline{z}$ genoteerd.

Omdat bij de complexe conjugatie het reële deel van een getal hetzelfde blijft en het imaginaire deel op zijn negatieve waarde gaat, is de complexe conjugatie gewoon de spiegeling in de x-as van het complexe vlak.

Men ziet rechtstreeks in dat z en $\overline{z}$ dezelfde absolute waarde $\sqrt{a^2 + b^2}$ hebben en dat het argument van $\overline{z}$ het negatieve van het argument van z is, dus $\arg(\overline{z}) = -\arg(z)$.

Verder geldt $z \cdot \overline{z} = |z|^2 \in \mathbb{R}$, want $(x + i \cdot y)(x - i \cdot y) = x^2 + y^2$. Dit geeft een handige methode, om complexe getallen te inverteren:

$$z^{-1} = \frac{1}{z} = \frac{\overline{z}}{z \cdot \overline{z}} = \frac{1}{|z|^2}\,\overline{z}$$

de inverse van z is dus de complex geconjugeerde gedeeld door het kwadraat van de absolute waarde. Dat de inverse van z een veelvoud van $\overline{z}$ moet zijn, hadden we natuurlijk ook al uit de argumenten kunnen aflezen, want uit $\arg(1) = 0$ volgt $\arg(\frac{1}{z}) = -\arg(z) = \arg(\overline{z})$.

We kunnen nu ook een willekeurige breuk van complexe getallen op standaardvorm brengen, want

$$\frac{z_1}{z_2} = \frac{z_1\overline{z_2}}{|z_2|^2} = \frac{1}{|z_2|^2}\,z_1\overline{z_2}.$$

Voorbeeld: Er geldt $\frac{2i}{1-i} = \frac{2i(1+i)}{2} = \frac{1}{2}(-2+2i) = -1+i$.

Met behulp van de complex geconjugeerde kunnen we ook reëel en imaginair deel van een getal z makkelijk uitdrukken:

$$\Re(z) = \frac{1}{2}(z+\overline{z}) \quad \text{en} \quad \Im(z) = \frac{1}{2i}(z-\overline{z}).$$

In het bijzonder is een getal $z \in \mathbb{C}$ een reëel getal, dan en slechts dan als $\overline{z} = z$.

Met behulp van de complexe conjugatie vinden we ook een belangrijke eigenschap van de complexe nulpunten van veeltermen met reële coëfficiënten. Stel $f(z) = c_n z^n + \ldots + c_1 z + c_0$ is een veelterm met $c_i \in \mathbb{R}$ en stel dat $a \in \mathbb{C}$ met $f(a) = 0$. Dan is natuurlijk ook $\overline{f(a)} = 0$, dus $\overline{c_n a^n} + \ldots + \overline{c_1 a} + \overline{c_0} = c_n \overline{a}^n + \ldots + c_1 \overline{a} + c_0 = 0$ en dus is ook $\overline{a}$ een nulpunt van $f(z)$. De niet-reële nulpunten van $f(z)$ komen dus in paren van complex geconjugeerden.

5.5 Machtsverheffen

We hebben inmiddels alle bewerkingen en operaties op de reële getallen kunnen uitbreiden tot de complexe getallen, met uitzondering van het machtsverheffen met complexe exponenten.

Om te beginnen moeten we zeker iets kunnen zeggen over $e^{i \cdot y}$ waarbij $y \in \mathbb{R}$. Als dit lukt, kunnen we ook voor $z = x + i \cdot y$ de e-macht definiëren, namelijk door $e^z = e^{x+i \cdot y} = e^x \cdot e^{i \cdot y}$. Uiteindelijk zullen we dan (net als voor de reële getallen) a^z definiëren door $a^z = e^{\log(a) z}$, maar zo ver zijn we nog niet.

In de volgende les zullen we de complexe exponentiële functie en logaritme nader bekijken die een zuivere motivatie voor de volgende definitie van $e^{i\varphi}$ geven, die bekend staat als

Formule van Euler:

$$e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi).$$

Volgens deze formule is het getal $e^{i\varphi}$ juist het complexe getal met absolute waarde 1 en argument φ. Als φ van 0 tot 2π loopt, loopt $e^{i\varphi}$ één keer rond de eenheidscirkel.

Zonder de complexe exponentiële functie kunnen we alvast twee redenen aangeven, die de Formule van Euler plausibel maken. Als we i (net als $\sqrt{2}$ of π) als een constante beschouwen, is de functie $f(\varphi) := e^{i\varphi}$ een functie van een reële veranderlijke die we kunnen afleiden, en dit geeft

$$\begin{aligned} f'(\varphi) &= (e^{i\varphi})' = (\cos(\varphi) + i \cdot \sin(\varphi))' \\ &= -\sin(\varphi) + i \cdot \cos(\varphi) = i \cdot (\cos(\varphi) + i \cdot \sin(\varphi)) = i \cdot e^{i\varphi} = i \cdot f(\varphi). \end{aligned}$$

Maar dit is precies wat we volgens de kettingregel van de exponentiële functie zouden verwachten.

Verder geldt volgens onze definitie dat

$$e^{i\varphi_1} \cdot e^{i\varphi_2} = e^{i(\varphi_1+\varphi_2)}$$

omdat we getallen op de eenheidscirkel vermenigvuldigen door hun argumenten op te tellen, dus lijkt onze definitie ook met deze eigenschappen van de reële exponentiële functie overeen te komen.

Met behulp van de relatie $e^{i\varphi} := \cos(\varphi) + i \cdot \sin(\varphi)$ en de symmetrieeigenschappen $\cos(-x) = \cos(x)$ en $\sin(-x) = -\sin(x)$ kunnen we nu $\sin(x)$ en $\cos(x)$ alleen maar met de exponentiële functie uitdrukken, want er geldt:
$e^{i\varphi} + e^{-i\varphi} = \cos(\varphi) + \cos(-\varphi) + i \cdot (\sin(\varphi) + \sin(-\varphi)) = 2\cos(\varphi)$,
$e^{i\varphi} - e^{-i\varphi} = \cos(\varphi) - \cos(-\varphi) + i \cdot (\sin(\varphi) - \sin(-\varphi)) = 2i\sin(\varphi)$, en dus

$$\cos(\varphi) = \frac{e^{i\varphi} + e^{-i\varphi}}{2} \text{ en } \sin(\varphi) = \frac{e^{i\varphi} - e^{-i\varphi}}{2i}.$$

We komen nu nog een keer op de regel van de Moivre terug, met onze nieuwe definitie ziet die er namelijk heel eenvoudig uit:

$$(e^{i\varphi})^n = e^{i(n\varphi)}.$$

Ook de optieltheorema's $\cos(\varphi_1 + \varphi_2) = \cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2)$ en $\sin(\varphi_1 + \varphi_2) = \cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2)$ kunnen we meteen uit de reële en imaginaire delen van $e^{i(\varphi_1+\varphi_2)} = e^{i\varphi_1} \cdot e^{i\varphi_2}$ aflezen.

Als toegift een beroemde formule, die de meest belangrijke constanten 0, 1, i, e en π in een relatie brengt:

$$e^{i\pi} + 1 = 0.$$

5.6 Toepassingen van de complexe getallen

Op grond van de samenhang $e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi)$ zijn complexe getallen in alle toepassingen belangrijk die met golven te maken hebben. Voorbeelden hiervoor zijn:

- Het berekenen van het overlappen van twee of meer golven (bijvoorbeeld watergolven, maar ook elektromagnetische golven).
- Kwantummechanica: een deeltje wordt door een golf-functie beschreven, waarvan de absolute waarde de kans aangeeft, het deeltje in een zeker gebied te vinden.
- Spraakherkenning: een spraaksignaal wordt beschreven door een som van *sinus*-functies voor verschillende frequenties, waarbij het patroon van frequenties (formanten) karakteristiek voor de klinkers is. Het bepalen van dit patroon uit een signaal wordt met behulp van de Fourieranalyse bepaald, die we later gaan behandelen.

- Beeldherkenning: een plaatje wordt gezien als een bron van lichtgolven, waarbij verschillende kleuren met verschillende frequenties corresponderen en de intensiteit met de amplitude van de golven.

Belangrijke begrippen in deze les

- complexe getallen
- reëel deel, imaginair deel
- poolcoördinaten, absolute waarde, argument
- complexe conjugatie
- relatie $e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi)$

Opgaven

43. Schrijf de volgende complexe getallen in de vorm $a + i \cdot b$ en in poolcoördinaten:

$$\text{(i) } (1 - i\sqrt{3})^2 \qquad \text{(ii) } \frac{1+i}{i-1} \qquad \text{(iii) } \frac{3+4i}{2-i}$$

Hoe kan men absolute waarde en argument van deze getallen bepalen, zonder de getallen eerst in de vorm $a + i \cdot b$ te brengen?

44. Teken een punt $z \in \mathbb{C}$ op de eenheidscirkel (d.w.z. met $|z| = 1$). Construeer de punten z^2, z^3, z^{-1}, $-z$, $\overline{z}$, $i \cdot z$, $-i \cdot z$. Ga in de figuur na dat $z + z^{-1}$ reëel is.

45. Bereken de (complexe) oplossingen van de vergelijking $z^2 + 3z + 4 = 0$.

46. Vind de oplossingen $z \in \mathbb{C}$ voor de volgende vergelijkingen:

(i) $z^3 = i$, (ii) $z^2 - 2z + 2 = 0$, (iii) $z^4 = -1$, (iv) $(3+4i)z^2 + 5z + (2-4i) = 0$.

Teken de wortels in het complexe vlak.

47. Beschrijf de volgende verzamelingen van complexe getallen in het complexe vlak:

(i) $z \in \mathbb{C}$ met $\Re(z^2) > 0$;

(ii) $z \in \mathbb{C}$ met $\Re\left(\dfrac{z+i}{z-2i}\right) = 1$.

48. Druk met behulp van de regel van de Moivre:

(i) $\cos(4\varphi)$ uit in $\cos(\varphi)$ en $\sin(\varphi)$,

(ii) $\sin(3\varphi)$ uit in $\sin(\varphi)$ (zonder $\cos(\varphi)$).

49. Maak een schets van de verzamelingen van complexe getallen z die aan de aangegeven voorwaarden voldoen:

(i) $|z| \leq 2$;

(ii) $|z - 2i| \leq 3$;

(iii) $|z - 3 + 4i| \leq 5$;

(iv) $\arg(z) = \frac{\pi}{3}$;

(v) $\pi \leq \arg(z) \leq \frac{7\pi}{4}$.

50. Zij $L_1 \subset \mathbb{C}$ de lijn met $\Re(z) = \Im(z)$ en $L_2 \subset \mathbb{C}$ de lijn met $\Im(z) = 1$. Wat zijn de beelden van deze lijnen onder de afbeelding $z \mapsto z^{-1}$ (d.w.z. de verzamelingen $\{z^{-1} \mid z \in L_1(L_2)\}$?

51. Welke baan doorloopt $w := \dfrac{z^2 - z + 1}{2z}$ als z de eenheidscirkel doorloopt (d.w.z. $z = e^{ix}$ met $x \in [0, 2\pi]$)?

52. Welke baan doorloopt $z := \dfrac{x - i}{x + i}$ als x langs de reële as loopt (in positieve richting)?

53. Beschrijf de baan die

$$z := 149597887\, e^{2\pi i \frac{t}{365.257}} + 384403\, e^{2\pi i \frac{t}{27.321}}$$

doorloopt, als t langs de positieve reële as loopt. Welk hemelsfenomeen wordt hierdoor weergegeven?

Hoe zou de baan veranderen, als de constante 384403 door 38440300 vervangen wordt?

Les 6 Complexe functies

Nadat we de complexe getallen hebben leren kennen, is het een voor de hand liggende vraag of hiervoor net als voor de reële getallen ook functies bestaan. Met een functie bedoelen we hierbij een *voorschrijft* die aan elk complex getal $z \in B$ uit een deelverzameling $B \subseteq \mathbb{C}$ een eenduidige waarde $f(z) \in \mathbb{C}$ toewijst. Het gebied $B \subseteq \mathbb{C}$ het dan het *domein* van de functie $f(z)$.

Bij reële functies hebben we veel over een functie $f(x)$ kunnen zeggen, door de grafiek $(x, f(x))$ de bekijken. Dit is bij complexe functies echter moeilijk, want voor het domein $\mathbb{C}$ (waar een functie op gedefinieerd is) hebben we al een 2-dimensionaal vlak nodig, en voor de functiewaarden ook nog eens een 2-dimensionaal vlak, zo dat we voor de grafiek een 4-dimensionaal plaatje nodig hebben.

Maar we kunnen wel een redelijk indruk van een complexe functie krijgen door de volgende methoden:

(1) Bekijk de reële en imaginaire delen van de functie apart. Dit betekent dat we een complexe functie $f(z) : \mathbb{C} \to \mathbb{C}$ opsplitsen in twee functies met reële waarden, namelijk

$$u(z) : \mathbb{C} \to \mathbb{R},\ z \mapsto \Re(f(z)) \quad \text{en} \quad v(z) : \mathbb{C} \to \mathbb{R},\ z \mapsto \Im(f(z)).$$

Als we nu z schrijven als $z = x+iy$ met $x, y \in \mathbb{R}$, kunnen we $u(z) = u(x, y)$ en $v(z) = v(x, y)$ opvatten als functies van de twee reële variabelen x en y met waarden in $\mathbb{R}$. Maar voor dit soort functies hebben we al eerder gezien dat we als grafiek een 3-dimensionaal plaatje krijgen, door de punten $(x, y, u(x, y))$ te bekijken.

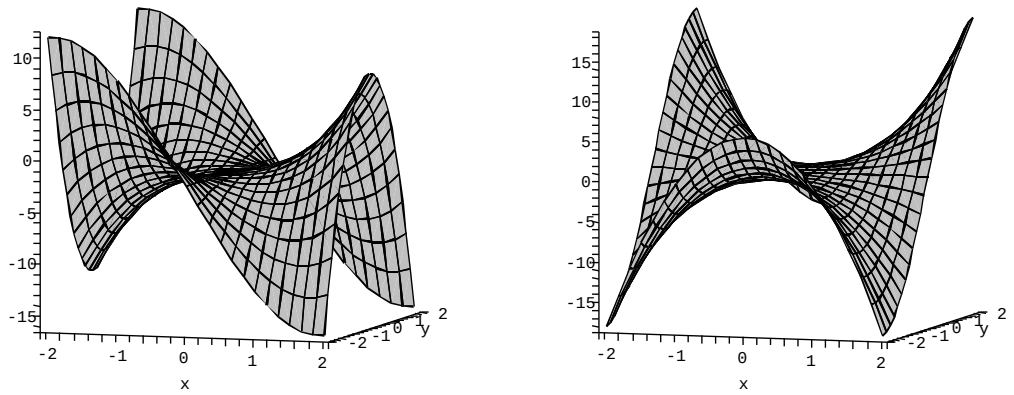
(2) We kunnen kijken hoe een functie $f(z)$ zekere lijnen afbeeldt, bijvoorbeeld de lijnen parallel met de x-as (dus de complexe getallen met hetzelfde imaginaire deel), de lijnen parallel met de y-as (de complexe getallen met hetzelfde reële deel), lijnen door de oorsprong (de complexe getallen met hetzelfde argument). We kunnen ook kijken wat met cirkels rond de oorsprong gebeurt, dus met complexe getallen met dezelfde absolute waarde.

Als voorbeeld laat Figuur I.19 de reële en imaginaire delen van de derdegraads veelterm $f(z) = z^3 + z - 2$ zien. Met $z = x + iy$ geldt $u(z) = u(x, y) = \Re(f(z)) = x^3 - 3xy^2 + x - 2$ en $v(z) = v(x, y) = \Im(f(z)) = -y^3 + 3x^2y + y$.

Opdracht 22 *Bepaal voor $z = x+iy$ de reële en imaginaire delen van de functie $f(z) = \frac{1}{1-z} = (1 - z)^{-1}$, d.w.z. bepaal reële functies $u(x, y)$ en $v(x, y)$ van twee variabelen zo dat $f(x + iy) = u(x, y) + iv(x, y)$.*

Omdat we weten hoe we complexe getallen optellen en vermenigvuldigen, hebben we met complexe functies die door een veelterm

$$f(z) = c_n z^n + c_{n-1} z^{n-1} + \ldots + c_1 z + c_0$$



Figuur I.19: Reëel en imaginair deel van $f(z) = z^3 + z - 2$

gegeven zijn helemaal geen moeite. Hierbij mogen de coëfficiënten natuurlijk ook zelfs complexe getallen zijn, dit geeft complexe functies zo als $f(z) := z^2 + 2iz + \sqrt{-3}$.

Maar natuurlijk kunnen we niet verwachten dat alle complexe functies veeltermfuncties zijn, de vraag is echter, hoe we aan andere complexe functies zouden kunnen komen. De oplossing hiervoor is verrassend eenvoudig. We hadden gezien dat we een reële functie $f(x)$ door Taylor polynomen kunnen benaderen en dat (in een kleine omgeving van een punt x_0) de functie gegeven is door de Taylor reeks. Dit brengt algemene functies op veeltermen en machtreeksen terug, en die kunnen we ook voor complexe getallen uitwerken.

We zullen ons dus in deze cursus beperken tot complexe functies $f(z)$ die door een machtreeks

$$f(z) = \sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + \ldots + a_n z^n + \ldots$$

gegeven zijn, waarbij we steeds veronderstellen dat de reeks voor de waarden z die we nodig hebben convergeert.

Een diepere analyse van complexe functies laat zien, dat de beperking tot complexe functies die door een machtreeks gegeven zijn helemaal geen sterke beperking is. Er laat zich namelijk aantonen dat een complexe functie die in een cirkel van straal R rond een punt z_0 *complex differentieerbaar* is (we komen hier in de Appendix voor deze les op terug) automatisch een convergente Taylor reeks rond z_0 heeft. Anders dan bij reële functies volgt namelijk uit het bestaan van de *eerste* afgeleide $f'(z)$ op een gebied B dat ook alle hogere afgeleiden $f^{(n)}(z)$ op B bestaan en continu zijn. De reden hiervoor is dat complexe differentieerbaarheid een veel sterkere eigenschap is dan reële differentieerbaarheid.

6.1 Complexe exponentiële functie

We hebben gezien dat we de (reële) exponentiële functie door de Taylor reeks $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ kunnen beschrijven, en omdat deze reeks voor alle x naar $\exp(x)$ convergeert mogen we zelfs zeggen, dat de twee gelijk zijn, dus dat

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Als we nu over een definitie voor de exponentiële functie op de complexe getallen nadenken, willen we natuurlijk dat die op de reële getallen met de reële exponentiële functie overeen komt. Het is nu een enigszins voor de hand liggende gedachte, de complexe exponentiële functie erdoor te definiëren, dat we complexe getallen in de Taylor reeks van de reële exponentiële functie invullen. Dan weten we in ieder geval dat voor reële getallen inderdaad de functiewaarden hetzelfde blijven.

Algemeen noemt men het invullen van waarden $z \in B \subseteq \mathbb{C}$ in een machtreeks van een functie $f(z)$ die op een kleiner domein $B_1 \subseteq B$ gedefinieerd is, het *voortzetten* van $f(z)$ op B. In ons geval zetten we de reële exponentiële functie van de reële lijn $\mathbb{R}$ op het hele complexe vlak voort.

> We hadden gezien dat er problemen met de Taylor reeks voor de functie $\exp(-\frac{1}{x^2})$ zijn, omdat de reeks de 0-functie is en (behalve voor $x = 0$) niet tegen de goede functiewaarden convergeert. Als we deze functie op de complexe getallen voortzetten, zien we dat we in het punt $z = 0$ helemaal geen continue voortzetting meer kunnen vinden (wat voor de reële getallen wel nog het geval was). Als we namelijk met $z = ix$ langs de imaginaire as lopen, hebben we $\exp(-\frac{1}{(ix)^2}) = \exp(\frac{1}{x^2})$ en dit gaat voor $x \to 0$ naar oneindig. Het feit dat we de functie in het complexe vlak niet continu in het punt 0 kunnen voortzetten hangt nauw samen met het feit dat de Taylor reeks op de reële getallen niet tegen de goede functie convergeert.

Het voortzetten van de reële exponentiële functie op het complexe vlak geeft (als definitie!) voor de *complexe exponentiële functie*:

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \dots$$

Om te rechtvaardigen dat deze definitie zinvol is, merken we het volgende op:

(1) Om te zien dat de reeks van $\exp(z)$ convergent is, is het voldoende dat de reeks over de absolute waarden van de termen convergent is, men zegt hiervoor dat de reeks *absoluut convergent* is. Maar

$$\sum_{n=0}^{\infty} \left| \frac{z^n}{n!} \right| = \sum_{n=0}^{\infty} \frac{|z|^n}{n!} = \exp(|z|),$$

dus volgt de absolute convergentie van de reeks voor $\exp(z)$ uit de convergentie van de Taylor reeks voor de reële exponentiële functie.

(2) We kunnen $\exp(z_1 + z_2)$ (in principe) uitrekenen door $z_1 + z_2$ in de reeks in te vullen, dus $\exp(z_1 + z_2) = \sum_{n=0}^{\infty} \frac{(z_1+z_2)^n}{n!}$. Aan de andere kant berekent men $\exp(z_1) \cdot \exp(z_2)$ door de reeksen $\sum_{n=0}^{\infty} \frac{z_1^n}{n!}$ en $\sum_{n=0}^{\infty} \frac{z_2^n}{n!}$ te vermenigvuldigen. Door de coëfficiënten van $z_1^j z_2^k$ in de uitdrukkingen voor $\exp(z_1 + z_2)$ en voor $\exp(z_1) \cdot \exp(z_2)$ te vergelijken, ziet men dat inderdaad

$$\exp(z_1 + z_2) = \exp(z_1) \cdot \exp(z_2),$$

net als we dat van de reële exponentiële functie gewend zijn. (In feite berust het bewijs dat $\exp(x + y) = \exp(x) \cdot \exp(y)$ voor $x, y \in \mathbb{R}$ precies op hetzelfde idee.)

(3) In Wiskunde 1 hadden we gezien dat de exponentiële functie gekarakteriseerd is door de eigenschappen dat

$$\exp(x)' = \exp(x) \quad \text{en} \quad \exp(0) = 1.$$

We zullen straks nader op het differentiëren van complexe functies ingaan, maar voor een functie die door een reeks gegeven is zou men hopen de afgeleide te vinden door de reeks termsgewijs af te leiden. Voor functies met een absoluut convergente reeks is dit inderdaad juist, dus hebben we voor de complexe exponentiële functie:

$$\exp(z)' = (\sum_{n=0}^{\infty} \frac{z^n}{n!})' = \sum_{n=1}^{\infty} n \cdot \frac{z^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \exp(z).$$

De complexe exponentiële functie heeft dus ook de eigenschappen die de reële exponentiële functie karakteriseren.

Met onze definitie van de complexe exponentiële functie kunnen we nu eindelijk ook de Formule van Euler

$$e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi)$$

uit de vorige les rechtvaardigen. Hiervoor vullen we $z = i\varphi$ in de reeks voor $\exp(z)$ in, waarbij we rekening ermee houden dat $i^2 = -1$, $i^3 = -i$ en $i^4 = 1$. We krijgen:

$$\begin{aligned}
e^{i\varphi} &= 1 + i \cdot \varphi - \frac{\varphi^2}{2!} - i \cdot \frac{\varphi^3}{3!} + \frac{\varphi^4}{4!} + i \cdot \frac{\varphi^5}{5!} - \frac{\varphi^6}{6!} - i \cdot \frac{\varphi^7}{7!} + \ldots \\
&= (1 - \frac{\varphi^2}{2!} + \frac{\varphi^4}{4!} - \frac{\varphi^6}{6!} + \ldots) + i \cdot (\varphi - \frac{\varphi^3}{3!} + \frac{\varphi^5}{5!} - \frac{\varphi^7}{7!} + \ldots) \\
&= \left(\sum_{n=0}^{\infty} (-1)^n \frac{\varphi^{2n}}{(2n)!} \right) + i \cdot \left(\sum_{n=0}^{\infty} (-1)^n \frac{\varphi^{2n+1}}{(2n+1)!} \right) \\
&= \cos(\varphi) + i \cdot \sin(\varphi).
\end{aligned}$$

In de laatste stap hebben we hierbij gebruik gemaakt van de (lang bekende) Taylor reeksen voor $\cos(x)$ en $\sin(x)$.

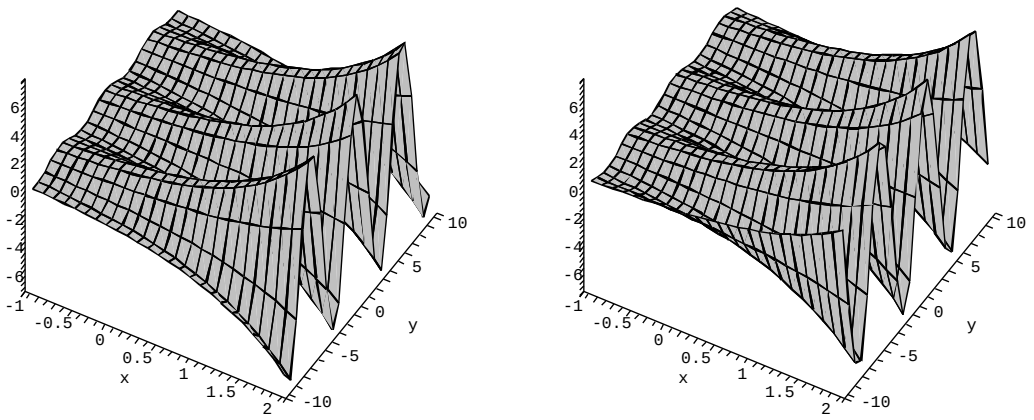
Om een beter idee van de complexe exponentiële functie te krijgen, is het verstandig naar de reële en imaginaire delen te kijken. Voor $z \in \mathbb{C}$ met $z = x + iy$, dus $x = \Re(z)$ en $y = \Im(z)$, geldt

$$\exp(x + iy) = \exp(x)\exp(iy) = \exp(x)(\cos(y) + i\sin(y))$$

waarbij we de formule van Euler $e^{iy} = \cos(y) + i\sin(y)$ hebben toegepast. Hieruit volgt:

$$\Re(\exp(x + iy)) = \exp(x)\cos(y) \quad \text{en} \quad \Im(\exp(x + iy)) = \exp(x)\sin(y).$$

In Figuur I.20 zijn de reële en imaginaire delen van de complexe exponentiële functie te zien.



Figuur I.20: Reëel en imaginair deel van $\exp(z)$

Zo zeer de complexe exponentiële functie in veel aspecten op de reële exponentiële functie lijkt, moeten we toch bij de overgang van de reële naar de complexe exponentiële functie afscheid nemen van sommige vertrouwde eigenschappen van de exponentiële functie. Een voorbeeld hiervan is dat de complexe exponentiële functie niet meer injectief is en daarom ook geen globale omkeerfunctie heeft:

Voor $x_1, x_2 \in \mathbb{R}$ geldt dat $e^{x_1} = e^{x_2}$ dan en slechts dan als $x_1 = x_2$, want de reële exponentiële functie is strikt stijgend en dus injectief. Er geldt namelijk: $e^{x_1} = e^{x_2} \Leftrightarrow e^{x_1}/e^{x_2} = e^{x_1 - x_2} = 1$, en dit is alleen maar het geval als $x_1 - x_2 = 0$, dus $x_1 = x_2$.

Als we hetzelfde argument op de complexe exponentiële functie toepassen, beleven we een kleine verrassing. Er geldt weer dat $e^{z_1} = e^{z_2} \Leftrightarrow e^{z_1 - z_2} = 1$. Maar voor een getal $z = x + iy \in \mathbb{C}$ geldt $e^z = e^x \cdot (\cos(y) + i\sin(y))$ en $|e^z| = e^x$, dus geldt $e^z = 1 \Leftrightarrow e^x = 1, \cos(y) = 1, \sin(y) = 0$ en dit is precies het geval voor $x = 0$ en $y = 2\pi k$ met $k \in \mathbb{Z}$. Er geldt dus

$$e^z = 1 \Longleftrightarrow z = 2\pi i k \text{ met } k \in \mathbb{Z}.$$

Merk op: Hieruit volgt in het bijzonder dat voor alle $z \in \mathbb{C}$ geldt dat

$$e^z = e^{z+2\pi i} = e^{z+2\pi i \cdot k} \text{ voor alle } k \in \mathbb{Z}$$

en we zeggen daarom dat de complexe exponentiële functie $2\pi i$-*periodiek* is.

Toepassing: Gedempte trilling

De kracht F die een massa m die een spiraalveer hangt, ervaart, is proportioneel met de afwijking x van de massa tegenover de evenwichtspositie, dus $F = -kx$. Het minteken betekent dat de kracht altijd naar de evenwichtspositie terugtrekt.

Een kracht F die op m werkt leidt tot een versnelling $x''(t)$ van de massa met $F = mx''(t)$. Zonder verdere invloed van buiten zou de tijdelijke beweging $x(t)$ van de massa dus voldoen aan de differentiaalvergelijking $mx''(t) = -kx(t)$. We hadden in Wiskunde 1 al gezien, dat de oplossingen van deze vergelijking sinus- en cosinusfuncties zijn, namelijk $x(t) = \sin(\omega t)$ of $x(t) = \cos(\omega t)$ (of een lineaire combinatie hiervan) met $\omega = \sqrt{\frac{k}{m}}$. De massa zou dus in een sinustrilling bewegen.

Als we nu ook met wrijving rekening willen houden, moeten we hierover een aanname maken. Meestal wordt verondersteld dat de wrijving proportioneel met de snelheid van de massa is, dus geldt voor de wrijvingskracht F_w dat $F_w = -\gamma x'(t)$. Ook hier houdt het minteken rekening ermee dat de wrijvingskracht de massa remt. In totaal geldt nu $mx''(t) = -kx - \gamma x'(t)$ of te wel

$$x''(t) + \beta x'(t) + \omega^2 x(t) = 0 \text{ met } \beta = \frac{\gamma}{m},\ \omega^2 = \frac{k}{m}.$$

Om een oplossing voor deze differentiaalvergelijking te vinden, proberen we een functie $x(t)$ van de vorm $x(t) = Ce^{at}$. Als we dit invullen, krijgen we $Ce^{at}(a^2 + \beta a + \omega^2) = 0$, dus moet a een oplossing van de kwadratische vergelijking $X^2 + \beta X + \omega^2 = 0$ zijn. Met de *abc*-formule (of anders) vinden we de oplossingen

$$a_{1,2} = -\frac{\beta}{2} \pm \frac{1}{2}\sqrt{\beta^2 - 4\omega^2}.$$

Als $\beta^2 > 4\omega^2$, dus $\beta > 2\omega$, zijn er twee reële oplossingen en we krijgen voor $x(t)$ een functie van de vorm

$$x(t) = C_1 e^{a_1 t} + C_2 e^{a_2 t} \text{ met } a_{1,2} = -\frac{\beta}{2} \pm \frac{1}{2}\sqrt{\beta^2 - 4\omega^2}.$$

Deze beweging beschrijft een steeds langzamer wordend terugvallen in de evenwichtspositie, waarbij het mogelijk is dat de beweging een keer door de evenwichtspositie doorheen gaat.

Als $\beta^2 < 4\omega^2$, dus als de wrijving zwakker is, zijn de oplossingen complex. We definiëren

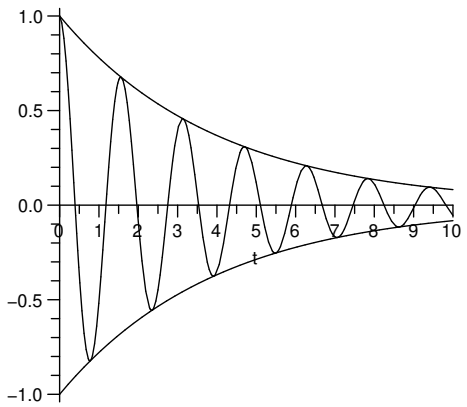
$$\overline{\omega} := \frac{1}{2}\sqrt{4\omega^2 - \beta^2} = \omega\sqrt{1 - \frac{\beta^2}{4\omega^2}} < \omega,$$

dan zijn de oplossingen $a_{1,2} = -\frac{\beta}{2} \pm i\,\overline{\omega}$. We krijgen dus $x(t)$ van de vorm

$$x(t) = C_1 e^{-\frac{\beta}{2}} e^{i\,\overline{\omega} t} + C_2 e^{-\frac{\beta}{2}} e^{-i\,\overline{\omega} t} = e^{-\frac{\beta}{2}}(c_1 \cos(\overline{\omega} t) + c_2 \sin(\overline{\omega} t)).$$

Deze functie geeft dus een sinus-vormige trilling met de frequentie $\overline{\omega} < \omega$ aan die gedempt is met de functie $e^{-\frac{\beta}{2}}$. De wrijving heeft dus naast het dempen van de trilling ook het effect dat de trilling om de factor $\sqrt{1 - \frac{\beta^2}{4\omega^2}}$ langzamer wordt dan in het vrije geval zonder wrijving.

In Figuur I.21 is de grafiek van een gedempte trilling $x(t) = e^{-\frac{\beta}{2}} \cos(\overline{\omega} t)$ te zien. Naast de functie $x(t)$ zijn ook de grensfuncties $\pm e^{-\frac{\beta}{2}}$ geschetst die de demping aangeven.



Figuur I.21: Gedempte trilling

6.2 Complexe sinus en cosinus functies

Nu dat we hebben gezien dat het voortzetten van de Taylor reeks van $\exp(x)$ op de complexe getallen een succes was, is het voor de hand liggend hetzelfde principe ook op de sinus en cosinus functies toe te passen. We definiëren dus de complexe sinus functie door

$$\sin(z) := \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} = z - \frac{1}{6}z^3 + \frac{1}{120}z^5 - \frac{1}{5040}z^7 + \ldots$$

en de complexe cosinus functie door

$$\cos(z) := \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} = 1 - \frac{1}{2}z^2 + \frac{1}{24}z^4 - \frac{1}{720}z^6 + \ldots$$

Als we nu nog een keer naar de berekening kijken waarmee we net hebben aangetoond dat $e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi)$, zien we dat we nergens iets speciaals

over φ verondersteld hebben. Als we dus precies hetzelfde opschrijven met z in plaats van φ waarbij $z \in \mathbb{C}$ een willekeurig complex getal is, vinden we de relatie

$$e^{iz} = \cos(z) + i\sin(z) \text{ voor alle } z \in \mathbb{C}.$$

Maar voor de boven aangegeven definities van $\cos(z)$ en $\sin(z)$ geldt net als op de reële getallen, dat

$$\cos(-z) = \cos(z) \quad \text{en} \quad \sin(-z) = -\sin(z)$$

want $(-z)^{2n} = (-1)^{2n} \cdot z^{2n} = z^{2n}$ en $(-z)^{2n+1} = (-z) \cdot (-z)^{2n} = (-z) \cdot z^{2n} = -z^{2n+1}$. Hieruit volgt

$$\begin{aligned} e^{iz} + e^{-iz} &= \cos(z) + \cos(-z) + i \cdot (\sin(z) + \sin(-z)) = 2\cos(z) \text{ en} \\ e^{iz} - e^{-iz} &= \cos(z) - \cos(-z) + i \cdot (\sin(z) - \sin(-z)) = 2i\sin(z) \end{aligned}$$

en we zien dus dat

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2} \quad \text{en} \quad \sin(z) = \frac{e^{iz} - e^{-iz}}{2i},$$

dus precies dezelfde relaties die we voor reële waarden van z al in de vorige les hadden verkregen.

> In de zuivere wiskunde wordt eigenlijk alleen maar de complexe exponentiële functie $\exp(z)$ door een reeks gedefinieerd, $\cos(z)$ en $\sin(z)$ worden vervolgens door de relaties $\cos(z) = \frac{e^{iz}+e^{-iz}}{2}$ en $\sin(z) = \frac{e^{iz}-e^{-iz}}{2i}$ *gedefinieerd*. Maar voor de toepassingen is uiteindelijk alleen maar de samenhang tussen deze functies belangrijk.

Voor de complexe cosinus en sinus is het uitwerken van de reële en imaginaire delen met iets meer rekenwerk verbonden. We zullen hierbij de hyperbolische functies $\sinh(x)$ en $\cosh(x)$ tegen komen, die we in Wiskunde 1 hebben leren kennen. Voor deze functies geldt:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \text{en} \quad \sinh(x) = \frac{e^x - e^{-x}}{2}.$$

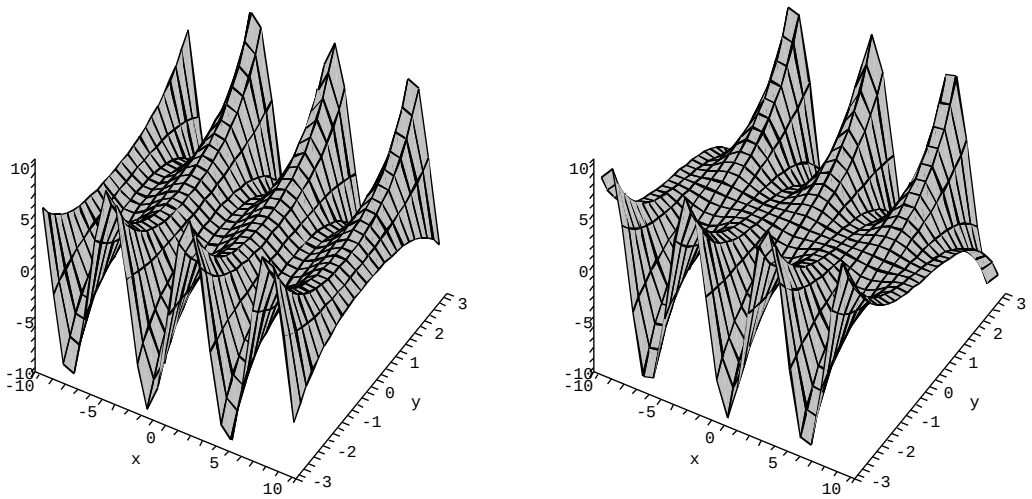
We hebben nu:

$$\begin{aligned} \sin(x+iy) &= \frac{e^{i(x+iy)} - e^{-i(x+iy)}}{2i} = \frac{e^{ix} \cdot e^{-y} - e^{-ix} \cdot e^{y}}{2i} \\ &= \frac{1}{2i}(\cos(x) + i\sin(x))e^{-y} - \frac{1}{2i}(\cos(-x) + i\sin(-x))e^{y} \\ &= \frac{1}{2}(-i\cos(x) + \sin(x))e^{-y} - \frac{1}{2}(-i\cos(-x) + \sin(-x))e^{y} \\ &= \frac{1}{2}(-i\cos(x) + \sin(x))e^{-y} + \frac{1}{2}(i\cos(x) + \sin(x))e^{y} \\ &= \sin(x)\frac{1}{2}(e^{-y} + e^{y}) + i\cos(x)\frac{1}{2}(-e^{-y} + e^{y}) \\ &= \sin(x)\cosh(y) + i\cos(x)\sinh(y). \end{aligned}$$

Er geldt dus:

$$\Re(\sin(x+iy)) = \sin(x)\cosh(y) \quad \text{en} \quad \Im(\sin(x+iy)) = \cos(x)\sinh(y).$$

In Figuur I.22 zijn de reële en imaginaire delen van de complexe sinus functie te zien.



Figuur I.22: Reëel en imaginair deel van $\sin(z)$

Een soortgelijke berekening voor $\cos(z)$ levert het volgende op:

$$\begin{aligned}
\cos(x+iy) &= \frac{e^{i(x+iy)} + e^{-i(x+iy)}}{2} = \frac{e^{ix}\cdot e^{-y} + e^{-ix}\cdot e^{y}}{2} \\
&= \frac{1}{2}(\cos(x) + i\sin(x))e^{-y} + \frac{1}{2}(\cos(-x) + i\sin(-x))e^{y} \\
&= \frac{1}{2}(\cos(x) + i\sin(x))e^{-y} + \frac{1}{2}(\cos(x) - i\sin(x))e^{y} \\
&= \cos(x)\frac{1}{2}(e^{-y} + e^{y}) + i\sin(x)\frac{1}{2}(e^{-y} - e^{y}) \\
&= \cos(x)\cosh(y) - i\sin(x)\sinh(y).
\end{aligned}$$

Er geldt dus:

$$\Re(\cos(x+iy)) = \cos(x)\cosh(y) \quad \text{en} \quad \Im(\cos(x+iy)) = \sin(x)\sinh(y).$$

Let op: We zijn gewend dat de reële sinus en cosinus *begrensde* functies zijn, de waarden liggen gewoon tussen -1 en 1. Voor de complexe versies van deze functies geldt dit echter niet meer: Als we in $\cos(z)$ met z langs de imaginaire as lopen, d.w.z. $z = ix$ invullen, hebben we $\cos(ix) = \frac{e^{-x}+e^{x}}{2} = \cosh(x)$ en dit is een onbegrensde functie.

De reden voor dit ongemaak is dat $\sin(z)$ en $\cos(z)$ zich langs de imaginaire as zo gedragen als de exponentiële functie langs de reële as en andersom.

Voor het gemak vatten we de formules voor reëel en imaginair deel van $\exp(z)$, $\cos(z)$ en $\sin(z)$ nog eens samen:

$$\begin{aligned}\exp(x+iy) &= \exp(x)\cos(y) + i\exp(x)\sin(y);\\ \sin(x+iy) &= \sin(x)\cosh(y) + i\cos(x)\sinh(y);\\ \cos(x+iy) &= \cos(x)\cosh(y) - i\sin(x)\sinh(y).\end{aligned}$$

Voor de volledigheid vermerken we nog, dat ook de hyperbolische functies een voortzetting naar de complexe getallen hebben. Dit gebeurt heel makkelijk door de relaties $\cosh(x) = \frac{e^x+e^{-x}}{2}$ en $\sinh(x) = \frac{e^x-e^{-x}}{2}$ van reële x naar complexe z uit te breiden, we definiëren dus:

$$\cosh(z) := \frac{e^z + e^{-z}}{2} \quad \text{en} \quad \sinh(z) := \frac{e^z - e^{-z}}{2}.$$

Ook voor deze functies kunnen we uit de Taylor reeks voor de exponentiële functie machtreeksen afleiden die overeen komen met de reële Taylor reeksen van $\cosh(x)$ en $\sinh(x)$, er geldt:

$$\cosh(z) = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \frac{z^2}{2} + \frac{z^4}{24} + \frac{z^6}{720} + \dots$$

$$\sinh(z) = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \frac{z^3}{6} + \frac{z^5}{120} + \frac{z^7}{5040} + \dots$$

Opdracht 23 *Laat zien dat* $\cosh(iz) = \cos(z)$ *en* $\sinh(iz) = i\sin(z)$.

6.3 Complexe logaritme

Als we een complexe logaritme willen definiëren hebben we (minstens) twee mogelijkheden om hieraan te beginnen. Aan de ene kant hebben we de Taylor reeks voor de reële logaritme en na onze goede ervaringen met deze aanpak zou het gek zijn als we deze reeks niet naar de complexe getallen zouden kunnen voortzetten. De Taylor reeks voor de reële logaritme is $\log(x+1) = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}$, dus kunnen we de complexe logaritme definiëren door

$$\log(z+1) := \sum_{n=1}^{\infty}(-1)^{n+1}\frac{z^n}{n} = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots$$

Het probleem is dat deze reeks niet voor alle waarden van z convergent is, maar alleen maar voor z met $|z| < 1$. We kunnen dus met deze machtreeks de waarden van de complexe logaritme alleen maar in een cirkel van straal 1 rond $z_0 = 1$ uitrekenen.

Aan de andere kant willen natuurlijk dat de complexe logaritme de omkeerfunctie van de complexe exponentiële functie is, dus dat $\log(e^z) = z$ en $e^{\log(z)} = z$. Beide mogelijkheden leiden uiteindelijk tot hetzelfde resultaat dat we nu vanuit het perspectief van de logaritme als omkeerfunctie van $\exp(z)$ gaan bekijken.

Voor een complex getal $z = re^{i\varphi}$ volgt uit de eis $e^{\log(z)} = z$ dat we $\log(z)$ noodzakelijk moeten definiëren door

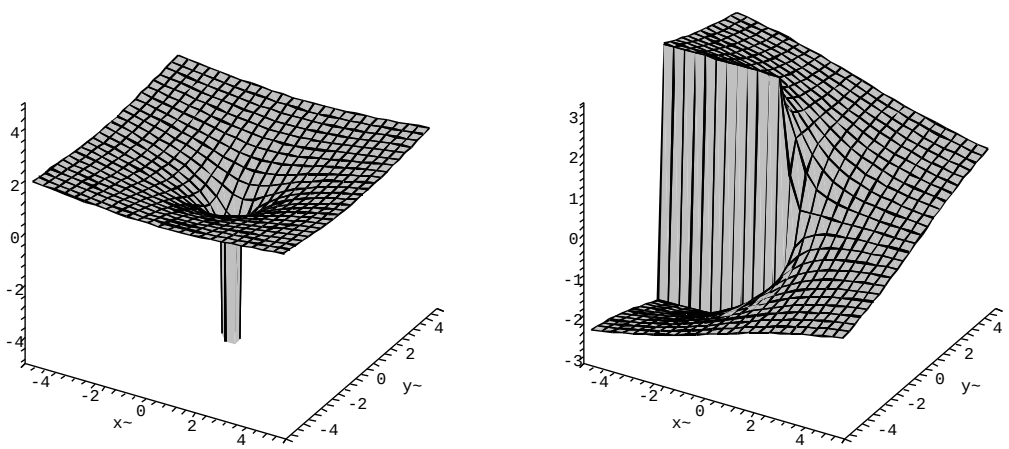
$$\log(re^{i\varphi}) := \log(r) + i\varphi$$

want $\log(z) = x + iy$ moet voldoen aan $re^{i\varphi} = z = e^{\log(z)} = e^{x+iy} = e^x \cdot e^{iy}$, dus hebben we nodig dat $e^x = r$ en $e^{iy} = e^{i\varphi}$ en dus $x = \log(r)$ en $y = \varphi$.

Er is wel een kleine complicatie bij deze definitie: Omdat de complexe exponentiële functie $2\pi i$-periodiek is, geldt ook voor $w = \log(z) + 2\pi i$ dat $e^w = z$, het imaginaire deel van $\log(z)$ is dus alleen maar tot op veelvouden van 2π na bepaald. De exponentiële functie beeld namelijk elke streep $S_a = \{z \in \mathbb{C} \mid \Im(z) \in (a, a+2\pi]\}$ op $\mathbb{C} \setminus \{0\}$ af en in principe is elke streep even goed. De conventie is echter, dat het imaginaire deel van $\log(z)$ in het interval $(-\pi, \pi]$ ligt. We hebben dus

$$\log(z) = \begin{cases} \log(|z|) + i\arg(z) & \text{als } \arg(z) \in [0, \pi] \\ \log(|z|) + i(\arg(z) - 2\pi) & \text{als } \arg(z) \in (\pi, 2\pi). \end{cases}$$

In Figuur I.23 zien we de reële en imaginaire delen van de complexe logaritme. Het is duidelijk dat het imaginaire deel op de negatieve reële as niet continu is, maar een sprong om 2π heeft.



Figuur I.23: Reëel en imaginair deel van $\log(z)$

Omdat we voor de complexe logaritme een keuze moeten maken in welke streep van breedte 2π het imaginaire deel van $\log(z)$ ligt, krijgen we een probleem dat we bij de reële logaritme niet kennen. Kijken we bijvoorbeeld naar $z_1 = z_2 = e^{i\frac{2}{3}\pi}$, dan is duidelijk $\log(z_1) = \log(z_2) = i\frac{2}{3}\pi$ en dus $\log(z_1) + \log(z_2) = i\frac{4}{3}\pi$. Maar $z_1 \cdot z_2 = e^{i(\frac{2}{3}+\frac{2}{3})\pi} = e^{i\frac{4}{3}\pi} = e^{i(-\frac{2}{3})\pi}$ en daarom is $\log(z_1 \cdot z_2) = -\frac{2}{3}\pi$. De relatie $\log(z_1 \cdot z_2) = \log(z_1) + \log(z_2)$ geldt dus niet

meer in elk geval, want de imaginaire delen aan de rechter en linker kant kunnen om veelvouden van 2π verschillen. We zeggen daarom, dat

$$\log(z_1 \cdot z_2) = \log(z_1) + \log(z_2) \text{ modulo veelvouden van } 2\pi i.$$

Omdat n-de machten slechts een speciaal geval van producten zijn, geldt ook de regel $\log(z^n) = n\log(z)$ alleen maar modulo veelvouden van $2\pi i$.

Het voordeel van onze keuze van de streep $\Im(z) \in (-\pi, \pi)$ van breedte 2π is, dat de relatie

$$\log(\frac{1}{z}) = -\log(z)$$

wel nog altijd geldt, want deze streep wordt onder de afbeelding $z \to -z$ op zich zelf afgebeeld.

Opdracht 24 *Schrijf* $\log(1+i)$, $\log(-i)$ *en* $\log(\frac{2+i}{2-i})$ *in de vorm* $x + iy$.

6.4 Differentiëren via Taylor reeksen

We zullen in de Appendix voor deze les de vraag nagaan, wanneer een functie complex differentieerbaar is. Voor het moment nemen we genoegen ermee dat we zeggen, dat een complexe functie $f(z)$ in het punt z_0 complex differentieerbaar is als de limiet

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

bestaat en onafhankelijk van het traject waarop z tegen z_0 aan loopt dezelfde waarde heeft. We zullen zien dat complexe differentieerbaarheid een veel sterkere eigenschap is dan de gewone differentieerbaarheid bij reële functies.

Definitie: Een functie $f(z)$ die in elk punt $z \in B$ van zijn domein complex differentieerbaar is, heet een (op B) *holomorfe* functie.

De *ontwikkelingsstelling van Cauchy-Taylor* zegt nu dat we een holomorfe functie altijd als een absoluut convergente Taylor reeks kunnen schrijven, dus dat

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \text{ waarbij } \sum_{n=0}^{\infty} |a_n z^n| \text{ convergent is.}$$

Omgekeerd geeft een absoluut convergente Taylor reeks steeds een holomorfe functie aan.

In de wereld van complex differentieerbare functies gaat eigenlijk alles goed, wat we zo maar zouden kunnen hopen, daarom is er ook een stelling die zegt dat we de afgeleide van een holomorfe functie krijgen door de (absoluut convergente) Taylor reeks van de functie termsgewijs af te leiden.

Stelling: Voor de holomorfe functie $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is de afgeleide $f'(z)$ gegeven door

$$f'(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}.$$

We weten dus dat een complexe functie $f(z)$ die door een absoluut convergente Taylor reeks gegeven is, complex differentieerbaar is en dat we de afgeleide $f'(z)$ vinden door de Taylor reeks termsgewijs af te leiden. Maar om een term in een Taylor reeks af te leiden, hebben we alleen maar de afgeleide van z^n nodig, en we zullen in de Appendix laten zien dat net als bij reële functies geldt dat

$$(z^n)' = nz^{n-1}.$$

Met behulp hiervan kunnen we nu de complexe functies afleiden die we tot nu toe hebben gezien:

$$(1) \qquad \exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \dots$$

$$\Rightarrow \exp'(z) = 1 + 2 \cdot \frac{z}{2!} + 3 \cdot \frac{z^2}{3!} + 4 \cdot \frac{z^3}{4!} + \dots = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$$

$$= \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{z^n}{n!} = \exp(z).$$

$$(2) \qquad \sin(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \dots$$

$$\Rightarrow \sin'(z) = 1 - 3 \cdot \frac{z^2}{3!} + 5 \cdot \frac{z^4}{5!} - 7 \cdot \frac{z^6}{7!} + \dots = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \dots$$

$$= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} = \cos(z).$$

$$(3) \qquad \cos(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \dots$$

$$\Rightarrow \cos'(z) = -2 \cdot \frac{z}{2!} + 4 \cdot \frac{z^3}{4!} - 6 \cdot \frac{z^5}{6!} + \dots = -z + \frac{z^3}{3!} - \frac{z^5}{5!} + \dots$$

$$= -\sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} = -\sin(z).$$

$$(4) \qquad \log(z+1) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{z^n}{n} = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots$$

$$\Rightarrow \log'(z+1) = 1 - 2 \cdot \frac{z}{2} + 3 \cdot \frac{z^2}{3} - 4 \cdot \frac{z^3}{4} + \dots = 1 - z + z^2 - z^3 + \dots$$

$$= \sum_{n=0}^{\infty} (-1)^n z^n = \frac{1}{1+z} \text{ (meetkundige reeks).}$$

De meetkundige reeks $\sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \dots$ is convergent als $|x| < 1$ en heeft in dit geval de waarde $\frac{1}{1-x}$. Dit ziet men in door uit te werken dat $(1 + x + x^2 + x^3 + \dots + x^n)(1 - x) = 1 - x^{n+1}$. Voor $|x| < 1$ gaat x^{n+1} voor $n \to \infty$ naar 0, dus is $(1 + x + x^2 + x^3 + \dots)(1 - x) = 1$.

We zien dus dat de afgeleiden precies zo zijn als we dat volgens ons kennis van de reële functies zouden verwachten.

6.5 Appendix: Complexe differentieerbaarheid

Bij reële functies hadden we de afgeleide $f'(x_0)$ in een punt x_0 gedefinieerd als de stijging van de raaklijn in het punt x_0 aan de grafiek van $f(x)$. Voor een complexe functie hebben we al gezien, dat we de reële en imaginaire delen van de functie apart als driedimensionale landschappen (grafieken) kunnen representeren. In een punt van zo'n landschap kunnen we wel een raak*vlak* definiëren, maar het is onduidelijk hoe we uit de raakvlakken voor reëel en imaginair deel van de functie een complex getal zullen maken die we als afgeleide van de functie in dit punt definiëren.

Maar de eigenschap dat de afgeleide de stijging van de raaklijn aangeeft kunnen we ook nog iets anders formuleren: De functie $f(x)$ wordt in een kleine omgeving van een punt goed door de raaklijn benaderd, we noemen daarom de afgeleide ook de *linearisering* van de functie. Dit volgt uit de definitie dat

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \text{ als deze limiet bestaat.}$$

Als we namelijk de definitie van de afgeleide voor kleine waarden van $\Delta x = h$ (en zonder limiet) bekijken, hebben we

$$f'(x) \approx \frac{f(x+\Delta x)-f(x)}{\Delta x} \text{ en dus } f(x+\Delta x) \approx f(x) + f'(x)\Delta x.$$

In een kleine omgeving van x wordt de functie dus goed beschreven door een vermenigvuldiging met $f'(x)$. Preciezer gezegd beeld de functie het punt x naar $f(x)$ af en een afwijking Δx van x wordt door de functie met $f'(x)$ vermenigvuldigd en op $f(x)$ opgeteld.

Deze interpretatie nemen we nu over als definitie van de complexe afgeleide: De afgeleide $f'(z)$ geeft aan, dat we in een (kleine) omgeving van z de functiewaarden van $f(z)$ kunnen benaderen door een afwijking Δz van z met $f'(z)$ te vermenigvuldigen en bij $f(z)$ op te tellen:

$$f(z+\Delta z) \approx f(z) + f'(z)\Delta z.$$

Dit kunnen we ook zuiver meetkundig interpreteren, want we weten wat vermenigvuldiging met een complex getal $f'(z) = re^{i\varphi}$ betekent, namelijk een schaling met een factor r en een draaiing om φ. In een omgeving van z wordt een complex differentieerbare functie $f(z)$ dus beschreven door een schaling gecombineerd met een draaiing.

De interpretatie van de complexe afgeleide als linearisering van $f(z)$ in een kleine omgeving maakt het noodzakelijk dat we de definitie van de reële afgeleide via de limiet $\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$ letterlijk overnemen voor complexe functies, waarbij we (volgens de conventies) de reële variabel x door een complexe variabel z vervangen. We krijgen dus de volgende definitie:

Definitie: Een complexe functie $f(z)$ heet in het punt z_0 (complex) *differentieerbaar*, als de limiet

$$\lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h}$$

bestaat. In dit geval noteren we de afgeleide in het punt z_0 door $f'(z_0)$. Een functie $f(z)$ die in elk punt van zijn domein differentieerbaar is, heet ook een *holomorfe functie* of een *analytische functie*.

Het cruciale punt bij deze definitie is het *bestaan* van de limiet. Voor reële functies kan h alleen maar van links of van rechts naar 0 toe lopen. Dan is het voldoende als de limiet van links en van rechts bestaat en deze twee limieten hetzelfde zijn. Zo zien we bijvoorbeeld dat de functie $f(x) = |x|$ in het punt 0 niet differentieerbaar is omdat de limiet van $\frac{f(x+h)-f(x)}{h}$ voor $h > 0$ (dus van rechts) gelijk is aan 1 terwijl de limiet voor $h < 0$ (dus van links) gelijk is aan -1. Maar we hoeven inderdaad niet meer te doen dan van links en van rechts te kijken.

Voor complexe functies is dit een heel ander verhaal, want h kan van rechts of links op de reële as naar 0 lopen, maar ook van boven of beneden op de imaginaire as of langs een willekeurige lijn met $\Im(z) = a \cdot \Re(z)$. En h mag zelfs langs een heel kromme lijn lopen, bijvoorbeeld langs een spiraal die zich om het nulpunt wikkelt. En voor elk van de mogelijke trajecten van h moet de limiet bestaan en steeds dezelfde waarde hebben. Het feit dat h op een willekeurig traject naar 0 toe mag lopen maakt van de complexe differentieerbaarheid een heel sterke eigenschap die vergaande consequenties heeft.

> De complexe differentieerbaarheid heeft een aantal indrukwekkende consequenties. Bijvoorbeeld volgt uit de samenhang tussen holomorfe functies en hun Taylor reeksen de *Stelling van Liouville* die zegt dat een op $\mathbb{C}$ differentieerbare functie alleen maar begrensd kan zijn als hij constant is. We hadden al gezien dat de complexe sinus en cosinus functies langs de imaginaire as tegen oneindig gaan. De stelling van Liouville zegt nu dat globaal begrensde functies zo als de reële sinus of cosinus functies op het complexe vlak niet kunnen bestaan.
>
> Er is echter nog een veel sterker resultaat: De complexe exponentiële functie heeft alle complexe getallen als waarden behalve van 0. Dit is inderdaad voor alle holomorfe functies zo, want een van de stellingen van Picard (Charles Emile Picard, niet Jean-Luc) zegt dat een holomorfe functie die twee complexe getallen niet als waarde heeft noodzakelijk een constante functie $f(z) = c$ is.

Het voordeel ervan, de afgeleide van een complexe functie net zo te definiëren als voor reële functies, is dat de *rekenregels* voor de afgeleide hetzelfde blijven.

Als $f(z)$ en $g(z)$ complex differentieerbare functies zijn, geldt dus:

$$\begin{aligned}
&(f+g)'(z) = f'(z) + g'(z) \\
&(f \cdot g)'(z) = f'(z)g(z) + f(z)g'(z) \text{ (productregel)} \\
&\left(\frac{f}{g}\right)'(z) = \frac{f'(z)g(z) - f(z)g'(z)}{g(z)^2} \text{ (quotiëntregel)} \\
&(f \circ g)'(z) = f(g(z))' = f'(g(z))g'(z) \text{ (kettingregel)}
\end{aligned}$$

Tot nu toe hebben we nog geen enkele complex differentieerbare functie gezien. Maar we hebben wel al een gok op de afgeleide van $f(z) = z^n$ gedaan, namelijk dat hiervoor $f'(z) = nz^{n-1}$ is, net als we dat van de reële functies gewend zijn. Voor de reële functies hebben we dit in Wiskunde 1 met behulp van de productregel per volledige inductie bewezen. Dit is wel elegant, maar een rechtstreekse berekening doet ook geen kwaad. Hierbij hebben we de binomische formule voor $(z+h)^n$ nodig, te weten

$$(z+h)^n = \sum_{k=0}^{n} \binom{n}{k} z^{n-k} h^k = z^n + nz^{n-1}h + \frac{n(n-1)}{2} z^{n-2}h^2 + \ldots + nzh^{n-1} + h^n.$$

Voor $f(z) = z^n$ geldt dus

$$\begin{aligned}
&\frac{f(z+h) - f(z)}{h} = \frac{(z+h)^n - z^n}{h} \\
&= \frac{1}{h}(z^n + nz^{n-1}h + \frac{n(n-1)}{2} z^{n-2}h^2 + \ldots + nzh^{n-1} + h^n - z^n) \\
&= nz^{n-1} + \frac{n(n-1)}{2} z^{n-2}h + \ldots + nzh^{n-2} + h^{n-1}.
\end{aligned}$$

In de laatste som bevat elke term vanaf de tweede een macht van h, en als we de limiet $h \to 0$ bekijken gaat dus ieder van deze termen naar 0. Merk op dat dit onafhankelijk van het traject is waarop h naar 0 loopt. Daarom bestaat de limiet $\lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$ en we hebben:

$$\text{Voor } f(z) = z^n \text{ is } f'(z) = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h} = nz^{n-1}.$$

Contrastvoorbeeld

Dat er bij complexe functies snel iets mis kan gaan zien we bij een heel eenvoudige, onschuldige functie, de complexe conjugatie

$$f(z) := \overline{z}.$$

We laten h eerst langs de reële as lopen, het maakt niet uit of van rechts of links. Voor $h \in \mathbb{R}$ geldt $\frac{\overline{z+h}-\overline{z}}{h} = \frac{\overline{z}+h-\overline{z}}{h} = \frac{h}{h} = 1$, dus is ook de limiet $h \to 0$ gelijk aan 1 als we langs de reële as lopen. Dit is natuurlijk geen verrassing, want op de reële as doet complexe conjugatie niets, en moet dus dezelfde afgeleide hebben als de reële functie $f(x) = x$.

Nu laten we h langs de imaginaire as lopen, hiervoor nemen we $h = i\varepsilon$ met $\varepsilon \in \mathbb{R}$. Er geldt $\frac{\overline{z+i\varepsilon}-\overline{z}}{i\varepsilon} = \frac{\overline{z}-i\varepsilon-\overline{z}}{i\varepsilon} = \frac{-i\varepsilon}{i\varepsilon} = -1$, dus is de limiet gelijk aan -1 als we langs de imaginaire as lopen.

We kunnen zelfs een willekeurig complex getal op de eenheidscirkel als limiet produceren. Als we langs de lijn vanuit het getal $e^{i\varphi}$ naar 0 lopen, hebben we $h = e^{i\varphi} \cdot \varepsilon$ met $\varepsilon \in \mathbb{R}$. Dan is

$$\frac{\overline{z+h}-\overline{z}}{h} = \frac{\overline{z+e^{i\varphi}\varepsilon}-\overline{z}}{e^{i\varphi}\varepsilon} = \frac{\overline{z}+e^{-i\varphi}\varepsilon-\overline{z}}{e^{i\varphi}\varepsilon} = \frac{e^{-i\varphi}\varepsilon}{e^{i\varphi}\varepsilon} = e^{-2i\varphi},$$

dus is de limiet op dit traject gelijk aan $e^{-2i\varphi}$. Maar we kunnen elk getal op de eenheidscirkel als $e^{-2i\varphi}$ schrijven, want als φ van 0 naar $-\pi$ loopt, loopt $e^{-2i\varphi}$ een keer langs de eenheidscirkel.

Cauchy-Riemann differentiaalvergelijkingen

Het voorbeeld van de complexe conjugatie is een beetje verontrustend, want het is toch heel omslachtig om steeds te testen of de limiet voor alle mogelijke trajecten waarop h naar 0 gaat hetzelfde is. Gelukkig is dit echter niet nodig, er is een stelling die zegt dat het voldoende is om langs de reële en langs de imaginaire as te kijken. Deze stelling gaan we hier niet bewijzen, maar we kunnen wel een motivatie geven.

We hebben al eerder gezien dat het handig is om apart naar de reële en imaginaire delen van een complexe functie te kijken, want hiervoor kunnen we 3-dimensionale plaatjes maken. Als we een complex getal $z \in \mathbb{C}$ in de vorm $z = x + iy$ met $x, y \in \mathbb{R}$ schrijven, kunnen we een complexe functie $f(z)$ zien als een functie van twee reële variabelen, namelijk van $\Re(z)$ en $\Im(z)$. We kunnen dus een complexe functie $f(z)$ beschrijven door twee reële functies van de twee reële variabelen $x = \Re(z)$ en $y = \Im(z)$, namelijk

$$f(z) = \Re(f(z)) + i \cdot \Im(f(z)) = u(x,y) + i \cdot v(x,y).$$

Als we nu ervan uitgaan dat $f(z)$ een complex differentieerbare functie met afgeleide $f'(z)$ is, kunnen we kijken wat dit voor de (reële) functies $u(x,y)$ en $v(x,y)$ betekent. In het bijzonder gaan we bekijken wat er gebeurt als we langs de reële as en langs de imaginaire as naar 0 lopen. Eerst lopen we langs de reële as met $h \to 0$ voor $h \in \mathbb{R}$, dan is

$$\begin{aligned}
f'(z) &= \lim_{h\to 0} \frac{f(z+h)-f(z)}{h} \\
&= \lim_{h\to 0} \frac{(u(x+h,y)+i\cdot v(x+h,y)) - (u(x,y)+i\cdot v(x,y))}{h} \\
&= \lim_{h\to 0} \frac{u(x+h,y)-u(x,y)}{h} + i \cdot \lim_{h\to 0} \frac{v(x+h,y)-v(x,y)}{h} \\
&= \frac{\partial u(x,y)}{\partial x} + i \cdot \frac{\partial v(x,y)}{\partial x}.
\end{aligned}$$

Het afleiden langs de reële as komt dus overeen met de partiële afgeleide naar het reële deel x van z, waarbij we y als een constante beschouwen.

Nu lopen we langs de imaginaire as met $ih \to 0$ voor $h \in \mathbb{R}$, dan geldt

$$\begin{aligned} f'(z) &= \lim_{h\to 0} \frac{f(z+ih)-f(z)}{ih} = (-i)\cdot \lim_{h\to 0} \frac{f(z+ih)-f(z)}{h} \\ &= (-i)\cdot \lim_{h\to 0} \frac{(u(x,y+h)+i\cdot v(x,y+h))-(u(x,y)+i\cdot v(x,y))}{h} \\ &= (-i)\cdot \lim_{h\to 0} \frac{u(x,y+h)-u(x,y)}{h} + (-i)\cdot i \cdot \lim_{h\to 0} \frac{v(x,y+h)-v(x,y)}{h} \\ &= -i\cdot \frac{\partial u(x,y)}{\partial y} + \frac{\partial v(x,y)}{\partial y}. \end{aligned}$$

Maar als $f(z)$ differentieerbaar is weten we dat de twee limieten gelijk moeten zijn, daarom hebben we de noodzakelijke voorwaarde dat de reële en imaginaire delen van de limieten hetzelfde zijn, dus

$$\frac{\partial u(x,y)}{\partial x} = \frac{\partial v(x,y)}{\partial y} \quad \text{en} \quad \frac{\partial u(x,y)}{\partial y} = -\frac{\partial v(x,y)}{\partial x}.$$

Deze twee noodzakelijke voorwaarden heten de *Cauchy-Riemann differentiaalvergelijkingen.*

Het belangrijke (en misschien iets verrassende) punt is nu dat deze voorwaarde ook voldoende is, d.w.z. een complexe functie waarvoor de reële en imaginaire delen aan de Cauchy-Riemann differentiaalvergelijkingen voldoen is complex differentieerbaar.

Stelling: Een complexe functie $f(z) = f(x+iy) = u(x,y) + iv(x,y)$ is complex differentieerbaar dan en slechts dan als $u(x,y)$ en $v(x,y)$ continu differentieerbaar zijn en voldoen aan de Cauchy-Riemann differentiaalvergelijkingen

$$\frac{\partial u(x,y)}{\partial x} = \frac{\partial v(x,y)}{\partial y} \text{ en } \frac{\partial u(x,y)}{\partial y} = -\frac{\partial v(x,y)}{\partial x}.$$

Omdat we de limiet van $\frac{f(z+h)-f(z)}{h}$ langs de reële as en langs de imaginaire as al in de partiële afgeleiden van $u(x,y)$ en $v(x,y)$ hebben uitgedrukt, kunnen we de waarde van de afgeleide $f'(z)$ expliciet aangeven, namelijk door

$$f'(z) = \frac{\partial u(x,y)}{\partial x} + i\cdot \frac{\partial v(x,y)}{\partial x} = \frac{\partial v(x,y)}{\partial y} - i\cdot \frac{\partial u(x,y)}{\partial y}.$$

Voorbeeld: We passen het criterium van de Cauchy-Riemann differentiaalvergelijkingen op de complexe exponentiële functie toe. We hebben gezien dat voor $z = x+iy$ geldt dat

$$\exp(z) = \exp(x)\cos(y) + i\cdot \exp(x)\sin(y),$$

dus hebben we:

$$\exp(x+iy) = u(x,y) + i\cdot v(x,y) \text{ met}$$

$$u(x,y) = \exp(x)\cos(y) \quad \text{en} \quad v(x,y) = \exp(x)\sin(y).$$

Om de Cauchy-Riemann differentiaalvergelijkingen te testen moeten we nu de partiële afgeleiden van $u(x,y)$ en $v(x,y)$ berekenen. Er geldt:

$$\frac{\partial u(x,y)}{\partial x} = \exp(x)\cos(y), \quad \frac{\partial u(x,y)}{\partial y} = -\exp(x)\sin(y),$$
$$\frac{\partial v(x,y)}{\partial x} = \exp(x)\sin(y), \quad \frac{\partial v(x,y)}{\partial y} = \exp(x)\cos(y)$$

en we zien dat inderdaad $\frac{\partial u(x,y)}{\partial x} = \frac{\partial v(x,y)}{\partial y}$ en $\frac{\partial u(x,y)}{\partial y} = -\frac{\partial v(x,y)}{\partial x}$. We concluderen dat de complexe exponentiële functie complex differentieerbaar is met afgeleide $\exp'(z) = \frac{\partial u(x,y)}{\partial x} + i \cdot \frac{\partial v(x,y)}{\partial x} = \exp(x)\cos(y) + i \cdot \exp(x)\sin(y) = \exp(z)$.

Opdracht 25 *Ga na dat* $\sin(z)$ *en* $\cos(z)$ *aan de Cauchy-Riemann differentiaalvergelijkingen voldoen en dus complex differentieerbaar zijn. (Hint: De reële en imaginaire delen van de complexe sinus en cosinus functies hebben we bepaald, voor de afgeleiden van de reële hyperbolische functies geldt* $\cosh'(x) = \sinh(x)$ *en* $\sinh'(x) = \cosh(x)$*.)*

Merk op: Met onze kennis over functies van meerdere veranderlijken kunnen we de Cauchy-Riemann differentiaalvergelijkingen ook op een iets andere manier afleiden: Als we het complexe getal $z = x + iy$ als 2-dimensionale vector $z = \begin{pmatrix} x \\ y \end{pmatrix}$ schrijven, wordt de vermenigvuldiging van z met een complex getal $a + ib$ beschreven door de 2×2-matrix $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, want $(a+ib)(x+iy) = (ax - by) + i(bx + ay)$ en er geldt

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax - by \\ bx + ay \end{pmatrix}.$$

Vermenigvuldiging met een complex getal wordt dus beschreven door de speciale 2×2-matrices van de vorm $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$.

We kunnen nu een complexe functie $f(z)$ opvatten als functie van de twee reële variabelen x en y met $z = x + iy$, die gegeven is door de twee componenten $u(x,y) = \Re(f(z))$ en $v(x,y) = \Im(f(z))$, dus als functie van de vorm

$$f(x,y) = (u(x,y), v(x,y)).$$

Maar voor zo'n functie hadden we gezien, dat de afgeleide van $f(x,y)$ een lineaire afbeelding $\mathbb{R}^2 \to \mathbb{R}^2$ gegeven door een 2×2-matrix is, namelijk door de Jacobi matrix J met de partiële afgeleiden:

$$J = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}.$$

In het algemeen is J een willekeurige lineaire afbeelding $\mathbb{R}^2 \to \mathbb{R}^2$, en alleen maar in het geval dat J van de vorm $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ is, is de afgeleide de vermenigvuldiging met een complex getal en dit is juist het geval als $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ en $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$, dus als $u(x,y)$ en $v(x,y)$ aan de Cauchy-Riemann differentiaalvergelijkingen voldoen.

Niet elke functie $u(x,y)$ kan reëel of imaginair deel van een holomorfe functie $f(z)$ zijn. Door toepassen van de Cauchy-Riemann differentiaalvergelijkingen op $u(x,y) = \Re(f(z))$ vinden we namelijk: $\frac{\partial^2 u(x,y)}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial u(x,y)}{\partial x} = \frac{\partial}{\partial x}\frac{\partial v(x,y)}{\partial y} = \frac{\partial^2 v(x,y)}{\partial x \partial y}$. Volgens de stelling van Schwarz mogen we partiële afgeleiden verruilen, dus is $\frac{\partial^2 v(x,y)}{\partial x \partial y} = \frac{\partial}{\partial y}\frac{\partial v(x,y)}{\partial x} = -\frac{\partial}{\partial y}\frac{\partial u(x,y)}{\partial y} = -\frac{\partial^2 u(x,y)}{\partial y^2}$. We hebben dus $\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0$. Met een analoge berekening vinden dezelfde relatie ook voor $v(x,y) = \Im(f(z))$, dus $\frac{\partial^2 v(x,y)}{\partial x^2} + \frac{\partial^2 v(x,y)}{\partial y^2} = 0$.

Algemeen heten functies met de eigenschap $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0$ *harmonische functies* of, gemotiveerd door de natuurkunde, *potentieelfuncties*. We hebben dus gezien dat alleen maar harmonische functies reëel of imaginair deel van een holomorfe functie $f(z)$ kunnen zijn.

Opdracht 26

(i) In welke punten $z \in \mathbb{C}$ is $f(z) := \Re(z)^2 + i \cdot \Im(z)^2$ complex differentieerbaar? Bepaal in deze punten $f'(z)$.

(ii) In welke punten $z \in \mathbb{C}$ is $f(z) := \overline{z}(3z^2 + \overline{z}^2)$ complex differentieerbaar? Bepaal in deze punten $f'(z)$.

Belangrijke begrippen in deze les

- complexe exponentiële functie $\exp(z)$
- $2\pi i$-periodiciteit van de complexe exponentiële functie
- complexe sinus en cosinus functies $\sin(z)$ en $\cos(z)$
- reële en imaginaire delen van $\exp(z)$, $\sin(z)$, $\cos(z)$
- complexe logaritme $\log(z)$
- termsgewijs afleiden van Taylor reeksen
- complexe differentieerbaarheid
- holomorfe functies
- Cauchy-Riemann differentiaalvergelijkingen

Opgaven

54. Bepaal voor de afbeelding $f(z) := z^2$ de beelden van de lijnen

(i) $L_1 := \{z = x + iy \in \mathbb{C} \mid x = 2y\}$,
(ii) $L_2 := \{z = x + iy \in \mathbb{C} \mid x = 2\}$,
(iii) $L_3 := \{z = x + iy \in \mathbb{C} \mid y = -1\}$.

Teken de beelden van de lijnen in het complexe vlak.

55. Bepaal het beeld van de rechthoek $R = \{x + iy \in \mathbb{C} \mid x \in [-1, 1], y \in [\frac{1}{2}, 1]\}$ onder de complexe exponentiële functie. Maak een schets. Kan je algemeen aangeven wat het beeld van een rechthoek $\{z \in \mathbb{C} \mid \Re(z) \in [a, b], \Im(z) \in [c, d]\}$ met $a, b, c, d \in \mathbb{R}$ is?

56. Laat zien dat de nulpunten van $\sin(z)$ alle reëel zijn, d.w.z. dat $\sin(z) \neq 0$ als $\Im(z) \neq 0$. Ga na dat hetzelfde ook voor $\cos(z)$ geldt.

57. Laat zien dat voor $z = x+iy$ geldt dat $\Re(\cosh(z)) = \cosh(x)\cos(y)$ en $\Im(\cosh(z)) = \sinh(x)\sin(y)$.

    Bepaal ook $\Re(\sinh(z))$ en $\Im(\sinh(z))$.

58. Vind de oplossingen in $\mathbb{C}$ voor de volgende vergelijkingen:

$$\text{(i) } e^z = i, \qquad \text{(ii) } e^z = 1 + i \qquad \text{(iii) } \cos(z) = -3.$$

59. We bekijken de afbeelding $f(z) := e^{iz}$.

    (i) Bepaal voor een vaste $w \in \mathbb{C}$ de waarden van z met $f(z) = w$.

    (ii) Bepaal een deel $D \subseteq \mathbb{C}$ van het complexe vlak zo dat $f(z)$ op D een omkeerfunctie heeft. Geef de omkeerfunctie aan.

60. Gebruik de relaties $\cos(z) = \frac{e^{iz}+e^{-iz}}{2}$ en $\sin(z) = \frac{e^{iz}-e^{-iz}}{2i}$ om de afgeleiden $\cos'(z) = -\sin(z)$ en $\sin'(z) = \cos(z)$ rechtstreeks uit de afgeleide van $\exp(z)$ te berekenen (zonder Taylor reeksen of partiële afgeleiden). Let op dat volgens de kettingregel $(e^{iz})' = i \cdot e^{iz}$.

61. De *arcustangens* functie heeft in $z_0 = 0$ de Taylor reeks

$$\arctan(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{2n+1} = z - \frac{z^3}{3} + \frac{z^5}{5} - \frac{z^7}{7} + \dots.$$

    Laat zien dat $\arctan'(z) = \frac{1}{1+z^2}$.

62. Schrijf voor $z = x + iy$ de volgende functies in de vorm $f(z) = u(x, y) + iv(x, y)$ waarbij $u(x, y)$ en $v(x, y)$ reële functies zijn:

    (i) $f(z) = z^2 + 2iz$;

    (ii) $f(z) = \frac{z}{3+z}$;

    (iii) $f(z) = \exp(z^2)$;

    (iv) $f(z) = \log(1 + z)$.

Deel II

Fourier theorie

Les 7 Fourier analyse

Veel gewone fenomenen hebben iets met golven te maken, zo is bijvoorbeeld geluid een golvende verandering van de luchtdruk en is licht een elektromagnetische golf.

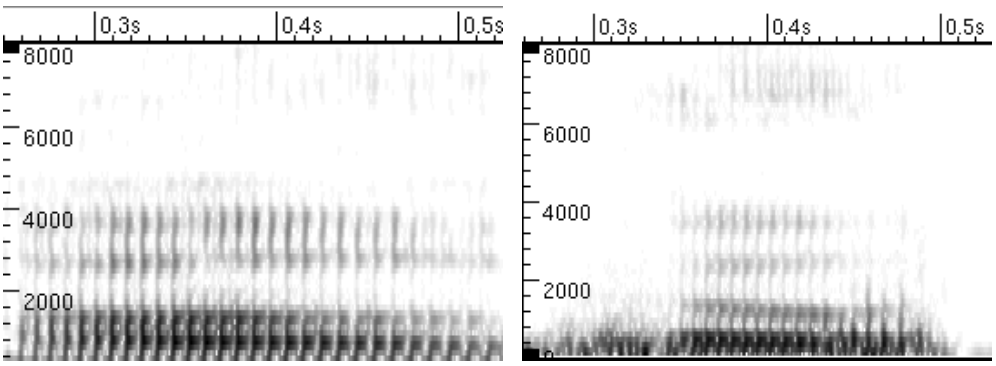
Als we nu eens kijken hoe bij een viool (of elk ander snaarinstrument) het geluid wordt geproduceerd, dan is het duidelijk dat de snaar aan de eindpunten vast zit, maar daartussen een golvende beweging uitvoert. De eenvoudigste mogelijkheid hiervoor is natuurlijk dat in het midden een buik is, waar de snaar de grootste amplitude heeft. Maar we kunnen ook precies in het midden een vinger op de snaar zetten, dan krijgen we twee half zo lange golven en de toon klinkt een octaaf hoger. Net zo kunnen we de vinger op een derde van de snaar plaatsen, de toon klinkt dan nog een kwint hoger, ook al heeft de intensiteit behoorlijk afgenomen. De tonen die we op deze manier produceren heten *boventonen* en klinken harmonisch met de grondtoon samen.

> Dit was al aan Pythagoras bekend en ons gewoon systeem van twaalf halftonen in een octaaf berust op het delen van een snaar in twee stukken met een eenvoudige verhouding: 2 : 1 octaaf, 3 : 2 kwint, 4 : 3 kwart, 5 : 4 grote terts, 6 : 5 kleine terts, 9 : 8 grote seconde, 16 : 15 kleine seconde (halftoon). Uiteindelijk moet men met sommige intervallen iets schuiven, omdat bij deze verhoudingen twaalf kwinten een iets groter interval geven dan acht octaven: $1.5^{12} \approx 129.75$, $2^8 = 128$. De verhouding $\frac{1.5^{12}}{2^8} \approx 1.01364$ noemt men ook het *Pythagoraeïsch komma*. Om dit probleem te ontsnappen zijn er verschillende *stemmingen* uitgevonden, bekende stemmingen zijn de *gelijkzwevende* en verschillende soorten van *Wohltemperierung*.

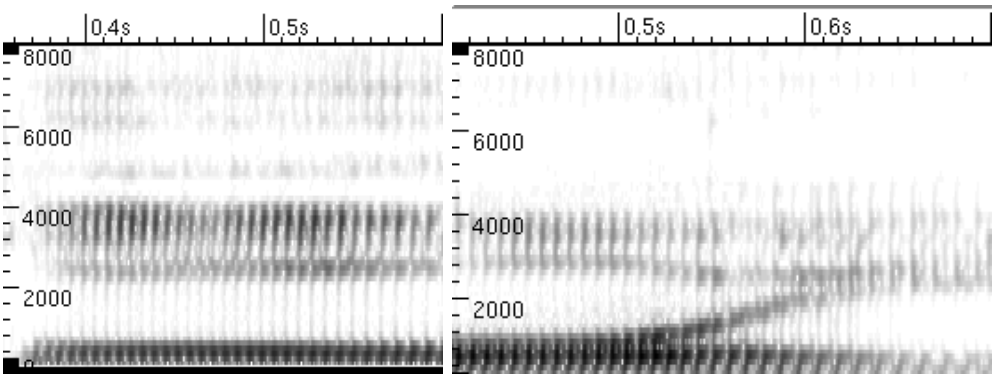
Als we naar verschillende instrumenten luisteren die dezelfde toon spelen, zullen we nog steeds makkelijk het verschil tussen een trompet, een viool en een piano kunnen horen (maar let wel: als je van een toon het begingeruis afplakt wordt dit veel moeilijker). De reden hiervoor ligt in de intensiteiten die de boventonen hebben, bij een trompet zijn het er veel meer en ook bij een viool zijn de boventonen nog relatief sterk.

Het idee is nu, de klank van een toon te beschrijven door naar de intensiteiten van de verschillende boventonen te kijken. De verdeling van de intensiteiten (waarbij we de grondtoon bijvoorbeeld op 1 normeren) geeft dan een karakterisering van de klank. Als we de hoogte van een toon door een grondfrequentie ω_0 beschrijven zo dat de frequenties van alle boventonen een veelvoud van ω_0 zijn, kunnen we de toon door de intensiteiten $a_1, a_2, \ldots, a_n$ beschrijven, waarbij a_k de intensiteit van de boventoon met frequentie $k\omega_0$ aangeeft (dit noemen we ook de boventoon van orde k). Bij de meeste instrumenten zijn de intensiteiten van de boventonen van orde 10 of meer erg klein, maar het menselijk oor is verbazingwekkend gevoelig voor heel kleine verschillen. Soms is het ook zo dat de grondfrequentie niet te hoogste intensiteit heeft, bij sommige instrumenten kan de speler dit zelfs bewust veranderen, bijvoorbeeld met de *flageolet* tonen bij een viool of fluit.

Het doel van de *Fourier analyse* is in principe, voor een gegeven 'klank' de intensiteiten $a_0, \ldots, a_n$ te bepalen, die een karakteristiek patroon voor de klank moeten geven. Dit past men bijvoorbeeld in de spraakherkenning toe, waar verschillende klinkers duidelijk verschillende patronen van intensiteiten voor het frequentie spectrum hebben. De frequenties met de hoogste intensiteiten heten *formanten* en bijvoorbeeld de afstand tussen de twee laagste formanten is een belangrijk kenmerk om klinkers te onderscheiden. Bij tweeklanken laat zich goed zien hoe de formanten over de tijd veranderen.



Figuur II.1: Formant spectra voor de klinkers `/a/` en `/oe/`



Figuur II.2: Formant spectra voor de klinker `/i/` en de tweeklank `/o-i/`

7.1 Periodieke functies

Om 'golvende fenomenen' (zo als trillingen) door een model te kunnen beschrijven, hebben we 'golvende functies' nodig, en daarbij denken we natuurlijk aan zo iets als de cosinus of sinus functies. Maar bij de sinus en cosinus heeft de golf een bepaalde vorm, om ook naar anders gevormde golven te kunnen kijken, spreken we algemeen van *periodieke functies*. Hiermee bedoelen we dat een functie zich naar een zeker interval weer herhaald.

Definitie: Een functie $f(t)$ heet *periodiek met periode* L als $f(t+L) = f(t)$ voor alle t.

Bij periodieke functies heet de variabel meestal t omdat we hierbij aan de *tijd* denken.

Bijvoorbeeld zijn $\cos(t)$ en $\sin(t)$ functies met periode 2π. Maar ook $\sin(2t)$ heeft periode 2π, de golven zijn bij deze functie half zo lang en de eigenlijke periode is π, maar dan is de functie natuurlijk ook periodiek met periode 2π. Algemener zijn alle functies $\cos(kt)$, $\sin(kt)$ met $k = 0, 1, 2 \ldots$ periodiek met periode 2π. Deze functies kunnen we natuurlijk ook nog met factoren (de amplitude) vermenigvuldigen en bij elkaar optellen, dit geeft dan periodieke functies zo als

$$f(t) = 5\sin(t) + 3\cos(t) - 2\sin(3t) + \sin(4t).$$

Nu komt er een roekeloze gedachte aan: Bij gewone functies hebben we gezien dat we deze goed door veeltermen kunnen benaderen, bijvoorbeeld door de Taylor veelterm van zekere orde of door interpolatie. We hebben dus ingewikkelde functies beschreven door een lineaire combinatie van de heel eenvoudige functies 1, x, x^2, x^3 enzovoorts. Het idee is nu of we niet periodieke functies goed kunnen benaderen door een lineaire combinatie van $\cos(kt)$ en $\sin(kt)$. Het antwoord hierop is een duidelijk 'ja' en we zullen nu toelichten dat dit eigenlijk een vraagstelling uit de Lineaire Algebra is.

7.2 Trigonometrische benadering

In Wiskunde 1 hebben we naar *orthogonale projecties* gekeken om de beste benadering van een punt in een deelruimte te vinden. Bijvoorbeeld wilden we een punt P in het 2-dimensionale vlak benaderen door een punt op een gegeven lijn, en het was bijna vanzelfsprekend dat de beste benadering (het punt op de lijn het dichtst bij P) de orthogonale projectie van P op de lijn was. Dit concept gaan we nu op de periodieke functies toepassen en het pakt nu goed uit dat we toen ook naar algemenere vectorruimten dan 2- en 3-dimensionale hebben gekeken.

Merk op: In deze les gaan we alleen maar periodieke functies met periode 2π bekijken. De uitbreiding van de theorie tot functies met een willekeurige periode L is echter geen probleem en leidt uiteindelijk tot de Fourier transformatie. Dit gaan we in de volgende les behandelen.

De periodieke functies met periode 2π vormen een vectorruimte V met optelling $(f+g)(t) = f(t)+g(t)$ en vermenigvuldiging met factoren $(cf)(t) = c\cdot f(t)$. De periodieke functies zijn dus de *vectoren* in V. We hebben boven al een paar vectoren in deze vectorruimte opgenoemd, namelijk $\cos(kt)$ en $\sin(kt)$ voor $k \in \mathbb{N}$. De nulvector is de 0-functie $\sin(0\cdot t)$ en men vindt alle constante functies als $c \cdot \cos(0 \cdot t)$.

Een belangrijk feit is nu dat de genoemde 'vectoren' lineair onafhankelijk zijn, d.w.z.:

$$a_0 + \sum_{k=1}^{n}(a_k \cos(kt) + b_k \sin(kt)) = 0 \text{ voor alle } t \;\Rightarrow a_k = b_k = 0 \text{ voor alle } k.$$

Het bewijs hiervan is niet erg moeilijk, maar we slaan het even over. Later zullen we namelijk aantonen dat deze vectoren *loodrecht* op elkaar staan en hieruit volgt in het bijzonder dat ze lineair onafhankelijk zijn.

> Omdat de functies $\cos(kt)$ en $\sin(kt)$ voor $k \geq 0$ lineaire onafhankelijk zijn, hebben we het bij de vectorruimte V met een vectorruimte van oneindige dimensie te maken, maar daar hoeven we niet van te schrikken. Ook de veeltermfuncties $1, x, x^2, x^3, \ldots$ vormen de basis van een oneindig-dimensionale vectorruimte, en die lijkt heel gewoon.

Het plan is nu, de vectoren uit V te benaderen door lineaire combinaties van $\cos(kt)$ en $\sin(kt)$, dus door vectoren in de deelruimte

$$U := \langle \cos(kt), \sin(lt) \mid k, l \in \mathbb{N}, l > 0 \rangle.$$

Omdat $\cos(kt)$ en $\sin(kt)$ trigonometrische functies zijn, noemt men dit ook een *trigonometrische benadering*.

We weten uit Wiskunde 1 dat we de beste benadering van een vector in een deelruimte vinden door een orthogonale projectie, maar hiervoor moeten we wel kunnen zeggen, wanneer twee periodieke functies *loodrecht* op elkaar staan. Dit hadden we altijd met behulp van een *inproduct* uitgedrukt en voor de periodieke functies definiëren we een inproduct als volgt:

$$\Phi(f(t), g(t)) := \int_{-\pi}^{\pi} f(t) \cdot g(t)\, dt.$$

We moeten natuurlijk na gaan dat dit inderdaad een inproduct is, maar gelukkig volgt de *bilineariteit* rechtstreeks uit de eigenschappen van de integraal:

(i) $\Phi(f(t), g(t)) = \Phi(g(t), f(t))$ (symmetrie)

(ii) $\Phi(f(t) + g(t), h(t)) = \Phi(f(t), h(t)) + \Phi(g(t), h(t))$ (optellen)

(iii) $\Phi(cf(t), g(t)) = c \cdot \Phi(f(t), g(t))$ (vermenigvuldigen met een factor).

Verder moet het inproduct positief definiet zijn, d.w.z. er moet gelden:

(i) $\Phi(f(t), f(t)) \geq 0$ voor alle periodieke functies $f(t)$;

(ii) $\Phi(f(t), f(t)) = 0$ alleen maar voor de 0-functie $f(t)$ met $f(t) = 0$ voor alle t.

Het eerste punt zien we makkelijk in: $\Phi(f(t), f(t)) := \int_{-\pi}^{\pi} f(t)^2\, dt \geq 0$, want $f(t)^2 \geq 0$ voor alle t en de integraal over een niet-negatieve functie is niet negatief.

Om echter te kunnen concluderen dat alleen maar voor de 0-functie geldt dat $\Phi(f(t), f(t)) = 0$, moeten we nog iets over de periodieke functies veronderstellen. We hebben bijvoorbeeld problemen met functies $f(t)$ die overal 0 zijn behalve in een paar geïsoleerde punten. Voor dit soort functies is de integraal over $f(t)^2$ namelijk wel 0, terwijl het niet de 0-functies zijn. Om dit

soort pathologische gevallen uit te sluiten, veronderstellen we dat onze functie *stuksgewijs continu* zijn:

Definitie: Een functie $f(t)$ heet *stuksgewijs continu* op het interval $[a, b]$ als het interval zich in eindig veel stukken laat opsplitsen waarop de functie continu is, d.w.z. als er punten $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ zijn, zo dat $f(t)$ continu op elk van de intervallen $[x_{i-1}, x_i]$ is.

Een stuksgewijs continue periodieke functie is dus (op een eindig interval) continu tot op een eindig aantal sprongen na.

> Als men de voorwaarde dat de periodieke functies stuksgewijs continu moeten zijn te sterk vindt, moet men een alternatieve aanpak kiezen. De functies $f(t)$ waarvoor de integraal over $f(t)^2$ gelijk aan 0 is, worden dan tot de 0-functie *verklaart*. Dit geeft de theorie van *Lebesgue integralen* waarbij men functies met elkaar identificeert die *bijna overal* gelijk zijn. De term 'bijna overal' heeft hierbij een precies gedefinieerde betekenis, namelijk dat de uitzonderingen een verzameling van *maat* 0 zijn. Voor een verzameling (bijvoorbeeld een interval) met een gelijkverdeelde kansverdeling heeft een deelverzameling maat 0 als de kans voor deze deelverzameling 0 is. Op het interval $[-\pi, \pi]$ geldt dit bijvoorbeeld voor alle eindige verzamelingen van punten, maar ook voor de verzameling van rationale getallen die in dit interval liggen.

Om goed naar orthogonale projecties te kunnen kijken, hebben we een orthogonale, of beter nog een orthonormale basis nodig. Een basis heet *orthogonaal* als $\Phi(v, w) = 0$ voor elk paar $v \neq w$ van basis vectoren. Als verder ook nog $\Phi(v, v) = 1$ voor alle basis vectoren, heet de basis *orthonormaal*. Algemeen definiëren we de *lengte* van een vector als $\sqrt{\Phi(v, v)}$. Voor een periodieke functie $f(t)$ is de lengte dus gedefinieerd als

$$\|f(t)\| = \sqrt{\Phi(f(t), f(t))} = \left(\int_{-\pi}^{\pi} f(t)^2 \, dt \right)^{\frac{1}{2}}.$$

We hebben inmiddels gezien hoe handig de complexe exponentiële functie is om uitspraken over de cosinus en sinus functies te bewijzen. Dit geldt ook voor het berekenen van de inproducten $\Phi(\cos(kt), \sin(lt))$. We zullen zien dat het stelsel $(\cos(kt), \sin(lt))$ al een orthogonaal stelsel is en dat de inproducten er als volgt uit zien:

$$\int_{-\pi}^{\pi} \cos(kt) \sin(lt) \, dt = 0 \qquad \text{voor alle } k, l \in \mathbb{N}$$

$$\int_{-\pi}^{\pi} \cos(kt) \cos(lt) \, dt = \begin{cases} 0 & \text{als } k \neq l \\ \pi & \text{als } k = l > 0 \\ 2\pi & \text{als } k = l = 0 \end{cases}$$

$$\int_{-\pi}^{\pi} \sin(kt) \sin(lt) \, dt = \begin{cases} 0 & \text{als } k \neq l \\ \pi & \text{als } k = l > 0 \end{cases}$$

Bewijs: Uit het feit dat de complexe exponentiële functie een periode van

$2\pi i$ heeft, volgt dat

$$\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt = \int_{-\pi}^{\pi} e^{i(k-l)t}\ dt = \begin{cases} 0 & \text{als } k \neq l \\ 2\pi & \text{als } k = l \end{cases}$$

want voor $k \neq l$ is $\frac{1}{i(k-l)}e^{i(k-l)t}$ een primitieve van $e^{i(k-l)t}$ en er geldt dus $\int_{-\pi}^{\pi} e^{i(k-l)t}\ dt = \frac{1}{i(k-l)}e^{i(k-l)t}\big|_{-\pi}^{\pi} = \frac{1}{i(k-l)}e^{i(k-l)\pi} - \frac{1}{i(k-l)}e^{i(k-l)(-\pi)} = 0$ (want $e^{i\pi} = e^{i(-\pi)}$). Voor $k = l$ is $e^{i(k-l)t} = 1$ en $\int_{-\pi}^{\pi} 1\ dt = 2\pi$.

Aan de andere kant is

$$\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt = \int_{-\pi}^{\pi} (\cos(kt) + i\sin(kt))(\cos(-lt) + i\sin(-lt))\ dt$$
$$= \int_{-\pi}^{\pi} (\cos(kt) + i\sin(kt))(\cos(lt) - i\sin(lt))\ dt$$
$$= \int_{-\pi}^{\pi} \cos(kt)\cos(lt) + \sin(kt)\sin(lt)\ dt + i\int_{-\pi}^{\pi} \cos(lt)\sin(kt) - \cos(kt)\sin(lt)\ dt.$$

We kijken eerst naar het reële deel $\int_{-\pi}^{\pi} \cos(kt)\cos(lt) + \int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt$ hiervan:
Er geldt $\cos(kt) = \sin(kt + \frac{\pi}{2})$, dus

$$\int_{-\pi}^{\pi} \cos(kt)\cos(lt)\ dt = \int_{-\pi}^{\pi} \sin(kt + \frac{\pi}{2})\sin(lt + \frac{\pi}{2})\ dt$$
$$= \int_{-\frac{\pi}{2}}^{\frac{3\pi}{2}} \sin(kt)\sin(lt)\ dt = \int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt$$

omdat we over een volle periode integreren. Hieruit volgt

$$\int_{-\pi}^{\pi} \cos(kt)\cos(lt) + \int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt = 2\int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt$$
$$= 2\int_{-\pi}^{\pi} \cos(kt)\cos(lt)\ dt.$$

(1) Voor $k \neq l$ is $\Re(\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt) = 0$, dus

$$0 = \int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt = \int_{-\pi}^{\pi} \cos(kt)\cos(lt)\ dt.$$

(2) Voor $k = l \neq 0$ volgt uit $\Re(\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt) = 2\pi$ dat

$$\int_{-\pi}^{\pi} \sin(kt)\sin(lt)\ dt = \int_{-\pi}^{\pi} \cos(kt)\cos(lt)\ dt = \pi.$$

(3) Voor $k = l = 0$ berekenen we heel eenvoudig dat $\int_{-\pi}^{\pi} \cos(0\cdot t)\cos(0\cdot t)\ dt = \int_{-\pi}^{\pi} 1\ dt = 2\pi$.

Hetzelfde trucje passen we nu op het imaginaire deel van $\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt$ toe, dus op $\int_{-\pi}^{\pi}\cos(lt)\sin(kt) - \int_{-\pi}^{\pi}\cos(kt)\sin(lt)\ dt$:
Met $\cos(kt) = \sin(kt + \frac{\pi}{2})$ en $\sin(lt) = -\cos(lt + \frac{\pi}{2})$ volgt

$$\begin{aligned}\int_{-\pi}^{\pi}\cos(kt)\sin(lt)\ dt &= -\int_{-\pi}^{\pi}\sin(kt+\frac{\pi}{2})\cos(lt+\frac{\pi}{2})\ dt\\ &= -\int_{-\frac{\pi}{2}}^{\frac{3\pi}{2}}\sin(kt)\cos(lt)\ dt = -\int_{-\pi}^{\pi}\sin(kt)\cos(lt)\ dt\end{aligned}$$

omdat we weer over een volle periode integreren. We hebben dus

$$\begin{aligned}\int_{-\pi}^{\pi}\cos(lt)\sin(kt)\ dt - \int_{-\pi}^{\pi}\cos(kt)\sin(lt)\ dt &= 2\int_{-\pi}^{\pi}\cos(lt)\sin(kt)\ dt\\ &= -2\int_{-\pi}^{\pi}\cos(kt)\sin(lt)\ dt.\end{aligned}$$

Maar $\Im(\int_{-\pi}^{\pi} e^{ikt}e^{-ilt}\ dt) = 0$, dus hebben we

$$0 = \int_{-\pi}^{\pi}\cos(lt)\sin(kt)\ dt \text{ voor alle } k, l.$$

We hebben dus bewezen dat de verzameling

$$B := \{1(=\cos(0\cdot t)), \cos(kt), \sin(lt) \mid k, l \geq 1\}$$

een orthogonaal stelsel is met

$$\Phi(1,1) = 2\pi, \quad \Phi(\cos(kt),\cos(kt)) = \pi, \quad \Phi(\sin(kt),\sin(kt)) = \pi.$$

We kunnen ook met behulp van een paar handige optelthеorema's zien dat de functies een orthogonaal stelsel vormen. Deze bewijzen we wederom het makkelijkst met behulp van de complexe exponentiële functie. Bijvoorbeeld is $\cos(kt)\cos(lt) = \frac{e^{ikt}+e^{-ikt}}{2}\cdot\frac{e^{ilt}+e^{-ilt}}{2} = \frac{1}{4}(e^{i(k+l)t} + e^{i(k-l)t} + e^{i(-k+l)t} + e^{i(-k-l)t}) = \frac{1}{2}(\frac{e^{i(k+l)t}+e^{-i(k+l)t}}{2} + \frac{e^{i(k-l)t}+e^{-i(k-l)t}}{2}) = \frac{1}{2}(\cos((k+l)t) + \cos(k-l)t)$. Maar de integraal $\int_{-\pi}^{\pi}\cos((k\pm l)t)\ dt$ kunnen we natuurlijk heel eenvoudig uitrekenen. Op dezelfde manier vindt men de optelthеorema's $\sin(kt)\sin(lt) = \frac{1}{2}(\cos((k-l)t) + \cos(k+l)t)$ en $\sin(kt)\cos(lt) = \frac{1}{2}(\sin((k+l)t) + \sin(k-l)t)$.

De Fourier reeks

Nu dat we weten dat de elementen van de basis B een orthogonaal stelsel vormen (dus dat ze alle loodrecht op elkaar staan) kunnen we ook projecties in de deelruimte U opgespannen door $\cos(kt)$ en $\sin(lt)$ berekenen. In Wiskunde 1 hadden we gezien, dat de projectie van een vector v in een deelruimte met orthogonale basis $(v_1, \ldots, v_n)$ gegeven is door

$$v_{\|} = c_1v_1 + \ldots + c_nv_n = \sum_{k=1}^{n} c_kv_k,$$

waarbij de coëfficiënten c_k gegeven zijn door

$$c_k = \frac{\Phi(v, v_k)}{\|v_k\|^2}.$$

Voor een ortho*normaal* stelsel zou $\|v_k\|^2 = 1$ en dus $c_k = \Phi(v, v_k)$ gelden, maar onze basis vectoren hebben lengte $\sqrt{\pi}$ of $\sqrt{2\pi}$.

Dat onze basis oneindig veel elementen bevat, heeft tot gevolg dat we de projectie als een oneindige reeks moeten schrijven. Dit is verder geen probleem, we zien deze reeks (net zo als de Taylor reeks) als de limiet $n \to \infty$ van de som over de eerste n termen. We moeten dan (in principe) wel na gaan of de reeks inderdaad convergeert.

De conclusie is nu dat we een periodieke functie $f(t)$ met periode 2π kunnen benaderen door de projectie

$$f_{\|}(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)$$

met

$$a_k = \frac{\Phi(f(t), \cos(kt))}{\|\cos(kt)\|^2} = \frac{1}{\pi}\int_{\pi}^{\pi} f(t)\cos(kt)\, dt \text{ voor } k = 0, 1, 2, 3, \ldots$$
$$b_k = \frac{\Phi(f(t), \sin(kt))}{\|\sin(kt)\|^2} = \frac{1}{\pi}\int_{\pi}^{\pi} f(t)\sin(kt)\, dt \text{ voor } k = 1, 2, 3, \ldots$$

De conventie de eerste coëfficiënt als $\frac{a_0}{2}$ te schrijven, zorgt ervoor dat de algemene formule voor de a_k ook voor a_0 geldt.

Definitie: De reeks $f_{\|}(t)$ heet de *Fourier reeks* van $f(t)$ en de coëfficiënten a_k, b_k heten de *Fourier coëfficiënten* van $f(t)$. De naam wijst op Jean Baptiste Joseph Fourier (1768-1830), die (in het kader van de zogeheten *hittevergelijking* als eerste de wiskundige theorie van trigonometrische reeksen voor periodieke functies heeft ontwikkeld.

7.3 Eigenschappen van de Fourier reeks

We moeten nu twee belangrijke vragen over de Fourier reeks van een functie beantwoorden:

(1) Wanneer is de Fourier reeks een convergente reeks?

(2) Als de Fourier reeks convergeert, is de limiet dan ook de goede functie $f(t)$?

Het antwoord op beide vragen geeft de volgende beroemde stelling:

Stelling van Dirichlet: Voor een periodieke functie $f(t)$ met periode 2π convergeert de Fourier reeks, als $f(t)$ aan de volgende voorwaarden voldoet:

(a) Het interval $[-\pi, \pi]$ laat zich in eindig veel deelintervallen splitsen waarop $f(t)$ continu en monotoon (stijgend of dalend) is.

(b) In een punt t_0 waar $f(t)$ niet continu is, bestaan de rechtszijdige limiet $f_+(t_0) := \lim_{t\to t_0^+} f(t)$ en de linkszijdige limiet $f_-(t_0) := \lim_{t\to t_0^-} f(t)$ (maar ze zijn niet gelijk).

In de punten waar $f(t)$ continu is, convergeert onder deze voorwaarden de Fourier reeks inderdaad tegen de goede waarde $f(t)$ en in een punt t_0 waar $f(t)$ niet continu is (dus een sprong heeft) convergeert de Fourier reeks tegen het gemiddelde van de rechts- en de linkszijdige limiet, dus tegen $\frac{1}{2}(f_+(t_0)+f_-(t_0))$.

Deze opmerkelijke stelling zegt in het bijzonder dat de projectie van een periodieke functie in de deelruimte U opgespannen van de cosinus en sinus functies in de limiet weer de functie geeft. We zeggen daarom, dat de deelruimte *dicht* ligt in de hele vectorruimte van periodieke functies. Dit is analoog met het feit, dat je elk reëel getal willekeurig goed kunt benaderen met rationale getallen (breuken), men zegt ook hier dat de rationale getallen dicht in de reële getallen liggen. Dit betekent echter niet dat alle periodieke functies in de deelruimte U liggen, want hiervoor zijn alleen maar eindige lineaire combinaties toegestaan en geen oneindige reeksen of limieten.

We kunnen zelfs de fout afschatten, die we maken als we de Fourier reeks naar een aantal termen afbreken (net zo als we de Taylor reeks naar een paar termen afbreken en een functie door een Taylor veelterm benaderen). Er geldt namelijk de **Parseval identiteit** die in principe uit het feit volgt dat we het kwadraat van de lengte van een lineaire combinatie van een orthogonaal stelsel berekenen als

$$\begin{aligned}\|c_1v_1+\ldots+c_nv_n\|^2 &= \Phi(c_1v_1+\ldots+c_nv_n, c_1v_1+\ldots+c_nv_n)\\ &= c_1^2\|v_1\|^2+\ldots+c_n^2\|v_n\|^2.\end{aligned}$$

Als we dit toepassen op een periodieke functie $f(t)$ met Fourier reeks $\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k\cos(kt)+b_k\sin(kt)$ krijgen we

$$\frac{a_0^2}{2}+\sum_{k=1}^{\infty}(a_k^2+b_k^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)^2\,dt.$$

Als we de reeks na n termen afbreken, kunnen we dus de kwadratische fout afschatten door $\sum_{k=n+1}^{\infty}(a_k^2+b_k^2)$.

We merken nog op dat de voorwaarden in de stelling van Dirichlet geen noodzakelijke voorwaarden zijn, d.w.z. er zijn ook functies die niet aan deze voorwaarden voldoen, maar waarvoor de Fourier reeks wel tegen de goede functie convergeert. Aan de andere kant zijn er zelfs continue functies waarvoor de Fourier reeks niet tegen de juiste functie convergeert. Het probleem om een precieze karakterisatie van de functies te geven, waarvoor de Fourier reeks tegen de goede functie convergeert, is nog steeds open!

Uit de symmetrie eigenschappen van cosinus en sinus kunnen we heel eenvoudig een aantal belangrijke conclusies trekken. We weten dat $\cos(t)$ een even

functie is, dus $\cos(-t) = \cos(t)$. Evenzo is $\sin(t)$ een oneven functie, want $\sin(-t) = -\sin(t)$. De integraal $\int_{-\pi}^{\pi} f(t)\cos(kt)\ dt$ kunnen we daarom makkelijk iets anders schrijven, namelijk als

$$\int_{-\pi}^{\pi} f(t)\cos(kt)\ dt = \int_0^{\pi} f(t)\cos(kt)\ dt + \int_0^{\pi} f(-t)\cos(-kt)\ dt$$
$$= \int_0^{\pi} f(t)\cos(kt)\ dt + \int_0^{\pi} f(-t)\cos(kt)\ dt.$$

Net zo is

$$\int_{-\pi}^{\pi} f(t)\sin(kt)\ dt = \int_0^{\pi} f(t)\sin(kt)\ dt + \int_0^{\pi} f(-t)\sin(-kt)\ dt$$
$$= \int_0^{\pi} f(t)\sin(kt)\ dt - \int_0^{\pi} f(-t)\sin(kt)\ dt.$$

Als nu $f(t)$ een even functie is, dan is $f(-t) = f(t)$ en dus

$$\int_{-\pi}^{\pi} f(t)\sin(kt)\ dt = \int_0^{\pi} f(t)\sin(kt)\ dt - \int_0^{\pi} f(t)\sin(kt)\ dt = 0.$$

Hieruit volgt dat in de Fourier reeks van $f(t)$ alle coëfficiënten b_k gelijk aan 0 zijn en $f(t)$ dus een lineaire combinatie van alleen maar cosinus functies is (die precies de even functies in de basis van de deelruimte U zijn). Dit is analoog met het feit dat de Taylor reeks van een even functie alleen maar even machten z^{2n} bevat.

Omgekeerd geldt voor een oneven functie $f(t)$ dat $f(-t) = -f(t)$. In dit geval is

$$\int_{-\pi}^{\pi} f(t)\cos(kt)\ dt = \int_0^{\pi} f(t)\cos(kt)\ dt - \int_0^{\pi} f(t)\cos(kt)\ dt = 0$$

en zijn dus alle coëfficiënten a_k in de Fourier reeks gelijk aan 0. Ook dit is een analogie met de Taylor reeksen voor oneven functies, want deze bevatten alleen maar oneven termen z^{2n+1}.

We hebben dus de volgende belangrijke stelling ingezien:

(i) Een even functie $f(t)$ met $f(-t) = f(t)$ heeft een Fourier reeks van de vorm
$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kt).$$

(ii) Een oneven functie $f(t)$ met $f(-t) = -f(t)$ heeft een Fourier reeks van de vorm
$$\sum_{k=1}^{\infty} b_k \sin(kt).$$

7.4 Fase verschuivingen

We kunnen ons afvragen hoe het komt, dat we alleen maar functies $\cos(kt)$ die in het nulpunt een maximum hebben, en functies $\sin(kt)$ die in het nulpunt een nulpunt hebben, nodig hebben om algemene periodieke functies te kunnen beschrijven. Hoe zit het bijvoorbeeld met een zuivere sinus functie die langs de x-as verschoven is, dus met $f(t) = \sin(kt + \varphi)$? Dit is een belangrijk punt, want we kunnen niet ervan uitgaan dat iedere golvende beweging op het tijdstip $t = 0$ of een nuldoorgang of een maximum heeft. De verschuiving φ noemt men ook de *fase* van de functie. De (misschien verrassende) oplossing is dat we de functie $f(t) = \sin(kt + \varphi)$ kunnen schrijven als lineaire combinatie van $\sin(kt)$ en $\cos(kt)$.

Uit het vergelijken van de imaginaire delen van $e^{i(kt+\varphi)}$ en $e^{ikt} \cdot e^{i\varphi}$ hadden we al eerder het optelthеorema $\sin(kt + \varphi) = \sin(\varphi)\cos(kt) + \cos(\varphi)\sin(kt)$ gevonden. Maar dit zegt precies dat we een functie $A\sin(kt+\varphi)$ kunnen schrijven als lineaire combinatie van $\cos(kt)$ en $\sin(kt)$, namelijk als:

$$A\sin(kt + \varphi) = a\cos(kt) + b\sin(kt) \text{ met } a = A\sin(\varphi) \text{ en } b = A\cos(\varphi).$$

Met behulp van de relaties $A = \sqrt{a^2 + b^2}$ en $\tan(\varphi) = \frac{a}{b}$ kunnen we makkelijk tussen de twee schrijfwijzen hen en weer gaan.

We hebben dus gezien dat het equivalent is een functie als lineaire combinatie $a\cos(kt) + b\sin(kt)$ te schrijven of als $A\sin(kt+\varphi)$, in beide gevallen zijn er drie parameters nodig: de amplituden a en b en de frequentie k óf de amplitude A, de fase φ en de frequentie k.

Dit betekent, dat we ook de Fourier reeks van een periodieke functie $f(t)$ met Fourier coëfficiënten a_k, b_k op een andere manier kunnen schrijven, namelijk als

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} A_k \sin(kt + \varphi_k) \text{ met } A_k = \sqrt{a_k^2 + b_k^2} \text{ en } \tan(\varphi_k) = \frac{a_k}{b_k}.$$

Dit noemt men ook de *spectrale* schrijfwijze van de Fourier reeks van $f(t)$. De rij $A_1, A_2, \ldots$ heet dan het *amplitude spectrum* en de rij $\varphi_1, \varphi_2, \ldots$ het *fase spectrum* van $f(t)$.

7.5 Complexe schrijfwijze

We hebben nu al een paar keer gezien dat het soms handig is de cosinus en sinus functies gezamenlijk door de complexe exponentiële functie te beschrijven. Dit geldt ook voor de Fourier reeks!

We weten dat

$$\begin{aligned} a_k\cos(kt) &= a_k\frac{e^{ikt} + e^{-ikt}}{2} = \frac{a_k}{2}e^{ikt} + \frac{a_k}{2}e^{-ikt} \quad \text{en} \\ b_k\sin(kt) &= b_k\frac{e^{ikt} - e^{-ikt}}{2i} = \frac{b_k}{2}(-i)e^{ikt} + i\frac{b_k}{2}e^{-ikt}. \end{aligned}$$

Hieruit volgt

$$a_k \cos(kt) + b_k \sin(kt) = \frac{1}{2}(a_k - ib_k)e^{ikt} + \frac{1}{2}(a_k + ib_k)e^{-ikt}.$$

We definiëren nu $c_0 := \frac{a_0}{2}$ en voor $k \geq 1$ definiëren we

$$c_k := \frac{1}{2}(a_k - ib_k) \quad \text{en} \quad c_{-k} := \overline{c_k} = \frac{1}{2}(a_k + ib_k).$$

Dan kunnen we de Fourier reeks van $f(t)$ herschrijven als

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} c_k e^{ikt} + c_{-k} e^{-ikt} = \sum_{k=-\infty}^{\infty} c_k e^{ikt}.$$

Voor de coëfficiënten c_k met $k \geq 1$ geldt:

$$\begin{aligned} c_k &= \frac{1}{2}(a_k - ib_k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)(\cos(kt) - i\sin(kt))\, dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)(\cos(-kt) + i\sin(-kt))\, dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}\, dt. \end{aligned}$$

Maar de relatie $c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}\, dt$ geldt ook voor $k < 0$, want

$$\begin{aligned} c_{-k} &= \frac{1}{2}(a_k + ib_k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)(\cos(kt) + i\sin(kt))\, dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{ikt}\, dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-i(-k)t}\, dt. \end{aligned}$$

We definiëren dus algemeen voor $k \in \mathbb{Z}$:

$$c_k := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}\, dt$$

en noemen dit de k-de *complexe Fourier coëfficiënt* van $f(t)$. Merk op dat deze formule ook voor $k = 0$ geldt, want $c_0 = \frac{a_0}{2}$ en a_0 was gedefinieerd door $a_0 := \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\, dt$.

De complexe schrijfwijze van de Fourier reeks van een periodieke functie $f(t)$ is dus:

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{ikt} \text{ met } c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}\, dt.$$

Voor reële functies $f(t)$ hebben we gezien dat $c_{-k} = \overline{c_k}$, maar we kunnen de complexe vorm van de Fourier reeks net zo goed op complexe functies $f(z)$ toepassen die langs de reële as periodiek zijn. De reële functies zijn echter de meest belangrijke toepassingen van de Fourier reeksen.

We hadden de complexe schrijfwijze van de Fourier reeks ook rechtstreeks middels het concept van de projectie op een deelruimte kunnen afleiden: De functies e^{ikt} met $k \in \mathbb{Z}$ zijn orthogonaal ten opzichte van het inproduct $\Psi(f(z), g(z)) := \int_{-\pi}^{\pi} f(z)\overline{g(z)}\, dz$. Merk op dat de complexe conjugatie bij de tweede factor nodig is om het inproduct positief definiet te hebben, want dan wordt $\Psi(f(z), f(z)) = \int_{-\pi}^{\pi} |f(z)|^2\, dz$. Met betrekking tot dit inproduct geldt $\Psi(e^{ikt}, e^{ikt}) = 2\pi$, dus vinden we de coëfficiënt c_k van e^{ikt} als $c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\overline{e^{ikt}}\, dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}\, dt$.

7.6 Belangrijke voorbeelden

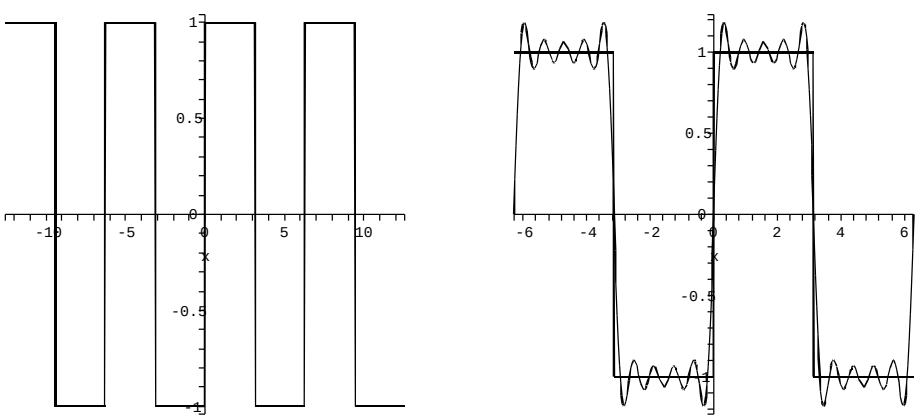
We zullen de theorie van Fourier reeksen nu op een aantal belangrijke periodieke functies toepassen. Hierbij berekenen we voor een gegeven functie de Fourier coëfficiënten en vergelijken de functie met de benadering door de eerste termen van de Fourier reeks. Het produceren van een zekere golfvorm middels lineaire combinaties van cosinus en sinus functies noemt men ook *Fourier synthese*. Dit principe wordt bijvoorbeeld in synthesizers toegepast, waar zuivere sinus-golven elektronisch geproduceerd en vervolgens tot periodieke functies met ingewikkeldere golfvormen gecombineerd worden.

Omdat we het over functies met periode 2π hebben, hoeven we de functies alleen maar voor het interval $[-\pi, \pi]$ te definiëren, door verschuiving van dit interval om veelvouden van 2π overdekken we de hele reële as.

De stapfunctie

De stapfunctie is gegeven door

$$f(t) := \begin{cases} -1 & \text{als} \quad -\pi \le t < 0 \\ 1 & \text{als} \quad 0 \le t < \pi. \end{cases}$$



Figuur II.3: Stapfunctie en benadering door Fourier reeks

Omdat $f(t)$ een oneven functie is, zijn de coëfficiënten a_k van $\cos(kt)$ alle gelijk aan 0 en voor de coëfficiënten b_k van $\sin(kt)$ geldt $b_k = 2\int_0^\pi f(t)\sin(kt)\ dt$. Hieruit volgt:

$$\begin{aligned} b_k &= \frac{2}{\pi}\int_0^\pi 1\cdot\sin(kt)\ dt = \frac{2}{\pi}\frac{1}{k}(-\cos(kt)\big|_0^\pi = \frac{2}{\pi}\frac{1}{k}(-\cos(k\pi)+1) \\ &= \begin{cases} \frac{4}{\pi}\frac{1}{k} & \text{als } k \text{ oneven} \\ 0 & \text{als } k \text{ even} \end{cases} \end{aligned}$$

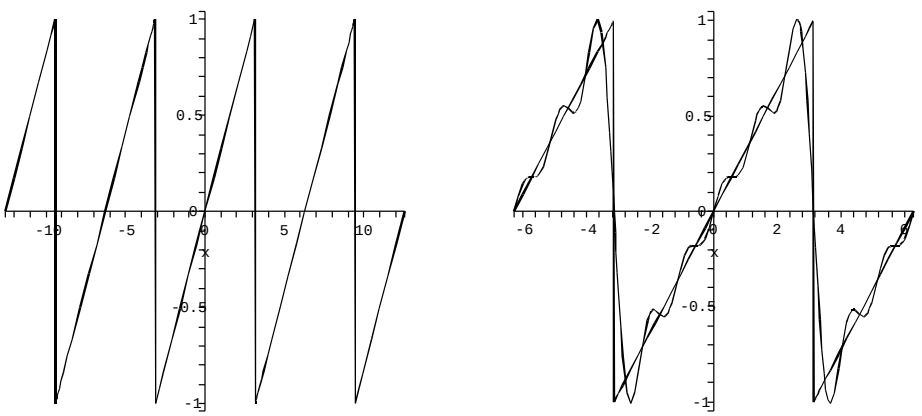
De ontwikkeling van $f(t)$ in een Fourier reeks is dus

$$f(t) = \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{\sin((2k+1)t)}{2k+1} = \frac{4}{\pi}\left(\sin(t) + \frac{\sin(3t)}{3} + \frac{\sin(5t)}{5} + \frac{\sin(7t)}{7} + \dots\right)$$

De zaagfunctie

De zaagfunctie is gedefinieerd door

$$f(t) := \frac{1}{\pi}t.$$



Figuur II.4: Zaagfunctie en benadering door Fourier reeks

Ook de zaagfunctie is een oneven functie, zus zijn ook hier de coëfficiënten van $\cos(kt)$ gelijk aan 0. We moeten de integraal $b_k = \frac{2}{\pi}\int_0^\pi \frac{1}{\pi}t\cdot\sin(kt)\ dt$ berekenen, en hiervoor bepalen we eerst met behulp van partiële integratie een primitieve van $t\cdot\sin(kt)$. Er geldt

$$\int t\sin(kt)\ dt = t\frac{1}{k}(-\cos(kt)) + \int\frac{1}{k}\cos(kt)\ dt = -t\frac{1}{k}\cos(kt) + \frac{1}{k^2}\sin(kt).$$

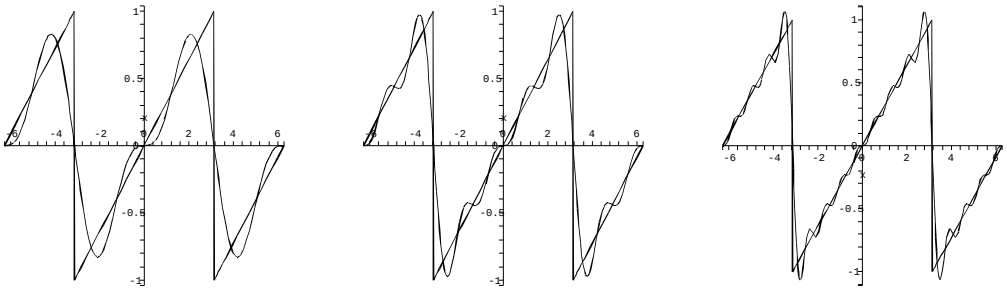
Hieruit volgt:

$$\begin{aligned} b_k &= \frac{2}{\pi}\int_0^\pi \frac{1}{\pi}t\cdot\sin(kt)\ dt = \frac{2}{\pi^2}(-t\frac{1}{k}\cos(kt)\big|_0^\pi + \frac{1}{k^2}\sin(kt)\big|_0^\pi) \\ &= \frac{2}{\pi^2}\frac{1}{k}(-\pi\cos(k\pi)) = \frac{2}{\pi}\frac{(-1)^k}{k} \end{aligned}$$

De ontwikkeling van $f(t)$ in een Fourier reeks is dus

$$f(t) = \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\sin(kt) = \frac{2}{\pi}\left(\sin(t) - \frac{\sin(2t)}{2} + \frac{\sin(3t)}{3} - \frac{\sin(4t)}{4} + \ldots\right)$$

Voor de zaagfunctie laten we in de volgende drie plaatjes zien hoe de benadering verbeterd door meer termen van de Fourier reeks erbij te pakken. De drie benaderingen breken de Fourier reeks naar 2, 4 en 8 termen af. Het is duidelijk dat vooral het punt waar de functie niet continu is problemen bij de benadering veroorzaakt.



Figuur II.5: Afbreken van de Fourier reeks naar 2, 4 en 8 termen

Als we de Fourier reeks van de zaagfunctie voor $t = \frac{\pi}{2}$ toepassen, krijgen we $\frac{1}{2} = \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\sin(k\frac{\pi}{2}) = \frac{2}{\pi}(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\ldots)$, want voor even k is $\sin(k\frac{\pi}{2}) = 0$ en voor oneven k is $\sin(k\frac{\pi}{2})$ afwisselend 1 en -1. Hieruit volgt de opmerkelijke formule

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

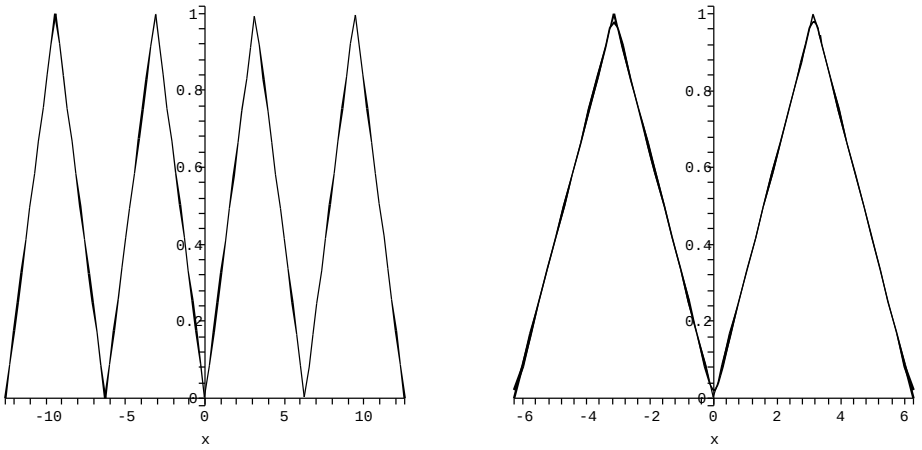
De zigzagfunctie

De zigzagfunctie krijgen we als we de absolute waarde van de zaagfunctie nemen. We hebben dus

$$f(t) := \frac{1}{\pi}|t|.$$

De zigzagfunctie is een even functie, daarom zijn de coëfficiënten b_k van $\sin(kt)$ gelijk aan 0. Ook hier gebruiken we de symmetrie om de integraal te vereenvoudigen, we krijgen dan $a_k = \frac{2}{\pi}\int_0^{\pi}\frac{1}{\pi}t\cdot\cos(kt)\,dt$. We moeten dus een primitieve van $t\cdot\cos(kt)$ bepalen:

$$\int t\cos(kt)\,dt = t\frac{1}{k}\sin(kt) - \int\frac{1}{k}\sin(kt)\,dt = t\frac{1}{k}\sin(kt) + \frac{1}{k^2}\cos(kt).$$



Figuur II.6: Zigzagfunctie en benadering door Fourier reeks

Hieruit volgt:

$$\begin{aligned} a_k &= \frac{2}{\pi}\int_0^\pi \frac{1}{\pi} t\cdot\cos(kt)\,dt = \frac{2}{\pi^2}(t\frac{1}{k}\sin(kt)\big|_0^\pi + \frac{1}{k^2}\cos(kt)\big|_0^\pi) \\ &= \frac{2}{\pi^2}\frac{1}{k^2}(\cos(k\pi)-1) = \begin{cases} -\frac{4}{\pi^2}\frac{1}{k^2} & \text{als } k \text{ oneven} \\ 0 & \text{als } k \text{ even} \end{cases} \end{aligned}$$

Voor a_0 hebben we

$$a_0 = \frac{2}{\pi}\int_0^\pi \frac{1}{\pi} t\,dt = \frac{2}{\pi^2}\frac{1}{2}t^2\big|_0^\pi = 1.$$

De ontwikkeling van $f(t)$ in een Fourier reeks is dus

$$\begin{aligned} f(t) &= \frac{1}{2} - \frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\cos((2k-1)t) \\ &= \frac{1}{2} - \frac{4}{\pi^2}\left(\cos(t) + \frac{\cos(3t)}{3^2} + \frac{\cos(5t)}{5^2} + \dots\right) \end{aligned}$$

Merk op dat men hier veel sneller een goede benadering van $f(t)$ krijgt, omdat de noemers in de Fourier reeks met k^2 groeien.

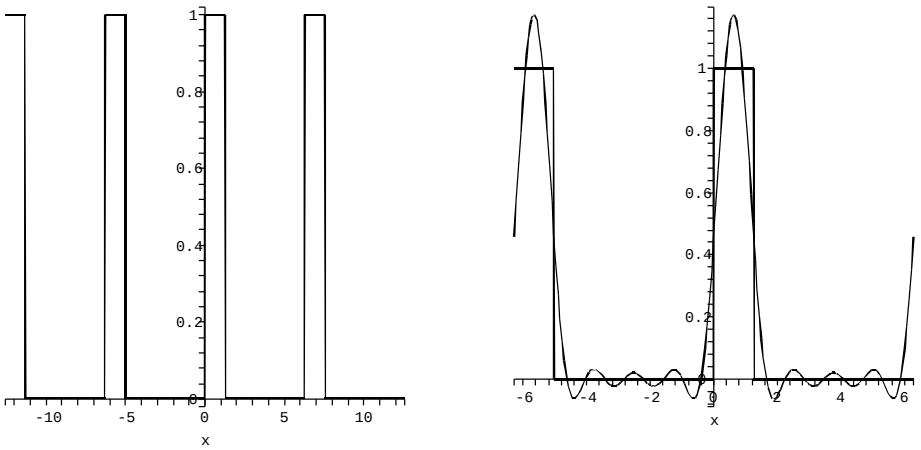
Als we de Fourier reeks van de zigzagfunctie voor $t = 0$ toepassen, krijgen we $0 = \frac{1}{2} - \frac{4}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\cos((2k-1)\cdot 0) = \frac{1}{2} - \frac{4}{\pi^2}(1 + \frac{1}{3^2} + \frac{1}{5^2} + \dots)$ en hieruit volgt de formule

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots$$

De impulsfunctie

De impulsfunctie $f(t)$ heeft op het tijdstip 0 een impuls van lengte a en is verder 0 op het interval $[-\pi, \pi]$. De functie is dus gegeven door

$$f(t) := \begin{cases} 1 & \text{als} \quad 0 \leq t \leq a \\ 0 & \text{als} \quad t \notin [0, a]. \end{cases}$$



Figuur II.7: Impulsfunctie en benadering door Fourier reeks

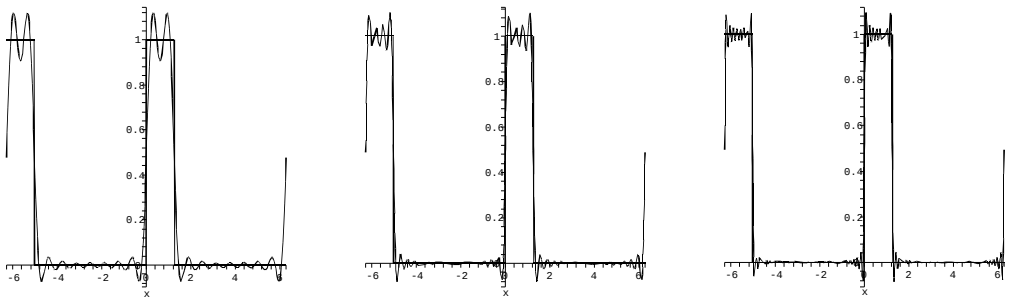
Omdat de functie buiten het interval $[0, a]$ gelijk aan 0 is, hoeven we ook alleen maar over dit deelinterval te integreren. Er geldt:

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_0^a 1 \; dt = \frac{1}{\pi} a \\ a_k &= \frac{1}{\pi} \int_0^a \cos(kt) \; dt = \frac{1}{\pi} \frac{1}{k} \sin(kt)\big|_0^a = \frac{1}{\pi} \frac{\sin(ka)}{k} \\ b_k &= \frac{1}{\pi} \int_0^a \sin(kt) \; dt = \frac{1}{\pi} \frac{1}{k} (-\cos(kt))\big|_0^a = \frac{1}{\pi} \frac{1 - \cos(ka)}{k} \end{aligned}$$

De ontwikkeling van $f(t)$ in een Fourier reeks is dus

$$\begin{aligned} f(t) &= \frac{a}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \left(\frac{\sin(ka)}{k} \cos(kt) + \frac{1 - \cos(ka)}{k} \sin(kt) \right) \\ &= \frac{a}{2\pi} + \frac{1}{\pi} \Bigg(\sin(a) \cos(t) + (1 - \cos(a)) \sin(t) \\ &\qquad + \frac{\sin(2a)}{2} \cos(2t) + \frac{1 - \cos(2a)}{2} \sin(2t) + \dots \Bigg) \end{aligned}$$

Bij de impulsfunctie geven we nog drie benaderingen van hogere graden aan, om te laten zien dat ook een korte impuls goed benaderd wordt.



Figuur II.8: Benaderingen van hogere orde voor de impulsfunctie

Belangrijke begrippen in deze les

- periodieke functies
- trigonometrische benadering
- Fourier reeks, Fourier coëfficiënten
- spectrale schrijfwijze, complexe schrijfwijze
- stapfunctie, zaagfunctie, impulsfunctie

Opgaven

63. We bekijken de functie

$$f(t) := \begin{cases} 0 & \text{als} \quad -\pi \leq t < 0 \\ 1 & \text{als} \quad 0 \leq t < \frac{\pi}{2} \\ -1 & \text{als} \quad \frac{\pi}{2} \leq t < \pi \end{cases}$$

die we door verschuiven om veelvouden van 2π tot een 2π-periodieke functie op $\mathbb{R}$ voortzetten.

(i) Maak een schets van de functie.

(ii) Bepaal de Fourier reeks van de functie

(iii) Maak een schets van de eerste twee benaderingen van $f(t)$ door afbreken van de Fourier reeks.

64. Bereken de Fourier reeks van $f(t) := |\sin(t)|$.

Hint: Met partiële integratie vindt men een primitieve van $\sin(t)\cos(kt)$ als volgt:
$\int \sin(t)\cos(kt)\ dt = -\cos(t)\cos(kt) - \int(-\cos(t))(-k\sin(kt))\ dt$
$= -\cos(t)\cos(kt) - k\int \cos(t)\sin(kt)\ dt$
$= -\cos(t)\cos(kt) - k(\sin(t)\sin(kt) - \int \sin(t)k\cos(kt)\ dt)$
$= -\cos(t)\cos(kt) - k\sin(t)\sin(kt) + k^2\int \sin(t)\cos(kt)\ dt$
$= \frac{1}{k^2-1}(\cos(t)\cos(kt) + k\sin(t)\sin(kt))$.

65. Bereken de Fourier reeks van de functie $f(t) := t^2$ die we van het interval $[-\pi, \pi]$ door verschuiven om veelvouden van 2π op de hele reële as voortzetten.

66. We bekijken de functie

$$f(t) := \begin{cases} 0 & \text{als} \quad -\pi \leq t \leq 0 \\ \sin(t) & \text{als} \quad 0 \leq t \leq \pi \end{cases}$$

die we door verschuiven om veelvouden van 2π tot een 2π-periodieke functie op $\mathbb{R}$ voortzetten.

(i) Maak een schets van de functie.

(ii) Laat zien dat $f(t)$ de Fourier reeks

$$\begin{aligned} &\frac{1}{\pi} + \frac{1}{2}\sin(t) - \frac{2}{\pi}\left(\sum_{k=1}^{\infty} \frac{\cos(2kt)}{(2k)^2 - 1}\right) \\ = &\frac{1}{\pi} + \frac{1}{2}\sin(t) - \frac{2}{\pi}\left(\frac{\cos(2t)}{1 \cdot 3} + \frac{\cos(4t)}{3 \cdot 5} + \frac{\cos(6t)}{5 \cdot 7} + \frac{\cos(8t)}{7 \cdot 9} + \ldots\right) \end{aligned}$$

heeft.

(iii) Concludeer dat

$$\frac{\pi}{4} = \frac{1}{2} + \frac{1}{1 \cdot 3} - \frac{1}{3 \cdot 5} + \frac{1}{5 \cdot 7} - \frac{1}{7 \cdot 9} + \ldots = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{(-1)^k}{(2k)^2 - 1}.$$

(Hint: Vul $t = \frac{\pi}{2}$ in de Fourier reeks in.)

67. Bereken de Fourier reeks van de functie

$$f(t) := \cos(\alpha t) \text{ met } \alpha \neq 0, \pm 1, \pm 2, \ldots$$

die van het interval $[-\pi, \pi]$ door verschuiven tot een periodieke functie met periode 2π voortgezet wordt.

Hint: Er geldt $2\cos(x)\cos(y) = \cos(x - y) + \cos(x + y)$.

Les 8 Fourier transformatie

8.1 Periodieke functies met perioden verschillend van 2π

In de vorige les hebben we naar de Fourier reeksen voor periodieke functies met periode 2π gekeken. De reden hiervoor was, dat we voor deze periode met de cosinus en sinus functies goed bekende voorbeelden hadden. Maar de beperking tot functies met periode 2π is natuurlijk erg kunstmatig, en we zullen de theorie van Fourier reeksen nu uitbreiden op functies met een willekeurige periode L.

Het idee dat hier achter zit is heel eenvoudig: Door een schaling (van de x-as) maken we uit een interval van lengte L een interval van lengte 2π, hiervoor moeten we L met de factor $\frac{2\pi}{L}$ vermenigvuldigen. Als $f(t)$ een periodieke functie met periode L is, dan definiëren we

$$\omega := \frac{2\pi}{L} \quad \text{en} \quad x := \omega t.$$

Loopt nu t over een volledige periode, d.w.z. over een interval van lengte L, dan loopt x over een interval van lengte $\omega L = 2\pi$. We definiëren nu een nieuwe functie $g(t)$ door

$$g(t) := f(\frac{1}{\omega}t), \text{ dus } f(t) = g(\omega t) = g(x)$$

en voor de nieuwe variabel x is $g(x)$ een functie met periode 2π.

Op de functie $g(x)$ kunnen we nu de theorie van Fourier reeksen voor functies met periode 2π toepassen, we krijgen dus

$$g(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \quad \text{met} \quad c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)e^{-ikx}\,dx.$$

Maar omdat $f(t) = g(x)$ en $x = \omega t$, kunnen we dit ook schrijven als

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{i\omega kt}.$$

Let op: In de coëfficiënten c_k mogen we x niet zo maar door ωt vervangen, omdat x hier een integratie variabel is. Dit is dus een echte *substitutie*:

$$x = \omega t, \quad dx = \omega\,dt,$$

De integratie over x loopt van $-\pi$ tot π, dus loopt de integratie voor t van $\frac{1}{\omega}(-\pi) = -\frac{L}{2}$ tot $\frac{1}{\omega}\pi = \frac{L}{2}$. We krijgen dus

$$\begin{aligned} c_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)e^{-ikx}\,dx = \frac{1}{2\pi}\int_{-\frac{L}{2}}^{\frac{L}{2}} g(\omega t)e^{-i\omega kt}\omega\,dt = \frac{\omega}{2\pi}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)e^{-i\omega kt}\,dt \\ &= \frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)e^{-i\omega kt}\,dt. \end{aligned}$$

De Fourier reeks van een functie $f(t)$ met periode L is dus:

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{i\omega kt} \quad \text{met} \quad c_k = \frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)e^{-i\omega kt}\, dt.$$

Op een soortgelijke manier wordt ook de reële versie van de Fourier reeks aangepast, want we schrijven

$$\begin{aligned} f(t) = g(x) &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx) + b_k \sin(kx) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(\omega kt) + b_k \sin(\omega kt) \end{aligned}$$

en weer met de substitutie $x = \omega t$, $dx = \omega\, dt$ vinden we:

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} g(x)\cos(kx)\, dx = \frac{2}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)\cos(\omega kt)\, dt \text{ voor } k \geq 0 \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} g(x)\sin(kx)\, dx = \frac{2}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)\sin(\omega kt)\, dt \text{ voor } k \geq 1. \end{aligned}$$

Hieruit volgt de reële versie van de Fourier reeks voor een functie $f(t)$ met periode L:

$$\begin{aligned} f(t) &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(\omega kt) + b_k \sin(\omega kt) \quad \text{met} \\ a_k &= \frac{2}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)\cos(\omega kt)\, dt, \qquad b_k = \frac{2}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)\sin(\omega kt)\, dt. \end{aligned}$$

De Fourier reeksen voor periodieke functies met periode L kan men ook direct afleiden door de methode van orthogonale projecties aan de periode L aan te passen:

Het inproduct is $\Phi(f(t), g(t)) = \int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)g(t)\, dt$, het orthogonale stelsel is $\{\cos(\omega kt), \sin(\omega lt) \mid k \geq 0, l \geq 0\}$ met $\omega = \frac{2\pi}{L}$ en er geldt $\Phi(\cos(0), \cos(0)) = L$, $\Phi(\cos(\omega kt), \cos(\omega kt)) = \Phi(\sin(\omega kt), \sin(\omega kt)) = \frac{L}{2}$ voor $k \geq 1$.

8.2 Van Fourier reeks naar Fourier integraal

Bij de ontwikkeling van een periodieke functie $f(t)$ met periode L in zijn Fourier reeks is $\omega = \frac{2\pi}{L}$ de basis frequentie en de coëfficiënt c_k geeft de intensiteit van de trilling met frequentie $k\omega$ in de functie $f(t)$ aan. In het bijzonder geven naburige coëfficiënten c_k en c_{k+1} de intensiteiten voor frequenties met een verschil van ω aan. Als we nu naar periodieke functies met verschillende periodes kijken, hangen de afstanden tussen de frequenties die aan de functies bijdragen van de periode af. Als we de periode verdubbelen, moeten we naar frequenties met een

half zo grote afstand kijken. Hoe langer de periode van een functie, hoe meer frequenties in een bepaald interval spelen dus een rol.

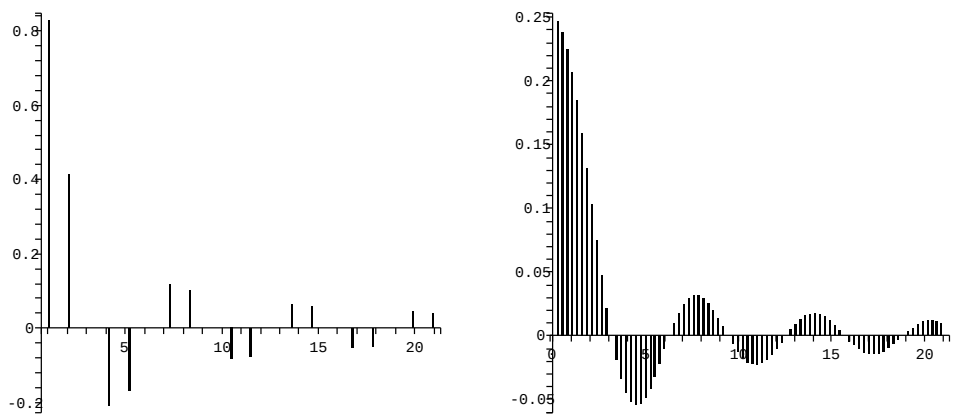
Voorbeeld: We kijken naar een impuls van lengte $2a$ en intensiteit 1 tussen $t = -a$ en $t = a$, die met een periode van L herhaald. Omdat de functie even is, hoeven we alleen maar de coëfficiënten a_k te bepalen, en er geldt:

$$\begin{aligned} a_k &= \frac{2}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)\cos(\omega k t)\,dt = \frac{2}{L}\int_{-a}^{a}\cos(\omega k t)\,dt = \frac{2}{L}\frac{1}{\omega k}\sin(\omega k t)\Big|_{-a}^{a} \\ &= \frac{4}{L}\frac{\sin(\omega k a)}{\omega k} = \frac{4a}{L}\frac{\sin(\omega k a)}{\omega k a}. \end{aligned}$$

Als we nu de afstand tussen de impulsen verdubbelen, dus de periode van L naar $L' = 2L$ verdubbelen, krijgen we nieuwe Fourier coëfficiënten a'_k voor veelvouden van de nieuwe grondfrequentie $\omega' = \frac{1}{2}\omega$, namelijk

$$a'_k = \frac{4a}{L'}\frac{\sin(\omega' k a)}{\omega' k a} = \frac{2a}{L}\frac{\sin(\omega \frac{k}{2} a)}{\omega \frac{k}{2} a}.$$

De coëfficiënt a_k geeft de intensiteit van de trilling met frequentie $\omega k = 2\omega' k = \omega' \cdot 2k$ aan, dus hoort bij de frequentie ωk in de nieuwe Fourier reeks de coëfficiënt a'_{2k}. Er geldt $a'_{2k} = \frac{2a}{L}\frac{\sin(\omega k a)}{\omega k a} = \frac{1}{2}\,a_k$, dus zijn de intensiteiten voor dezelfde frequentie tot op een factor $\frac{1}{2}$ na hetzelfde. Maar tussen twee naburige frequenties ωk en $\omega(k+1)$ voor de impuls met periode L ligt er nu nog de frequentie $\omega\frac{k+(k+1)}{2} = \omega'(2k+1)$ omdat de afstanden tussen de frequenties gehalveerd zijn.



Figuur II.9: Fourier coëfficiënten voor rechthoek impuls met periode L en $4L$.

In Figuur II.9 is het effect van het vergrootten van de periode te zien. De x-as geeft de frequenties aan, en voor de frequenties die aan de Fourier reeks bijdragen is de waarde van de bijhorende coëfficiënt a_k door een verticale lijn aangegeven. Het is duidelijk te zien dat bij de overgang van een periode L naar $4L$ de bijdragende frequenties 4 keer dichter bij elkaar liggen.

Als we de lengte van de periode steeds verder laten groeien, verliest de functie uiteindelijk zijn periodieke karakter en in de limiet kunnen we iedere functie als periodieke functie met periode $L = \infty$ opvatten. Maar de limiet $L \to \infty$ correspondeert met de limiet $\omega \to 0$, dus moeten we voor dit geval frequenties met oneindig kleine afstand bekijken en een intensiteit voor elke frequentie op een continue lijn bepalen. In het rechter plaatje van Figuur II.9 kan men zich dit al goed voorstellen, de verticale lijnen geven al bijna de contour van een continue functie aan.

We gaan nu de overgang van discrete frequenties voor een periodieke functie naar een continu spectrum van frequenties voor een niet-periodieke functie nader bekijken.

Hiervoor schrijven we een functie $f(t)$ met periode L als Fourier reeks:

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{i\omega kt} \text{ met } c_k = \frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(t)e^{-i\omega kt}\, dt \text{ waarbij } \omega = \frac{2\pi}{L}.$$

Eerst vullen we de coëfficiënten c_k in de Fourier reeks in, dit geeft:

$$\begin{aligned} f(t) &= \sum_{k=-\infty}^{\infty} \left(\frac{1}{L}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(\tau)e^{-i\omega k\tau}\, d\tau \right) e^{i\omega kt} \\ &= \sum_{k=-\infty}^{\infty} \left(\frac{1}{2\pi}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(\tau)e^{-i\omega k\tau}\, d\tau \right) e^{i\omega kt}\omega. \end{aligned}$$

Als we nu de limiet $L \to \infty$ bekijken, gaat $\omega \to 0$, dus loopt ωk in steeds kleinere stappen ω van $-\infty$ naar ∞. Uiteindelijk wordt ωk een continue variabel die we u noemen en die in stappen van $\Delta u = \omega$ van $-\infty$ naar ∞ loopt. Elke term in de som wordt dan een term van de vorm

$$\left(\frac{1}{2\pi}\int_{-\frac{L}{2}}^{\frac{L}{2}} f(\tau)e^{-iu\tau}\, d\tau \right) e^{iut}\Delta u$$

en uiteindelijk gaat de som over deze termen over in een integraal en we krijgen een van de centrale stellingen van de Fourier theorie:

Fourier integraal identiteit

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(\tau)e^{-iu\tau}\, d\tau \right) e^{iut}\, du.$$

Merk op dat in de binnenste integraal τ de integratie variabel is, terwijl u hier als constante behandeld wordt.

Om de Fourier integraal identiteit iets overzichtelijker te schrijven definiëren we de binnenste integraal als een aparte functie van u:

$$F(u) := \int_{-\infty}^{\infty} f(t)e^{-iut}\, dt \quad \text{dan is} \quad f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{iut}\, du.$$

Als we de formule $f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{iut}\,du$ met de Fourier reeks $f(t) = \sum_{k=-\infty}^{\infty} c_k e^{i\omega kt}$ van een periodieke functie vergelijken, zien we dat we de functie $F(u)$ kunnen opvatten als de continue versie van de Fourier coëfficiënten c_k.

Definitie:

(i) De functie
$$F(u) := \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt$$
heet de *Fourier getransformeerde* of *Fourier transformatie* van $f(t)$ en wordt genoteerd met $F(u) = \mathcal{F}[f(t)]$ of $F(u) = \hat{f}$.

(ii) De afbeelding $\mathcal{F} : f(t) \to \mathcal{F}[f(t)]$ die een functie $f(t)$ op zijn Fourier getransformeerde $F(u)$ afbeeldt, wordt zelf ook *Fourier transformatie* genoemd.

(iii) Omgekeerd komen we van de functie $F(u)$ terug naar $f(t)$ door
$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{iut}\,du$$
en we noemen dit de *inverse Fourier transformatie* van $F(u)$, genoteerd met $\mathcal{F}^{-1}[F(u)]$.

De Fourier transformatie levert een 'tweede taal' om over functies te praten: De Fourier transformatie vertaald een functie vanuit het *tijdsdomein* naar het *frequentiedomein*, de inverse Fourier transformatie is de vertaling de andere kant op. Omdat sommige eigenschappen van een functie beter in het tijdsdomein, andere beter in het frequentiedomein beschreven worden, is het heen en weer schakelen tussen de twee talen een fundamenteel hulpmiddel in vele gebieden van signaalverwerking en patroonherkenning.

De factor $\frac{1}{2\pi}$ in de Fourier integraal identiteit geeft aanleiding tot verschillende formuleringen van de Fourier transformatie. Bij de Fourier transformatie en de inverse Fourier transformatie moeten er namelijk factoren voor de integraal staan, die met elkaar vermenigvuldigd $\frac{1}{2\pi}$ opleveren. Drie voor de hand liggende mogelijkheden zijn:

(i) 1 bij de Fourier transformatie en $\frac{1}{2\pi}$ bij de inverse Fourier transformatie (zo als aangegeven),

(ii) $\frac{1}{2\pi}$ bij de Fourier transformatie en 1 bij de inverse Fourier transformatie,

(iii) $\frac{1}{\sqrt{2\pi}}$ bij de Fourier transformatie en ook bij de inverse Fourier transformatie.

De laatste versie benadrukt de symmetrie tussen Fourier transformatie en inverse Fourier transformatie, maar wij zullen bij de eerste optie blijven.

8.3 Schrijfwijzen van de Fourier transformatie

Net als voor de Fourier reeks zijn er ook voor de Fourier transformatie verschillende schrijfwijzen.

Amplitude en fase spectrum

Met behulp van absolute waarde en argument kunnen we $F(u)$ schrijven als

$$F(u) = |F(u)|e^{i\Phi(u)}$$

dan heet $|F(u)|$ het *amplitude spectrum* en $\Phi(u)$ het *fase spectrum* van $f(t)$.

Uit $e^{-i(-u)t} = e^{iut} = \overline{e^{-iut}}$ volgt voor reële functies $f(t)$:

$$F(-u) = \int_{-\infty}^{\infty} f(t)e^{-i(-u)t}\,dt = \int_{-\infty}^{\infty} f(t)\overline{e^{-iut}}\,dt = \int_{-\infty}^{\infty} \overline{f(t)e^{-iut}}\,dt = \overline{F(u)}.$$

Hieruit volgt in het bijzonder dat het amplitude spectrum $|F(u)|$ een even functie en het fase spectrum $\Phi(u)$ een oneven functie is.

Reële schrijfwijze

In analogie met de reële schrijfwijze van Fourier reeksen is er ook een reële schrijfwijze voor de Fourier transformatie van een reële functie $f(t)$. We schrijven

$$\begin{aligned} F(u) &= \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt = \int_{-\infty}^{\infty} f(t)(\cos(ut) - i\sin(ut))\,dt \\ &= \int_{-\infty}^{\infty} f(t)\cos(ut)\,dt - i\cdot\int_{-\infty}^{\infty} f(t)\sin(ut)\,dt \end{aligned}$$

en definiëren

$$a(u) := \int_{-\infty}^{\infty} f(t)\cos(ut)\,dt, \qquad b(u) := \int_{-\infty}^{\infty} f(t)\sin(ut)\,dt$$

dan is $F(u) = a(u) - ib(u)$. Er geldt $a(-u) = a(u)$ omdat $\cos(-ut) = \cos(ut)$ en $b(-u) = -b(u)$ omdat $\sin(-ut) = -\sin(ut)$, dus:

$$a(u) \text{ is een even functie,} \quad b(u) \text{ is een oneven functie.}$$

Als we de Fourier integraal identiteit met behulp van reële functies uitschrijven, krijgen we:

$$\begin{aligned} f(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{iut}\,du = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)(\cos(ut) + i\sin(ut))\,du \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (a(u) - ib(u))(\cos(ut) + i\sin(ut))\,du \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (a(u)\cos(ut) + b(u)\sin(ut))\,du \\ &+ i\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty} (a(u)\sin(ut) - b(u)\cos(ut))\,du. \end{aligned}$$

Maar $f(t)$ is een reële functie, daarom verdwijnt het imaginaire deel en we hebben

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (a(u)\cos(ut) + b(u)\sin(ut))\, du.$$

Nu weten we dat $a(u)$ even en $b(u)$ oneven is, hieruit volgt $a(-u)\cos(-ut) = a(u)\cos(ut)$ en $b(-u)\sin(-ut) = -b(u)(-\sin(ut)) = b(u)\sin(ut)$ en daarom levert de integratie van $-\infty$ tot 0 hetzelfde op als de integratie van 0 tot ∞. We hebben dus uiteindelijk gevonden:

$$f(t) = \frac{1}{\pi} \int_{0}^{\infty} (a(u)\cos(ut) + b(u)\sin(ut))\, du \quad \text{met}$$
$$a(u) = \int_{-\infty}^{\infty} f(t)\cos(ut)\, dt \quad \text{en} \quad b(u) = \int_{-\infty}^{\infty} f(t)\sin(ut)\, dt.$$

Als $f(t)$ een even functie is, is $f(t)\sin(ut)$ een oneven functie, maar dan is $b(u) = \int_{-\infty}^{\infty} f(t)\sin(ut)\, dt = 0$. Voor een oneven functie $f(t)$ is $f(t)\cos(ut)$ een oneven functie, dus volgt in dit geval met hetzelfde argument dat $a(u) = 0$ is.

Fourier cosinus en Fourier sinus transformatie

Voor reële functies $f(t)$ die alleen maar voor $0 \leq t < \infty$ gedefinieerd zijn, worden vaak ook de *Fourier cosinus transformatie* en de *Fourier sinus transformatie* toegepast. Het idee hierbij is, de functie door $f(-t) := f(t)$ tot een even of door $f(-t) := -f(-t)$ tot een oneven functie op de hele reële as voort te zetten. Voor de voortgezette functie is dan bij de Fourier transformatie of de functie $b(u)$ de 0-functie (even geval) of de functie $a(u)$ is de 0-functie (oneven geval).

Voor de *Fourier cosinus transformatie* stellen we ons voor dat $f(t)$ door $f(-t) := f(t)$ tot een even functie $f_e(t)$ op de hele reële as voortgezet wordt en berekenen hiervoor de reële versie van de Fourier transformatie.

Voor de voortgezette functie $f_e(t)$ krijgen we

$$a(u) = \int_{-\infty}^{\infty} f_e(t)\cos(ut)\, dt = 2\int_{0}^{\infty} f(t)\cos(ut)\, dt$$

omdat $f(t)\cos(ut)$ een even functie is.

Definitie: De integraal

$$F_c(u) := \int_{0}^{\infty} f(t)\cos(ut)\, dt$$

heet de *Fourier cosinus transformatie* van $f(t)$. Er geldt:

$$f(t) = \frac{2}{\pi} \int_{0}^{\infty} F_c(u)\cos(ut)\, dt \quad \text{met} \quad F_c(u) = \int_{0}^{\infty} f(t)\cos(ut)\, dt.$$

Net zo kunnen we $f(t)$ door $f(-t) := -f(t)$ tot een oneven functie $f_o(t)$ op de hele reële as voortzetten. Voor de voortgezette functie $f_o(t)$ krijgen we

$$b(u) = \int_{-\infty}^{\infty} f_o(t)\sin(ut)\, dt = 2\int_{0}^{\infty} f(t)\sin(ut)\, dt$$

omdat nu $f(t)\sin(ut)$ een even functie is.

Definitie: De integraal

$$F_s(u) := \int_0^\infty f(t)\sin(ut)\ dt$$

heet de *Fourier sinus transformatie* van $f(t)$. Hiervoor geldt:

$$f(t) = \frac{2}{\pi}\int_0^\infty F_s(u)\sin(ut)\ dt \quad \text{met} \quad F_s(u) = \int_0^\infty f(t)\sin(ut)\ dt.$$

8.4 Eigenschappen van de Fourier transformatie

De meeste eigenschappen van de Fourier transformatie volgen uit eigenschappen van de integraal. Het probleem is dat we het hier met oneindige integralen te maken hebben, waar soms dingen mis kunnen gaan. Omdat we ons hier niet met wiskundige details willen bemoeien, veronderstellen we nu dat het voor de functies waarin wij geïnteresseerd zijn nooit mis gaat en onderdrukken twijfels die bij sommige stappen misschien op komen dagen.

Bestaan

Omdat de Fourier getransformeerde middels een integratie over de hele reële as gedefinieerd is, moeten we wel een opmerking kwijt, wanneer de integraal überhaupt bestaat. Een integraal $\int_{-\infty}^{\infty} f(x)\ dx$ is namelijk gedefinieerd als de limiet $L \to \infty$ van $\int_{-L}^{L} f(x)\ dx$ en die limiet hoeft helemaal niet te bestaan. Voldoende voorwaarden waaronder de Fourier getransformeerde van $f(t)$ wel bestaat, zijn:

(i) $f(t)$ en $f'(t)$ zijn stuksgewijs continu op eindige intervallen;

(ii) $\int_{-\infty}^{\infty} |f(t)|\ dt < \infty$.

Net als bij de stelling van Dirichlet over de Fourier reeksen is in het bijzonder (ii) geen noodzakelijke voorwaarde. Sommige functies waarvoor (ii) niet geldt zullen we zelfs in de volgende les bekijken, omdat ze bijzonder belangrijk zijn.

Fourier transformatie $\leftrightarrow$ inverse Fourier transformatie

Als $F(u) = \mathcal{F}[f(t)]$ de Fourier getransformeerde van $f(t)$ is, geldt $f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{iut}\ du$ en dus $f(-t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u)e^{-iut}\ du$. Maar het laatste is tot op de factor $\frac{1}{2\pi}$ na de Fourier getransformeerde van $F(u)$, dus is

$$f(-t) = \frac{1}{2\pi}\mathcal{F}[F(u)] = \frac{1}{2\pi}\mathcal{F}[\mathcal{F}[f(t)]].$$

Er geldt dus

$$\mathcal{F}[\mathcal{F}[f(t)]] = 2\pi f(-t)$$

en als we dit vergelijken met de toepassing

$$\mathcal{F}^{-1}[\mathcal{F}[f(t)]] = f(t)$$

van de inverse Fourier transformatie zien we, dat de inverse Fourier transformatie tot op een vermenigvuldiging met de factor 2π na hetzelfde is als de Fourier transformatie gecombineerd met *tijdsomkeer*. In het bijzonder zijn voor even functies $f(t)$ Fourier transformatie en inverse Fourier transformatie tot op de factor 2π na hetzelfde.

Lineariteit

Uit de lineariteit van de integraal volgt meteen de lineariteit van de Fourier transformatie, want voor twee functies $f(t)$ en $g(t)$ is $\int_{-\infty}^{\infty}(f(t)+g(t))e^{-iut}\,dt = \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt + \int_{-\infty}^{\infty} g(t)e^{-iut}\,dt$. Net zo geldt voor de vermenigvuldiging met een factor c dat $\int_{-\infty}^{\infty} c\cdot f(t)e^{-iut}\,dt = c\cdot\int_{-\infty}^{\infty} f(t)e^{-iut}\,dt$. Dit geeft

$$\mathcal{F}[f(t)+g(t)] = \mathcal{F}[f(t)] + \mathcal{F}[g(t)] \quad \text{en} \quad \mathcal{F}[c\cdot f(t)] = c\cdot\mathcal{F}[f(t)]$$

dat wil zeggen de Fourier transformatie is een lineaire afbeelding van functies.

Verschuiving

Als we de Fourier transformatie van een functie $f(t)$ hebben berekend is het handig als we hieruit de Fourier transformatie van een verschuiving van $f(t)$ langs de reële as af kunnen leiden. Als we bijvoorbeeld de Fourier transformatie van een rechthoek impuls rond $t=0$ kennen, willen we hieruit graag de Fourier transformatie van een rechthoek impuls rond een tijdstip t_0 kunnen berekenen. Hiervoor moeten we de Fourier transformatie van de functie

$$g(t) := f(t-t_0)$$

berekenen. Als we de Fourier getransformeerden van $g(t)$ en $f(t)$ met $G(u) := \mathcal{F}[g(t)] = \mathcal{F}[f(t-t_0)]$ en $F(u) := \mathcal{F}[f(t)]$ noteren, vinden we:

$$\begin{aligned} G(u) &= \int_{-\infty}^{\infty} f(t-t_0)e^{-iut}\,dt = \int_{-\infty}^{\infty} f(t-t_0)e^{-iu(t-t_0)}e^{-iut_0}\,dt \\ &= e^{-iut_0}\int_{-\infty}^{\infty} f(\tau)e^{-iu\tau}\,d\tau = e^{-iut_0}F(u) \end{aligned}$$

waarbij we in de stap van de eerste naar de tweede rij de substitutie $\tau := t-t_0$, $d\tau = dt$ hebben toegepast. Voor een om t_0 langs de reële as verschoven functie geldt dus

$$\mathcal{F}[f(t-t_0)] = e^{-iut_0}\mathcal{F}[f(t)].$$

Dit betekent dat we alleen maar de fase van de Fourier getransformeerde veranderen, maar niet de amplitude $|\mathcal{F}[f(t-t_0)]|$, omdat e^{-iut_0} een getal op de eenheidscirkel is. Een verschuiving in het tijdsdomein resulteert dus in een fase-verschuiving in het frequentiedomein.

Schaling

Naast een *verschuiving* in het tijdsdomein is ook een *schaling* van de tijd een eenvoudige maar belangrijke transformatie die we vaak tegen komen. Zij $g(t) := f(at)$ en noteer de Fourier getransformeerden met $G(u) := \mathcal{F}[g(t)] = \mathcal{F}[f(at)]$ en $F(u) := \mathcal{F}[f(t)]$. We veronderstellen eerst dat $a > 0$, dan krijgen we met de substitutie $\tau = at$, $d\tau = a\ dt$ (dus $dt = \frac{1}{a}\ d\tau$):

$$G(u) = \int_{-\infty}^{\infty} f(at)e^{-iut}\ dt = \int_{-\infty}^{\infty} f(\tau)e^{-i\frac{u}{a}\tau}\frac{1}{a}\ d\tau = \frac{1}{a}F(\frac{u}{a}).$$

Als $a < 0$ werkt de substitutie hetzelfde, maar als t van $-\infty$ naar ∞ loopt, loopt in dit geval $\tau = at$ van ∞ naar $-\infty$. We moeten dus de grenzen van de integratie omdraaien en daarom het resultaat met -1 vermenigvuldigen, dus krijgen we in dit geval $G(u) = -\frac{1}{a}F(\frac{u}{a})$. Als we de twee gevallen combineren, volgt hieruit

$$\mathcal{F}[f(t)] = F(u) \quad \Rightarrow \quad \mathcal{F}[f(at)] = \frac{1}{|a|}F(\frac{u}{a}).$$

Dit betekent dat een schaling in het tijdsdomein correspondeert met de inverse schaling in het frequentiedomein, dus een rekking wordt een comprimering en andersom.

Afgeleiden

We kunnen ons afvragen, of er een eenvoudige manier is om van de Fourier getransformeerde van een functie $f(t)$ naar de Fourier getransformeerde van de afgeleide $f'(t)$ te komen. Laten $F(u) := \mathcal{F}[f(t)]$ en $G(u) := \mathcal{F}[f'(t)]$, dan geldt met partiële integratie

$$\begin{aligned} G(u) &= \int_{-\infty}^{\infty} f'(t)e^{-iut}\ dt = f(t)e^{-iut}\big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(t)(-iu)e^{-iut}\ dt \\ &= f(t)e^{-iut}\big|_{-\infty}^{\infty} + iuF(u). \end{aligned}$$

Als we nu veronderstellen dat $f(t) \to 0$ voor $t \to \pm\infty$, valt de eerste term weg en er geldt

$$\mathcal{F}[f'(t)] = iu\mathcal{F}[f(t)].$$

Ook over de afgeleide van de Fourier getransformeerde $F(u) = \mathcal{F}[f(t)]$ in het frequentiedomein kunnen we iets zeggen. Er geldt $F(u) = \int_{-\infty}^{\infty} f(t)e^{-iut}\ dt$ en als we dit naar u gaan afleiden, mogen we de differentiatie onder zekere voorwaarden (die we hier als gegeven veronderstellen) met de integratie verruilen.

> Denk hierbij aan de integraal als een oneindige som: Ook voor functies die door een oneindige reeks gegeven zijn, hadden we gezien dat we de afgeleide krijgen door de reeks termsgewijs af te leiden.

We hebben dus

$$F'(u) = \int_{-\infty}^{\infty} f(t)(e^{-iut})'\ dt = \int_{-\infty}^{\infty} f(t)(-it)e^{-iut}\ dt = -i\int_{-\infty}^{\infty} tf(t)e^{-iut}\ dt$$

en de laatste integraal is de Fourier getransformeerde van $tf(t)$. Er geldt dus

$$\mathcal{F}[f(t)]' = F'(u) = -i\mathcal{F}[tf(t)],$$

waarbij we veronderstellen dat de functie $tf(t)$ een Fourier getransformeerde *heeft*, dus dat in het bijzonder de integraal $\int_{-\infty}^{\infty} tf(t)\, dt$ bestaat.

8.5 Het convolutieproduct

De meest belangrijke operatie bij Fourier transformaties is de vermenigvuldiging van de getransformeerde functies in het frequentiedomein. Het idee hierbij is, bepaalde frequentie intervallen te versterken of af te zwakken door de Fourier getransformeerde met een functie te vermenigvuldigen die voor deze frequenties een grote of kleine waarde heeft. We zullen hier in de volgende les concrete voorbeelden van zien.

Laten $f(t)$ en $g(t)$ twee functies in het tijdsdomein zijn met Fourier getransformeerden $\mathcal{F}[f(t)] = F(u)$ en $\mathcal{F}[g(t)] = G(u)$. Dan is

$$\begin{aligned} F(u) \cdot G(u) &= (\int_{-\infty}^{\infty} f(\tau)e^{-iu\tau}\, d\tau) \cdot (\int_{-\infty}^{\infty} g(\tau')e^{-iu\tau'}\, d\tau') \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(\tau)e^{-iu\tau}\, d\tau \right) g(\tau')e^{-iu\tau'}\, d\tau' \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\tau)g(\tau')e^{-iu(\tau+\tau')}\, d\tau d\tau'. \end{aligned}$$

Als de lezer twijfels heeft over de manier hoe de integralen gemanipuleerd worden, is het verstandig om de integralen als oneindige sommen te interpreteren. Hiervoor zijn de omvormingen redelijk voor de hand liggend.

We substitueren nu $t = \tau + \tau'$, dan is $\tau' = t - \tau$ en $d\tau' = dt$. Dit geeft

$$\begin{aligned} F(u) \cdot G(u) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(\tau)g(t-\tau)e^{-iut}\, d\tau \right) dt \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(\tau)g(t-\tau)\, d\tau \right) e^{-iut} dt. \end{aligned}$$

We zien dat het resultaat ook weer de Fourier getransformeerde van een functie is, namelijk van de functie

$$h(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\, d\tau$$

die door de binnenste integraal gegeven is.

Definitie: De functie

$$h(t) := f(t) * g(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\, d\tau$$

heet het *convolutieproduct*, de *convolutie* of de *vouwing* van $f(t)$ en $g(t)$ en wordt genoteerd met een sterretje voor het product.

De cruciale eigenschap van het convolutieproduct $f(t)*g(t)$ is dat $F(u)\cdot G(u)$ de Fourier getransformeerde van deze functie is, dus dat

$$\mathcal{F}[f(t)] \cdot \mathcal{F}[g(t)] = \mathcal{F}[f(t) * g(t)].$$

Merk op: Het puntsgewijs product $F(u) \cdot G(u)$ van twee functies in het frequentiedomein correspondeert via de Fourier transformatie met het convolutieproduct $\mathcal{F}^{-1}[F(u)] * \mathcal{F}^{-1}[G(u)]$ van de inverse Fourier getransformeerden in het tijdsdomein. Deze samenhang is in feite de hoofdreden om überhaupt naar een zo rare constructie als als het convolutieproduct te kijken.

Belangrijke begrippen in deze les

- Fourier reeksen voor periodieke functies met willekeurige periode
- Fourier integraal identiteit
- Fourier transformatie, inverse Fourier transformatie
- amplitude spectrum, fase spectrum
- Fourier cosinus transformatie, Fourier sinus transformatie
- eigenschappen van de Fourier transformatie
- convolutieproduct

Opgaven

68. Zij $f(t)$ een functie met Fourier getransformeerde $F(u) = \mathcal{F}[f(t)]$.
    (i) Toon aan dat $\mathcal{F}[f(-t)] = F(-u)$. Dit betekent dat een spiegeling in het tijdsdomein ook een spiegeling in het frequentiedomein tot gevolg heeft.
    (ii) Neem nu aan dat $f(t)$ een *reële* functie is. Laat zien dat $F(-u) = \overline{F(u)}$.
69. Toon aan dat voor $F(u) = \mathcal{F}[f(t)]$ een verschuiving in het frequentiedomein gegeven is door de formule $F(u - u_0) = \mathcal{F}[f(t)e^{iu_0 t}]$.
70. Zij $F(u) = \mathcal{F}[f(t)]$ de Fourier getransformeerde van $f(t)$. Bepaal de Fourier getransformeerde van $f(t)\cos(\omega t)$. (De functie $f(t)\cos(\omega t)$ noemt men een *modulatie*.)
71. Ga na dat het convolutieproduct *commutatief* is, d.w.z. dat $f(t) * g(t) = g(t) * f(t)$.
72. We hebben gezien dat voor de Fourier getransformeerden $F(u) = \mathcal{F}[f(t)]$ en $G(u) = \mathcal{F}[g(t)]$ geldt, dat $F(u) \cdot G(u) = \mathcal{F}[f(t) * g(t)]$.
    (i) Laat zien dat omgekeerd geldt dat

$$\mathcal{F}^{-1}[F(u) * G(u)] = 2\pi f(t) \cdot g(t) \text{ en dus } \mathcal{F}[f(t) \cdot g(t)] = \frac{1}{2\pi} F(u) * G(u).$$

(ii) Bewijs hiermee dat

$$\int_{-\infty}^{\infty} f(t) \cdot g(t)\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(u) \cdot G(-u)\, du.$$

(iii) Concludeer dat voor een *reële* functie $f(t)$ geldt dat

$$\int_{-\infty}^{\infty} |f(t)|^2\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(u)|^2\, du.$$

Deze relatie heet de *Parseval identiteit*.

Les 9 Voorbeelden en toepassingen van de Fourier transformatie

We hebben in de vorige les de theorie van de Fourier transformatie behandeld en een aantal eigenschappen van de Fourier transformatie bekeken. Tot nog toe is de Fourier transformatie echter een abstracte procedure die aan een functie in het tijdsdomein een functie in het frequentiedomein toewijst. Om enigszins begrip van de Fourier transformatie te krijgen, zullen we in deze les de Fourier getransformeerden van een aantal belangrijke functies expliciet berekenen.

Verder leren we een nieuwe *functie* kennen, die geen functie van de gebruikelijke soort is, namelijk de *Dirac δ-functie*. Deze functie is overal nul, behalve in een enkele punt, waar ze zo groot is (oneindig) dat de integraal over de hele reële as niet de waarde 0 maar 1 geeft. We zullen zien dat deze rare functie in het kader van de Fourier transformatie heel nuttig is.

Ten slotte gaan we in deze les als toepassing van de Fourier transformatie bekijken, hoe high-pass en low-pass filters werken, die de contrasten van een plaatje of geluid aanscherpen of verzachten.

9.1 Belangrijke voorbeelden

We zullen voor een aantal elementaire functies de Fourier transformaties bepalen. Met behulp van verschuiven, schalen en het optellen van functies laten zich uit deze functies natuurlijk ingewikkeldere functies opbouwen.

Rechthoek impuls

Zij $f(t)$ een rechthoek impuls van sterkte 1 tussen de tijden $-a$ en a, dus

$$f(t) := \begin{cases} 1 & \text{als } |t| \leq a \\ 0 & \text{als } |t| > a \end{cases}$$

dan berekenen we de Fourier getransformeerde $F(u) := \mathcal{F}[f(t)]$ als volgt:

$$\begin{aligned} F(u) &= \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt = \int_{-a}^{a} e^{-iut}\,dt = \frac{1}{-iu}\,e^{-iut}\Big|_{-a}^{a} = \frac{1}{-iu}\,(e^{-iua} - e^{iua}) \\ &= \frac{2}{u}\,\frac{e^{iua} - e^{-iua}}{2i} = \frac{2}{u}\sin(au) = 2\,\frac{\sin(au)}{u} \end{aligned}$$

Om de Fourier getransformeerden voor verschillende breedten van de impuls goed te kunnen vergelijken, is het handig de integraal $\int_{-\infty}^{\infty} f(t)\ dt$ op 1 te normeren. Hiervoor moeten we naar een impuls van sterkte $\frac{1}{2a}$ in plaats van 1 kijken.

Omdat de Fourier transformatie een lineaire afbeelding is, hoeven we nu niet opnieuw te rekenen, we moeten het resultaat alleen maar met $\frac{1}{2a}$ vermenigvuldigen. De functie $g(t)$ met

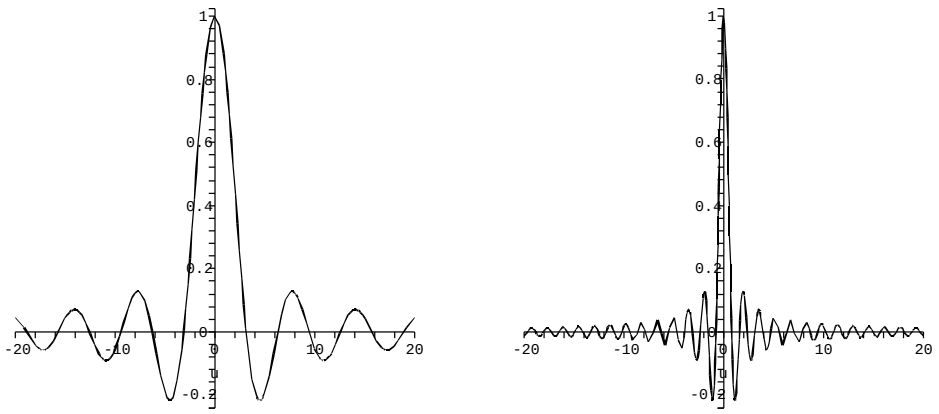
$$g(t) := \begin{cases} \frac{1}{2a} & \text{als } |t| \leq a \\ 0 & \text{als } |t| > a \end{cases}$$

heeft dus de Fourier getransformeerde

$$G(u) = \mathcal{F}[g(t)] = \frac{1}{2a} \cdot 2\frac{\sin(au)}{u} = \frac{\sin(au)}{au}.$$

We zien dat de parameter a van de rechthoek impuls bij de Fourier getransformeerde de rol van een schaling van de u-as speelt: Als we a verdubbelen, moeten we de u-as met een factor 2 samen schuiven. Dit betekent in het bijzonder dat voor een grotere waarde van a, dus een langere impuls, de Fourier getransformeerde met stijgende frequenties sneller afneemt als voor en kleinere waarde van a. Dit zouden we ook zo verwachten, want een korte impuls geeft een snelle verandering en heeft dus met hogere frequenties te maken.

In Figuur II.10 zijn de Fourier getransformeerden van een rechthoek impuls voor $a = 1$ en $a = 4$ te zien. Zo als verwacht neemt de getransformeerde van de langere impuls met $a = 4$ duidelijk sneller af dan de getransformeerde van de kortere impuls met $a = 1$.



Figuur II.10: Fourier getransformeerden van rechthoek impulsen met $a = 1$ (links) en $a = 4$ (rechts).

Driehoek impuls

Zij $f(t)$ een driehoek impuls die tussen de tijden $-a$ en 0 lineair van 0 tot 1 groeit en tussen 0 en a weer lineair naar 0 daalt, dus

$$f(t) := \begin{cases} 1 - \frac{|t|}{a} & \text{als } |t| \leq a \\ 0 & \text{als } |t| > a \end{cases}$$

De Fourier getransformeerde $F(u)$ van $f(t)$ is

$$F(u) = \int_{-\infty}^{\infty} f(t)e^{-iut}\, dt = \int_{-a}^{0} (1 + \frac{t}{a})e^{-iut}\, dt + \int_{0}^{a} (1 - \frac{t}{a})e^{-iut}\, dt.$$

Dit lossen we met partiële integratie op, de primitieve van e^{-iut} is $\frac{1}{-iu}e^{-iut}$ en de afgeleide van $1 \pm \frac{t}{a}$ is $\pm\frac{1}{a}$. Voor de eerste integraal volgt hieruit

$$\begin{aligned}\int_{-a}^{0}(1+\frac{t}{a})e^{-iut}\,dt &= (1+\frac{t}{a})\frac{1}{-iu}e^{-iut}\Big|_{-a}^{0} - \int_{-a}^{0}\frac{1}{a}\cdot\frac{1}{-iu}e^{-iut}\,dt\\ &= -\frac{1}{iu} - 0 + \frac{1}{a}\cdot\frac{1}{iu}\int_{-a}^{0}e^{-iut}\,dt = -\frac{1}{iu} - \frac{1}{a}\cdot\frac{1}{(iu)^2}e^{-iut}\Big|_{-a}^{0}\\ &= -\frac{1}{iu} + \frac{1}{a}\cdot\frac{1}{u^2} - \frac{1}{a}\cdot\frac{1}{u^2}e^{iua}\end{aligned}$$

Net zo krijgen we voor de tweede integraal

$$\begin{aligned}\int_{0}^{a}(1-\frac{t}{a})e^{-iut}\,dt &= (1-\frac{t}{a})\frac{1}{-iu}e^{-iut}\Big|_{0}^{a} - \int_{0}^{a}-\frac{1}{a}\cdot\frac{1}{-iu}e^{-iut}\,dt\\ &= 0 + \frac{1}{iu} - \frac{1}{a}\cdot\frac{1}{iu}\int_{0}^{a}e^{-iut}\,dt = \frac{1}{iu} + \frac{1}{a}\cdot\frac{1}{(iu)^2}e^{-iut}\Big|_{0}^{a}\\ &= \frac{1}{iu} - \frac{1}{a}\cdot\frac{1}{u^2}e^{-iua} + \frac{1}{a}\cdot\frac{1}{u^2}\end{aligned}$$

Als we de twee integralen bij elkaar optellen, krijgen we dus

$$\begin{aligned}F(u) &= -\frac{1}{iu} + \frac{1}{a}\cdot\frac{1}{u^2} - \frac{1}{a}\cdot\frac{1}{u^2}e^{iua} + \frac{1}{iu} - \frac{1}{a}\cdot\frac{1}{u^2}e^{-iua} + \frac{1}{a}\cdot\frac{1}{u^2}\\ &= \frac{1}{au^2}(2 - e^{iua} - e^{-iua}).\end{aligned}$$

Met een klein trucje kunnen we dit nog iets eenvoudiger schrijven, er geldt namelijk

$$(e^{iu\frac{a}{2}} - e^{-iu\frac{a}{2}})^2 = e^{iua} - 2 + e^{-iua} = -(2 - e^{iua} - e^{-iua})$$

Hieruit volgt

$$\sin^2(\frac{a}{2}u) = (\frac{e^{iu\frac{a}{2}} - e^{-iu\frac{a}{2}}}{2i})^2 = -\frac{1}{4}(e^{iu\frac{a}{2}} - e^{-iu\frac{a}{2}})^2 = \frac{1}{4}(2 - e^{iua} - e^{-iua})$$

en als we dit in de gevonden formule van $F(u)$ invullen, krijgen we uiteindelijk:

$$F(u) = \frac{4}{au^2}\cdot\sin^2(\frac{a}{2}u) = \frac{a}{(\frac{a}{2}u)^2}\cdot\sin^2(\frac{a}{2}u) = a\cdot\left(\frac{\sin(\frac{a}{2}u)}{\frac{a}{2}u}\right)^2.$$

Als we ook bij deze functie de integraal $\int_{-\infty}^{\infty} f(t)\,dt$ op 1 normeren, moeten we $f(t)$ met $\frac{1}{a}$ vermenigvuldigen, dit geeft de functie

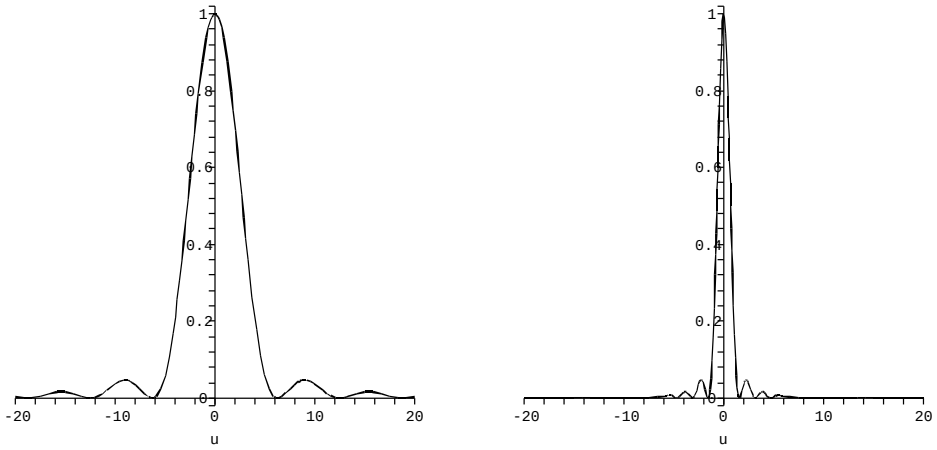
$$g(t) := \begin{cases} \frac{1}{a} - \frac{|t|}{a^2} & \text{als } |t| \le a \\ 0 & \text{als } |t| > a \end{cases}$$

Deze functie $g(t)$ heeft als Fourier getransformeerde

$$\mathcal{F}[g(t)] = \left(\frac{\sin(\frac{a}{2}u)}{\frac{a}{2}u}\right)^2$$

dus speelt ook hier de parameter a de rol van een schaling van de u-as.

Ook de plaatjes van de Fourier getransformeerden van de driehoek impuls in Figuur II.11 laten duidelijk zien dat bij de kortere impuls voor $a = 1$ de hogere frequenties een grotere rol spelen dan bij de impuls voor $a = 4$.



Figuur II.11: Fourier getransformeerden van driehoek impulsen met $a = 1$ (links) en $a = 4$ (rechts).

Gauss functie

Een belangrijke rol bij de Fourier transformaties speelt de Gauss functie $f(t) := e^{-at^2}$, die we in verband met de normale verdeling ook in de kansrekening en statistiek tegen komen. De opmerkelijke eigenschap van deze functie ten opzichte van de Fourier transformatie is, dat de Fourier getransformeerde weer een functie van dezelfde soort is. Om dit na te gaan, hebben we de integraal van $-\infty$ tot ∞ van de Gauss functie nodig, die we in het kader van de integratie van functies van meerdere veranderlijken hebben bepaald:

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi} \quad \text{en dus} \quad \int_{-\infty}^{\infty} e^{-at^2}\, dt = \sqrt{\frac{\pi}{a}}$$

waarbij de tweede integraal met de substitutie $\sqrt{a} \cdot t = x$, $\sqrt{a}\, dt = dx$ uit de eerste volgt.

Voor de Gauss functie $f(t) := e^{-at^2}$ met $a > 0$ geldt voor de Fourier getransformeerde $F(u) := \mathcal{F}[f(t)]$:

$$\begin{aligned} F(u) &= \int_{-\infty}^{\infty} e^{-at^2} \cdot e^{-iut}\, dt = \int_{-\infty}^{\infty} e^{-a(t^2 + i\frac{u}{a}t)}\, dt = \int_{-\infty}^{\infty} e^{-a(t+i\frac{u}{2a})^2} \cdot e^{a(i\frac{u}{2a})^2}\, dt \\ &= e^{-\frac{u^2}{4a}} \cdot \int_{-\infty}^{\infty} e^{-a(t+i\frac{u}{2a})^2}\, dt = e^{-\frac{u^2}{4a}} \cdot \int_{-\infty}^{\infty} e^{-a\tau^2}\, d\tau = \sqrt{\frac{\pi}{a}} \cdot e^{-\frac{u^2}{4a}} \end{aligned}$$

waarbij we in de voorlaatste stap de substitutie $\tau = t + i\frac{u}{2a}$, $d\tau = dt$ toepassen. We hebben dus ingezien:

Stelling: De Fourier getransformeerde van een Gauss functie is ook weer een Gauss functie, er geldt:

$$\mathcal{F}[e^{-at^2}] = \sqrt{\frac{\pi}{a}} \cdot e^{-\frac{u^2}{4a}}.$$

De dichtheidsfunctie van een normale verdeling met verwachtingswaarde 0 en standaardafwijking σ is de Gauss functie

$$f(t) = e^{-\frac{t^2}{2\sigma^2}},$$

dus moeten we $a = \frac{1}{2\sigma^2}$ in ons resultaat invullen. Dit geeft $\sqrt{\frac{\pi}{a}} = \sqrt{2\pi}\,\sigma$ en $\frac{u^2}{4a} = \frac{\sigma^2 u^2}{2} = \frac{u^2}{2\sigma^{-2}}$, dus geldt:

$$f(t) = e^{-\frac{t^2}{2\sigma^2}} \quad \Rightarrow \quad F(u) = \mathcal{F}[f(t)] = \sqrt{2\pi}\,\sigma \cdot e^{-\frac{u^2}{2\sigma^{-2}}}.$$

De Fourier getransformeerde van een normale verdeling met standaardafwijking σ is dus een normale verdeling met standaardafwijking $\sigma^{-1} = \frac{1}{\sigma}$. In het bijzonder is de Fourier getransformeerde van een standaard-normale verdeling (met standaardafwijking 1) zelf ook een standaard-normale verdeling.

Eindige cosinus golf

We kijken naar een cosinus functie op het eindige interval $[-a, a]$, dus

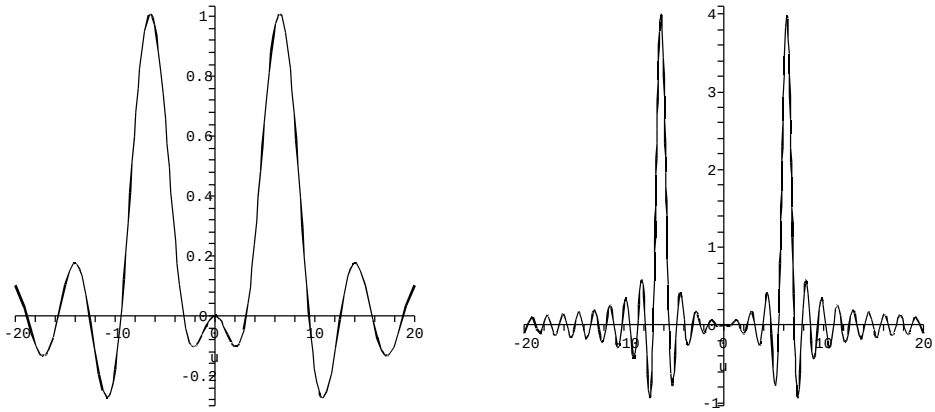
$$f(t) := \begin{cases} \cos(\omega t) & \text{als } |t| \le a \\ 0 & \text{als } |t| > a \end{cases}$$

De Fourier getransformeerde is

$$\begin{aligned} F(u) &= \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt = \int_{-a}^{a} \cos(\omega t)e^{-iut}\,dt = \int_{-a}^{a} \frac{1}{2}(e^{i\omega t} + e^{-i\omega t})e^{-iut}\,dt \\ &= \frac{1}{2}\int_{-a}^{a} e^{-i(u-\omega)t}\,dt + \frac{1}{2}\int_{-a}^{a} e^{-i(u+\omega)t}\,dt \\ &= \frac{1}{-2i(u-\omega)} e^{-i(u-\omega)t}\Big|_{-a}^{a} + \frac{1}{-2i(u+\omega)} e^{-i(u+\omega)t}\Big|_{-a}^{a} \\ &= -\frac{1}{u-\omega}\cdot\frac{1}{2i}\left(e^{-i(u-\omega)a} - e^{i(u-\omega)a}\right) - \frac{1}{u+\omega}\cdot\frac{1}{2i}\left(e^{-i(u+\omega)a} - e^{i(u+\omega)a}\right) \\ &= \frac{\sin(a(u-\omega))}{u-\omega} + \frac{\sin(a(u+\omega))}{u+\omega} \end{aligned}$$

Als we dit resultaat met de Fourier getransformeerde van een rechthoek impuls vergelijken, zien we dat de Fourier getransformeerde van een eindige cosinus golf de som van twee Fourier getransformeerden van rechthoek impulsen is, waarvan één om ω naar rechts en de andere om ω naar links verschoven is. Dit is in feite niet erg moeilijk om in te zien:

Net zo als een verschuiving om t_0 in het tijdsdomein met een vermenigvuldiging van de Fourier getransformeerde met e^{iut_0} in het frequentiedomein correspondeert, correspondeert een verschuiving van een Fourier getransformeerde om ω in het frequentiedomein met een vermenigvuldiging met $e^{i\omega t}$ in het tijdsdomein. Maar als we een rechthoek impuls met $e^{i\omega t}$ vermenigvuldigen, krijgen we de functie $e^{i\omega t}$ op een eindig interval, en door één keer om ω en één keer om $-\omega$ te verschuiven, krijgen we $e^{i\omega t} + e^{-i\omega t} = 2\cos(\omega t)$ op een eindig interval, en dit is (tot op de factor 2 na) precies waarmee we zijn begonnen.



Figuur II.12: Fourier getransformeerden van eindige cosinus golven met frequentie $\omega = 2\pi$ op het interval $[-a, a]$ voor $a = 1$ (links) en $a = 4$ (rechts).

Exponentiële afname

Een exponentiële afname wordt door de functie

$$f(t) := \begin{cases} e^{-at} & \text{als } t \geq 0 \\ 0 & \text{als } t < 0 \end{cases} \qquad \text{met } a > 0$$

beschreven. Hierbij is de afname om zo sneller hoe groter de parameter a is.

De Fourier getransformeerde van deze functie $f(t)$ berekenen we als volgt:

$$\begin{aligned} F(u) &= \int_{-\infty}^{\infty} f(t)e^{-iut}\,dt = \int_0^{\infty} e^{-at} \cdot e^{-iut}\,dt = \int_0^{\infty} e^{-(iu+a)t}\,dt \\ &= -\frac{1}{iu+a} e^{-(iu+a)t}\Big|_0^{\infty} = \frac{1}{iu+a} = \frac{a-iu}{u^2+a^2} = \frac{a}{u^2+a^2} - i\frac{u}{u^2+a^2} \end{aligned}$$

want voor $t \to \infty$ gaat $e^{-at} \to 0$ en dus ook $e^{-(iu+a)t} = e^{-iut}e^{-at} \to 0$.

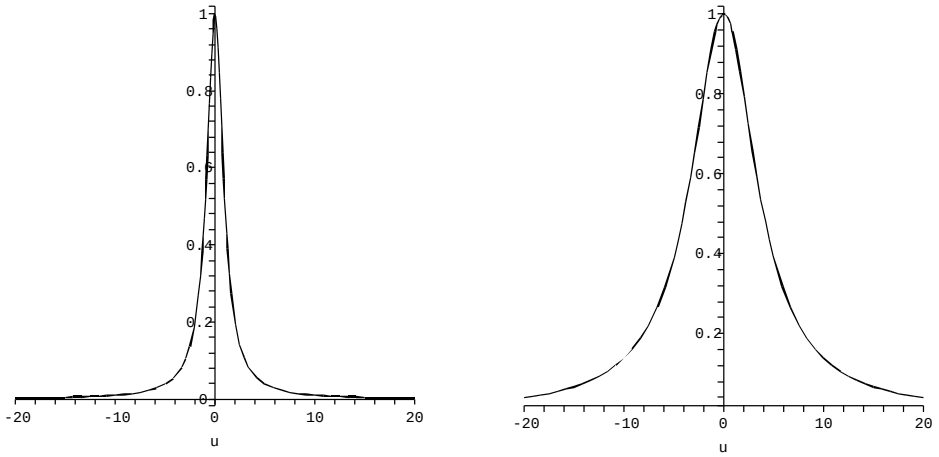
Om ook in dit geval verschillende parameters a goed te kunnen vergelijken, normeren we $\int_0^{\infty} f(t)\,dt$ weer op 1, hiervoor moeten we e^{-at} met a vermenigvuldigen, want $\int_0^{\infty} e^{-at}\,dt = \frac{-1}{a}e^{-at}\Big|_0^{\infty} = \frac{1}{a}$. De functie

$$g(t) := \begin{cases} ae^{-at} & \text{als } t \geq 0 \\ 0 & \text{als } t < 0 \end{cases}$$

heeft de Fourier getransformeerde

$$G(u) := \mathcal{F}[g(t)] = \frac{a^2 - iau}{a^2+u^2} \text{ met reëel deel } \Re(G(u)) = \frac{a^2}{a^2+u^2}.$$

Ook hier is in de plaatjes van Figuur II.13 duidelijk te zien dat de snellere afname voor $a = 4$ een grotere bijdrage van hoge frequenties tot gevolg heeft.



Figuur II.13: Fourier getransformeerden van exponentiële afname met $a = 1$ (links) en $a = 4$ (rechts).

Omdat we het hier eigenlijk met een functie te maken hebben, die alleen maar voor positieve tijden gedefinieerd is, is dit een typisch geval om de Fourier cosinus transformatie

$$F_c(u) = \int_0^\infty f(t)\cos(ut)\ dt$$

toe te passen. Hiervoor hebben we weer een partiële integratie nodig:

$$\begin{aligned}
&\int e^{-at}\cos(ut)\ dt = e^{-at}\frac{1}{u}\sin(ut) - \int(-a)e^{-at}\frac{1}{u}\sin(ut)\ dt \\
&= \frac{1}{u}e^{-at}\sin(ut) + \frac{a}{u}\int e^{-at}\sin(ut)\ dt \\
&= \frac{1}{u}e^{-at}\sin(ut) + \frac{a}{u}\left(e^{-at}(-\frac{1}{u})\cos(ut) - \int(-a)e^{-at}(-\frac{1}{u})\cos(ut)\ dt\right) \\
&= \frac{1}{u}e^{-at}\sin(ut) - \frac{a}{u^2}e^{-at}\cos(ut) - \frac{a^2}{u^2}\int e^{-at}\cos(ut)\ dt
\end{aligned}$$

Met $1 + \frac{a^2}{u^2} = \frac{u^2+a^2}{u^2}$ volgt hieruit

$$\begin{aligned}
F_c(u) = \int_0^\infty e^{-at}\cos(ut)\ dt &= \frac{u^2}{a^2+u^2}\left(\frac{1}{u}e^{-at}\sin(ut) - \frac{a}{u^2}e^{-at}\cos(ut)\right)\Big|_0^\infty \\
&= \frac{1}{a^2+u^2}\left(ue^{-at}\sin(ut) - ae^{-at}\cos(ut)\right)\Big|_0^\infty = \frac{a}{a^2+u^2}
\end{aligned}$$

want voor $t \to \infty$ gaat $ue^{-at}\sin(ut) - ae^{-at}\cos(ut) \to 0$.

De Fourier cosinus transformatie levert dus inderdaad het reële deel van de complexe Fourier transformatie op (maar eist wel behoorlijk meer rekenwerk).

9.2 De Dirac δ-functie

Een kunstmatige maar uiterst belangrijke en nuttige functie is de *Dirac δ-functie* of *δ-impuls*. Deze functie is overal 0, behalve van een oneindige spits in het nulpunt en is gedefinieerd door de eigenschap dat de integraal van $-\infty$ tot ∞ de waarde 1 heeft. We kunnen dit zien als limiet van een rechthoek impuls op het interval $[-a, a]$ van sterkte $\frac{1}{2a}$, waarbij men $a \to 0$ laat gaan. Deze functie wordt aangegeven met $\delta(x)$. De wezenlijke eigenschappen van de δ-impuls zijn:

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1 \quad \text{en} \quad \int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0).$$

De tweede eigenschap gaat men na door de δ-functie door een rechthoek impuls van breedte $2a$ te benaderen en hiervan de limiet $a \to 0$ te bekijken: Een rechthoek impuls van breedte $2a$ rond een punt x_0 is gegeven door de functie

$$r_a(x) := \begin{cases} \frac{1}{2a} & \text{als } x \in [x_0 - a, x_0 + a] \\ 0 & \text{als } x \notin [x_0 - a, x_0 + a] \end{cases}$$

We veronderstellen nu dat we een primitieve van $f(x)$ kennen, dus een functie $F(x)$ met $F'(x) = f(x)$. Er geldt nu

$$\int_{-\infty}^{\infty} f(x) r_a(x)\, dx = \int_{x_0-a}^{x_0+a} f(x) \frac{1}{2a}\, dx = \frac{1}{2a} F(x)\Big|_{x_0-a}^{x_0+a} = \frac{F(x_0 + a) - F(x_0 - a)}{2a}$$

Maar de laatste quotiënt is juist de differentiaalquotiënt waardoor de afgeleide van $F(x)$ in het punt $x_0 - a$ gedefinieerd is, dus gaat deze quotiënt voor $a \to 0$ naar $F'(x_0 - a) = F'(x_0) = f(x_0)$.

Wat de eigenschap $\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0)$ in feite betekent is dat het *convolutieproduct* van een functie $f(x)$ met de δ-functie $\delta(x)$ in het punt x_0 juist de waarde $f(x_0)$ oplevert: De δ-functie is een even functie, dus geldt $\delta(x - x_0) = \delta(x_0 - x)$ en dus is

$$f(x_0) * \delta(x_0) = \int_{-\infty}^{\infty} f(x)\delta(x_0 - x)\, dx = \int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0).$$

Merk op: Het convolutieproduct van een functie $f(x)$ met de δ-functie $\delta(x)$ levert juist de waarde $f(x_0)$ op.

Uit de eigenschappen van de δ-functie volgt meteen wat de Fourier getransformeerde $\mathcal{F}[\delta(t)]$ is, namelijk:

$$F(u) = \mathcal{F}[\delta(t)] = \int_{-\infty}^{\infty} \delta(t) e^{-iut}\, dt = e^{-iu \cdot 0} = 1.$$

Dit zegt dat de Fourier getransformeerde de constante functie $\mathbf{1}$ met waarde 1 voor alle u is (deze functie noteren we met een vet gedrukte $\mathbf{1}$).

De inverse Fourier transformatie geeft nu een alternatieve schrijfwijze voor de δ-functie, namelijk

$$\delta(t) = \mathcal{F}^{-1}[\mathbf{1}] = \frac{1}{2\pi} \int_{-\infty}^{\infty} 1 \cdot e^{iut}\, du = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iut}\, du.$$

Voor de δ-functie $\delta(t-t_0)$ met spits in t_0 vinden we met de formule voor een verschoven functie (of direct) dat

$$\mathcal{F}[\delta(t-t_0)] = e^{-iut_0}.$$

De samenhang $\mathcal{F}[\mathcal{F}[f(t)]] = 2\pi f(-t)$ tussen Fourier transformatie en inverse Fourier transformatie geeft nu ook de mogelijkheid om de Fourier transformatie van constante functies te berekenen. We hadden gezien dat $\mathcal{F}[\delta(t)] = \mathbf{1}$, en als we de variabel t tot u hernoemen, volgt hieruit:

$$\mathcal{F}[\mathbf{1}] = \mathcal{F}[\mathcal{F}[\delta(u)]] = 2\pi\,\delta(-u) = 2\pi\,\delta(u)$$

omdat de δ-functie met spits in $u=0$ een even functie is.

Als we in de relatie $\mathcal{F}[\delta(t-t_0)] = e^{-iut_0}$ voor de verschoven δ-functie de variabelen hernoemen tot $t \to u$, $t_0 \to -\omega$, $u \to t$ luidt deze $\mathcal{F}[\delta(u+\omega)] = e^{i\omega t}$ en met hetzelfde argument als boven volgt:

$$\mathcal{F}[e^{i\omega t}] = \mathcal{F}[\mathcal{F}[\delta(u+\omega)]] = 2\pi\,\delta(-u+\omega) = 2\pi\,\delta(u-\omega),$$

waarbij we weer uit symmetrie redenen weten dat $\delta(-u+\omega) = \delta(u-\omega)$.

Samenvattend geeft dit de twee paren van relaties:

$$\begin{aligned} \mathcal{F}[\delta(t)] &= \mathbf{1} & \mathcal{F}[\delta(t-t_0)] &= e^{-iut_0} \\ \mathcal{F}[\mathbf{1}] &= 2\pi\,\delta(u) & \mathcal{F}[e^{i\omega t}] &= 2\pi\,\delta(u-\omega). \end{aligned}$$

Met de gewone theorie van integralen is het eigenlijk onzin naar een integraal als $\int_{-\infty}^{\infty} e^{iut}\,du$ te kijken, want de limiet $L \to \infty$ van $\int_{-L}^{L} e^{iut}\,du$ bestaat niet. Om hier wel na te kunnen kijken, moeten we noodzakelijk een nieuwe functie met de eigenschappen van de δ-functie definiëren. Dit is een verder voorbeeld van een definitie die door de wiskundigen (en natuurkundigen) is verzonnen, om uit een doodlopende straat te ontsnappen.

We hadden gezegd dat we de δ-functie als limiet van een rechthoek impuls met oppervlakte 1 kunnen zien. Daarom zouden we kunnen verwachten, dat ook de net gevonden Fourier getransformeerde van de δ-functie interpreteerbaar is als limiet van de Fourier getransformeerden van de rechthoek impulsen.

De Fourier getransformeerde van een rechthoek impuls van breedte $2a$ rond 0 was $F(u) = \frac{\sin(au)}{au}$. Deze functie heeft voor $u=0$ de waarde 1 (want $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$) en de nulpunten zijn gegeven door $au = k\pi$ met $k \in \mathbb{Z}$, dus door $u = \frac{k\pi}{a}$. In het bijzonder ligt het kleinste positieve nulpunt bij $u = \frac{\pi}{a}$. Als we nu de limiet $a \to 0$ bekijken, betekent dit dat de eerste nulpunt naar oneindig gaat, dus de heuvel rond $u=0$ wordt steeds breder en in de limiet wordt de heuvel helemaal plat en wordt de functie $F(u)$ de constante functie 1.

Omgekeerd laten we nu eens voor een rechthoek impuls van sterkte 1 en breedte $2a$ de parameter $a \to \infty$ lopen. Deze rechthoek impuls heeft de Fourier

getransformeerde $F(u) = 2\frac{\sin(au)}{u} = 2a\frac{\sin(au)}{au}$ en voor $a \to \infty$ wordt de functie $\frac{\sin(au)}{au}$ en spits van hoogte 1 in het nulpunt, dus wordt $F(u)$ inderdaad een δ-functie. (Dat de constanten hierbij kloppen moet men met een integratie nagaan.)

We zien dus dat de definitie van de δ-functie als limiet van rechthoek impulsen ten opzichte van de Fourier transformatie de gewenste eigenschappen heeft.

Periodieke functies

Met behulp van de δ-functie kunnen we nu ook de Fourier transformaties van periodieke functies bepalen. Een periodieke functie $f(t)$ met periode L en grondfrequentie $\omega = \frac{2\pi}{L}$ kunnen we schrijven als Fourier reeks

$$f(t) = \sum_{-\infty}^{\infty} c_k e^{i\omega kt}.$$

Uit de lineariteit van de Fourier transformatie en de resultaten voor de δ-functie volgt

$$F(u) = \mathcal{F}[f(t)] = \sum_{-\infty}^{\infty} c_k \mathcal{F}[e^{i\omega kt}] = \sum_{-\infty}^{\infty} c_k \delta(u - \omega k)$$

(waarbij we ons geen zorgen over de oneindige som maken). Dit betekent dat we een periodieke functie beschrijven door een som van δ-functies met spitsen op de veelvouden van de grondfrequentie en geschaald volgens de Fourier coëfficiënten. In feite is deze beschrijving van een periodieke functie niets anders dan de beschrijving door de Fourier reeks.

Als we nu nog eens terug kijken naar de Fourier transformaties van eindige cosinus golven in Figuur II.12, zien we dat de functie voor $a = 4$ al twee duidelijke spitsen heeft. Als we het interval $[-a, a]$ van de cosinus golf nu laten groeien, worden deze spitsen steeds geprononceerder en voor $a \to \infty$ krijgen we uiteindelijk de som van twee δ-functies met spitsen in ω en $-\omega$.

Sprongfunctie

Nu dat we de δ-functie kennen en de Fourier getransformeerde hiervan hebben bepaald, kunnen we ook naar functies met een sprong kijken. Het eenvoudigste voorbeeld hiervan is de *Heaviside functie* $H(t)$, gegeven door

$$H(t) := \begin{cases} 0 & \text{als } t < 0 \\ 1 & \text{als } t \geq 0 \end{cases}$$

Omdat voor de Dirac δ-functie geldt dat $\int_{-\infty}^{t} \delta(x)\ dx = 0$ als $t < 0$ en $\int_{-\infty}^{t} \delta(x)\ dx = 1$ als $t > 0$, is $H(t)$ een primitieve van de δ-functie. Omgekeerd betekent dit dat $H'(t) = \delta(t)$. Dit is een verdere motivatie om een functie zo als de δ-functie te definiëren.

We hadden in de vorige les gezien dat $\mathcal{F}[f'(t)] = iu\mathcal{F}[f(t)]$, maar hierbij hadden we verondersteld dat $\lim f(t) = 0$ voor $t \to \pm\infty$ en dit is voor de Heaviside functie $H(t)$ zeker niet het geval. Als we de formule niettemin op $H(t)$ toepassen, krijgen we $\mathcal{F}[H(t)] = \frac{1}{iu}\mathcal{F}[\delta(t)] = \frac{1}{iu}$. Maar we zien hier ook waar het probleem zit: Als we op de Heaviside functie een constante C optellen, is de afgeleide nog steeds $(H(t)+C)' = \delta(t)$, maar nu is $\mathcal{F}[H(t)+C] = \mathcal{F}[H(t)]+\mathcal{F}[C] = \mathcal{F}[H(t)]+C\delta(u)$. Het optellen van een constante leidt dus tot een verschil om een veelvoud van de δ-functie bij de Fourier getransformeerde.

Uit deze discussie volgt dat we hooguit kunnen verwachten dat $\mathcal{F}[H(t)]$ van de vorm $\mathcal{F}[H(t)] = \frac{1}{iu} + c\delta(u)$ is.

Dat dit inderdaad het geval is, kunnen we op een andere manier onderbouwen: We hebben eerder in deze les aangetoond dat de functie

$$f(t) := \begin{cases} e^{-at} & \text{als } t \geq 0 \\ 0 & \text{als } t < 0 \end{cases} \qquad \text{met } a > 0$$

de Fourier getransformeerde $\mathcal{F}[f(t)] = \frac{a}{u^2+a^2} - i\frac{u}{u^2+a^2}$ heeft. Maar in de limiet $a \to 0$ geeft $f(t)$ juist de Heaviside functie weer, want hoe kleiner a is, hoe langzamer neemt de functie af, en in de limiet $a \to 0$ neemt de functie helemaal niet meer af.

Als we nu voor $u \neq 0$ de limiet $a \to 0$ van $\mathcal{F}[f(t)]$ bepalen, krijgen we $\lim_{a\to 0}\mathcal{F}[f(t)] = \frac{0}{u^2} - i\frac{u}{u^2} = \frac{1}{iu}$ en dit klopt inderdaad met onze gok dat $\mathcal{F}[H(t)] = \frac{1}{iu} + c\delta(u)$.

We moeten dus alleen maar nog de waarde van $\mathcal{F}[H(t)]$ voor $u = 0$ bepalen. Dit doen we als volgt: Er geldt $H(t)+H(-t) = \mathbf{1}$ (behalve voor $t = 0$). Hieruit volgt dat

$$\mathcal{F}[H(t)] + \mathcal{F}[H(-t)] = \mathcal{F}[\mathbf{1}] = 2\pi\delta(u).$$

Maar voor $F(u) = \mathcal{F}[H(t)]$ geldt dat $\mathcal{F}[H(-t)] = F(-u)$, dus hebben we $F(u) + F(-u) = 2\pi\delta(u)$ en dus $F(0) = \pi\delta(u)$.

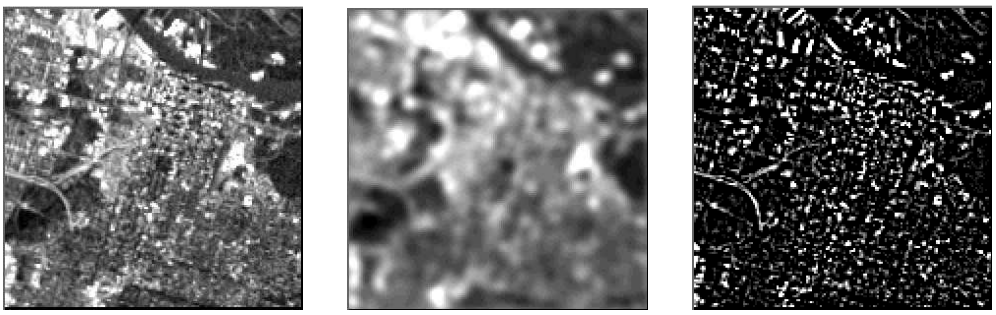
Bij elkaar genomen hebben we dus voor de sprong functie $H(t)$ aangetoond dat

$$\mathcal{F}[H(t)] = \pi\delta(u) + \frac{1}{iu}.$$

9.3 Toepassing: Filters

Het idee bij een filter is dat we zekere frequentieaandelen in een signaal willen versterken en andere onderdrukken. Dit gebeurt typisch door de Fourier getransformeerde met een geschikte functie te vermenigvuldigen en het resultaat met de inverse Fourier transformatie terug naar het tijdsdomein te transformeren:

Principe van een filter: Een functie $f(t)$ met $\mathcal{F}[f(t)] = F(u)$ wordt in het frequentiedomein met behulp van de filter functie $H(u)$ tot de nieuwe functie $G(u) = F(u)\cdot H(u)$ veranderd. Hieruit wordt met behulp van de inverse Fourier transformatie het *gefilterde signaal* $g(t) = \mathcal{F}^{-1}[G(u)] = f(t) * h(t)$ verkregen, waarbij $h(t) = \mathcal{F}^{-1}[H(u)]$ de inverse Fourier getransformeerde van $H(u)$ is.



Figuur II.14: Origineel, low-pass en high-pass gefilterd versie van een satelliet foto.

In Figuur II.14 is een voorbeeld van de toepassing van een low-pass en een high-pass filter op een satelliet opname te zien. Bij een low-pass filter worden de lage frequenties doorgelaten en de hoge onderdrukt en bij een high-pass filter is het omgekeerd, de lage frequenties worden onderdrukt en de hoge doorgelaten. Het effect is, dat met een low-pass filter het grove patroon van een signaal hetzelfde blijft, maar de scherpe contrasten tot een meer geleidelijke overgang verzacht worden. Omgekeerd benadrukt een high-pass filter alleen maar de scherpe contrasten, terwijl bijvoorbeeld de informatie over de intensiteit van het signaal verloren gaat. Zo is in het rechterplaatje niet meer te achterhalen of een gebied in het origineel licht of donker grijs is, maar wel heel duidelijk waar de grijstonen sterk veranderen.

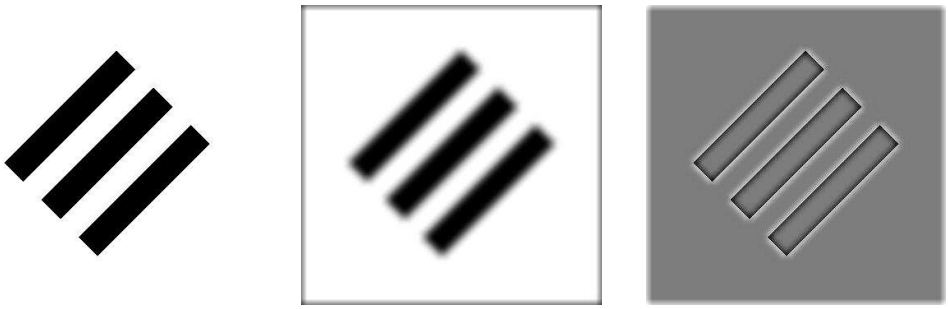
Belangrijke typen en voorbeelden van filters zijn:

(1) Low pass filter: De lage frequenties corresponderen met de grove kenmerken over grotere gebieden. Een filter die deze frequenties benadrukt en hogere frequenties onderdrukt heet een *low-pass filter*. In de beeldverwerking zal zo'n filter een zachter beeld tot gevolg hebben, waarbij de scherpe contrasten afgezwakt zijn.

(2) High pass filter: De hoge frequenties corresponderen met snelle veranderingen. Een filter die lage frequenties onderdrukt en hoge frequenties benadrukt heet *high-pass filter*. Bij de beeldverwerking worden hierdoor scherpe contrasten, zo als knikken benadrukt en hierdoor kan een beeld scherper lijken.

(3) Gauss filter: We hebben gezien dat een Gauss functie de mooie eigenschap heeft dat ook zijn Fourier getransformeerde weer een Gauss functie is. Dit maakt het omschakelen tussen tijd en frequentiedomein bijzonder eenvoudig.

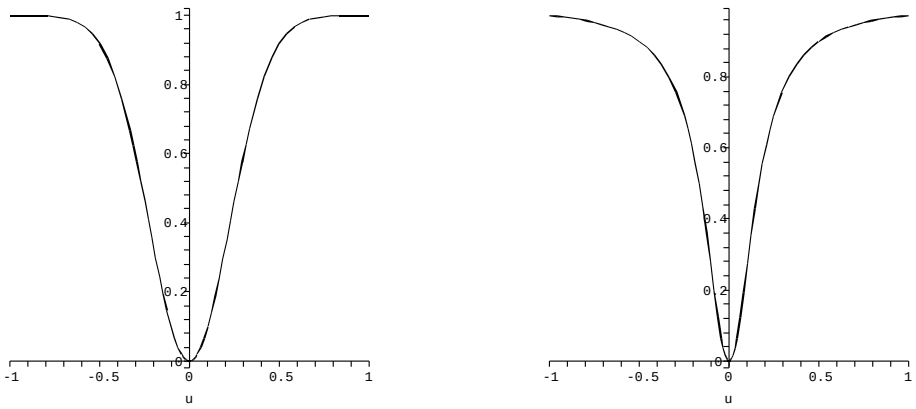
Met $H(u) = A \cdot e^{-\frac{u^2}{2\sigma^2}}$ krijgt men een low-pass filter waarbij de parameter σ (de standaardafwijking) aangeeft, hoe breed de filter is, d.w.z. voor een grotere σ worden meer frequenties doorgelaten, terwijl een kleine σ bijna alle frequenties wegdraait.



Figuur II.15: Origineel, low-pass en high-pass gefilterd versie van een eenvoudig plaatje.

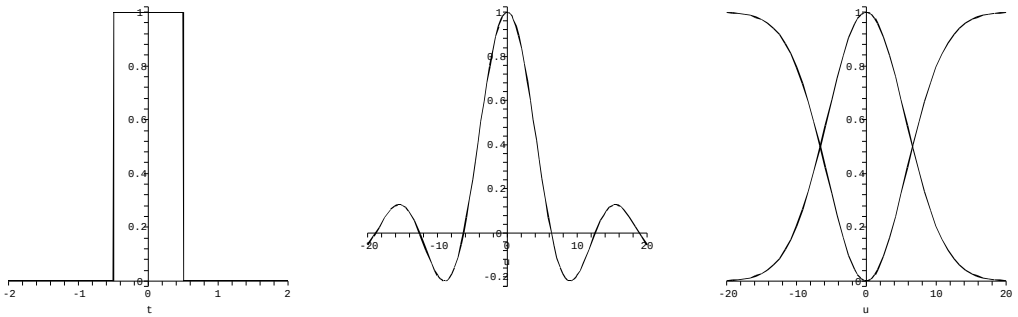
Een Gauss high-pass filter krijgt men in het eenvoudigste geval als verschil van een constante functie en een Gauss low-pass filter, namelijk als $H(u) = A(1 - e^{-\frac{u^2}{2\sigma^2}})$. Algemener neemt men vaak het verschil van twee Gauss functies met parameters σ en ρ: $H(u) = A \cdot e^{-\frac{u^2}{2\sigma^2}} - B \cdot e^{-\frac{u^2}{2\rho^2}}$.

(4) Notch filter: In het punt $u = 0$ geeft de Fourier getransformeerde het gemiddelde $F(0) = \int_{-\infty}^{\infty} f(t)\, dt$ van de functie $f(t)$ aan. Door dit op 0 te zetten, wordt bijvoorbeeld de helderheid van plaatjes genormeerd. De hiervoor benodigde filter heeft in het frequentiedomein de functie $H(0) = 0$ en $H(u) = 1$ voor $u \neq 0$, maar deze functie is natuurlijk niet continu. Om wel met een continue functie te kunnen werken moeten we rond $u = 0$ een scherpe dip hebben. In principe kunnen we hiervoor een high-pass filter nemen, die alleen maar de frequenties heel dicht bij $u = 0$ onderdrukt. Een mogelijkheid hiervoor is een Gauss high-pass filter $H(u) = 1 - e^{-\frac{u^2}{2\sigma^2}}$ met een kleine waarde van σ. Maar ook een functie zo als $H(u) = \frac{(au)^2}{1+(au)^2}$ met een grote waarde van a is geschikt.



Figuur II.16: Notch filters: links Gauss high-pass filter, rechts high-pass filter met functie $\frac{(au)^2}{1+(au)^2}$.

Voorbeeld: Rechthoek impuls



Figuur II.17: Rechthoek impuls, Fourier getransformeerde en Gauss low-pass en high-pass filters

We kijken naar een rechthoek impuls $f(t)$ van breedte 1 en sterkte 1. Voor deze functie hebben we in het begin van deze les al de Fourier getransformeerde bepaald, namelijk

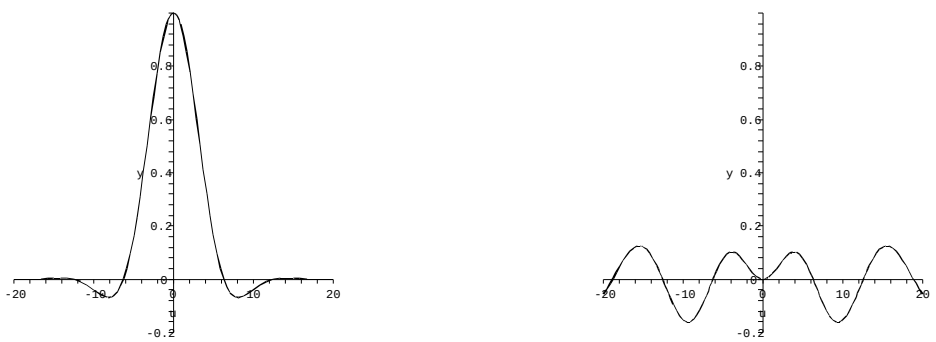
$$F(u) = \frac{\sin(\frac{1}{2}u)}{\frac{1}{2}u}.$$

De kleinste positieve nulpunt van $F(u)$ is 2π, zo als ook in het middelste plaatje van Figuur II.17 te zien is.

We gaan nu een Gauss low-pass en high-pass filter op dit signaal toepassen en kiezen hiervoor (enigszins willekeurig) de filter functie

$$H_l(u) = e^{-\frac{u^2}{20\pi}}.$$

De breedte van de Gauss functie $H_l(u)$ is zo gekozen, dat bij het minimum van $F(u)$ op $u = 3\pi$ de filter functie ongeveer bij 0.5 zit.



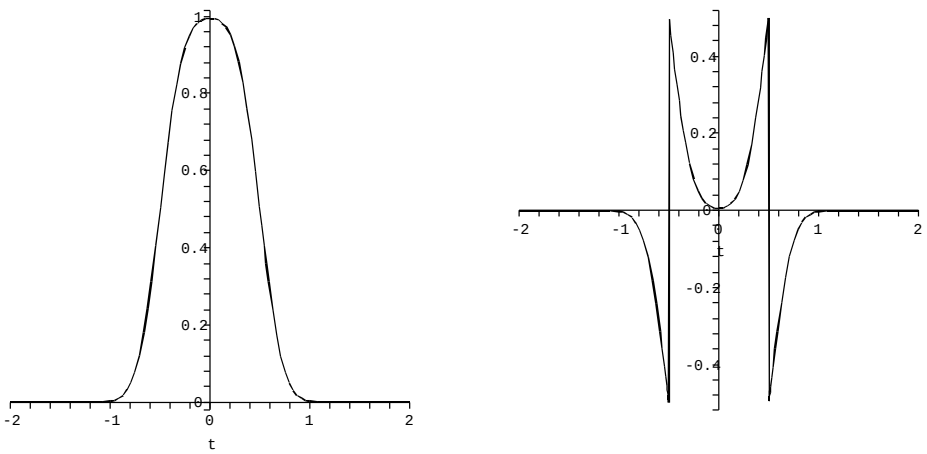
Figuur II.18: Product van Fourier getransformeerde met Gauss low-pass en high-pass filters

Als high-pass filter kiezen we

$$H_h(u) = 1 - H_l(u) = 1 - e^{-\frac{u^2}{20\pi}}.$$

De twee filter functies zijn in het rechterplaatje van Figuur II.17 te zien.

Zo als gezegd is het idee bij een filter dat we zekere frequenties onderdrukken en andere gewoon doorlaten, hiervoor vermenigvuldigen we de Fourier getransformeerde van het signaal met de filter functie. Voor de low-pass filter krijgen we zo de functie $G_l(u) = F(u) \cdot H_l(u)$ en voor de high-pass filter $G_h(u) = F(u) \cdot H_h(u)$. De functies $G_l(u)$ en $G_h(u)$ zijn in Figuur II.18 te zien en het is duidelijk hoe de low-pass filter de heuvels in de Fourier getransformeerde $F(u)$ onderdrukt, behalve van de hoofdspits in $u = 0$. Omgekeerd verdwijnt bij de high-pass filter deze hoofdspits helemaal, terwijl de andere maxima en minima voor hogere frequenties duidelijk zichtbaar blijven.



Figuur II.19: Inverse Fourier transformatie van het product met de filter functies geeft de gefilterde signalen

Het gefilterde signaal krijgen we nu door de inverse Fourier transformatie op het product van de Fourier getransformeerde en de filter functie in het frequentiedomein toe te passen, het low-pass gefilterde signaal is dus

$$f_l(t) = \mathcal{F}^{-1}[F(u) \cdot H_l(u)]$$

en het high-pass gefilterde signaal is

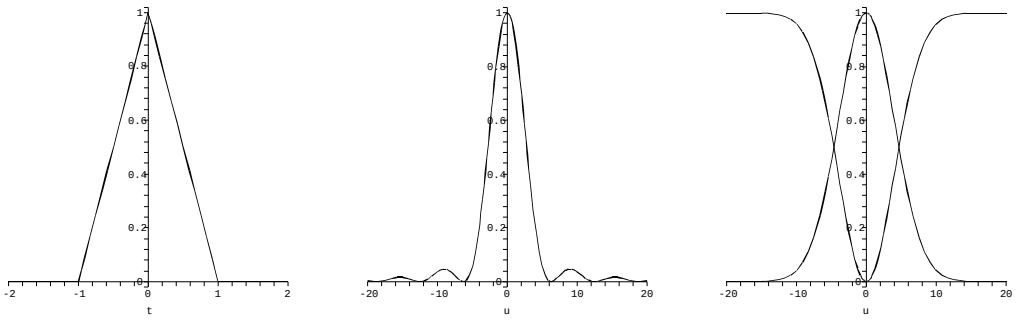
$$f_h(t) = \mathcal{F}^{-1}[F(u) \cdot H_h(u)].$$

Deze signalen zijn in Figuur II.19 te zien. Het is duidelijk dat de low-pass filter het signaal grofweg bewaart, maar de scherpe knikken tot ronde bochten verzacht. In tegenstelling hiermee geeft het high-pass gefilterde signaal aan, waar het signaal sterk verandert, namelijk in de punten $t = \pm 0.5$ van de knikken. In het punt $t = 0$ waar het signaal de grootste intensiteit heeft, geeft de high-pass filter zelf de waarde 0, want hier veranderd het signaal niet. Dit is in

overeenstemming met de waarnemingen bij Figuur II.14: Als de grijs-intensiteit op een gebied nauwelijks veranderd, geeft de high-pass filter steeds eenzelfde grijs-kleuring, onafhankelijk of het origineel in dit gebied licht of donker grijs is.

Het feit dat in dit voorbeeld $H_l(u) + H_h(u) = 1$ heeft tot gevolg dat $F(u) = F(u) \cdot H_l(u) + F(u) \cdot H_h(u)$ en hieruit volgt dat $f(t) = \mathcal{F}^{-1}[F(u)] = \mathcal{F}^{-1}[F(u) \cdot H_l(u)] + \mathcal{F}^{-1}[F(u) \cdot H_h(u)]$. Dit betekent dat in dit voorbeeld het originele signaal de som van het low-pass gefilterde signaal en het high-pass gefilterde signaal is.

Voorbeeld: Driehoek impuls



Figuur II.20: Driehoek impuls, Fourier getransformeerde en Gauss low-pass en high-pass filters

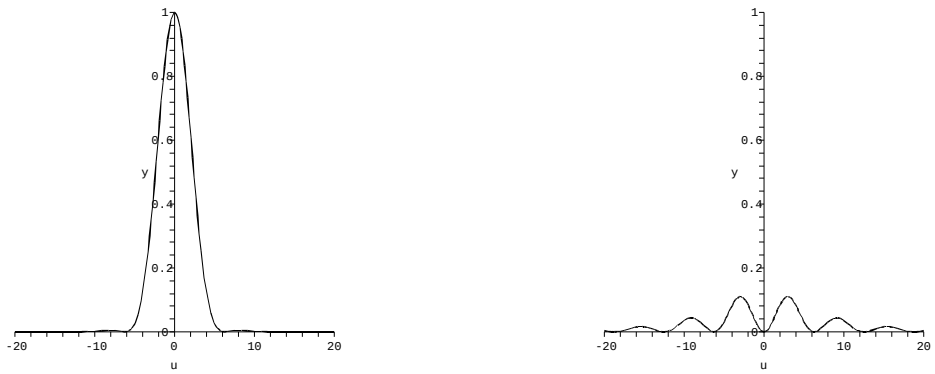
We kijken op een soortgelijke manier naar het voorbeeld van een driehoek impuls $f(t)$ tussen -1 en 1 die in $t = 0$ de sterkte 1 heeft. Zo als eerder in deze les gevonden, heeft $f(t)$ de Fourier getransformeerde

$$\mathcal{F}[f(t)] = F(u) = \left(\frac{\sin(\frac{1}{2}u)}{\frac{1}{2}u} \right)^2 .$$

In dit geval passen we een scherpere low-pass filter toe, die al op de frequentie $u = 2\pi$ waar $F(u)$ het kleinste positieve nulpunt heeft een waarde van ongeveer 0.5 heeft. De high-pass filter definiëren we weer als verschil $H_h(u) := 1 - H_l(u)$:

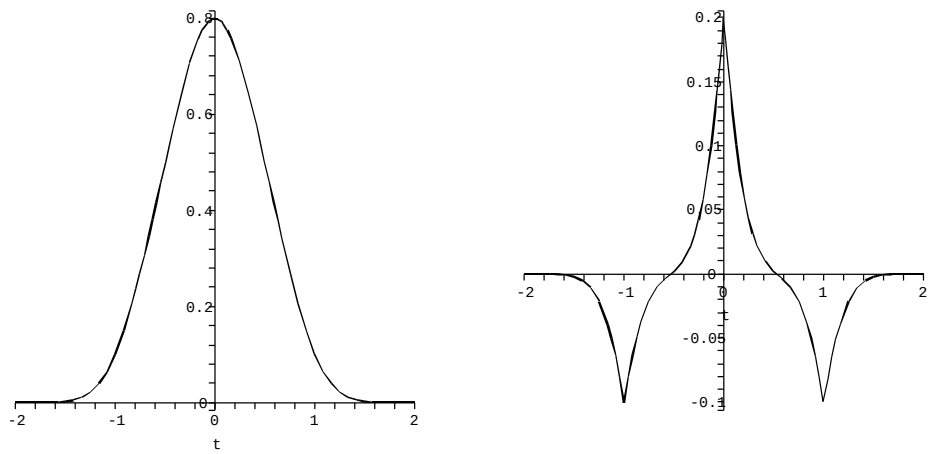
$$H_l(u) = e^{-\frac{u^2}{10\pi}}, \quad H_h(u) = 1 - H_l(u) = 1 - e^{-\frac{u^2}{10\pi}}.$$

Als we in dit geval naar de producten van de Fourier getransformeerde $F(u)$ met de filter functies kijken, zien we duidelijk dat de low-pass filter alleen maar de hoofdspits doorlaat en de rest van de Fourier getransformeerde onderdrukt. Interessant is hier het product met de high-pass filter. De functie $F(u)$ heeft het eerste positieve maximum bij $u = 3\pi$, maar in het rechterplaatje van Figuur II.21 is duidelijk te zien dat het product $F(u) \cdot H_h(u)$ in $u = \pi$ al een maximum heeft. Met het blote oog is dit verschil tussen de hoofdspitsen van $F(u)$ en $F(u) \cdot H_l(u)$ nauwelijks te zien omdat de absolute hoogte van de spits overweegt.



Figuur II.21: Product van Fourier getransformeerde met Gauss low-pass en high-pass filters

Maar de inverse Fourier transformatie van de producten in het frequentiedomein maakt duidelijk dat de subtiele verschillen een belangrijke effect op het signaal hebben.



Figuur II.22: Inverse Fourier transformatie van het product met de filter functies geeft de gefilterde signalen

Bij de gefilterde signalen in Figuur II.22 zien we in het linkerplaatje dat de low-pass filter de vorm van de driehoek impuls ongeveer bewaart, maar wel de knikken behoorlijk verzacht. Dit is het gevolg van onze keuze van een relatief scherpe low-pass filter. Het high-pass gefilterde signaal in het rechterplaatje laat duidelijk de drie knikken van de driehoek impuls bij $t = -1$, $t = 0$ en $t = 1$ zien.

Belangrijke begrippen in deze les

- Fourier getransformeerde van rechthoek impuls

- Fourier getransformeerde van Gauss functies zijn Gauss functies
- Dirac δ-functie
- low-pass en high-pass filters

Opgaven

73. Zij $f(t)$ de functie gegeven door

$$f(t) := \begin{cases} 1 - t^2 & \text{als} \quad |t| \leq 1 \\ 0 & \text{als} \quad |t| > 1. \end{cases}$$

(i) Maak een schets van de functie.

(ii) Bepaal de Fourier getransformeerde $F(u) = \mathcal{F}[f(t)]$ van $f(t)$.

(iii) Maak een schets van de Fourier getransformeerde $F(u)$.

74. Bepaal de Fourier getransformeerde van de functie

$$f(t) := e^{-a|t|} \quad \text{met } a > 0.$$

75. Bepaal de Fourier getransformeerden van

(i) $f(t) := \cos(\omega t)$;

(ii) $f(t) := \sin(\omega t)$.

76. Bepaal de Fourier getransformeerden van

(i) $f(t) := e^{iat^2}$;

(ii) $f(t) := \cos(at^2)$;

(iii) $f(t) := \sin(at^2)$.

(Hint: Zonder afleiding mag je de oneindige integralen $\int_{-\infty}^{\infty} \sin(x^2)\, dx = \sqrt{\frac{\pi}{2}}$ en $\int_{-\infty}^{\infty} \cos(x^2)\, dx = \sqrt{\frac{\pi}{2}}$ gebruiken.)

77. Laat zien dat voor de *signum functie*, gegeven door

$$f(t) := \begin{cases} -1 & \text{als} \quad t < 0 \\ 1 & \text{als} \quad t > 0 \end{cases}$$

geldt, dat $\mathcal{F}[f(t)] = \dfrac{2}{iu}$.

Les 10 Discrete Fourier transformatie

We hebben in de vorige lessen gezien hoe we met behulp van de Fourier transformatie een alternatieve beschrijving van een signaal in het frequentiedomein kunnen berekenen. Helaas hebben we het in de praktijk bijna nooit met signalen te maken die zich door eenvoudige combinaties van simpele continue functies zo als $\cos(x)$ of $\sin(x)$ of door rechthoek of driehoek impulsen laten beschrijven. Meestal kunnen we zelfs de functiewaarden $f(t)$ van een signaal alleen maar door een meting bepalen omdat het signaal a priori onbekend is. Dit maakt het natuurlijk onmogelijk, de Fourier getransformeerde van $f(t)$ volgens de (theoretische) formule

$$F(u) = \int_{-\infty}^{\infty} f(t)e^{-iut}\, dt$$

uit te rekenen, waarbij we het nog niet eens over de oneindige grenzen hebben.

Het idee om uit dit dilemma te ontsnappen is, de functie door voldoende metingen zo goed te beschrijven, dat we niettemin belangrijke informatie over de functie in het tijds- en frequentiedomein krijgen.

10.1 Discretisering

Het beschrijven van een functie door in zeker afstanden de functiewaarde te meten noemt men *sampling*, het resultaat van een sampling is een *discretisering* van de functie.

Als men een tijdsinterval Δt en het aantal N van metingen kiest, beschrijft men een functie $f(t)$ op de tijdstippen $t_k := k \cdot \Delta t$ voor $k = 0, 1, \ldots, N-1$ door de N discrete waarden $f_0, f_1, \ldots, f_{N-1}$ gegeven door

$$f_0 := f(0 \cdot \Delta t), \ldots, f_k := f(k \cdot \Delta t), \ldots, f_{N-1} := f((N-1) \cdot \Delta t).$$

> **Merk op:** Het feit dat we steeds met $t_0 = 0$ beginnen is geen echte beperking, want we kunnen door een verschuiving in het tijdsdomein steeds ervoor zorgen dat de eerste meting op het tijdstip $t = 0$ plaats vindt.

Met betrekking tot de Fourier transformatie kijkt men nu ook naar een discrete versie van de Fourier getransformeerde in het frequentiedomein. We hebben N metingen met een afstand van Δt, dus metingen over een tijdsinterval van $T = N \cdot \Delta t$. Het beste dat we zouden kunnen verwachten is dat de functie $f(t)$ periode T heeft, want dan kunnen we elke functiewaarde $f(t)$ identificeren met een functiewaarde in het interval $[0, T]$ en vervolgens deze functiewaarde door de dichtstbij liggende meting benaderen.

Als we nu eens veronderstellen, dat $f(t)$ inderdaad periodiek met periode T is, dan kunnen we $f(t)$ in een Fourier reeks $f(t) = \sum_{k=-\infty}^{\infty} c_k e^{ik\omega t}$ ontwikkelen waarbij $\omega = \frac{2\pi}{T}$ de grondfrequentie is.

De naburige frequenties die een rol in de Fourier reeks van $f(t)$ spelen, hebben dus een verschil van $\frac{2\pi}{T}$ en dit interpreteren we als *afstand* $\Delta\omega$ van de frequenties waarover we informatie uit $f(t)$ kunnen halen.

Dit idee veralgemenen we nu naar functies $f(t)$ die niet periodiek zijn en definiëren de discrete frequenties ω_j door $\omega_j := j \cdot \Delta\omega = j \cdot \frac{2\pi}{T}$. Tenslotte beslissen we nog dat het aantal discrete waarden in het frequentiedomein even groot moet zijn als het aantal tijdstippen in het tijdsdomein, dus juist N. Het interval in het frequentiedomein dat we zo overdekken is $N \cdot \Delta\omega = \frac{2\pi N}{T} = \frac{2\pi}{\Delta t}$.

Definitie: Een *discretisering* van een functie $f(t)$ op N tijdstippen t_k met afstand Δt is gegeven door de functiewaarden

$$f_k := f(t_k) = f(k \cdot \Delta t) \text{ met } t_k := k \cdot \Delta t \text{ voor } k = 0, 1, \ldots, N-1.$$

Voor de discretisering in het frequentiedomein zij

$$T := N \cdot \Delta t \text{ en } \Delta\omega := \frac{2\pi}{T} = \frac{1}{N} \cdot \frac{2\pi}{\Delta t},$$

dan zijn de discrete frequenties ω_j gegeven door

$$\omega_j := j \cdot \Delta\omega = j \cdot \frac{2\pi}{T} = \frac{j}{N} \cdot \frac{1}{\Delta t} \text{ voor } j = 0, 1, \ldots, N-1.$$

Door deze definities krijgen we in het bijzonder de relatie

$$N \cdot \Delta t \cdot \Delta\omega = 2\pi$$

die zegt dat we bij een constant aantal N van metingen een kleinere tijdelijke afstand Δt in het tijdsdomein moeten compenseren door een grotere afstand $\Delta\omega$ in het frequentiedomein, en andersom.

10.2 De discrete Fourier transformatie

In analogie met de Fourier reeks en de Fourier transformatie proberen we nu de waarden $f(t_k)$ op de discrete tijdstippen $t_k = k \cdot \Delta t$ te beschrijven door informatie voor de frequenties $\omega_j = j \cdot \Delta\omega$ in het frequentie domein. Als $f(t)$ een periodieke functie met periode T en $\Delta\omega = \frac{2\pi}{T}$ was, konden we $f_k = f(t_k)$ schrijven als

$$f_k = f(t_k) = \sum_{j=-\infty}^{\infty} c_j e^{ij\Delta\omega t_k}.$$

Aan de andere kant geldt voor de Fourier transformatie dat

$$f_k = f(t_k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(u) e^{iut_k}\, du.$$

Bij de discretisering moeten we ons beperken tot informatie over de discrete frequenties $\omega_j = j \cdot \Delta\omega$ voor $j = 0, 1, \ldots, N-1$, dus moeten we de Fourier reeks en de Fourier transformatie als volgt veranderen:

- In de Fourier reeks kan j alleen maar van 0 tot $N-1$ lopen.

- De integraal in de Fourier transformatie moet vertaald worden naar een som over de termen e^{iut_k} met $u = \omega_j = j\Delta\omega$ voor $j = 0, 1, \ldots, N-1$.

Uit beide invalshoeken komen we tot de conclusie dat we $f_k = f(t_k)$ moeten schrijven als lineaire combinatie van de termen

$$e^{i(j\Delta\omega)t_k} = e^{i(j\Delta\omega)(k\Delta t)} = e^{ijk\Delta\omega\Delta t}$$

met *geschikte coëfficiënten* F_j voor $j = 0, 1, \ldots, N-1$. Dit geeft al de formele opzet voor de discrete Fourier transformatie, preciezer gezegd voor de inverse discrete Fourier transformatie, namelijk

$$f_k = f(t_k) := f(k \cdot \Delta t) = \frac{1}{N} \sum_{j=0}^{N-1} F_j \, e^{ij\Delta\omega t_k} = \frac{1}{N} \sum_{j=0}^{N-1} F_j \, e^{ijk\Delta\omega\Delta t}.$$

Hierbij is de factor $\frac{1}{N}$ net als bij de Fourier reeks en de Fourier transformatie een normeringsfactor die enigszins willekeurig op de Fourier transformatie en de inverse Fourier transformatie opgedeeld zal moeten worden.

De vraag is nu hoe we de coëfficiënten F_j kunnen bepalen. Hiervoor zijn er verschillende mogelijkheden, die gelukkig alle tot hetzelfde resultaat lijden namelijk:

$$F_j = \sum_{l=0}^{N-1} f(t_l) \, e^{-i\,j\Delta\omega\,l\Delta t}.$$

Als we dit in de boven aangegeven formule voor de inverse discrete Fourier transformatie invullen, krijgen we de discrete versie van de Fourier integraal formule en dit geeft aanleiding tot het volgende resultaat over de discrete Fourier transformatie:

Discrete Fourier transformatie

De *discrete Fourier integraal formule* luidt

$$f_k = f(t_k) = \frac{1}{N} \sum_{j=0}^{N-1} \left(\sum_{l=0}^{N-1} f(t_l) \, e^{-i\,j\Delta\omega\,t_l} \right) e^{i\,j\Delta\omega\,t_k}.$$

De *discrete Fourier transformatie* is gegeven door

$$F_j = \sum_{l=0}^{N-1} f(t_l) \, e^{-i\,j\Delta\omega\,t_l} = \sum_{l=0}^{N-1} f_l \, e^{-i\,jl\,\Delta\omega\,\Delta t}$$

en de *discrete inverse Fourier transformatie* door

$$f_k = f(t_k) = \frac{1}{N} \sum_{j=0}^{N-1} F_j \, e^{i\,j\Delta\omega\,t_k} = \frac{1}{N} \sum_{j=0}^{N-1} F_j \, e^{i\,jk\,\Delta\omega\,\Delta t}.$$

Merk op: Vaak wordt een gesampelde functie en zijn discrete Fourier transformatie gewoon als rij van waarden aangegeven, bijvoorbeeld in de vorm

$$\{f_0, \ldots, f_k, \ldots, f_{N-1}\} \quad \text{en} \quad \{F_0, \ldots, F_j, \ldots, F_{N-1}\}.$$

Voor de geïnteresseerde lezer geven we nu drie manieren aan, hoe dit resultaat afgeleid kan worden:

(1) middels projecties op een orthogonaal stelsel van functies;

(2) middels trigonometrische interpolatie;

(3) middels de Dirac δ-functie.

Afleiding middels een orthogonaal stelsel van functies

Het idee bij de Fourier reeks en bij de Fourier transformatie was in principe, een orthogonaal stelsel van functies (met betrekking tot een geschikt inproduct) te kiezen en de orthogonale projecties van de functie $f(t)$ op deze functies te berekenen. Iets soortgelijks gebeurt in principe ook bij de discrete Fourier transformatie. Als de tijdstippen $t_0, t_1, \ldots, t_{N-1}$ vast gekozen zijn, kunnen we een inproduct voor complexe functies definiëren door

$$\langle f(t), g(t)\rangle := \sum_{k=0}^{N-1} f(t_k) \cdot \overline{g(t_k)}.$$

We laten nu zien dat de functies $e^{ij\Delta\omega t}$ voor $j = 0, 1, \ldots, N-1$ een orthogonaal stelsel met betrekking tot dit inproduct vormen.

Hulpstelling: Door uitschrijven van het product ziet men algemeen dat

$$(1-a)(1+a+a^2+\ldots+a^{N-1}) = 1-a^N$$

en hieruit volgt dat

$$\sum_{k=0}^{N-1} a^k = 1+a+a^2+\ldots+a^{N-1} = \begin{cases} \frac{1-a^N}{1-a} & \text{als} \quad a \neq 1 \\ N & \text{als} \quad a = 1 \end{cases}$$

Deze hulpstelling passen we nu op het inproduct van de functies $e^{ij\Delta\omega t}$ en $e^{il\Delta\omega t}$ toe, er geldt:

$$\langle e^{ij\Delta\omega t}, e^{il\Delta\omega t}\rangle = \sum_{k=0}^{N-1} e^{ij\Delta\omega t_k} \cdot e^{-il\Delta\omega t_k} = \sum_{k=0}^{N-1} e^{i(j-l)\Delta\omega t_k} = \sum_{k=0}^{N-1} (e^{i(j-l)\Delta\omega\Delta t})^k.$$

Merk op dat we $\Delta\omega$ zo hebben gekozen dat $\Delta\omega\Delta t \cdot N = 2\pi$, dus $\Delta\omega\Delta t = \frac{2\pi}{N}$. Hieruit volgt dat

$$(e^{i(j-l)\Delta\omega\Delta t})^N = e^{i(j-l)\Delta\omega\Delta t N} = (e^{2\pi i})^{j-l} = 1.$$

Aan de andere kant is

$$e^{i(j-l)\Delta\omega\Delta t} = e^{i(j-l)\frac{2\pi}{N}} = e^{2\pi i \frac{j-l}{N}} \begin{cases} \neq 1 & \text{als} \quad j \neq l \\ = 1 & \text{als} \quad j = l \end{cases}$$

omdat j en l van 0 tot $N-1$ lopen en dus $j-l$ nooit een veelvoud van N kan zijn. Bij elkaar genomen volgt hieruit met de hulpstelling dat

$$\langle e^{ij\Delta\omega t}, e^{il\Delta\omega t}\rangle = \sum_{k=0}^{N-1} (e^{i(j-l)\Delta\omega\Delta t})^k = \begin{cases} 0 & \text{als} \quad j \neq l \\ N & \text{als} \quad j = l. \end{cases}$$

Net als bij de Fourier reeks vinden we nu coëfficiënten c_j met $f(t_k) = \sum_{j=0}^{N-1} c_k e^{ij\Delta\omega t_k}$ door

$$c_j = \frac{\langle f(t), e^{ij\Delta\omega t}\rangle}{\langle e^{ij\Delta\omega t}, e^{ij\Delta\omega t}\rangle} = \frac{1}{N}\langle f(t), e^{ij\Delta\omega t}\rangle$$

dus is de coëfficiënt F_j van $e^{ij\Delta\omega t_k}$ in $f(t_k) = \frac{1}{N}\sum_{j=0}^{N-1} F_j e^{ij\Delta\omega t_k}$ gegeven door

$$F_j = \langle f(t), e^{ij\Delta\omega t}\rangle = \sum_{k=0}^{N-1} f(t_k) e^{-ij\Delta\omega t_k} = \sum_{k=0}^{N-1} f_k\, e^{-ijk\,\Delta\omega\,\Delta t}.$$

Als controle vullen we dit eens in:

$$\begin{aligned}
\sum_{j=0}^{N-1} F_j e^{ij\Delta\omega t_k} &= \sum_{j=0}^{N-1} \left(\sum_{l=0}^{N-1} f(t_l) e^{-ij\Delta\omega t_l} \right) e^{ij\Delta\omega t_k} \\
&= \sum_{j=0}^{N-1} \left(\sum_{l=0}^{N-1} f(t_l) e^{-ij\Delta\omega l\Delta t} \right) e^{ij\Delta\omega k\Delta t} \\
&= \sum_{l=0}^{N-1} f(t_l) \left(\sum_{j=0}^{N-1} e^{ij\Delta\omega(k-l)\Delta t} \right) \\
&= \sum_{l=0}^{N-1} f(t_l) \left(\sum_{j=0}^{N-1} (e^{i(k-l)\Delta\omega\Delta t})^j \right) \\
&= f(t_k) \cdot N.
\end{aligned}$$

In de laatste stap hebben we weer de hulpstelling toegepast die zegt dat de som $\sum_{j=0}^{N-1}(e^{i(k-l)\Delta\omega\Delta t})^j$ alleen maar voor $l = k$ ongelijk aan 0 is en in dit geval de waarde N heeft. Van de som over l blijft dus alleen maar de term voor $l = k$ over en hiervoor krijgen we juist de waarde $f(t_k) \cdot N$.

Motivatie middels trigonometrische interpolatie

Een alternatieve formulering van deze toegang tot de discrete Fourier transformatie zit in de *trigonometrische interpolatie*. We weten dat er voor N verschillende x-waarden $x_0, x_1, \ldots, x_{N-1}$ met bijhorende y-waarden $y_0, y_1, \ldots, y_{N-1}$ een eenduidige veelterm $p(x) = a_0 + a_1 x + \ldots + a_{N-1}x^{N-1}$ van graad $N-1$ bestaat, zo dat $p(x_k) = y_k$ voor alle $k = 0, 1, \ldots, N-1$. Met noemt $p(x)$ de interpolatie van de gegeven punten, omdat de grafiek van $p(x)$ de punten (x_k, y_k) verbindt. Het idee achter het bewijs is simpel: De N paren van $x-y$-waarden geven N lineaire vergelijkingen voor de N coëfficiënten van $p(x)$, en omdat de functies

x^j lineair onafhankelijk zijn, heeft het bijhorende stelsel lineaire vergelijkingen een eenduidige oplossing.

Als men nu de functies $x^0, \ldots, x^j, \ldots$ door de functies $(e^{i\omega x})^0, \ldots, (e^{i\omega x})^j, \ldots$ vervangt die ook lineair onafhankelijk zijn, krijgt men een analoge uitspraak voor de interpolatie met behulp van deze functies. Wegens de relatie $e^{i\omega x} = \cos(\omega x) + i\sin(\omega x)$ spreekt men hierbij van *trigonometrische interpolatie*.

We noemen nu de variabel x weer t en vervangen ω door $\Delta\omega$, dan luidt het idee, de functie $f(t)$ te benaderen door de trigonometrische interpolatie $\tilde{f}(t) = \sum_{j=0}^{N-1} c_j e^{ij\Delta\omega t}$ die gedefinieerd is door de eigenschappen dat $\tilde{f}(t_k) = f(t_k)$ voor $k = 0, \ldots N-1$. Als men hierop de orthogonaliteitsrelaties voor de functies $e^{ij\Delta\omega t}$ toepast, vindt men weer dat de coëfficiënten c_j voldoen aan

$$c_j = \sum_{k=0}^{N-1} f(t_k) e^{-ijk\Delta\omega\Delta t}.$$

Motivatie middels Dirac δ-functie

Omdat we bij de Fourier transformatie de Dirac δ-functie behandeld hebben, kunnen we nog een andere motivatie voor de coëfficiënten van de discrete Fourier transformatie geven die hierop gebaseerd is.

We weten van de functie $f(t)$ alleen maar de waarden die we op de tijdstippen $t_k = k \cdot \Delta t$ voor $k = 0, 1, \ldots, N-1$ gemeten hebben, dus de functiewaarden $f_k = f(t_k)$. We vervangen de functie $f(t)$ nu door de gesampelde functie

$$f_s(t) := \sum_{k=0}^{N-1} f(t_k) \cdot \delta(t - t_k)$$

die uit (oneindige) spitsen op de tijdstippen t_k bestaat die de gemeten waarden $f(t_k)$ als gewichten hebben. De gesampelde functie heeft de eigenschap dat $\int_{-\infty}^{\infty} f_s(t)\ dt = \sum_{k=0}^{N-1} f(t_k)$ en dat de relatieve intensiteiten van de spitsen evenredig zijn met de gemeten functiewaarden.

Voor de functie $f_s(t)$ kunnen we de Fourier getransformeerde makkelijk berekenen, dit is

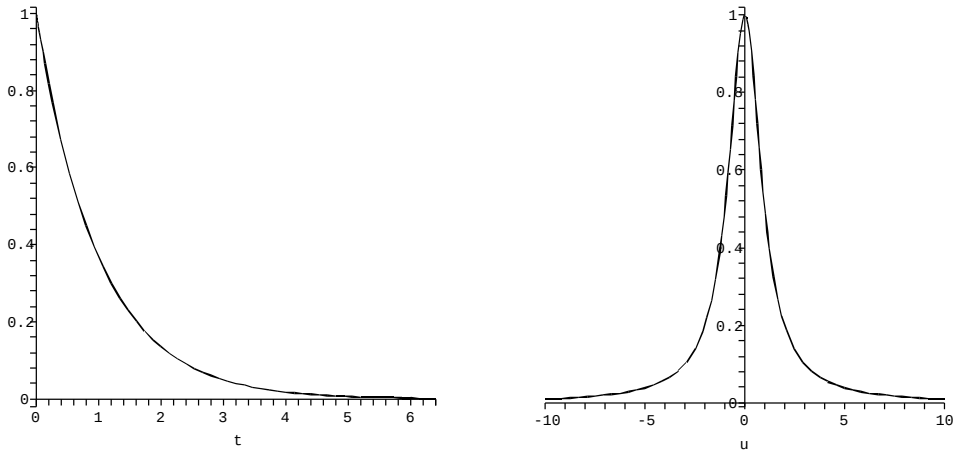
$$F_s(u) = \mathcal{F}[f_s(t)] = \sum_{k=0}^{N-1} f(t_k) e^{-iut_k}.$$

Als we de functiewaarden van $F_s(u)$ nu voor de discrete frequenties $u = \omega_j = j \cdot \Delta\omega$ bepalen, krijgen we

$$F_s(\omega_j) = f(t_k) e^{-ij\Delta\omega t_k} = F_j$$

dus de functiewaarden van de Fourier getransformeerde van de gesampelde functie op de discrete frequenties ω_j zijn precies de coëfficiënten van de discrete Fourier transformatie. Boven hebben we al gezien hoe we de discrete functiewaarden $f(t_k)$ uit de discrete waarden $F_s(\omega_j) = F_j$ van de Fourier getransformeerde kunnen reproduceren, namelijk juist met de discrete inverse Fourier transformatie.

10.3 Voorbeeld van een discrete Fourier transformatie



Figuur II.23: Exponentiële afname $f(t) = e^{-t}$ met Fourier getransformeerde.

We gaan nu eens een voorbeeld van een discrete Fourier transformatie bekijken. Hiervoor kiezen we de functie

$$f(t) := \begin{cases} e^{-t} & \text{als} \quad t \geq 0 \\ 0 & \text{als} \quad t < 0 \end{cases}$$

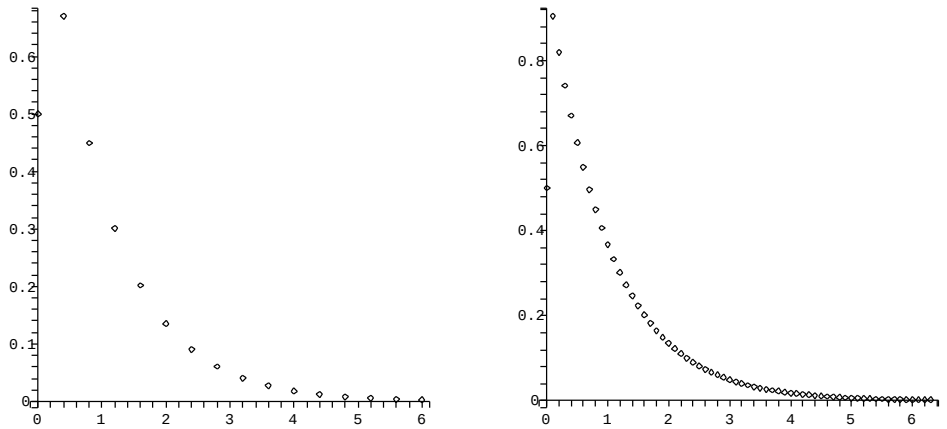
die een exponentiële afname beschrijft. De Fourier getransformeerde van $f(t)$ hebben we al eerder bepaald, er geldt

$$F(u) = \mathcal{F}[f(t)] = \frac{1}{1+iu} \quad \text{met} \quad \Re(F(u)) = \frac{1}{1+u^2}.$$

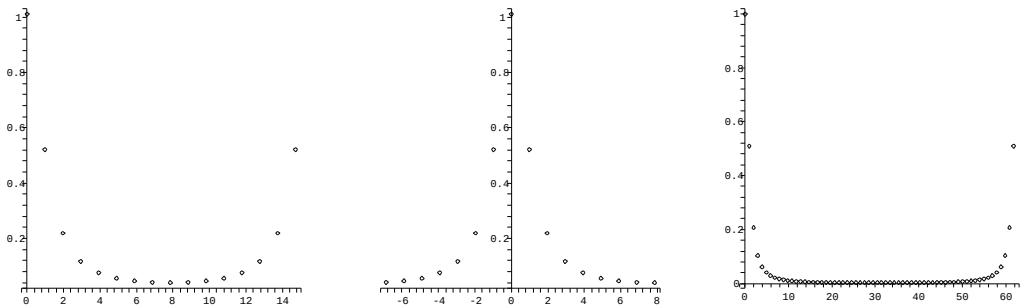
We sampeln de functie $f(t)$ op twee manieren: Een keer met $N = 16$ punten op een afstand van $\Delta t = 0.4$ en een keer met $N = 64$ punten op een afstand van $\Delta t = 0.1$. Dit is in Figuur II.24 te zien. Merk op dat de t-as net zo geschald is als bij de continue versie, dus met $k \cdot \Delta t$. In beide gevallen is $N \cdot \Delta t = 6.4$ en dus $\Delta\omega = \frac{2\pi}{6.4}$. In het eerste geval is $T = N \cdot \Delta\omega = 5\pi$, in het tweede geval is $N \cdot \Delta\omega = 20\pi$, we overdekken dus bij een sampling met vier keer zo veel punten over hetzelfde tijdsinterval $[0, T]$ een vier keer zo groot interval in het frequentie domein.

Als we het linker en het rechter plaatje in Figuur II.25 met de continue Fourier getransformeerde vergelijken, zien we dat we de discrete Fourier transformatie iets anders moeten interpreteren. Eigenlijk hoort namelijk (net zo als bij de Fourier reeks) bij elke coëfficiënt F_j met $j > 0$ ook een coëfficiënt met $j < 0$. Deze vinden we als volgt: We kunnen de definitie

$$F_j = \sum_{k=0}^{N-1} f(t_k) e^{-ij\Delta\omega k\Delta t}$$



Figuur II.24: Discretisering van een exponentiële afname met 16 (links) en 64 punten (rechts).



Figuur II.25: Discrete Fourier transformatie van een exponentiële afname gesampeld met 16 (links en midden) en 64 punten (rechts).

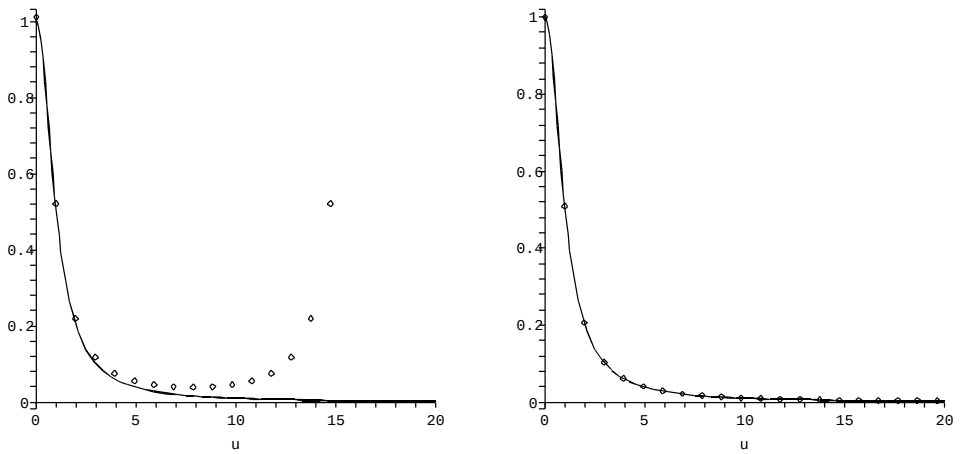
ook lezen voor j buiten het interval $[0 \dots N-1]$, en omdat $\Delta\omega\Delta t \cdot N = 2\pi$ en dus $e^{-iN\Delta\omega\Delta t} = 1$ is, geldt

$$F_{j+N} = \sum_{k=0}^{N-1} f(t_k) e^{-i(j+N)\Delta\omega k\Delta t} = \sum_{k=0}^{N-1} f(t_k) e^{-ij\Delta\omega k\Delta t} = F_j.$$

In het bijzonder betekent dit dat $F_{-j} = F_{N-j}$ en we moeten de coëfficiënten $F_{\frac{N}{2}}, \dots, F_{N-1}$ dus eigenlijk lezen als de coëfficiënten $F_{-\frac{N}{2}}, \dots, F_{-1}$.

Voor het gemak laten we in de plaatjes niettemin j van 0 tot $N-1$ lopen en denken ons bij het vergelijken met de continue Fourier transformatie het plaatje van de discrete Fourier getransformeerde in het midden door geknipt en de rechter helft van het plaatje links aan de linker helft geplakt. De zo veranderde versie van de discrete Fourier transformatie met 16 punten is in het middelste plaatje van Figuur II.25 te zien.

Ten slotte vergelijken we de discrete Fourier transformatie met de continue. In Figuur II.26 zijn de waarden van de coëfficiënten F_j als punten naast de grafiek van de continue Fourier getransformeerde $F(u)$ te zien. Ook in dit geval is de u-as zo geschald, dat de coëfficiënt F_j op de frequentie $u = j \cdot \Delta\omega$ terecht komt. Het is duidelijk dat bij een sampling met te grote tijdsintervallen de discrete Fourier getransformeerde al snel (en duidelijk voor de helft van de sampling frequenties) behoorlijk van de continue Fourier getransformeerde afwijkt. Een verhoging van het aantal metingen op hetzelfde tijdsinterval lijdt echter tot een duidelijk verbeterde benadering. Merk op dat in het rechter plaatje niet eens de helft van de 64 discrete frequentie waarden afgebeeld zijn.



Figuur II.26: Vergelijk van continue en discrete Fourier transformatie van de exponentiële afname.

10.4 Eigenschappen van de discrete Fourier transformatie

In principe heeft de discrete Fourier transformatie dezelfde eigenschappen als de gewone Fourier transformatie. Om deze aan te geven, houden we de volgende notaties aan:

Zij $f(t)$ een functie, die we op N punten $t_k = k \cdot \Delta t$ met een tijdelijke afstand van Δt sampeln. We definiëren $f_k := f(t_k)$, dan zijn voor $j = 0, 1, \ldots, N-1$ de coëfficiënten F_j van de discrete Fourier getransformeerde van $f(t)$ voor de frequenties $j \cdot \Delta\omega$ gegeven door $F_j = \sum_{k=0}^{N-1} f_k \, e^{-ijk\Delta\omega\Delta t}$. Hierbij geldt $\Delta\omega = \frac{2\pi}{N\Delta t}$ en dus $\Delta\omega\Delta t = \frac{2\pi}{N}$.

Voor verdere functies $g(t)$ en $h(t)$ nemen we aan dat deze met dezelfde N en Δt gediscretiseerd zijn en de waarden $g_k = g(t_k)$ en $h_k = h(t_k)$ hebben. De discrete Fourier getransformeerden van deze functies geven we met G_j en H_j aan.

Lineariteit

Omdat we functies puntsgewijs optellen, is de discrete Fourier getransformeerde van $f_k + g_k$ gelijk aan $F_j + G_j$.

Analoog geldt voor de vermenigvuldiging met een factor, dat $a \cdot f_k$ de discrete Fourier getransformeerde $a \cdot F_j$ heeft.

Verschuiving

Als de sampling waarden g_k van $g(t)$ om l posities tegenover de sampling waarden van $f(t)$ verschoven zijn, dus als $g_k = f_{k-l}$ is, dan geldt

$$G_j = F_j\, e^{-ijl\Delta\omega\Delta t} = F_j\, e^{-i\frac{2\pi}{N}jl},$$

waarbij we met F_j en G_j de discrete Fourier getransformeerden van f_k en g_k noteren. Dit gaat men als volgt na:

$$\begin{aligned} G_j &= \sum_{k=0}^{N-1} f_{k-l}\, e^{-ij\Delta\omega k\Delta t} = \sum_{k=l}^{l+N-1} f_{k-l}\, e^{-ij\Delta\omega k\Delta t} = \sum_{k=0}^{N-1} f_k\, e^{-ij\Delta\omega(k+l)\Delta t} \\ &= \left(\sum_{k=0}^{N-1} f_k\, e^{-ij\Delta\omega k\Delta t}\right) e^{-ij\Delta\omega l\Delta t} = F_j\, e^{-ijl\Delta\omega\Delta t}. \end{aligned}$$

Op een soortgelijke manier laat zich aantonen (zie opgaven), dat voor een verschoven functie $G_j = F_{j-l}$ in het frequentiedomein geldt, dat de discrete inverse Fourier getransformeerde g_k van G_j uit de getransformeerde f_k van F_j wordt verkregen door

$$g_k = f_k\, e^{ikl\Delta\omega\Delta t} = f_k\, e^{i\frac{2\pi}{N}kl}.$$

Convolutie

Het convolutieproduct $h_k = f_k * g_k$ van twee gediscretiseerde functies f_k en g_k is juist zo gedefinieerd als of f_k en g_k de coëfficiënten van de term x^k in twee veeltermen zijn en h_k de coëfficiënt van de term x^k in het product van deze veeltermen is. Dit geeft voor het convolutieproduct de definitie

$$h_k := \sum_{l=0}^{N-1} f_l\, g_{k-l} =: f_k * g_k.$$

In deze formule moeten we voor de coëfficiënten g_{k-l} met index $k - l < 0$ de coëfficiënt g_{k-l} met g_{N+k-l} identificeren, dus we moeten N bij de index optellen. We waren namelijk bij de definitie van de discrete Fourier transformatie ervan uit gegaan dat de gediscretiseerde functie periodiek met periode $T = N \cdot \Delta t$ is.

De belangrijke eigenschap van het convolutieproduct is nu net als in het continue geval, dat de discrete Fourier getransformeerde H_j van het convolutieproduct $h_k = f_k * g_k$ gelijk is aan het gewone product $F_j \cdot G_j$ van de discrete Fourier getransformeerden van f_k en g_k, dus dat

$$H_j := \sum_{k=0}^{N-1} h_k\, e^{-ijk\Delta\omega\Delta t} = \sum_{k=0}^{N-1} \left(\sum_{l=0}^{N-1} f_l\, g_{k-l}\right) e^{-ijk\Delta\omega\Delta t} = F_j \cdot G_j$$

voor $F_j = \sum_{k=0}^{N-1} f_k\, e^{-ijk\Delta\omega\Delta t}$ en $G_j = \sum_{k=0}^{N-1} g_k\, e^{-ijk\Delta\omega\Delta t}$. Dit ziet men als volgt in:

$$\begin{aligned}
H_j &= \sum_{k=0}^{N-1} h_k\, e^{-ijk\Delta\omega\Delta t} = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} f_l\, g_{k-l}\, e^{-ijk\Delta\omega\Delta t} \\
&= \sum_{l=0}^{N-1} f_l\, e^{-ijl\Delta\omega\Delta t}(\sum_{k=0}^{N-1} g_{k-l}\, e^{-ij(k-l)\Delta\omega\Delta t}) \\
&=_{k'=k-l} (\sum_{l=0}^{N-1} f_l\, e^{-ijl\Delta\omega\Delta t})(\sum_{k'=0}^{N-1} g_{k'}\, e^{-ijk'\Delta\omega\Delta t}) \\
&= F_j \cdot G_j.
\end{aligned}$$

In de voorlaatste stap maken we hierbij gebruik ervan dat de gesampelde waarden periodiek met periode N zijn, dus dat $g_k = g_{k+N}$ is. Hiermee volgt dat met k ook $k-l$ over alle waarden van 0 t/m $N-1$ loopt.

Op een soortegelijke manier (zie opgaven) laat zich aantonen dat het gewone product in het tijdsdomein met het convolutieproduct in het frequentiedomein correspondeert, hierbij komt echter nog een factor N te voorschijn. Voor gediscretiseerde functies f_k en g_k met discrete Fourier transformaties F_j en G_j geldt dat het gewone product $h_k := f_k \cdot g_k$ de getransformeerde $H_j = \frac{1}{N} F_j * G_j$ heeft, dus dat

$$H_j = \sum_{k=0}^{N-1} f_k g_k\, e^{-ijk\Delta\omega\Delta t} = \frac{1}{N} F_j * G_j = \frac{1}{N}\sum_{l=0}^{N-1} F_l\, G_{j-l}.$$

10.5 Snelle (discrete) Fourier transformatie (FFT)

Om bij een gesampelde functie $(f_0, \ldots, f_{N-1})$ de coëfficiënten F_j van de discrete Fourier transformatie te berekenen, zijn er voor elke coëfficiënt N vermenigvuldigingen nodig, voor alle coëfficiënten dus N^2 vermenigvuldigingen. Bij een typische waarde van bijvoorbeeld $N = 1024$ is dit al behoorlijk veel rekenwerk, omdat bij veranderlijke signalen vaak opnieuw gesampeld moet worden.

Een nauwkeurige analyse van het rekenwerk laat immers zien, dat men door een slimme opzet het rekenwerk behoorlijk kan reduceren, namelijk tot $N \cdot {}^2\log(N)$ in plaats van N^2 vermenigvuldigingen. Deze manier om de discrete Fourier transformatie uit te rekenen noemt men *snelle Fourier transformatie*, afgekort met *FFT* voor *fast Fourier transformation*. Bij $N = 1024$ scheelt de FFT bijvoorbeeld een factor van $\frac{N}{{}^2\log(N)} = \frac{1024}{10} \approx 100$ in het rekenwerk.

We gaan vanaf nu ervan uit dat N even is (meestal is zelfs $N = 2^m$ een macht van 2) en definiëren

$$z := e^{-i\Delta\omega\Delta t} = e^{-i\frac{2\pi}{N}}.$$

Merk op: Wegens $z^N = e^{-2\pi i} = 1$ geldt dat $z^{jk} = z^{j(k+N)}$.

Met deze notatie ziet de discrete Fourier transformatie er zo uit:

$$F_j = \sum_{k=0}^{N-1} f_k z^{jk}$$

en dit laat zich ook met een matrix schrijven, namelijk als

$$\begin{pmatrix} F_0 \\ F_1 \\ \vdots \\ F_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & z & z^2 & \dots & z^{N-1} \\ \vdots & & & & \\ 1 & z^{N-1} & z^{2(N-1)} & \dots & z^{(N-1)^2} \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{pmatrix}$$

Een coëfficiënt F_{2j} met even index schrijven we nu iets anders, want voor $k' \geq \frac{N}{2}$ kunnen we k' schrijven als $k' = k + \frac{N}{2}$, dan is

$$f_{k'} z^{2jk'} = f_{k+\frac{N}{2}} z^{2j(k+\frac{N}{2})} = f_{k+\frac{N}{2}} z^{2jk} z^{jN} = f_{k+\frac{N}{2}} z^{2jk}.$$

Hieruit volgt voor coëfficiënten met even index:

$$F_{2j} = \sum_{k=0}^{N-1} f_k z^{2jk} = \sum_{k=0}^{\frac{N}{2}-1} (f_k + f_{k+\frac{N}{2}}) z^{2jk}.$$

Op een soortgelijke manier behandelen we ook de coëfficiënten F_{2j+1} met oneven index. Voor $k' \geq \frac{N}{2}$ schrijven we weer $k' = k + \frac{N}{2}$, dan is

$$f_{k'} z^{(2j+1)k'} = f_{k+\frac{N}{2}} z^{(2j+1)(k+\frac{N}{2})} = f_{k+\frac{N}{2}} z^{2jk} z^k z^{jN} z^{\frac{N}{2}} = -f_{k+\frac{N}{2}} z^k z^{2jk},$$

want $z^{\frac{N}{2}} = e^{-\pi i} = -1$. Hieruit volgt voor coëfficiënten met oneven index:

$$F_{2j+1} = \sum_{k=0}^{N-1} f_k z^{(2j+1)k} = \sum_{k=0}^{\frac{N}{2}-1} (f_k - f_{k+\frac{N}{2}}) z^k z^{2jk}$$

Met behulp van deze formules kunnen we de coëfficiënten met even en oneven indices apart door matrices beschrijven, het aardige daarbij is dat de matrices nu alleen maar nog half zo groot zijn, dus $\frac{N}{2} \times \frac{N}{2}$ in plaats van $N \times N$:

$$\begin{pmatrix} F_0 \\ F_2 \\ \vdots \\ F_{N-2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & z^2 & z^4 & \dots & (z^2)^{\frac{N}{2}-1} \\ \vdots & & & & \\ 1 & z^{N-2} & z^{2(N-2)} & \dots & (z^2)^{(\frac{N}{2}-1)^2} \end{pmatrix} \cdot \begin{pmatrix} f_0 + f_{\frac{N}{2}} \\ f_1 + f_{1+\frac{N}{2}} \\ \vdots \\ f_{\frac{N}{2}-1} + f_{N-1} \end{pmatrix}$$

en

$$\begin{pmatrix} F_1 \\ F_3 \\ \vdots \\ F_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & z^2 & z^4 & \dots & (z^2)^{\frac{N}{2}-1} \\ \vdots & & & & \\ 1 & z^{N-2} & z^{2(N-2)} & \dots & (z^2)^{(\frac{N}{2}-1)^2} \end{pmatrix} \cdot \begin{pmatrix} f_0 - f_{\frac{N}{2}} \\ z(f_1 - f_{1+\frac{N}{2}}) \\ \vdots \\ z^{\frac{N}{2}-1}(f_{\frac{N}{2}-1} + f_{N-1}) \end{pmatrix}$$

Om deze reductie herhaald toe te kunnen passen is het wenselijk dat N een macht van 2 is, dus van de vorm $N = 2^m$. Dit bereikt men meestal door N gewoon zo te kiezen, maar soms ook door aanvullen van de waarden met nullen. Er zijn ook versies van de FFT ontwikkeld, waarbij dit niet nodig is.

Om de methode beter toe te lichten gaan we eens een voorbeeld met $N = 8$ expliciet uitwerken. De berekening die we eigenlijk moeten uitvoeren is

$$\begin{pmatrix} F_0 \\ F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \\ F_7 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & z & z^2 & z^3 & z^4 & z^5 & z^6 & z^7 \\ 1 & z^2 & z^4 & z^6 & z^8 & z^{10} & z^{12} & z^{14} \\ 1 & z^3 & z^6 & z^9 & z^{12} & z^{15} & z^{18} & z^{21} \\ 1 & z^4 & z^8 & z^{12} & z^{16} & z^{20} & z^{24} & z^{28} \\ 1 & z^5 & z^{10} & z^{15} & z^{20} & z^{25} & z^{30} & z^{35} \\ 1 & z^6 & z^{12} & z^{18} & z^{24} & z^{30} & z^{36} & z^{42} \\ 1 & z^7 & z^{14} & z^{21} & z^{28} & z^{35} & z^{42} & z^{49} \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \end{pmatrix}$$

In de eerste reductiestap gaat dit over in de twee vergelijkingen

$$\begin{pmatrix} F_0 \\ F_2 \\ F_4 \\ F_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & z^2 & z^4 & z^6 \\ 1 & z^4 & z^8 & z^{12} \\ 1 & z^6 & z^{12} & z^{18} \end{pmatrix} \cdot \begin{pmatrix} f_0 + f_4 \\ f_1 + f_5 \\ f_2 + f_6 \\ f_3 + f_7 \end{pmatrix}$$

$$\begin{pmatrix} F_1 \\ F_3 \\ F_5 \\ F_7 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & z^2 & z^4 & z^6 \\ 1 & z^4 & z^8 & z^{12} \\ 1 & z^6 & z^{12} & z^{18} \end{pmatrix} \cdot \begin{pmatrix} f_0 - f_4 \\ z(f_1 - f_5) \\ z^2(f_2 - f_6) \\ z^3(f_3 - f_7) \end{pmatrix}$$

en in de tweede stap krijgen we de vier vergelijkingen

$$\begin{pmatrix} F_0 \\ F_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & z^4 \end{pmatrix} \cdot \begin{pmatrix} (f_0 + f_4) + (f_2 + f_6) \\ (f_1 + f_5) + (f_3 + f_7) \end{pmatrix}$$

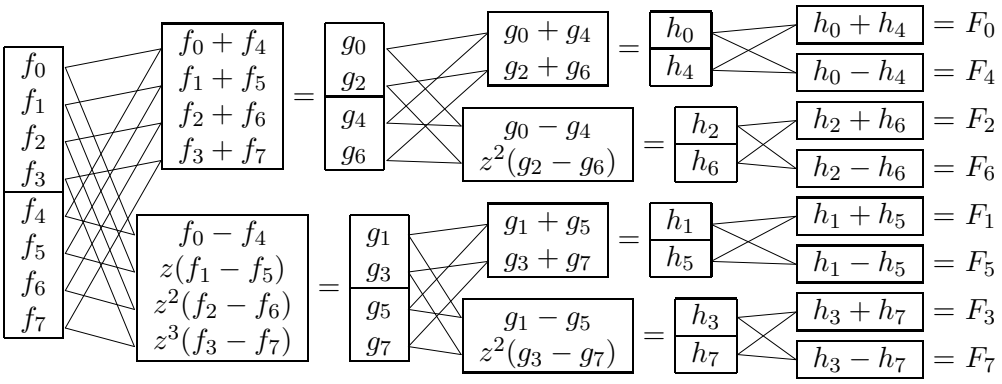
$$\begin{pmatrix} F_2 \\ F_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & z^4 \end{pmatrix} \cdot \begin{pmatrix} (f_0 + f_4) - (f_2 + f_6) \\ z^2((f_1 + f_5) - (f_3 + f_7)) \end{pmatrix}$$

$$\begin{pmatrix} F_1 \\ F_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & z^4 \end{pmatrix} \cdot \begin{pmatrix} (f_0 - f_4) + z^2(f_2 - f_6) \\ z(f_1 - f_5) + z^3(f_3 - f_7) \end{pmatrix}$$

$$\begin{pmatrix} F_3 \\ F_7 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & z^4 \end{pmatrix} \cdot \begin{pmatrix} (f_0 - f_4) - z^2(f_2 - f_6) \\ z^2(z((f_1 - f_5) - (f_3 - f_7))) \end{pmatrix}$$

De grap is nu dat we dit iets anders uitrekenen, namelijk beginnend met de vector $(f_0, \ldots, f_7)$ de stappen op één vector toepassen, die uiteindelijk (tot op volgorde na) de vector $(F_0, \ldots, F_7)$ wordt. We vermijden zo, dat we uitdrukkingen zo als $(f_0 + f_4)$ of $z(f_1 - f_5)$ die herhaald voorkomen meerdere keren berekenen. Dit gebeurt volgens het schema in Figuur II.27. In de laatste stap maken we hierbij gebruik ervan dat $z^{\frac{N}{2}} = z^4 = -1$.

Merk op dat we de indices van de g_i en h_i zo hebben aangepast dat deze met de goede index van F_i corresponderen. Dit is in de praktijk niet nodig, want de



Figuur II.27: Schema van FFT voor $N = 8$ punten

volgorde van de F_i kunnen we makkelijk achterhalen, als we de indices binair schrijven. De goede volgorde van de F_i krijgen we, door de binaire schrijfwijzen voor de getallen $0, \ldots, N-1$ te spiegelen (d.w.z. van rechts naar links te lezen). Hierbij schrijven we een getal n binair als een keten $n = a_m a_{m-1} \ldots a_1 a_0$ met $a_i \in \{0, 1\}$ zo dat $n = a_m \cdot 2^m + a_{m-1} \cdot 2^{m-1} + \ldots + a_1 \cdot 2 + a_0$. Bijvoorbeeld schrijven we het getal 42 binair als 101010. Om duidelijk te maken dat een getal een binaire schrijfwijze is, voegt men soms een index 2 aan het getal toe, bijvoorbeeld $42 = 101010_2$.

In het voorbeeld met $N = 8$ krijgen we op deze manier de volgorde van de F_i als volgt:

$$\text{indices} \quad \begin{array}{l} 0 = 000_2 \\ 1 = 001_2 \\ 2 = 010_2 \\ 3 = 011_2 \\ 4 = 100_2 \\ 5 = 101_2 \\ 6 = 110_2 \\ 7 = 111_2 \end{array} \quad \text{van } f_i \text{ worden gespiegeld} \quad \begin{array}{l} 000_2 = 0 \\ 100_2 = 4 \\ 010_2 = 2 \\ 110_2 = 6 \\ 001_2 = 1 \\ 101_2 = 5 \\ 011_2 = 3 \\ 111_2 = 7 \end{array} \quad \text{de indices van } F_i$$

In de praktijk zijn er verschillende manieren om de FFT te implementeren, de meest bekende zijn de *Cooley-Tukey* en de *Sande-Tukey* methode.

In principe heeft pas de ontwikkeling van de FFT de doorbraak van de Fourier transformatie in de signaalverwerking veroorzaakt, want eerder was het berekenen van de Fourier getransformeerde voor interessante toepassingen gewoon ondoenlijk. Inmiddels heeft men overigens achterhaald dat het idee van de FFT al door Gauss werd toegepast.

10.6 Shannon's aftast-theorema

Een belangrijke vraag die we ons bij het sampling van een signaal moeten stellen, is, hoe veel informatie we door het sampling eigenlijk verliezen. In de signaalverwerking komen we dit probleem bijvoorbeeld bij de digitalisering van

analoge signalen tegen, zo als bij een CD-opname van een concert. De vraag is of we het oorspronkelijke signaal uit de digitale informatie, die in de discrete samples zit, kunnen reconstrueren.

In het algemeen is dit natuurlijk onmogelijk, als we maar een keer per seconde een sample hebben kunnen we bijna niets erover zeggen, wat tussendoor gebeurd is. Maar als we met een hogere frequentie sampeln, kunnen we wel verwachten dat we meer informatie terug kunnen vinden.

Intuïtief kunnen we verwachten dat het resultaat ermee te maken heeft, welke frequenties in het signaal voorkomen. Om hoge frequenties te kunnen reconstrueren, moeten we zeker ook met een hogere frequentie sampeln. Hoe het hiermee precies zit, zegt het *aftast-theorema* dat door Claude E. Shannon 1949 werd bewezen. Hierbij gaat het om signalen met een begrensde *bandbreedte*, d.w.z. om signalen waarin alleen frequenties uit een begrensd interval een rol spelen. In de taal van de Fourier transformatie kunnen we dit zo uitdrukken, dat de Fourier getransformeerde van het signaal alleen maar op een begrensd interval ongelijk aan 0 is.

Het concept van begrensde bandbreedte is iets redelijk gewoons, het meest belangrijke voorbeeld is het menselijke oor, dat bij kinderen frequenties tot hooguit 25000 Hz kan verwerken en bij oudere mensen al bij frequenties van 12000 - 15000 Hz ophoudt. Om deze reden zijn ook de meeste HiFi-toestellen zo gebouwd, dat ze alleen maar frequenties tussen 20 Hz en 25 kHz verwerken. Nog beperkter is de bandbreedte van de telefoon, hier worden alleen maar frequenties tussen 300 Hz en 3500 Hz over gebracht.

Het *aftast-theorema* van Shannon beweert nu dat we een signaal minstens met de dubbele frequentie moeten sampeln die in het signaal een rol speelt:

Shannon's aftast-theorema (sampling/scanning theorem):
Zij $f(t)$ een signaal met Fourier getransformeerde $F(u) = \mathcal{F}[f(t)]$ waarvoor geldt dat $F(u) = 0$ voor alle u met $|u| > u_m = 2\pi f_m$.

Als voor het aftast-interval Δt en de aftast-frequentie $\omega = \frac{2\pi}{\Delta t}$ geldt dat

$$\omega \geq 2u_m \text{ en dus } \Delta t \leq \frac{1}{2f_m},$$

dan laat zich het signaal $f(t)$ volledig uit de discrete samples $f_k := f(k \cdot \Delta t)$ met $k = 0, \pm 1, \pm 2, \ldots$ reconstrueren.

De minimale frequentie $2f_m$ heet ook de *Nyquist-frequentie*, het interval $\Delta t = \frac{1}{2f_m}$ het *Nyquist-interval*. Let op dat er in de literatuur verschillende definities voor de Nyquist-frequentie en het Nyquist-interval gegeven worden, die soms om de factor 2 van elkaar afwijken.

Voor een CD-opname met een bandbreedte van 22 kHz hebben we dus een sampling frequentie van minstens 44 kHz nodig, terwijl voor een telefoon gesprek een sampling frequentie van 8 kHz voldoende is.

Reconstructie van een signaal van uit gesampelde waarden

De interessante vraag is nu, hoe we een continu signaal $f(t)$ uit de discrete waarden f_k kunnen reconstrueren. De oplossing hiervoor is heel simpel, we maken gewoon een interpolatie van continue functies van een bepaalde vorm, te weten functies van de vorm $\frac{\sin(x)}{x}$, die we als Fourier getransformeerde van een rechthoek impuls al eerder zijn tegengekomen.

Er laat zich aantonen dat men onder de voorwaarde van Shannon's aftasttheorema het oorspronkelijke signaal $f(t)$ terug vindt met behulp van de formule

$$f(t) = \sum_{k=-\infty}^{\infty} f(k\Delta t)\frac{\sin(\frac{\omega}{2}(t-k\Delta t))}{\frac{\omega}{2}(t-k\Delta t)} = \sum_{k=-\infty}^{\infty} f_k\,\frac{\sin(\frac{t}{\Delta t}\pi - k\pi)}{\frac{t}{\Delta t}\pi - k\pi}.$$

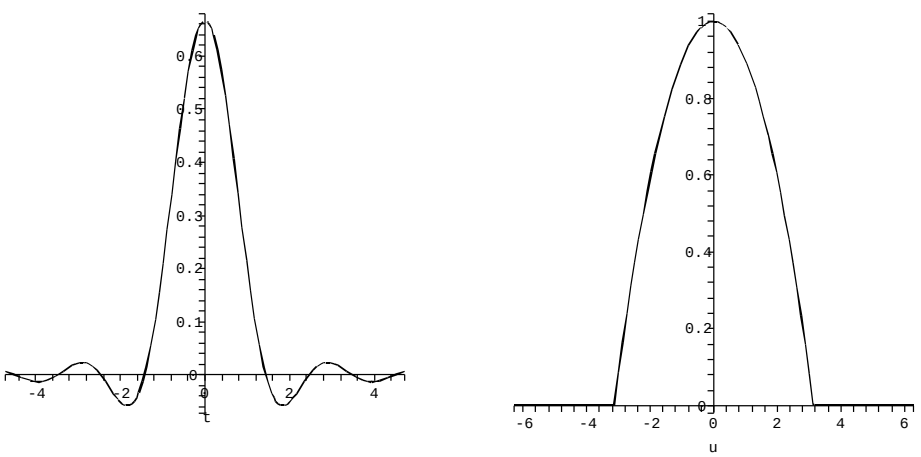
We gaan het proces van sampling en interpolatie aan een voorbeeld bekijken. Als functie nemen we

$$f(t) = \frac{2\sin(\pi t) - 2\pi t\cos(\pi t)}{\pi^3 t^3}$$

dan heeft $f(t)$ de Fourier getransformeerde

$$F(u) = \mathcal{F}[f(t)] = \begin{cases} 1-(\frac{u}{\pi})^2 & \text{als} \quad |u| \le \pi \\ 0 & \text{als} \quad |u| > \pi. \end{cases}$$

De functie $f(t)$ is natuurlijk verkregen als inverse Fourier transformatie van een functie die alleen maar op een eindig interval ongelijk aan 0 is. In Figuur II.28 zijn de grafieken van $f(t)$ en $\mathcal{F}[f(t)]$ te zien.



Figuur II.28: Functie met Fourier getransformeerde van begrensde bandbreedte.

Als we $f(t)$ met tijdelijke afstanden van Δt sampeln, krijgen we de discrete waarden $f_k = f(k \cdot \Delta t) = f(t_k)$. De discrete functiewaarden f_k beschrijven we nu door de *sampling functie* van $f(t)$, namelijk

$$f_s(t) := \sum_{k=-\infty}^{\infty} f_k\,\delta(t-t_k)$$

die op de tijdstippen t_k Dirac δ-functies met intensiteit $f_k = f(t_k)$ heeft.

We berekenen nu eerst de Fourier getransformeerde van de speciale functie $\text{Ш}(t) := \sum_{k=-\infty}^{\infty} \delta(t - t_k)$: De functie $\text{Ш}(t)$ is periodiek met periode Δt en heeft dus een Fourier reeks van de vorm

$$\text{Ш}(t) = \sum_{k=-\infty}^{\infty} c_k e^{ik\omega t}, \text{ waarbij } \omega = \frac{2\pi}{\Delta t} \text{ is.}$$

Voor de coëfficiënten c_k geldt

$$c_k = \frac{1}{\Delta t} \int_{-\frac{\Delta t}{2}}^{\frac{\Delta t}{2}} \text{Ш}(t) e^{-ik\omega t}\, dt = \frac{1}{\Delta t} \int_{-\frac{\Delta t}{2}}^{\frac{\Delta t}{2}} \delta(t) e^{-ik\omega t}\, dt = \frac{1}{\Delta t}.$$

Hieruit volgt dat

$$\text{Ш}(t) = \frac{1}{\Delta t} \sum_{k=-\infty}^{\infty} e^{ik\omega t}.$$

Uit de vorige les weten we dat de functie $e^{ik\omega t}$ de Fourier getransformeerde $2\pi\delta(u - k\omega)$ heeft, bij elkaar genomen volgt hieruit (met nog steeds $\omega = \frac{2\pi}{\Delta t}$):

$$\mathcal{F}[\text{Ш}(t)] = \frac{1}{\Delta t} \sum_{k=-\infty}^{\infty} 2\pi\delta(u - k\omega) = \omega \cdot \sum_{k=-\infty}^{\infty} \delta(u - k\omega).$$

Omdat $f_s(t) = \text{Ш}(t) \cdot f(t)$ is, kunnen we de Fourier getransformeerde $F_s(u)$ van de gesampelde functie $f_s(t)$ nu met behulp van een convolutieproduct berekenen, waarbij we met $F(u) := \mathcal{F}[f(t)]$ de Fourier getransformeerde van $f(t)$ noteren:

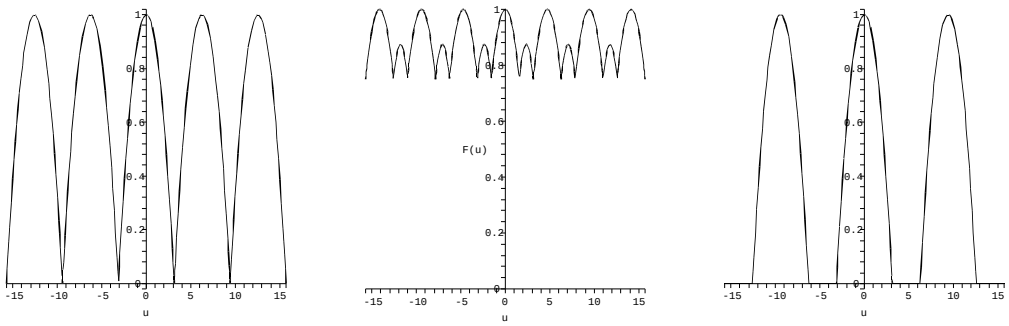
$$\begin{aligned} F_s(u) := \mathcal{F}[f_s(t)] &= \frac{1}{2\pi}\mathcal{F}[\text{Ш}(t)] * \mathcal{F}[f(t)] = \left(\frac{\omega}{2\pi} \cdot \sum_{k=-\infty}^{\infty} \delta(u - k\omega) \right) * F(u) \\ &= \frac{1}{\Delta t} \sum_{k=-\infty}^{\infty} \delta(u - k\omega) * F(u) = \frac{1}{\Delta t} \sum_{k=-\infty}^{\infty} F(u - k\omega) \end{aligned}$$

omdat $\delta(u - k\omega) * F(u) = F(u - k\omega)$.

Dit betekent dat zich bij $F_s(u)$ de functie $F(u)$ (geschaald met een factor $\frac{1}{\Delta t}$) periodiek met periode ω herhaald.

Als we nu $F(u)$ uit $F_s(u)$ zouden kunnen reconstrueren, dan kunnen we met behulp van de inverse Fourier transformatie ook $f(t)$ weer reconstrueren. Hier is het punt waar de Nyquist-frequentie in het spel komt: Als de periodieke herhalingen van $F(u)$ niet overlappen, dan kunnen we $F(u)$ door vermenigvuldiging met een rechthoek-filter functie terug vinden. Hiervoor hebben we nodig, dat $\omega \geq 2u_m$, want dit is de lengte van het interval waarop $F(u) \neq 0$ is, en dit is precies de voorwaarde die in Shannon's aftast-theorema gegeven wordt.

De plaatjes in Figuur II.29 laten de effecten van verschillende aftast frequenties op de Fourier getransformeerde $F_s(u)$ zien. In het linker plaatje wordt



Figuur II.29: Effect van verschillende aftast frequenties op de Fourier getransformeerde: aftasten met Nyquist-frequentie (links), undersampling (midden) en oversampling (rechts).

precies met de Nyquist-frequentie gesampeld, hier laat zich de oorspronkelijke Fourier getransformeerde terug vinden, door met een rechthoek functie te vermenigvuldigen. Hetzelfde geldt voor het rechter plaatje, waar met een frequentie hoger dan de Nyquist-frequentie gesampeld wordt, dit noemt men *oversampling*. In het middelste plaatje is daarentegen het geval van *undersampling* te zien, de verschillende kopieën van $F(u)$ overlappen en worden gedeeltelijk bij elkaar opgeteld. De oorspronkelijke Fourier getransformeerde $F(u)$ is uit deze functie niet meer terug te vinden.

In het geval dat met een voldoende hoge frequentie gesampeld is, dus met $\omega \geq 2u_m$, vermenigvuldigen we $F_s(u)$ met de rechthoek-filter functie

$$r(u) := \begin{cases} 1 & \text{als} \quad |u| \leq \frac{\omega}{2} \\ 0 & \text{als} \quad |u| > \frac{\omega}{2} \end{cases}$$

Dan is $F(u) = \Delta t \cdot F_s(u) \cdot r(u)$ en dus

$$f(t) = \mathcal{F}^{-1}[F(u)] = \mathcal{F}^{-1}[\Delta t \cdot F_s(u)] * \mathcal{F}^{-1}[r(u)].$$

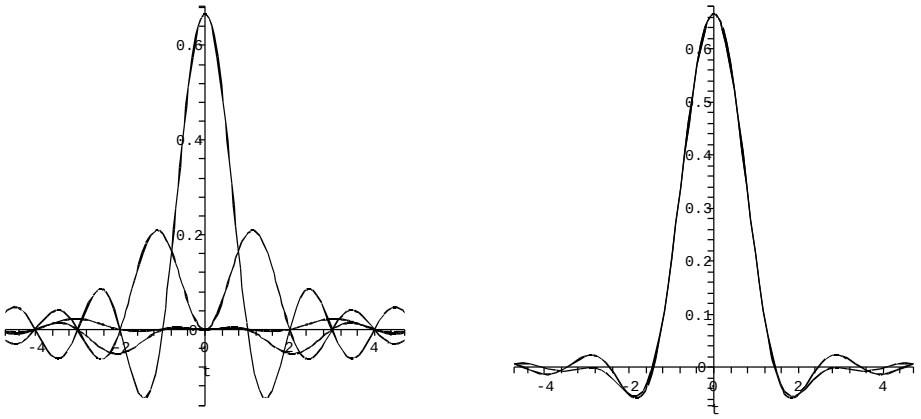
We hebben $F_s(u)$ zo gedefinieerd dat $\mathcal{F}^{-1}[F_s(u)] = \sum_{k=-\infty}^{\infty} f(t_k)\delta(t-t_k)$. Maar met de relatie $\mathcal{F}^{-1}[f(u)] = \frac{1}{2\pi}\mathcal{F}[f(-u)]$ tussen inverse Fourier transformatie en Fourier transformatie kunnen we ook de inverse Fourier transformatie van $r(u)$ makkelijk berekenen, dit hebben we in feite eerder al gedaan, het gaat om de Fourier getransformeerde van een rechthoek impuls van breedte $\omega = \frac{2\pi}{\Delta t}$:

$$\mathcal{F}^{-1}[r(u)] = \frac{2}{2\pi}\frac{\sin(\frac{\omega}{2}t)}{t} = \frac{\frac{\omega}{2}}{\pi}\frac{\sin(\frac{\omega}{2}t)}{\frac{\omega}{2}t} = \frac{1}{\Delta t}\frac{\sin(\frac{\omega}{2}t)}{\frac{\omega}{2}t}.$$

Hieruit krijgen we de interpolatie formule

$$f(t) = \sum_{k=-\infty}^{\infty} f(t_k)\left(\delta(t-t_k) * \frac{\sin(\frac{\omega}{2}t)}{\frac{\omega}{2}t}\right) = \sum_{k=-\infty}^{\infty} f_k \frac{\sin(\frac{\omega}{2}(t-t_k))}{\frac{\omega}{2}(t-t_k)}.$$

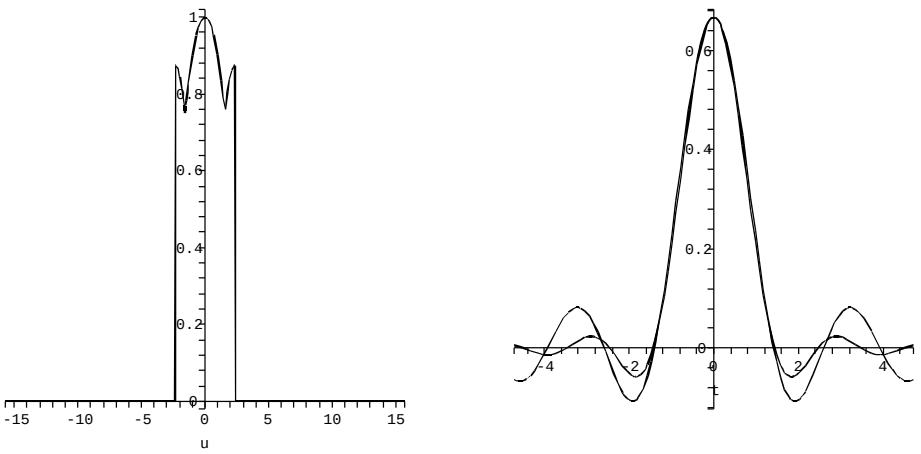
In Figuur II.30 zijn de interpolerende functies $f(t_k)\frac{\sin(\frac{\omega}{2}(t-t_k))}{\frac{\omega}{2}(t-t_k)}$ voor $k = -3, \ldots, 3$ te zien, waarbij de functies voor k en $-k$ bij elkaar opgeteld zijn



Figuur II.30: Interpolerende functies en reconstructie door som van interpolerende functies.

om symmetrische functies te krijgen. Het rechter plaatje laat de reconstructie van de originele functie $f(t)$ door de som van de interpolerende functies voor $k = -2, \ldots, 2$ zien. Een vergelijk met het linkerplaatje in Figuur II.28 maakt duidelijk dat dit al een erg goede benadering oplevert.

Wat gebeurt er nu als we een signaal met een te lage aftast-frequentie sampeln, dus bij undersampling? De bijdragen van de frequenties boven $\frac{\omega}{2} = \frac{\pi}{\Delta t}$ worden afgebeeld op lagere frequenties en dit resulteert in een verandering van de functie $F_s(u)$ en dus ook van het gereconstrueerde signaal. Dit effect noemt men *aliasing*. Om zo'n effect te vermijden, wordt een signaal voor het aftasten meestal door een low pass filter gestuurd die de te hoge frequenties wegsnijdt.



Figuur II.31: Aliasing in het frequentie domein en reconstructie bij undersampling.

In Figuur II.31 is de functie $f(t)$ uit het voorbeeld met de te lage frequentie $\frac{3}{2}u_m$ in plaats van de Nyquist frequentie $2u_m$ gesampeld. Dit heeft tot gevolg dat de functie $F_s(u)$ zo als in het middelste plaatje van Figuur II.29 verandert, en het vermenigvuldigen met de rechthoek-filter lijdt tot een functie $\Delta t F_s(u) r(u)$ met te hoge waarden voor de lage frequenties, omdat hogere frequenties op deze lage frequenties afgebeeld worden. Als effect hiervan levert de reconstructie van de oorspronkelijke functie middels inverse Fourier transformatie een functie die niet snel genoeg afneemt, wat we duidelijk aan het te grote tweede maximum van de gereconstrueerde functie in het rechter plaatje van Figuur II.31 kunnen zien.

Belangrijke begrippen in deze les

- sampling, discretisering
- discrete Fourier transformatie
- trigonometrische interpolatie
- FFT: snelle Fourier transformatie
- Shannon's aftast-theorema
- Nyquist frequentie, undersampling, oversampling

Opgaven

78. Bereken de coëfficiënten F_j van de discrete Fourier getransformeerde voor de volgende gesampelde waarden van de functie $f(t)$. In elk geval is $\Delta t = 1$.

(i)

k	0	1	2	3
$f(t_k)$	1	2	3	4

(ii)

k	0	1	2	3	4	5	6	7
$f(t_k)$	1	2	3	4	0	0	0	0

79. Een sampling van de functie $f(t)$ met $N = 8$ en $\Delta t = 1$ geeft de volgende waarden:

$$f_k = f(t_k) = \begin{cases} \frac{1}{2} & k = 0, 4 \\ 1 & k = 1, 2, 3 \\ 0 & k = 5, 6, 7 \end{cases}$$

(i) Bepaal de coëfficiënten F_j van de discrete Fourier getransformeerde van $f(t)$.

(ii) Stel dat $f(t)$ periodiek met periode 8 is en dat de functie $g(t)$ gegeven is door $g(t) := f(t-3)$. Wat zijn de sampling waarden $g_k = g(t_k)$ van $g(t)$ voor $k = 0, 1, \ldots, 7$?

Bepaal de coëfficiënten G_j van de discrete Fourier getransformeerde van $g(t)$.

80. Zij $\{f_0, \ldots, f_{N-1}\}$ een gesampelde functie en zij $\{F_0, \ldots, F_{N-1}\}$ de discrete Fourier getransformeerde hiervan.

(i) Laat zien dat de (in $t = 0$) gespiegelde functie $\{f_{N-0}, \ldots, f_{N-(N-1)}\}$ de (in $\omega = 0$) gespiegelde Fourier getransformeerde $\{F_{N-0}, \ldots, F_{N-(N-1)}\}$ heeft.

(ii) Laat zien dat de complex geconjugeerde functie $\{\overline{f_0}, \ldots, \overline{f_{N-1}}\}$ de complex geconjugeerde en gespiegelde Fourier getransformeerde $\{\overline{F_{N-0}}, \ldots, \overline{F_{N-(N-1)}}\}$ heeft.

81. Zij $\{f_0, \ldots, f_{N-1}\}$ een gesampelde functie met *reële* waarden en zij $\{F_0, \ldots, F_{N-1}\}$ de discrete Fourier getransformeerde hiervan. Laat zien dat voor de discrete Fourier transformatie de volgende versie van de *Parseval identiteit* geldt:

$$\sum_{k=0}^{N-1} f_k^2 = \frac{1}{N} \sum_{j=0}^{N-1} |F_j|^2.$$

82. Zij $\{f_0, \ldots, f_{N-1}\}$ een gesampelde functie en zij $\{F_0, \ldots, F_{N-1}\}$ de discrete Fourier getransformeerde hiervan. We verschuiven F_j in het frequentiedomein om l posities, dit geeft de functie $G_j := F_{j-l}$. Laat zien dat $\{G_0, \ldots, G_{N-1}\}$ de discrete Fourier getransformeerde van de functie $\{g_0, \ldots, g_{N-1}\}$ is met

$$g_k = f_k \, e^{ikl\Delta\omega\Delta t} = f_k \, e^{i\frac{2\pi}{N}kl}.$$

83. Laten de gesampelde functies $\{f_0, \ldots, f_{N-1}\}$ en $\{g_0, \ldots, g_{N-1}\}$ de discrete Fourier getransformeerden $\{F_0, \ldots, F_{N-1}\}$ en $\{G_0, \ldots, G_{N-1}\}$ hebben. Laat zien dat het convolutieproduct

$$H_j := F_j * G_j = \sum_{l=0}^{N-1} F_l \, G_{j-l}$$

de discrete Fourier getransformeerde van de functie $h_k := N \cdot f_k \, g_k$ is.

Deel III

Probabilistische Modellen

Les 11 Onzekerheid, entropie en informatie

Als we erover nadenken hoe we conclusies trekken, komen we er snel achter dat dit meestal met het verkrijgen en verwerken van informatie te maken heeft. Vaak stellen we hiervoor vragen of maken een meting, om de onzekerheid die we over benodigde gegevens hebben te overkomen of tenminste te verkleinen.

Als we nu op het gebied van de kunstmatige intelligentie een systeem willen bouwen, dat op grond van zekere informatie beslissingen neemt, moeten we voor de begrippen *informatie* of *onzekerheid* definities vinden, die het mogelijk maken om ook kwantitatieve uitspraken hierover te kunnen doen. Een cruciaal begrip in dit kader is de *entropie* van een kansverdeling, die in principe aangeeft hoeveel bits we minstens nodig hebben, om de uitkomsten van een kansexperiment te beschrijven.

11.1 Onzekerheid

Als we een experiment of gebeurtenis door een kansverdeling beschrijven, drukken we hiermee uit dat we niet zeker over de uitkomst zijn. Maar we hebben ook een intuïtieve idee dat de onzekerheid soms groter is dan in andere gevallen. Bijvoorbeeld zijn we onzekerder over de uitkomst bij het werpen van een dobbelsteen dan bij het werpen van een munt, omdat er in het ene geval 6 mogelijke uitkomsten zijn, maar in het andere geval slechts 2. Ook bij een sportwedstrijd hangt onze onzekerheid ervan af hoe we de kansen voor de uitkomst inschatten: Als alleen maar de KI-studenten onderling een zwemwedstrijd uitvechten is de onzekerheid waarschijnlijk groter dan als Pieter van den Hoogenband ook meedoet.

Voorbeeld: Stel bij een paardenrace doen 8 paarden mee die niet even sterk zijn, maar waarvoor de kansen om te winnen gegeven zijn door

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{4}, \quad p_3 = \frac{1}{8}, \quad p_4 = \frac{1}{16}, \quad p_5 = p_6 = p_7 = p_8 = \frac{1}{64}.$$

Als we gewoon het nummer van het winnende paard door willen geven, hebben we hiervoor 3 bits nodig (want $2^3 = 8$). Maar omdat de kansen niet uniform verdeeld zijn, kunnen dit aantal reduceren, door de paarden met een hogere kans op een kortere manier te coderen. Hierbij moeten we er wel op letten, dat de beginstukken van de langere coderingen zelfs geen coderingen zijn. Een mogelijke codering voor de nummers 1 t/m 8 van de paarden in het voorbeeld is (aangegeven met strings van bits):

1: `0`, 2: `10`, 3: `110`, 4: `1110`, 5: `111100`, 6: `111101`, 7: `111110`, 8: `111111`.

Als we voor deze codering het gemiddeld benodigde aantal bits berekenen (dus de verwachtingswaarde van het aantal bits), krijgen we $\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{16} \cdot 4 + 4 \cdot \frac{1}{64} \cdot 6 = 2$. We hebben het aantal benodigde bits dus van 3 op 2 kunnen reduceren, door de uitkomsten die we vaker verwachten korter te coderen.

De onzekerheid bij een kansexperiment is natuurlijk bepaald door de kansen die we aan de mogelijke uitkomsten toewijzen. We kunnen ons dus afvragen

hoe we voor een discrete kansverdeling $P = (p_1, \ldots, p_n)$ een waarde voor de onzekerheid kunnen berekenen. Het idee dat we hiervoor hebben, is een functie

$$H(P) = H(p_1, \ldots, p_n)$$

te vinden, die de onzekerheid weergeeft. Omdat we intuïtief wel een idee van de onzekerheid bij een kansverdeling hebben, moet zo'n functie zekere eigenschappen hebben. In het jaar 1948 is hiervoor door C.E. Shannon (dezelfde Shannon als bij het sampling theorema) in het kader van de communicatietheorie een voorstel gedaan aan welke eisen zo'n functie $H(P)$ zou moeten voldoen. De link tussen communicatietheorie en kansrekening bestaat erin, dat communicatie als transmissie (van bit-strings, dus van ketens van 0en en 1en) via kanalen gemodelleerd wordt, waarbij er toevallig fouten kunnen optreden. De vraag is dan, hoe veel onzekerheid in het ontvangen signaal ligt.

Eisen aan een functie voor de onzekerheid van een kansverdeling

De eisen die Shannon heeft gesteld zijn als volgt:

(1) $H(P)$ is een continue functie in de argumenten $p_1, \ldots, p_n$, want als we de kansen maar heel weinig veranderen, verandert ook de onzekerheid nauwelijks.

(2) De onzekerheid hangt alleen maar van de kansen p_i, maar niet van hun volgorde af, dus geldt $H(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)})$ voor elke permutatie π van de indices.

(3) $H(P) \geq 0$ en $H(P) = 0$ alleen maar als één van de $p_i = 1$ is (en de anderen dus 0). Dit betekent dat we altijd onzeker zijn, behalve als een uitkomst kans 1 heeft en dus zeker gaat gebeuren.

(4) $H(p_1, \ldots, p_n) = H(p_1, \ldots, p_n, 0)$, dus de onzekerheid verandert niet, als we de kansverdeling uitbreiden tot meer mogelijke gebeurtenissen, maar de nieuwe opties kans 0 hebben en dus nooit kunnen gebeuren.

(5) $H(\frac{1}{n}, \ldots, \frac{1}{n}) \leq H(\frac{1}{n+1}, \ldots, \frac{1}{n+1})$, d.w.z. de onzekerheid bij een uniforme verdeling met $n+1$ mogelijke uitkomsten is groter dan bij n mogelijke uitkomsten.

(6) $H(\frac{1}{mn}, \ldots, \frac{1}{mn}) = H(\frac{1}{m}, \ldots, \frac{1}{m}) + H(\frac{1}{n}, \ldots, \frac{1}{n})$. Als we twee onafhankelijke experimenten met uniforme verdelingen tot een gezamenlijk experiment combineren, willen we dat de onzekerheid van het gecombineerde experiment juist de som van de onzekerheden bij de enkele experimenten is.

(7) We splitsen de verzameling $\Omega = \{1, \ldots, n\}$ op in de twee deelverzamelingen $\Omega_1 = \{1, \ldots, r\}$ en $\Omega_2 = \{r+1, \ldots, n\}$. De totale kans voor de uitkomsten in Ω_1 is $q_1 = p_1 + \ldots + p_r$ en de kans voor Ω_2 is $q_2 = p_{r+1} + \ldots + p_n$. De onzekerheid of een uitkomst in Ω_1 of Ω_2 ligt, is $H(q_1, q_2)$, de onzekerheid over een uitkomst in Ω_1 is $H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1})$, omdat $(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1})$ juist de

kansverdeling op Ω_1 is. Net zo is $H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2})$ de onzekerheid over een uitkomst in Ω_2. De totale onzekerheid over de uitkomst van P is samengesteld uit de onzekerheden in welke deelverzameling een uitkomst ligt en de onzekerheden van de twee deelverzamelingen, die met hun kansen q_1 en q_2 gewogen zijn, dus moet gelden:

$$H(p_1, \ldots, p_n) = H(q_1, q_2) + q_1 H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1}) + q_2 H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2}).$$

De meeste van deze punten zijn volstrekt intuïtief, alleen de punten (6) en (7) stellen inhoudelijke eisen, namelijk hoe de onzekerheden van verschillende gebeurtenissen gecombineerde moeten worden.

Het interessante (en misschien verrassende) is nu, dat deze eisen zo sterk zijn dat er in principe alleen maar een enkele functie $H(P)$ bestaat die aan de eisen voldoet, namelijk de functie:

$$H(P) = H(p_1, \ldots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log(p_i)$$

met $\lambda > 0$, waarbij de som alleen maar over de p_i met $p_i \neq 0$ loopt.

We zullen dit hier niet bewijzen, maar wel toelichten dat de functie $H(P)$ aan de eisen (1)-(7) voldoet. Hierbij zijn de punten (1)-(4) rechtstreeks duidelijk, de andere punten gaan we even na. Omdat de constante λ geen enkel verschil in de argumenten maakt, werken we voor het gemak met $\lambda = 1$.

(5) Voor een uniforme verdeling U_n op n punten geldt

$$H(U_n) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \sum_{i=1}^{n} \frac{1}{n} \log(n) = \log(n)$$

en omdat $\log(x)$ een strikt stijgende functie is, is $H(U_n) = \log(n) < \log(n+1) = H(U_{n+1})$.

(6) Dit volgt ook uit het feit dat $H(U_n) = \log(n)$, omdat $\log(mn) = \log(m) + \log(n)$.

(7) Uit $q_1 = \sum_{i=1}^{r} p_i$ en $q_2 = \sum_{i=r+1}^{n} p_i$ volgt

$$\begin{aligned}
&H(q_1, q_2) + q_1 H(\frac{p_1}{q_1}, \ldots, \frac{p_r}{q_1}) + q_2 H(\frac{p_{r+1}}{q_2}, \ldots, \frac{p_n}{q_2}) \\
&= -q_1 \log(q_1) - q_2 \log(q_2) - q_1 \sum_{i=1}^{r} \frac{p_i}{q_1} \log(\frac{p_i}{q_1}) - q_2 \sum_{i=r+1}^{n} \frac{p_i}{q_2} \log(\frac{p_i}{q_2}) \\
&= -\sum_{i=1}^{r} p_i \log(q_1) - \sum_{i=r+1}^{n} p_i \log(q_2) - \sum_{i=1}^{r} p_i(\log(p_i) - \log(q_1)) \\
&\quad - \sum_{i=r+1}^{n} p_i(\log(p_i) - \log(q_2)) \\
&= -\sum_{i=1}^{n} p_i \log(p_i) = H(p_1, \ldots, p_n).
\end{aligned}$$

Als we ons afvragen, bij welke kansverdeling met n mogelijke uitkomsten we de grootste onzekerheid hebben, ligt het voor de hand dat dit bij een uniforme verdeling het geval is, want in dit geval hebben we geen reden om een voorkeur aan een of andere uitkomst te geven. Als $H(P)$ een maat voor de onzekerheid is, zouden we dus verwachten dat de waarde van de functie $H(P)$ voor een uniforme verdeling maximaal is en dit laat zich inderdaad bewijzen.

Stelling: Van alle kansverdelingen P op n mogelijke uitkomsten geeft de uniforme verdeling met $p_i = \frac{1}{n}$ de maximale waarde van de functie $H(P)$.

Omdat het bewijs van deze stelling niet moeilijk is en belangrijke inzichten geeft, gaan we het even na:

In het punt $x = 1$ is $\log(x) = 0$ en $\log'(x) = 1$, dus is de lijn met vergelijking $y = x - 1$ de raaklijn aan de grafiek van de logaritme in het punt $x = 1$. Omdat $\log''(x) = -\frac{1}{x^2} < 0$, blijft de logaritme steeds onder deze raaklijn, daarom geldt

$$\log(x) \leq x - 1 \text{ met gelijkheid alleen maar voor } x = 1.$$

Voor twee kansverdelingen $P = (p_1, \ldots, p_n)$ en $Q = (q_1, \ldots, q_n)$ volgt hieruit dat

$$\sum_{i=1}^{n} p_i \log(\frac{q_i}{p_i}) \leq \sum_{i=1}^{n} p_i(\frac{q_i}{p_i} - 1) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0.$$

Wegens $\log(\frac{q_i}{p_i}) = \log(q_i) - \log(p_i)$ volgt hieruit dat

$$-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i).$$

Als we nu voor Q speciaal de uniforme verdeling U_n met $q_i = \frac{1}{n}$ kiezen, volgt hieruit aan de ene kant dat

$$H(P) \leq -\sum_{i=1}^{n} p_i \log(\frac{1}{n}) = \sum_{i=1}^{n} p_i \log(n) = \log(n).$$

Maar aan de andere kant is $H(U_n) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \log(n)$, dus is de waarde voor de uniforme verdeling inderdaad maximaal.

We hebben inmiddels twee belangrijke inzichten gewonnen, die we nog eens expliciet willen aangeven:

(I) Voor een uniforme verdeling U_n op n punten is $H(U_n) = \log(n)$.

(II) Voor twee kansverdelingen P en Q is $-\sum p_i \log(q_i) \geq H(P)$ en er geldt

$$D(P, Q) := \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}) \geq 0$$

want we hadden gezien dat $\sum_{i=1}^{n} p_i \log(p_i) \geq \sum_{i=1}^{n} p_i \log(q_i)$ en hieruit volgt $\sum_{i=1}^{n} p_i(\log(p_i) - \log(q_i)) = \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}) \geq 0$.

Omdat de ideeën voor het formaliseren van onzekerheid uit de communicatietheorie komen waar men het over bit-strings heeft, is het gebruikelijk de functie $H(P)$ niet met behulp van de natuurlijke logaritme (met basis e) maar met de logaritme met basis 2 te formuleren. Wegens ${}^2\log(x) = \frac{\log(x)}{\log(2)}$ geeft dit alleen maar een verschil van de constante factor $\log(2)$.

Definitie: De functie

$$H(P) = H(p_1, \ldots, p_n) := -\sum_{i=1}^{n} p_i \; {}^2\log(p_i)$$

heet de *entropie* van de kansverdeling P.

> Het begrip *entropie* speelt ook in de natuurkunde, vooral in de thermodynamica, een belangrijke rol. Hier geeft de entropie een maat voor de wanorde in een systeem. De tweede hoofdstelling van de thermodynamica zegt (in het grof) dat in een gesloten systeem de entropie nooit afneemt, d.w.z. dat zonder invloed van buiten de wanorde in een systeem steeds toeneemt. (Dit is natuurlijk ook een alledaagse ervaring.)

We hebben tot nu toe de entropie alleen maar voor een kansverdeling gedefinieerd. Vaak spreekt men immers ook van de entropie van een stochast X. Hiermee is de entropie van de kansverdeling van de mogelijke uitkomsten van X bedoelt. Stel een stochast X heeft de mogelijke uitkomsten $x_1, \ldots, x_n$, dan geeft $p_i := p(X = x_i)$ de kans op de i-de mogelijke uitkomst en de kansverdeling $P = (p_1, \ldots, p_n)$ beschrijft de kansen van de mogelijke uitkomsten van X. We definiëren dus de de entropie van een stochast X met mogelijke uitkomsten $x_1, \ldots, x_n$ door

$$H(X) := -\sum_{i=1}^{n} p(X = x_i) \; {}^2\log(p(X = x_i)).$$

Voorbeeld: Zij X de stochast van een Bernoulli experiment met kans p op succes (uitkomst 1) en kans $1 - p$ op mislukken (uitkomst 0). Er geldt

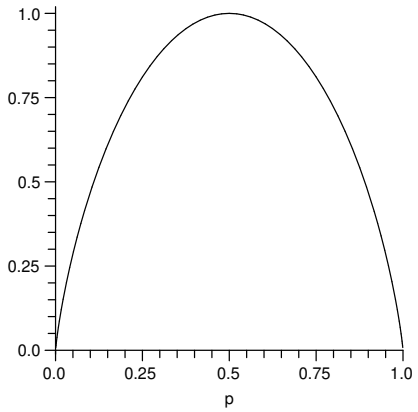
$$\begin{aligned} H(X) &= -p(X = 1) \; {}^2\log(p(X = 1)) - p(X = 0) \; {}^2\log(p(X = 0)) \\ &= -p \; {}^2\log(p) - (1 - p) \; {}^2\log(1 - p). \end{aligned}$$

In Figuur III.1 is duidelijk te zien dat de entropie maximaal wordt voor $p = 0.5$, dus voor een uniforme verdeling en dat in dit geval de entropie juist 1 *bit* is.

Relatieve entropie en de Kullback-Leibler afstand

We hebben gezien dat voor twee kansverdelingen $P = (p_1, \ldots, p_n)$ en $Q = (q_1, \ldots, q_n)$ geldt dat

$$D(P, Q) := \sum p_i ({}^2\log(p_i) - {}^2\log(q_i)) = \sum p_i \; {}^2\log(\frac{p_i}{q_i}) \geq 0$$



Figuur III.1: Entropie van een Bernoulli experiment afhankelijk van de kans p op succes.

met gelijkheid alleen maar als $p_i = q_i$ voor alle i. Men noemt $D(P,Q)$ de *relatieve entropie* of *Kullback-Leibler afstand* tussen P en Q.

De relatieve entropie $D(P,Q)$ geeft aan, hoe veel bits we gemiddeld extra nodig hebben, omdat we de codering van de gegevens op grond van de (verkeerde) kansverdeling Q in plaats van P hebben gekozen. Er geldt namelijk

$$H(P) + D(P,Q) = -\sum_{i=1}^{n} p_i \ {}^2\log(p_i) + \sum_{i=1}^{n} p_i \ {}^2\log(\frac{p_i}{q_i}) = -\sum_{i=1}^{n} p_i \ {}^2\log(q_i)$$

en dit is juist de verwachtingswaarde van het aantal benodigde bits op grond van de kansverdeling Q.

Merk op: De naam Kullback-Leibler *afstand* voor de relatieve entropie is een beetje misleidend, omdat we het niet met een afstand zo als de gewone Euclidische afstand in het vlak of in de ruimte te maken hebben.

Een echte afstandsfunctie moet namelijk de volgende drie eigenschappen hebben:

(i) $d(P,Q) \geq 0$ en $d(P,Q) = 0$ alleen maar als $P = Q$,

(ii) $d(P,Q) = d(Q,P)$ (symmetrie),

(iii) $d(P,Q) + d(Q,R) \geq d(P,R)$ (driehoeksongelijkheid).

De relatieve entropie heeft alleen maar de eerste van deze drie eigenschappen. Maar met een eenvoudig trucje kunnen we van de relatieve entropie wel een symmetrische functie maken, namelijk door

$$d_{KL}(P,Q) := \frac{1}{2}(D(P,Q) + D(Q,P)) = \frac{1}{2}\sum p_i \ {}^2\log(\frac{p_i}{q_i}) + q_i \ {}^2\log(\frac{q_i}{p_i}).$$

Ook dit heet meestal de Kullback-Leibler afstand van P en Q, soms iets duidelijker de *symmetrische Kullback-Leibler afstand*.

Ook al voldoet de Kullback-Leibler niet aan de driehoeksongelijkheid, zijn $D(P,Q)$ of $d_{KL}(P,Q)$ toch vaak handig om te kwantificeren hoe sterk verschillende kansverdelingen Q op een vaste (doel-)kansverdeling P lijken.

11.2 Entropie van continue kansverdelingen

We hebben ons tot nog toe tot discrete kansverdelingen beperkt. De overgang tot continue kansverdeling is echter geen probleem: In plaats van de kansen p_i krijgen we een dichtheidsfunctie $f(x)$ voor de kansverdeling en de som over de mogelijke uitkomsten wordt de integraal over de continue variabel x. Voor de entropie van een stochast X met dichtheidsfunctie $f(x)$ krijgt men zo:

$$H(X) := -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\ dx.$$

Om duidelijk te maken dat het om de entropie van een continue variabel gaat, spreekt men vaak ook van *differentiële entropie*. Het idee achter deze naam is, de variabel x te discretiseren door de waarden in het interval $[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}]$ aan de discrete waarde x_i toe te wijzen en de kans over dit interval als kans $p_i := \int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} f(x)\ dx$ te definiëren. Met de overgang $\Delta x \to 0$ komt men dan naar de continue versie van de entropie.

Ook de *relatieve entropie* of *Kullback-Leibler afstand* van twee stochasten X en Y met dichtheidsfuncties $f(x)$ en $g(x)$ wordt analoog met het discrete geval gedefinieerd, namelijk door

$$D(X,Y) := \int_{-\infty}^{\infty} f(x)\ {}^2\log(\frac{f(x)}{g(x)})\ dx.$$

Met hetzelfde argument als bij de discrete kansverdelingen geldt weer

$$D(X,Y) \geq 0 \quad \text{en} \quad \int_{-\infty}^{\infty} f(x)\ {}^2\log(g(x))\ dx = H(X) + D(X,Y).$$

Men ziet makkelijk in dat de entropie van een stochast X onafhankelijk van de verwachtingswaarde $\mu := E[X]$ is, want met de substitutie $x' = x + a$ volgt dat de verschoven stochast $X + a$ dezelfde entropie als X heeft. Aan de andere kant heeft de variantie $Var(X) = E[X^2 - E[X]^2]$ zeker een invloed op de entropie, want hoe sterker de resultaten vaan X verspreid zijn, hoe onzekerder zijn we over de uitkomsten van X.

Bij discrete kansverdelingen hadden we gezien, dat onder de verdelingen met n mogelijke uitkomsten de uniforme verdeling de hoogste entropie heeft. De equivalente vraag voor continue kansverdelingen is, welke verdeling met gegeven variantie σ^2 de grootste entropie heeft.

Het lijkt misschien enigszins verrassend dat we ook deze vraag kunnen beantwoorden, want we moeten een uitspraak over alle mogelijke dichtheidsfuncties maken. Maar er laat zich aantonen dat bij gegeven variantie de normale

verdeling de maximale entropie heeft, dus dat we bij de normale verdeling de grootste onzekerheid over de mogelijke uitkomsten hebben.

Stelling: Onder alle continue kansverdelingen met variantie σ^2 heeft de normale verdeling

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

de maximale entropie.

> Het idee onder zekere randvoorwaarden de kansverdeling met maximale entropie te bepalen, geeft aanleiding tot een alternatieve manier om parameters van een probabilistisch model te schatten. In Wiskunde 1 hadden we hiervoor al de *maximum likelihood* methode leren kennen, waarbij de parameters zo bepaald worden dat de kans op de waargenomen resultaten maximaal is. Bij de toegang middels maximale entropie worden de parameters zo gekozen, dat de entropie maximaal wordt, dus het wordt het meest algemene model verondersteld dat de waarnemingen verklaart.
>
> Vaak is een algemeen model ook eenvoudiger dan een speciaal model en heeft het voordeel enigszins robuust tegen uitschieters in het training materiaal te zijn. Het principe om onder gegeven randvoorwaarden het eenvoudigste model te kiezen staat ook bekend onder de naam *Ockham's razor* (na de filosoof William van Ockham (1285-1349): 'The simplest explanation is the best.').

Voor het bewijs van de stelling dat de normale verdeling de maximale entropie heeft, is een techniek namens *variatierekening* nodig. Hierbij gaat het om het vinden van extrema van functies, die niet van een of meerdere variabelen afhangen maar van een continue hoeveelheid variabelen, anders gezegd om maxima en minima van functies, die zelfs ook weer van functies afhangen. We zullen in deze cursus geen variatierekening behandelen, maar schetsen wel even het idee.

In ons geval willen we een maximum van de functie

$$H(f) = -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\ dx$$

vinden, die van de dichtheidsfunctie $f = f(x)$ afhangt. Hierbij moet $f(x)$ aan zekere randvoorwaarden voldoen, namelijk dat het een dichtheidsfunctie is, dat de variantie σ^2 is, en we mogen nog veronderstellen dat de verwachtingswaarde $\mu = 0$ is. We moeten dus een maximum van $H(f)$ vinden onder de randvoorwaarden:

(i) $f(x) \geq 0$;

(ii) $\int_{-\infty}^{\infty} f(x)\ dx = 1$;

(iii) $\int_{-\infty}^{\infty} x f(x)\ dx = 0$;

(iv) $\int_{-\infty}^{\infty} x^2 f(x)\, dx = \sigma^2$.

Dit is natuurlijk typisch een situatie voor Lagrange multiplicatoren, we definiëren daarom de Lagrange functie

$$\begin{aligned} L(f) = &- \int_{-\infty}^{\infty} f(x) \log(f(x))\, dx \\ &+ \lambda_0 (\int_{-\infty}^{\infty} f(x)\, dx - 1) + \lambda_1 (\int_{-\infty}^{\infty} x f(x)\, dx) + \lambda_2 (\int_{-\infty}^{\infty} x^2 f(x)\, dx - \sigma^2). \end{aligned}$$

Hierbij vergeten we even de randvoorwaarde $f(x) \geq 0$, die zal uiteindelijk van zelfs goed komen. We werken met de natuurlijke logaritme $\log(x)$ in plaats van de logaritme met basis 2, omdat dit voor het bepalen van de afgeleiden handiger is.

Om de kritieke punten van de Lagrange functie $L(f)$ te vinden, moeten we nu de partiële afgeleiden naar de variabelen bepalen, dus naar de functiewaarden $f(x)$ van de dichtheidsfunctie. Merk op dat x in dit geval een constante en geen variabel is, de variabelen zijn juist de functiewaarden op gegeven punten x. We moeten nu $L(f)$ voor een vaste x naar $f(x)$ afleiden en dit gelijk aan 0 zetten. Omdat we hierbij alleen maar naar een enkele waarde van x kijken, mogen we de integralen in $L(f)$ meteen weglaten. We krijgen

$$\begin{aligned} \frac{\partial L}{\partial f(x)} &= -\log(f(x)) - f(x) \cdot \frac{1}{f(x)} + \lambda_0 \cdot 1 + \lambda_1 \cdot x + \lambda_2 \cdot x^2 \\ &= -\log(f(x)) - 1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2. \end{aligned}$$

Uit $\frac{\partial L}{\partial f(x)} = 0$ volgt nu $\log(f(x)) = -1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2$ en dus

$$f(x) = e^{-1+\lambda_0+\lambda_1 x+\lambda_2 x^2}.$$

Maar dit betekent dat $f(x)$ juist een normale verdeling is en volgens de randvoorwaarden moeten de constanten λ_0, λ_1, λ_2 zo gekozen worden dat de verwachtingswaarde 0 en de variantie σ^2 wordt, en dit is juist voor

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

het geval.

We hebben tot nu toe alleen maar aangetoond dat de normale verdeling een kritieke waarde voor de Lagrange functie is. Maar als we nu veronderstellen dat $g(x)$ de dichtheidsfunctie van een stochast Y met verwachtingswaarde 0 en variantie σ^2 is, kunnen we aantonen dat $H(Y) \leq H(X)$ is, dus dat de entropie

voor de normale verdeling inderdaad een maximum aanneemt:

$$\begin{aligned}
H(Y) &= -\int_{-\infty}^{\infty} g(x)\log(g(x))\,dx = -\int_{-\infty}^{\infty} g(x)\log(\frac{g(x)}{f(x)}\cdot f(x))\,dx \\
&= -\int_{-\infty}^{\infty} g(x)\log(\frac{g(x)}{f(x)})\,dx - \int_{-\infty}^{\infty} g(x)\log(f(x))\,dx \\
&= -D(Y,X) - \int_{-\infty}^{\infty} g(x)\log(f(x))\,dx \\
&\leq_{(*)} -\int_{-\infty}^{\infty} g(x)(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2)\,dx \\
&=_{(**)} -\int_{-\infty}^{\infty} f(x)(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2)\,dx \\
&= -\int_{-\infty}^{\infty} f(x)\log(f(x))\,dx = H(X).
\end{aligned}$$

Bij $(*)$ hebben we toegepast dat de relatieve entropie $D(Y,X) \geq 0$ is, en bij $(**)$ dat X en Y kansverdelingen met dezelfde verwachtingswaarde en variantie hebben, dus dat $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} g(x)\,dx = 1$, $\int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{\infty} xg(x)\,dx = 0$ en $\int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 g(x)\,dx = \sigma^2$.

Als voorbeelden vergelijken we de entropie van een normale verdeling met variantie σ^2 met de entropie van een uniforme verdeling met dezelfde variantie.

Entropie van de normale verdeling

Zij X een normaal verdeelde stochast met verwachtingswaarde μ en variantie σ^2, dan heeft X de dichtheidsfunctie $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Voor de entropie van X geldt

$$\begin{aligned}
H(X) &= -\int_{-\infty}^{\infty} f(x)\ {}^2\log(f(x))\,dx \\
&= -\int_{-\infty}^{\infty} f(x)({}^2\log(\frac{1}{\sqrt{2\pi}\,\sigma}) + {}^2\log(e^{-\frac{(x-\mu)^2}{2\sigma^2}}))\,dx \\
&= -\ {}^2\log(\frac{1}{\sqrt{2\pi}\,\sigma})\underbrace{\int_{-\infty}^{\infty} f(x)\,dx}_{=1} - \frac{1}{\log(2)}\int_{-\infty}^{\infty} f(x)\log(e^{-\frac{(x-\mu)^2}{2\sigma^2}})\,dx \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) - \frac{1}{\log(2)}\int_{-\infty}^{\infty} f(x)(-\frac{(x-\mu)^2}{2\sigma^2})\,dx \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) + \frac{1}{\log(2)}\frac{1}{2\sigma^2}\underbrace{\int_{-\infty}^{\infty} f(x)(x-\mu)^2\,dx}_{=Var(X)=\sigma^2} \\
&= {}^2\log(\sqrt{2\pi}\,\sigma) + \frac{1}{2\log(2)} = {}^2\log(\sqrt{2\pi}\,\sigma) + {}^2\log(\sqrt{e}) \\
&= {}^2\log(\sqrt{2\pi e}\,\sigma)
\end{aligned}$$

Entropie van de uniforme verdeling

Zij X een stochast met uniforme verdeling op het interval $[-a, a]$, dus met dichtheidsfunctie $f(x) = \frac{1}{2a}$ voor $x \in [-a, a]$ en $f(x) = 0$ voor $x \notin [-a, a]$. We moeten eerst de variantie van X berekenen, hiervoor geldt

$$Var(X) = \int_{-a}^{a} \frac{1}{2a} x^2 \, dx = \frac{1}{2a} \cdot \frac{x^3}{3}\Big|_{-a}^{a} = \frac{1}{2a} \frac{2a^3}{3} = \frac{a^2}{3}.$$

De variantie is dus voor $a = \sqrt{3}\,\sigma$ gelijk aan σ^2.

Voor de entropie geldt nu

$$H(X) = -\int_{-a}^{a} \frac{1}{2a} \; {}^2\log(\frac{1}{2a}) \, dx = -\; {}^2\log(\frac{1}{2a}) = \; {}^2\log(2a).$$

Voor een uniforme stochast met variantie σ^2, dus met $a = \sqrt{3}\,\sigma$, krijgen we dus de entropie

$$H(X) = \; {}^2\log(\sqrt{12}\,\sigma).$$

Omdat $\sqrt{12} \approx 3.464 < 4.132 \approx \sqrt{2\pi e}$ is de entropie bij een uniforme verdeling inderdaad kleiner dan bij een normale verdeling met dezelfde variantie.

11.3 Voorwaardelijke entropie

Een belangrijke vraag is hoe zich de entropie van verschillende stochasten gedraagt als we deze combineren. We zouden verwachten, dat voor twee onafhankelijke stochasten X en Y de entropie van de combinatie van X en Y de som van de entropieën van X en Y is. Voor stochasten X, Y met uniforme verdelingen is dit juist eis (7) in onze lijst. Voor twee stochasten X en Y geldt inderdaad de stelling:

$H(X, Y) \leq H(X) + H(Y)$ en
$H(X, Y) = H(X) + H(Y)$ alleen maar als X en Y onafhankelijk zijn.

Dit zien we als volgt in: We definiëren de kansen voor de stochasten als $p_i := p(X = x_i)$ voor $1 \leq i \leq n$, $y_j := p(Y = y_j)$ voor $1 \leq j \leq m$ en de gecombineerde kans als $r_{ij} := p(X = x_i, Y = y_j)$. Als we voor vaste i de kansen r_{ij} voor alle j optellen, krijgen we de kans op x_i, dus geldt $p_i = \sum_{j=1}^{m} r_{ij}$ en evenzo $q_j = \sum_{i=1}^{n} r_{ij}$. We hebben dus

$$\begin{aligned}
H(X) + H(Y) &= -\sum_{i=1}^{n} p_i \; {}^2\log(p_i) - \sum_{j=1}^{m} q_j \; {}^2\log(q_j) \\
&= -\sum_{i=1}^{n} (\sum_{j=1}^{m} r_{ij}) \; {}^2\log(p_i) - \sum_{j=1}^{m} (\sum_{i=1}^{n} r_{ij}) \; {}^2\log(q_j) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} ({}^2\log(p_i) + {}^2\log(q_j)) = -\sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \; {}^2\log(p_i q_j) \\
&\geq -\sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \; {}^2\log(r_{ij}) = H(X, Y).
\end{aligned}$$

De ongelijkheid $-\sum\sum r_{ij}\ {}^2\log(p_iq_j) \geq -\sum\sum r_{ij}\ {}^2\log(r_{ij})$ volgt hierbij weer uit de eigenschap (II) die we boven hebben bewezen, omdat ook p_iq_j een kansverdeling op $\{1,\ldots,n\}\times\{1,\ldots,m\}$ is.

We zien dat $H(X)+H(Y)$ alleen maar geldt als $p_iq_j = r_{ij}$ voor alle paren (i,j), dus als $p(X=x_i)\cdot p(Y=y_j) = p(X=x_i, Y=y_j)$, maar dit is precies de uitspraak dat X en Y onafhankelijk zijn.

Als we stochasten combineren, moeten we het natuurlijk ook over voorwaardelijke kansen hebben. Maar voorwaardelijke kansen zijn ook gewoon kansverdelingen: Als we de kans op een uitkomst x_i voor de stochast X onder de voorwaarde A weer als $p_i := p(X=x_i \mid A)$ beschrijven, is $P=(p_1,\ldots,p_n)$ een kansverdeling en $\sum_{i=1}^n p_i = 1$. We definiëren daarom de *voorwaardelijke entropie* $H(X \mid A)$ door

$$H(X \mid A) := -\sum_{i=1}^n p(X=x_i \mid A)\ {}^2\log(p(X=x_i \mid A)).$$

Nog algemener kunnen we ook de voorwaardelijke entropie van een stochast X, gegeven een andere stochast Y definiëren. Het idee hierbij is, dat de uitkomsten van de stochast Y de onzekerheid over de stochast X kunnen veranderen. We lopen dus over alle mogelijke uitkomsten y_j van de stochast Y, berekenen voor deze uitkomsten de voorwaardelijke entropie $H(X \mid y_j)$ en tellen deze entropieën op, met de kansen op de enkele y_j als gewichten.

Definitie: De *voorwaardelijke entropie* van de stochast X onder de voorwaarde van de stochast Y is gedefinieerd door

$$\begin{aligned} H(X \mid Y) &:= \sum_{j=1}^m H(X \mid Y=y_j)p(Y=y_j) \\ &= -\sum_{j=1}^m\sum_{i=1}^n p(X=x_i \mid Y=y_j)\ {}^2\log(p(X=x_i \mid Y=y_j))\cdot p(Y=y_j). \end{aligned}$$

Dat deze definitie enigszins zinvol is, zien we aan de twee extreme gevallen $Y=X$ en X,Y onafhankelijk:

(1) Als $Y=X$ is, dan is $p(X=x_i \mid X=x_j)=1$ als $i=j$ en 0 als $i\neq j$. Maar dan geldt

$$\begin{aligned} H(X|X) &= -\sum_{j=1}^n\sum_{i=1}^n p(X=x_i|X=x_j)\ {}^2\log(p(X=x_i|X=x_j))p(X=x_j) \\ &= -\sum_{i=1}^n 1\cdot 0\cdot p(X=x_i) = 0. \end{aligned}$$

Er geldt dus

$$H(X \mid X) = 0.$$

Dit zegt dat er geen onzekerheid over X meer bestaat, als we de uitkomsten van X al kennen.

(2) Als X en Y onafhankelijk zijn, dan geldt $p(X = x_i \mid Y = y_j) = p(X = x_i)$, en hieruit volgt

$$\begin{aligned} H(X \mid Y) &= -\sum_{j=1}^{m}\sum_{i=1}^{n} p(X = x_i)\ {}^2\log(p(X = x_i))p(Y = y_j) \\ &= -\sum_{i=1}^{n} p(X = x_i)\ {}^2\log(p(X = x_i)) = H(X). \end{aligned}$$

Voor onafhankelijke stochasten X en Y geldt dus

$$H(X \mid Y) = H(X).$$

Dit betekent, dat de kennis over Y de onzekerheid bij X niet reduceert, en dat is precies wat we bij onafhankelijke stochasten zouden verwachten.

We kunnen nu ook de precieze samenhang tussen de voorwaardelijke entropie $H(X \mid Y)$ en de entropie van de combinatie van X en Y aangeven, er geldt namelijk

$$H(X, Y) = H(Y) + H(X \mid Y) \quad \text{of te wel} \quad H(X \mid Y) = H(X, Y) - H(Y).$$

Dit zien we als volgt in: We schrijven weer $r_{ij} := p(X = x_i, Y = y_j)$ voor de gecombineerde kans op x_i en y_j. Volgens de definitie van de voorwaardelijke kans geldt dat $p(X = x_i \mid Y = y_j) = \frac{r_{ij}}{q_j}$ en dus $r_{ij} = p(X = x_i \mid Y = y_j)q_j$, waarbij we weer $q_j := p(Y = y_j)$ schrijven. Er geldt dus:

$$\begin{aligned} H(X, Y) &= -\sum_{i,j} r_{ij}\ {}^2\log(r_{ij}) = -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)q_j) \\ &= -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)) - \sum_{i,j} r_{ij}\ {}^2\log(q_j) \\ &= -\sum_{i,j} r_{ij}\ {}^2\log(p(X = x_i \mid Y = y_j)) - \sum_{j=1}^{m} q_j\ {}^2\log(q_j) \\ &= -\sum_{i,j} p(X = x_i \mid Y = y_j)q_j\ {}^2\log(p(X = x_i \mid Y = y_j)) - H(Y) \\ &= H(X \mid Y) + H(Y). \end{aligned}$$

Hieruit volgt in het bijzonder dat

$$H(X \mid Y) \leq H(X),$$

want $H(X \mid Y) = H(X, Y) - H(Y) \leq H(X) + H(Y) - H(Y) = H(X)$, en dus is de voorwaardelijke entropie van een stochast nooit groter dan zijn absolute entropie. Ook dit is een eigenschap die we van een redelijke maat voor onzekerheid hadden kunnen verwachten, want door aanvullende informatie zouden we niet onzekerder over de uitkomsten van X worden.

11.4 Informatie

We hebben bij de voorwaardelijke entropie gezien, dat kennis over een stochast Y de onzekerheid over de stochast X kan reduceren. Het verschil van de entropieën $H(X) - H(X \mid Y)$ kunnen we dus zien als de informatie die Y aan onze kennis over X bijdraagt. Dit lijdt tot een precieze definitie van het begrip *informatie*, die we nu gaan behandelen.

Net als bij de entropie stellen we ook bij de informatie eisen aan een functie die de informatie van een gebeurtenis beschrijft. We schrijven $I(X = x_i)$ voor de informatie die de uitkomst x_i van de stochast X oplevert. Maar eigenlijk mag een abstracte definitie van informatie niet van de specifieke uitkomst afhangen, maar alleen maar van de kans op deze uitkomst. Dit geeft aanleiding tot de eerste eis die we aan een functie voor de informatie hebben:

(1) Er geldt $I(X = x_i) = I(p_i)$ voor $p_i = p(X = x_i)$.

Verder bekijken we de informatie van onafhankelijke gebeurtenissen: Als X en Y onafhankelijke stochasten zijn, geldt met $p_i = p(X = x_i)$ en $q_j = p(Y = y_j)$ dat $p(X = x_i, Y = y_j) = p_i q_j$. Maar het ligt voor de hand dat de informatie die in de uitkomst $X = x_i$ en $Y = y_j$ zit, de som van de informaties van de enkele uitkomsten is. Dit geeft de eis:

(2) Voor *onafhankelijke* stochasten X en Y met $p_i = p(X = x_i)$ en $q_j = p(Y = y_j)$ geldt $I(p_i q_j) = I(p_i) + I(q_j)$.

Met een soortgelijke (maar eenvoudigere) redenering als bij de entropie kan men nu aantonen dat de functie I noodzakelijk van de vorm $I(p) = -\lambda \log(p)$ is, en ook hier kiest men voor de logaritme met basis 2, dus definieert men:

Definitie: Voor een stochast X is de *informatie* van de uitkomst $X = x$ met $p(X = x) = p$ gegeven door

$$I(p) := -\,{}^2\log(p).$$

Deze definitie van informatie is in ieder geval ook in overeenstemming met onze intuïtie dat het optreden van een gebeurtenis met een kleine kans meer informatie oplevert dan een gebeurtenis met een grote kans, dus van *het gewone*.

Een belangrijke rechtvaardiging van deze definitie van informatie vinden we weer in de communicatietheorie: Als we een bit-string van lengte n produceren door toevallig n keer een 0 of 1 te kiezen, heeft elke bit van de string de informatie $I(\frac{1}{2}) = -\,{}^2\log(\frac{1}{2}) = {}^2\log(2) = 1$ en de totale informatie in de string is dus $-n\,{}^2\log(\frac{1}{n}) = n$, omdat de keuzes van de bits onafhankelijk zijn. Het is daarom ook gebruikelijk, informatie (en entropie) in *bits* aan te geven.

Verband tussen informatie en entropie

Met behulp van het begrip van informatie kunnen we nu de entropie herinterpreteren. Er geldt

$$H(X) = -\sum p_i \, {}^2\log(p_i) = \sum p_i \cdot I(p_i)$$

dus is de entropie het gemiddelde van de informatie in de enkele uitkomsten, gewogen met de kansen van de uitkomsten. Maar in de taal van de kansrekening is dit gewogen gemiddelde juist de verwachtingswaarde:

Merk op: De entropie $H(X)$ van een stochast X is de verwachtingswaarde van de informatie van de enkele uitkomsten van de stochast.

Dit kunnen we ook nog iets anders formuleren: Een uitkomst met informatie $I = {}^2\log(n)$ heeft kans $p = \frac{1}{n}$. Als de uitkomst bij een uniforme verdeling hoort, is $\frac{1}{p} = n$ het aantal mogelijke uitkomsten. Dit betekent dat we voor een uniforme verdeling het aantal mogelijke uitkomsten kunnen schrijven als $n = 2^I$, waarbij I de informatie is die in een enkele uitkomst zit. Maar we hebben net gezien dat de entropie de verwachtingswaarde van de informatie in de enkele uitkomsten is, dus kunnen we $2^{H(X)}$ interpreteren als het gemiddelde aantal alternatieven, dat we bij de stochast X kunnen verwachten. Dit kunnen we ook als volgt formuleren:

Merk op: De onzekerheid bij een stochast X is even groot als de onzekerheid bij een uniforme verdeling met $2^{H(X)}$ mogelijke uitkomsten. Anders gezegd is $2^{H(X)}$ het gemiddelde aantal alternatieven, dat we bij een kansexperiment voor de stochast X verwachten.

We hebben in het begin van deze sectie gesteld, dat het verschil van de entropieën $H(X) - H(X \mid Y)$ de informatie is, die Y over X onthult. Als notatie hiervoor gebruiken we

$$I(X \mid Y) := H(X) - H(X \mid Y).$$

Er geldt $I(X \mid X) = H(X)$, want $H(X \mid X) = 0$, en dit is ook zinvol omdat kennis van X de onzekerheid over X precies moet compenseren. Aan de andere kant geldt voor onafhankelijke stochasten X en Y dat $I(X \mid Y) = 0$, want $H(X \mid Y) = H(X) + H(Y)$. Ook dit is juist wat we nodig hebben, want onafhankelijke stochasten mogen onderling geen informatie onthullen.

Bij de definitie van $I(X \mid Y)$ kijken we naar de gemiddelde reductie die de enkele uitkomsten van Y voor de entropie van X opleveren. We kunnen natuurlijk ook naar de informatie kijken, die een bepaalde uitkomst $Y = y$ voor de stochast Y over X oplevert, deze is gedefinieerd door

$$I(X \mid Y = y) = H(X) - H(X \mid Y = y).$$

Er bestaat een iets verrassende symmetrie voor het onthullen van informatie van een stochast over de andere. We hebben namelijk

$$\begin{aligned} I(X \mid Y) &= H(X) - H(X \mid Y) = H(X) - (H(X,Y) - H(Y)) \\ &= H(Y) + (H(X) - H(X,Y)) = H(Y) - H(Y \mid X) \\ &= I(Y \mid X), \end{aligned}$$

dus onthult de stochast X net zo veel informatie over Y als de stochast Y over X onthult.

11.5 Toepassing: Automatische Taalherkenning

Als voorbeeld voor de toepassing van de concepten van entropie en informatie bekijken we het probleem van de automatische taalherkenning op geschreven tekst. Voor een mens is dit meestal nauwelijks een probleem, tenminste bij bekende talen of bij talen waar men iets over weet, maar de automatisering hiervan is al een stukje lastiger.

Onze aanpak is, de relatieve frequenties van de letters te gebruiken. Het is natuurlijk bekend dat de letters in het alfabet niet even vaak gebruikt worden, in het Nederlands is bijvoorbeeld de letter `E` de meest frequente. Het idee is dat de relatieve frequenties voor verschillende talen er verschillend uit zien en dat we hiermee de talen kunnen onderscheiden.

Vanaf de 16de eeuw zijn de relatieve frequenties in de cryptanalyse gebruikt om versleutelingen met monoalfabetische substitutie (elke letter wordt door een andere letter vervangen, maar één letter steeds door dezelfde) te kraken. Tot op die tijd dacht men eigenlijk dat zo'n versleuteling niet te kraken was, omdat er veel te veel sleutels bestaan ($26! \approx 4.03 \cdot 10^{26}$) om alle te proberen. Maar als men al weet dat de meest frequente letter in de versleuteling een `E` is en de volgende waarschijnlijk een `N` kan men al gauw verdere letters gokken.

Het idee dat de letters überhaupt verschillende frequenties hebben, is waarschijnlijk pas na de opkomst van de boekdrukkerij (door Gutenberg) ontdekt, omdat de loodletters verschillend snel versleten waren.

Voor een gegeven taal kan men op een grote achtergrondtekst de frequenties tellen en dit als kansverdeling van de stochast X die de letters beschrijft nemen. Men krijgt zo de kansen $p_1 := p(X = \texttt{A})$, $p_2 := p(X = \texttt{B})$, $\ldots$, $p_{26} := p(X = \texttt{Z})$, $p_{27} := p(X = \textit{spatie})$.

Tabel III.1 geeft deze kansverdelingen voor de vier talen *Nederlands*, *Engels*, *Duits* en *Fins* weer. De gebruikte achtergrondtekst is een tekst van de Europese Unie die in de verschillende talen vertaald is en ongeveer 50000 letters bevat. Uit deze tabel kan men concluderen dat de kansverdelingen voor Nederlands, Engels en Duits enigszins op elkaar lijken, terwijl de verdeling voor Fins er behoorlijk anders uit ziet. Bijvoorbeeld bepaalt de relatieve frequentie van de *spatie* de gemiddelde lengte van de woorden (namelijk door $l_{gem} = \frac{1}{p} - 1$) en men ziet dat de woorden in het Fins gemiddeld duidelijk langer zijn dan in de andere talen.

Een betere voorstelling van de frequentieverdelingen dan met de tabel krijgt men door de verdelingen als histogrammen te plotten, zo als in Figuur III.2 te zien. Hier valt bijvoorbeeld op, dat er in het Fins meer letters met een relatief hoge frequentie zijn, en dat in het Nederlands en Duits de letter `E` met duidelijke afstand de hoogste frequentie heeft.

letter	Nederlands	Engels	Duits	Fins
A	5.55%	6.37%	4.14%	9.57%
B	1.45%	0.99%	1.82%	0.10%
C	1.45%	3.20%	2.09%	0.05%
D	4.72%	2.56%	4.09%	1.40%
E	17.31%	9.93%	13.89%	8.50%
F	0.68%	1.95%	2.28%	0.07%
G	2.79%	1.41%	2.67%	0.19%
H	1.83%	3.00%	3.00%	1.77%
I	6.09%	7.62%	8.22%	9.90%
J	0.70%	0.10%	0.14%	1.57%
K	1.51%	0.27%	1.21%	4.74%
L	2.87%	2.93%	2.83%	3.75%
M	1.98%	2.52%	2.81%	2.65%
N	8.67%	7.63%	9.14%	8.08%
O	4.94%	7.73%	2.92%	6.68%
P	1.53%	2.78%	1.03%	1.78%
Q	0.01%	0.04%	0.01%	0.01%
R	5.81%	5.15%	6.69%	2.16%
S	3.44%	4.92%	5.10%	8.24%
T	5.63%	8.30%	5.40%	9.54%
U	2.01%	2.57%	3.85%	4.70%
V	2.77%	0.70%	0.80%	2.10%
W	0.67%	0.75%	0.77%	0.02%
X	0.05%	0.12%	0.05%	0.01%
Y	0.04%	0.84%	0.06%	1.71%
Z	0.55%	0.02%	1.36%	0.05%
spatie	14.94%	15.61%	13.63%	10.64%

Tabel III.1: Letter frequenties voor vier verschillende talen

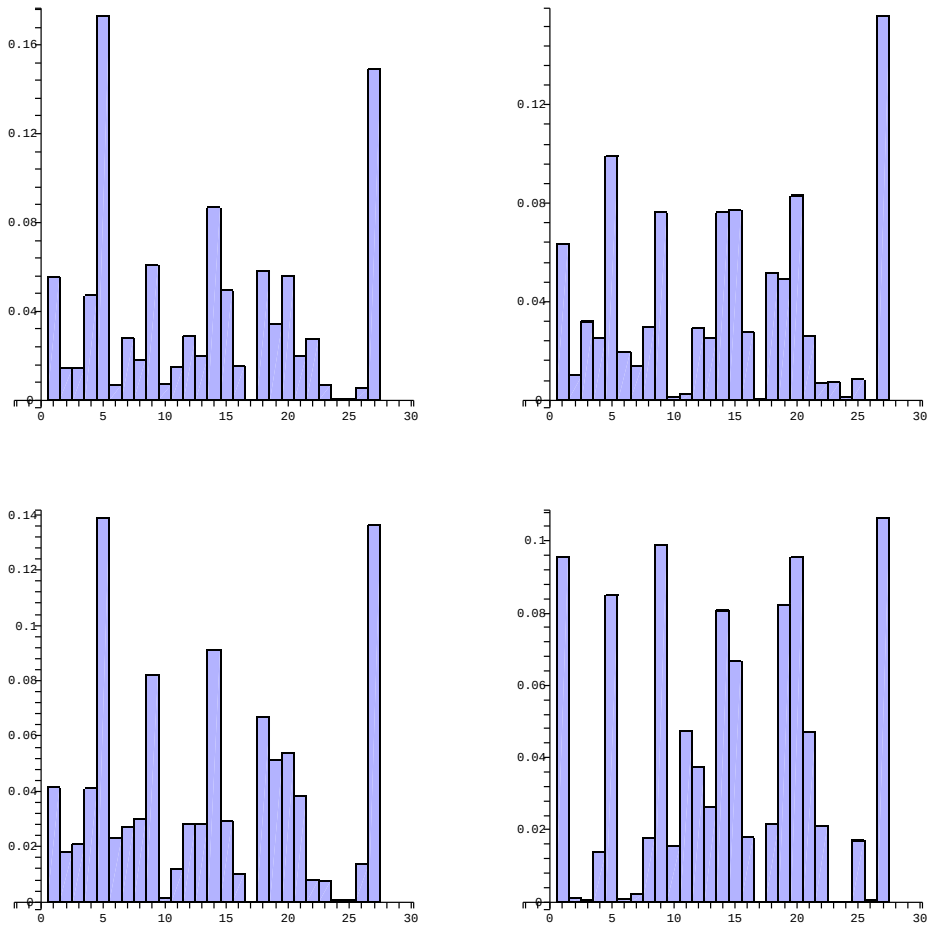
Als we de frequentieverdelingen als kansverdelingen opvatten, kunnen we voor de verschillende talen de entropieën van deze verdelingen uitrekenen, dit geeft de volgende waarden:

$$H(\text{Nederlands}) = 4.019, \qquad H(\text{Engels}) = 4.070,$$
$$H(\text{Duits}) = 4.109, \qquad H(\text{Fins}) = 3.982.$$

Met de interpretatie van de entropie met behulp van informatie geeft dit:

$$2^{H(\text{Nederlands})} = 16.21, \qquad 2^{H(\text{Engels})} = 16.80,$$
$$2^{H(\text{Duits})} = 17.26, \qquad 2^{H(\text{Fins})} = 15.80.$$

Het gemiddelde aantal alternatieven, dat we in de verschillende talen voor een letter verwachten, ligt dus tussen 15.80 voor Fins en 17.26 voor Duits, terwijl we bij een uniforme verdeling 27 alternatieven zouden hebben.



Figuur III.2: Letter-frequentieverdelingen voor Nederlands (links boven) en Engels (rechts boven), Duits (links onder) en Fins (rechts onder).

Classificatie van patronen

Een typisch probleem in de patroonherkenning is, gegeven een aantal klassen $K_1, \ldots, K_n$ van mogelijke patronen, een nieuw patroon aan een van de klassen K_i toe te wijzen. Denk bij de klassen bijvoorbeeld aan letters in de handschriftherkenning, aan woorden of fonemen in de spraakherkenning of objecten in de beeldherkenning. In ons voorbeeld van de automatische taalherkenning zijn de klassen natuurlijk de talen en het nieuwe patroon is een nieuwe tekst.

In het verleden is geprobeerd, regels te vinden waarmee de klasse van een nieuw patroon bepaald kan worden. Maar er is gebleken dat dit slechts zeer beperkt inzetbaar is en de beste methoden in de patroonherkenning gebruiken nu probabilistische modellen, bijvoorbeeld (hidden) Markov modellen of/en neuronale netwerken.

Er zijn verschillende mogelijkheden voor de rol die kansverdelingen bij het classificeren van patronen kunnen spelen:

- Het nieuwe patroon wordt door een vector (of een rij vectoren) in de *kenmerkruimte* (feature space) weergegeven. De klassen zijn gerepresenteerd door kansverdelingen op de kenmerkruimte die aangeven hoe groot de kans is dat een patroon met een zekere vector bij deze klasse hoort. Het patroon wordt dan aan de klasse toegewezen waarvoor deze kans maximaal is.

- Ook voor het patroon wordt een kansverdeling bepaald en er wordt de klasse gekozen, waarvoor deze kansverdeling het meeste op de eerder berekende kansverdeling van de klasse lijkt.

We zullen de tweede insteek nu eens nader bekijken, omdat die minder voor de hand liggend lijkt als de eerste. In het voorbeeld van de automatische taalherkenning zijn de kansverdelingen gegeven door de relatieve frequenties van de letters. Voor een nieuwe tekst waarvan we de taal willen bepalen moeten we daarom ook de frequentieverdeling berekenen en vervolgens deze kansverdeling met de bekende kansverdelingen van de verschillende talen vergelijken. De aanname is dan, dat de tekst bij die taal hoort waarvoor de kansverdelingen het meeste op elkaar lijken.

De vraag is nu hoe men objectief bepaald, dat een kansverdeling meer op een dan op een andere lijkt.

Afstanden tussen kansverdelingen

Om een eenvoudige notatie te krijgen, beschrijven we een discrete kansverdeling P op de verzameling $\Omega = \{1, \ldots, n\}$ door de vector van kansen $p_i := p(i)$, dus $P = (p_1, p_2, \ldots, p_n)$. Voor een tweede kansverdeling $Q = (q_1, q_2, \ldots, q_n)$ op dezelfde verzameling Ω willen we nu een afstand tussen P en Q definiëren.

Een voor de hand liggende idee is, de Euclidische afstand van de vectoren P en Q in de n-dimensionale ruimte te nemen, dit geeft

$$d_2(P, Q) = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{\frac{1}{2}}.$$

Maar net zo goed zouden we in plaats van de kwadraten van de verschillen tussen p_i en q_i ook de absolute waarden van de verschillen kunnen optellen:

$$d_1(P, Q) = \sum_{i=1}^{n} |p_i - q_i|.$$

We kunnen zelfs heel algemeen een macht van de verschillen tussen p_i en q_i optellen, dit geeft

$$d_r(P, Q) = \left(\sum_{i=1}^{n} |p_i - q_i|^r \right)^{\frac{1}{r}}.$$

Hierbij hoeft r niet eens een geheel getal te zijn, we kunnen een willekeurige r met $0 < r < \infty$ kiezen. De reden dat we bij een r-de macht ook weer een

r-de machtswortel trekken, heeft ermee te maken dat men graag wil dat een vermenigvuldiging van de vectoren met een constante factor tot een vermenigvuldiging van de afstand met dezelfde factor leidt.

Voor de volledigheid noemen we nog een verdere afstand, die we formeel kunnen krijgen als we bij $d_r(P,Q)$ de $r \to \infty$ laten lopen. Dan krijgen we namelijk de afstand

$$d_\infty(P,Q) = \max_i |p_i - q_i|$$

die gewoon het grootste verschil in een van de componenten aangeeft. Maar als we naar vectoren van kansverdelingen kijken, is dit meestal geen bijzonder nuttige afstand.

De vraag welke afstand nu een slimme keuze is, heeft helaas geen eenvoudig antwoord. Het hangt namelijk van het probleem af. Hoe groter de waarde van de parameter r is hoe groter is relatief het gewicht van de grotere verschillen en hoe kleiner de invloed van kleine verschillen. Als r heel groot wordt, speelt inderdaad alleen maar het grootste verschil nog een rol. In sommige problemen is het misschien wenselijk, kleine verschillen te onderdrukken, maar soms ligt de informatie juist in de componenten met kleine verschillen.

In een iets algemenere opzet zou men voor elke component een functie $d_i(p_i, q_i)$ definiëren, die de afstand in deze component aangeeft. Als afstand krijgt men dan

$$d(P,Q) = \sum_{i=1}^{n} d_i(p_i, q_i).$$

Hierbij kan de functie d_i aan de ene kant ervoor zorgen, dat componenten met belangrijkere informatie een hoog gewicht krijgen, maar ook dat afhankelijk van de kansen een hoger of lager gewicht toegewezen wordt.

Een eenvoudig voorbeeld hiervan is het toewijzen van gewichten aan de enkele componenten, dus bijvoorbeeld

$$d(P,Q) = \sum_{i=1}^{n} w_i |p_i - q_i| \quad \text{of} \quad d(P,Q) = \sum_{i=1}^{n} w_i p_i q_i.$$

Het laatste is een inproduct van de twee vectoren P en Q en geeft weer dat we in principe ook de hoek tussen twee vectoren als een soort afstand kunnen interpreteren, zeker als de lengte van de vectoren genormeerd is.

> Het idee de afstand tussen kansverdelingen met behulp van een inproduct te berekenen wordt bijvoorbeeld in (eenvoudige) zoekmachines gebruikt, de gewichten zijn dan bijvoorbeeld de negatieve logaritmen van de relatieve frequenties van de woorden. Zo houdt men rekening ermee, dat frequente woorden weinig informatie over een document geven, terwijl minder frequente woorden vaak een belangrijke hint zijn.

De afstanden die we tot nu toe hebben bekeken, hebben op zich weinig met kansverdelingen te maken, want we hebben eigenlijk alleen maar naar vectoren gekeken. Het enige wat van de kansverdelingen over blijft, is dat de som van de componenten 1 is, dus dat $\sum_{i=1}^{n} p_i = 1$.

Kullback-Leibler afstand

Maar natuurlijk hebben we eerder in deze les ook al een maat voor de afstand tussen kansverdelingen gezien, namelijk de Kullback-Leibler afstand (of relatieve entropie).

We hadden gezien dat de Kullback-Leibler afstand $D(P,Q)$ het verschil tussen $-\sum p_i\ {}^2\log(q_i)$ en de entropie $H(P)$ van de kansverdeling P aangeeft, dus dat

$$D(P,Q) = \sum_{i=1}^{n} p_i\ {}^2\log(\frac{p_i}{q_i}) = (-\sum_{i=1}^{n} p_i\ {}^2\log(q_i)) - H(P).$$

Als we nu $2^{H(P)}$ als het gemiddelde aantal alternatieven interpreteren, die we bij een stochast X met kansverdeling P verwachten, kunnen we ook de Kullback-Leibler afstand herinterpreteren: Er geldt

$$2^{H(P)+D(P,Q)} = 2^{H(P)} \cdot 2^{D(P,Q)},$$

dus is $2^{D(P,Q)}$ de factor waarmee we het gemiddelde aantal alternatieven moeten vermenigvuldigen, omdat we de *verkeerde* kansverdeling Q in plaats van P veronderstellen.

De volgende tabellen geven links de Kullback-Leibler afstanden tussen de talen uit het voorbeeld met de frequentieverdelingen en rechts de factoren $2^{D(P,Q)}$. Hierbij betekent bijvoorbeeld een factor 1.138 een afwijking van 13.8% van het aantal verwachte alternatieven bij de juiste kansverdeling. Merk op dat de tabellen niet symmetrisch zijn, omdat we de gewone Kullback-Leibler afstand $D(P,Q)$ en niet de symmetrische versie $d_{KL}(P,Q)$ toepassen.

taal	NL	EN	DU	FI
NL	-	0.186	0.091	0.471
EN	0.171	-	0.155	0.458
DU	0.090	0.177	-	0.610
FI	0.397	0.373	0.453	-

taal	NL	EN	DU	FI
NL	-	1.138	1.065	1.386
EN	1.126	-	1.114	1.373
DU	1.064	1.130	-	1.527
FI	1.317	1.295	1.368	-

Het is opvallend hoe sterk Duits en Fins van elkaar afwijken, terwijl Nederlands en Duits redelijk dicht bij elkaar liggen.

De Kullback-Leibler afstand speelt een belangrijke rol bij het bepalen van de parameters van probabilistische modellen. Het idee is dat op een zekere hoeveelheid training materiaal de kansen p_i worden bepaald en vervolgens een probabilistisch model gebouwd wordt, dat van enkele parameters afhangt. Dit kan bijvoorbeeld een normale verdeling zijn, met als parameters de verwachtingswaarde en de variantie. Deze parameters kunnen meestal niet rechtstreeks berekent worden, maar worden in een iteratief proces benadert, waarbij de Kullback-Leibler afstand stapsgewijs kleiner wordt. Als geen verbetering meer bereikt wordt, worden deze parameters voor het model gekozen.

Belangrijke begrippen in deze les

- onzekerheid, entropie
- relatieve entropie, Kullback-Leibler afstand
- entropie bij continue kansverdelingen
- maximale entropie bij normale verdeling
- voorwaardelijke entropie
- informatie
- afstanden tussen kansverdelingen

Opgaven

84. Er vinden twee paardenraces plaats, het eerste met 7 paarden en het tweede met 8 paarden. In de eerste race hebben 3 paarden kans $\frac{1}{6}$ om te winnen, de andere 4 hebben kans $\frac{1}{8}$. In de tweede race hebben 2 paarden kans $\frac{1}{4}$ om te winnen en de andere 6 kans $\frac{1}{12}$. Maak eerst een gok in welk van de races de uitkomst onzekerder is (en geef een reden hiervoor), en bereken dan de entropieën voor de twee races.

85. Er wordt met een eerlijke dobbelsteen gedobbeld. De stochast X geeft het aantal ogen dat gedobbeld wordt, de stochast Y heeft de waarde 0 of 1, afhankelijk of het aantal ogen even of oneven is. Bereken $H(X)$, $H(Y)$ en $H(X \mid Y)$.

86. Voor een geheel getal N neemt de stochast X volgens een uniforme verdeling de waarden $1, 2, \ldots, 2N$ aan. De stochast Y is 0 als de waarde van X even is en Y is 1 als de waarde van X oneven is. Laat zien dat $H(X \mid Y) = H(X) - 1$ en dat $H(Y \mid X) = 0$.

87. De uitkomsten van twee (eerlijke) dobbelstenen worden door de stochasten X en Y beschreven, de som van de twee dobbelstenen door de stochast Z. Ga na dat voor de combinatie van de stochasten X en Y geldt dat $H(X, Y) = H(X) + H(Y)$ en dat $H(Z) < H(X, Y)$.

88. Een stochast X heeft een binomiale verdeling met parameters n en p, d.w.z. de kans op de i-de uitkomst is $p(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$. Laat zien dat

$$H(X) = -n(p\ {}^2\log(p) + (1-p)\ {}^2\log(1-p)).$$

89. Laat zien dat de entropie $H(X)$ van een continue stochast X met een exponentiële verdeling met dichtheidsfunctie

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \text{ voor } x \geq 0$$

gegeven is door

$$H(X) = {}^2\log(\lambda\, e).$$

90. Bij een *best-of-five* tennis match is de speler de winnaar die als eerste drie sets gewonnen heeft. Stel dat de spelers A en B (ongeveer) even sterk zijn, zo dat een set met kans $\frac{1}{2}$ door A of B gewonnen wordt.

    Zij X de stochast die de mogelijke rijtjes van gewonnen sets beschrijft, dus bijvoorbeeld AAA, $ABBAA$ of $ABBB$. Verder zij Y de stochast die het aantal benodigde sets aangeeft (en dus een van de waarden 3, 4 of 5 heeft).

    Bepaal de entropieën $H(X)$ en $H(Y)$ en de voorwaardelijke entropieën $H(Y \mid X)$ en $H(X \mid Y)$.

91. Waar zit meer informatie in, in een string van 10 letters uit $\{A, \ldots, Z\}$ of in een string van 26 cijfers uit $\{0, \ldots, 9\}$?

92. Er wordt met een eerlijke dobbelsteen gedobbeld. Wat is de informatie, die de kennis dat het aantal ogen niet door 3 deelbaar is, over het aantal ogen onthult?

93. Uit onderzoek is gebleken dat 70% van de mannen donker haar hebben en 25% van de vrouwen blond zijn. Verder is bekend dat 80% van de blonde vrouwen met een donkerharig man trouwen. Hoeveel informatie over de haarkleur van de man onthult de haarkleur van zijn vrouw?

Les 12 Markov processen en Markov modellen

We hebben in het kader van de Lineaire Algebra in Wiskunde 1 een aantal voorbeelden van systemen gezien, die zich door overgangsmatrices laten beschrijven. Voorbeelden hiervan waren:

- Populaties die zich volgens overgangen tussen de verschillende generaties ontwikkelen.
- De verspreiding van de Euro munten over de verschillende landen.

Iets algemener gesproken hebben we het hierbij over systemen, die gekarakteriseerd zijn door: (1) mogelijke *toestanden* van het systeem; en (2) *overgangen* tussen deze toestanden.

We zullen het in deze (en de volgende) les over dit soort systemen hebben, waarbij we vooral naar het belangrijke geval kijken, dat de overgangen door kansverdelingen worden beschreven.

12.1 Markov processen

Als de overgangen tussen de mogelijke toestanden van een systeem door kansverdelingen gegeven zijn, spreekt men meestal van *Markov processen*.

Definitie: Een *Markov proces* wordt door de volgende gegevens gekarakteriseerd:

- een aantal mogelijke toestanden $S_1, S_2, \ldots, S_N$, die we *states* noemen;
- op elk tijdstip $t = 0, 1, 2, \ldots$ een state $q_t \in \{S_1, \ldots, S_N\}$, waarin het systeem zich op dit tijdstip bevindt;
- gegeven de states $q_0, q_1, \ldots, q_{t-1}$ op de tijdstippen $0, 1, \ldots, t-1$, de kansverdeling dat het systeem op tijdstip t in de state S_j terecht komt, d.w.z. de voorwaardelijke kansen

$$p(q_t = S_j \mid q_{t-1} = S_{i_{t-1}}, \ldots, q_1 = S_{i_1}, q_0 = S_{i_0}).$$

Het probleem om in de praktijk een proces als een Markov proces te beschrijven, zit in de exponentiële groei van het aantal mogelijke states op de tijdstippen $0, 1, \ldots, t-1$, dit zijn er namelijk N^t. Voor elke van deze mogelijkheden moeten we in principe een aparte kansverdeling voor de states op tijdstip t aangeven, maar dat is natuurlijk voor grotere waarden van t ondoenlijk.

Als we nog eens naar het voorbeeld van de taalherkenning middels letterfrequenties kijken, kunnen we dit zien als een Markov proces waarbij de states de verschillende letters zijn. In dit geval zouden voor elk beginstuk van t letters een kansverdeling voor de daarop volgende letter moeten bepalen. Voor een beginstuk van 8 letters zijn dit bijvoorbeeld $27^8 = 282429536481$ verdelingen, en die kunnen we nog bepalen nog opslaan.

Maar dit voorbeeld wijst ook al een mogelijke oplossing aan: We kunnen ervan uitgaan dat de kansverdeling voor de 9-de letter niet erg verandert als

we de eerste letter q_0 veranderen en waarschijnlijk zal ook de letter op tijdstip $t = 1$ nog geen grote betekenis voor de kansen van de verschillende waarden van q_8 hebben.

Dit leidt tot het idee, de kansverdeling voor de states op tijdstip t te benaderen door de kansverdeling die alleen maar met de k voorafgaande states rekening houdt, d.w.z. we nemen aan dat

$$p(q_t = S_j | q_{t-1} = S_{i_{t-1}}, \ldots, q_0 = S_{i_0}) \approx p(q_t = S_j | q_{t-1} = S_{i_{t-1}}, \ldots, q_{t-k} = S_{i_{t-k}})$$

een voldoende nauwkeurige benadering geeft.

Definitie: Een Markov proces, waarbij de kans op de states op tijdstip t alleen maar van de k voorafgaande states afhangt, heet een *Markov proces van orde k*. Hierbij wordt verondersteld dat de kansen niet van het tijdstip t afhangen, maar alleen maar van de rij voorafgaande states.

Voor een systeem met N mogelijke states wordt een Markov proces van orde k dus beschreven door de N^k kansverdelingen

$$p(q_t = S_j \mid q_{t-1} = S_{i_{t-1}}, \ldots, q_{t-k} = S_{i_{t-k}})$$

waarbij $(S_{i_{t-1}}, \ldots, S_{i_{t-k}})$ over alle mogelijke combinaties van states op de tijdstippen $t-1, \ldots, t-k$ loopt (onafhankelijk van t).

Bij een **Markov proces van orde** 0 speelt de geschiedenis helemaal geen rol, de states worden alleen maar volgens een kansverdeling op de states voortgebracht. Zo'n Markov proces krijgen we bijvoorbeeld, als we (zo als in de laatste les) alleen maar de relatieve frequenties van de letters in een taal bepalen en vervolgens letters volgens deze kansverdeling produceren. De relatieve frequentie van de letters zal dan wel kloppen, maar bijvoorbeeld de relatieve frequenties van paren van letters niet meer. Hiervoor hebben we een Markov proces van orde 1 nodig.

Een **Markov proces van orde** 1 is gekarakteriseerd door de overgangskansen $a_{ij} := p(q_t = S_j \mid q_{t-1} = S_i)$. Omdat we veronderstellen dat deze kansen onafhankelijk van t zijn, kunnen we de kansen in een *overgangsmatrix* $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ invullen. Voor deze overgangsmatrix A geldt dat $a_{ij} \geq 0$ en dat $\sum_{j=1}^{n} a_{ij} = 1$ voor alle $i = 1, \ldots, n$, omdat het systeem vanuit state S_i naar een van de states S_j moet overgaan.

Een handige eigenschap van de overgangsmatrix A is dat we met behulp van de machten van A makkelijk kunnen berekenen wat er over een aantal stappen gebeurt: Het element (i, j) van A^k geeft de kans aan, dat het systeem in (precies) k stappen van state S_i naar S_j gaat.

Een Markov proces van orde 1 laat zich ook overzichtelijk door een *graaf* of *state diagram* representeren: De states zijn punten en de overgangen zijn pijltjes tussen de states, met de kans voor de overgangen als labels.

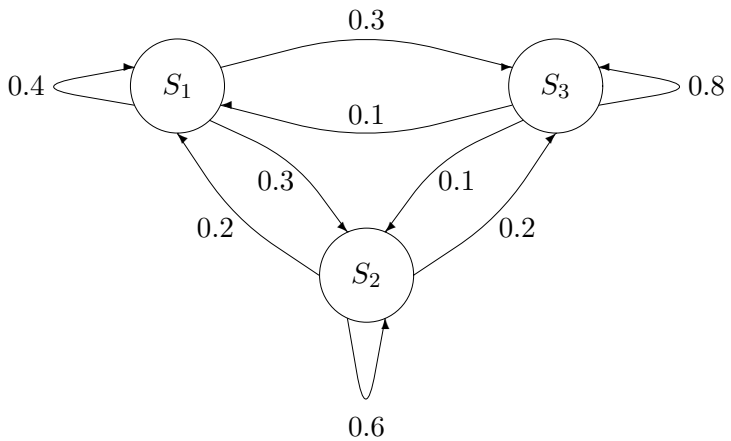
Als we bijvoorbeeld het weer als een (eenvoudige) Markov proces willen beschrijven, zouden we misschien de drie states

$$S_1 = \text{ regen}, \quad S_2 = \text{ bewolkt}, \quad S_3 = \text{ zonnig}$$

kunnen kiezen. Als overgangsmatrix veronderstellen we de (erg optimistische) matrix

$$A = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}.$$

Dan heeft deze Markov proces het state diagram uit Figuur III.3.



Figuur III.3: Markov proces van orde 1 voor het weer.

Aan de hand van de overgangskansen kunnen we een aantal vragen makkelijk beantwoorden:

(1) Wat is de kans op drie dagen zon gevolgd van een dag regen?

(2) Wat is de kans dat het weer precies d dagen hetzelfde blijft?

Bij vraag (1) willen we de kans op de rij $O = S_3S_3S_3S_1$ van states weten. Maar de overgang $S_3 \to S_3$ heeft kans $a_{33} = 0.8$ en de overgang $S_3 \to S_1$ heeft kans $a_{31} = 0.1$, dus is de kans op deze rij $0.8 \cdot 0.8 \cdot 0.1 = 0.064$. Hierbij veronderstellen we wel, dat de vraag op een dag gesteld wordt, waar het al zonnig is, dus waar we al in state S_3 zitten.

Vraag (2) gaat over een rij $O = \underbrace{S_iS_i \ldots S_i}_{d} S_j$ van states, waarbij er precies d keer de state S_i voorkomt en de state S_j verschillend is van S_i. Maar de kans dat we van state S_i naar een state verschillend van S_i gaan is $1 - a_{ii}$, dus is de kans $p(O)$ van deze rij states $p(O) = a_{ii}^{d-1} \cdot (1 - a_{ii})$.

We kunnen nu zelfs de verwachtingswaarde voor het aantal dagen d berekenen, die we in state S_i blijven, er geldt:

$$E[d] = \sum_{d=1}^{\infty} d \cdot a_{ii}^{d-1} \cdot (1 - a_{ii}) = \frac{1}{1 - a_{ii}}.$$

Dit zien we als volgt in: Voor de meetkundige reeks geldt $\sum_{d=0}^{\infty} x^d = \frac{1}{1-x}$ als $|x| < 1$. Maar $\sum_{d=1}^{\infty} dx^{d-1} = (\sum_{d=0}^{\infty} x^d)'$, omdat we in dit geval termsgewijs

mogen afleiden. Aan de andere kant is $(\frac{1}{1-x})' = \frac{1}{(1-x)^2}$, en dus is

$$\sum_{d=1}^{\infty} d \cdot a_{ii}^{d-1} \cdot (1 - a_{ii}) = \frac{1}{(1 - a_{ii})^2}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}.$$

In ons optimistisch model van het weer is de kans dat het blijft regenen $a_{11} = 0.4$, dus is de verwachtingswaarde voor het aantal regendagen achter elkaar gelijk aan $\frac{1}{1-0.4} = \frac{1}{0.6} \approx 1.67$. Net zo krijgen we voor het verwachte aantal bewolkte dagen achter elkaar de waarde $\frac{1}{1-0.6} = 2.5$ en voor het aantal zonnige dagen achter elkaar hebben we de verwachtingswaarde $\frac{1}{1-0.8} = 5$.

12.2 Stochastische automaten

Bij de Markov processen zijn we ervan uitgegaan dat het systeem op tijdstippen $t = 0, 1, \ldots$ van een state naar een andere state overgaat. In sommige samenhangen wordt zo'n overgang veroorzaakt door een input aan het systeem. Maar dan is het plausibel dat de overgangskansen ook van de mogelijke inputs kunnen afhangen. Dit betekent, dat er bij een Markov proces van orde 1 voor elke mogelijke input een aparte overgangsmatrix is. Zo'n systeem noemt men ook een *stochastische automaat*.

We bekijken dit aan de hand van het voorbeeld van een *emotionele robot*. Stel de robot heeft drie mogelijke states, namelijk

$$S_1 = \text{ gelukkig}, \quad S_2 = \text{ bedroefd}, \quad S_3 = \text{ mal}$$

en er zijn de twee mogelijke inputs

$$X = \text{ 'hallo schat'} \quad \text{en} \quad Y = \text{ 'oude roestdoos'}$$

dan horen bij deze twee inputs misschien de overgangsmatrices

$$A_X = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 1.0 & 0 & 0 \\ 0 & 0 & 1.0 \end{pmatrix} \quad \text{en} \quad A_Y = \begin{pmatrix} 0 & 0.9 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1.0 \end{pmatrix}.$$

Een state zo als S_3 waaruit een systeem niet meer kan ontsnappen, heet een *absorberende* state.

Ook voor een stochastische automaat laten zich de kansen over langere perioden door producten van de overgangsmatrices berekenen, als de rij van inputs bekend is. Als de robot bijvoorbeeld op de werkdagen input Y maar op het weekend input X te horen krijgt, zijn de overgangskansen van maandag tot maandag gegeven door

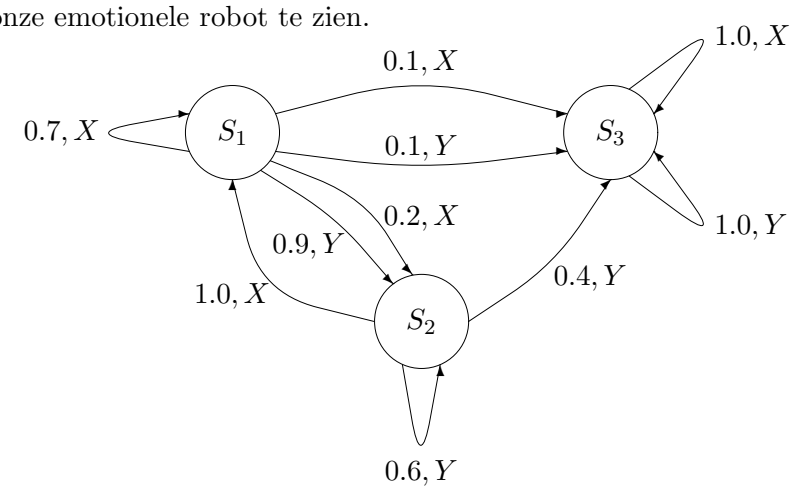
$$A_Y^5 \cdot A_X^2 = \begin{pmatrix} 0.08163 & 0.02332 & 0.8950 \\ 0.05443 & 0.01555 & 0.9300 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}$$

dus is zelfs een gelukkige robot na afloop van een week (en na een eigenlijk opbouwend weekend) met hoge kans mal. Nog erger is het van vrijdag tot vrijdag, dit geeft de overgangskansen

$$A_X^2 \cdot A_Y^5 = \begin{pmatrix} 0.0 & 0.09135 & 0.9086 \\ 0.0 & 0.09718 & 0.9028 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}$$

en alleen maar de robots die op vrijdag middag bedroefd en nog niet mal zijn, krijgen we tot maandag weer opgeknapt.

Een stochastische automaat laat zich analoog met een Markov proces van orde 1 door een state diagram beschrijven, waarbij de labels de input en de kans voor de overgang bij deze input bevatten. In Figuur III.4 is het state diagram voor onze emotionele robot te zien.



Figuur III.4: Stochastische automaat voor een emotionele robot.

Merk op dat in een state diagram voor elke state de som van de kansen op de uitgaande pijltjes voor eenzelfde input gelijk aan 1 moet zijn.

12.3 Markov modellen

We hebben tot nu toe het standpunt ingenomen, dat een rij states volgens de kansverdelingen van een Markov proces voortgebracht wordt. Maar we kunnen de opzet van een Markov proces ook opvatten als een *model* voor een niet verder gespecificeerd mechanisme dat de rij van states voortbrengt. Uit deze perspectief noemt men het stelsel van states en kansverdelingen voor de overgangen tussen de states een *Markov model*. Het idee hier achter is dat een onbekend proces de states produceert, maar dat we veronderstellen dat dit proces zich gedraagt als een Markov proces en de states en overgangskansen dus een model voor het proces zijn.

Om voor een onbekend proces een Markov model te maken, moeten we uit waarnemingen van rijen van states de overgangskansen tussen de states schatten. Hoe dit in zijn werk gaat, bekijken we aan het voorbeeld van de letterfrequenties:

Markov model van **orde 0**:
We hebben alleen maar de kansverdeling van de states nodig, dus de kansverdeling van de letters, en die krijgen we als relatieve frequenties van de letters in een (grote) achtergrond tekst (training tekst).

Markov model van **orde 1**:
We moeten de overgangskansen $a_{ij} := p(q_t = S_j \mid q_{t-1} = S_i)$ bepalen. Maar er geldt voor de voorwaardelijke kans a_{ij} dat

$$p(q_t = S_j \mid q_{t-1} = S_i) = \frac{p(q_t = S_j, q_{t-1} = S_i)}{p(q_{t-1} = S_i)},$$

dus kunnen we de a_{ij} op een training tekst bepalen als quotiënt van de frequentie f_{ij} van letterparen met S_i als eerste letter en de totale frequentie f_i van de letter S_i. Hierbij hoeven we frequenties f_i van de enkele letters niet eens expliciet te bepalen, want er geldt $f_i = \sum_j f_{ij}$ omdat we elk voorkomen van S_i hebben geteld als we alle paren met S_i in de eerste plaats hebben geteld. (Voor de letter op de laatste plaats in de training tekst klopt dit natuurlijk niet, maar deze fout kunnen we verwaarlozen). We krijgen dus de overgangskansen a_{ij} heel makkelijk als

$$a_{ij} = \frac{f_{ij}}{f_i}.$$

Markov model van **orde k** ≥ 2:
In principe passen we hier hetzelfde idee toe als bij een Markov model van orde 1 en berekenen de voorwaardelijke kansen $p(q_t = S_j \mid q_{t-1} = S_{i_k}, \ldots, q_{t-k} = S_{i_1})$ door

$$p(q_t = S_j \mid q_{t-1} = S_{i_k}, \ldots, q_{t-k} = S_{i_1}) = \frac{p(q_t = S_j, q_{t-1} = S_{i_k}, \ldots, q_{t-k} = S_{i_1})}{p(q_{t-1} = S_{i_k}, \ldots, q_{t-k} = S_{i_1})}.$$

De kans in de teller vinden we hierbij als relatieve frequentie van de rij van states $(S_{i_1}, \ldots, S_{i_k}, S_j)$ in alle rijen van $k+1$ states en de kans in de noemer als relatieve frequentie van de rij $(S_{i_1}, \ldots, S_{i_k})$ in alle rijen van k states.

In de aanpak met de relatieve frequenties bestaat er een klein probleem met de zogeheten *zeldzame gebeurtenissen*. Voor een rij van states met een lage kans kan het gebeuren dat deze rij in het training materiaal helemaal niet voorkomt. Maar in het algemeen is het niet verstandig om aan een overgang in het model de kans 0 toe te kennen, omdat dit betekent dat het model deze overgang nooit zou produceren en aan een rij states waarin deze overgang wel voorkomt de kans 0 geeft. De enige uitzondering zijn *verboden overgangen*, d.w.z. overgangen die uit inhoudelijke redenen inderdaad uitgesloten kunnen worden (bijvoorbeeld in een populatie kippen de overgang van een overleden kip tot een vruchtbaar kip).

Een simpele (maar vaak voldoende) oplossing van het probleem van de zeldzame gebeurtenissen is, de teller voor de frequenties van de rijen niet met 0 maar met 1 te initialiseren, dus te veronderstellen dat elk gebeurtenis wel een keer is gezien (zo iets als éénmaal is geen maal). Maar

er zijn ook ingewikkeldere, theoretisch beter onderbouwde oplossingen voor dit probleem bedacht, dit valt onder het begrip van *smoothing* (gladmaken) van kansverdelingen.

We hebben gezien dat Markov processen en Markov modellen in principe twee zijden van eenzelfde munt zijn: Uit waarnemingen van een onbekend proces maken we een Markov model, en we zeggen dat het Markov model het proces goed beschrijft als de Markov proces die bij het Markov model hoort een rij gebeurtenissen produceert dat goed met de waarnemingen overeen komt.

12.4 Toepassingen van Markov modellen

De twee perspectieven om naar Markov modellen te kijken geven ook de meest belangrijke typen van toepassingen: Simulatie met behulp van Markov processen en classificatie (of toetsing) met behulp van Markov modellen.

Simulatie

Hierbij gebruiken we een Markov model om een rij gebeurtenissen voort te brengen, waarop bijvoorbeeld andere modellen getoetst kunnen worden.

Als we bijvoorbeeld voor de rijen van letters in verschillende talen een Markov model van orde 1 bouwen, kunnen we (onzinnige) teksten produceren, die niettemin typische elementen van de taal laten zien.

Voor de talen Nederlands (NL), Engels (EN), Duits (DU) en Fins (FI) krijgen we zo bijvoorbeeld *teksten* als de volgende (met 160 letters):

```
NL EVEFOOE OVORER KET DESTS NDEFT MELL CEN HEN ET MEDE ENIJFEBE
   HEPGE G IN JEN VOONDEDE HE ESTETETE DE HE DER COROPEETLL
   NFFTE LENG MHT VOT HET EUDE DERANLODENGEMH

EN S COTHENN CHENCTHER BEN THXS INTHABJ IT EUPAUS ISTHANTEN
   CIOPE WAGESON IN M CONA ATHEDEDED AN JON DERENN T RTH THEPLE
   UES PTAD TIONTHAT ERO OR FFION TTUNEROCTHE

DU RELEMM FT DELLATIT APTZERKO TUN ASER WOPF KPEH RARINTOKEN IG
   W MT BURER MENGS URHEM ZICHAAT KAHED URIIENSP ENTEN ERT
   ZUNAUN SIONG D SE VERZUR HUMAN TSER DIE ASC

FI VA EN MMA LLLEN LILIOD TOS IHTORON ATUN MISA VUN KA
   OROLUSAMUJA POKUNITUSIM M DOSTOTA HAITTANEMINTISISON URECD
   KOMI HTI KUOHOONTOULI T OUUUSKARIS OP SSEHJOITAVU
```

Deze teksten laten verschillende elementen zien, die typisch voor de talen zijn, zo als de dubbele `OO` en de `IJ` in het Nederlands of de `TH` in het Engels.

Als men hier in plaats van een Markov model van orde 1 een Markov model van orde 2 toepast, dus de relatieve frequenties van tripels van letters telt, worden de verschillen nog veel duidelijker. Merk op dat er $27^3 = 19683$ verschillende tripels van letters zijn, om hiervoor een redelijke kansverdeling te krijgen, zou men een training tekst van een paar miljoen letters nodig hebben (voor de

$27^2 = 729$ paren van letters zijn training teksten van slechts ongeveer 50000 letters gebruikt). Maar natuurlijk zijn zo grote teksten voor alle soorten van talen beschikbaar, en als een Markov model van orde 2 een tekst als

```
IBUS CENT IPITIA IPSE CUM VIVIVS SE ACETITI DEDENTUR
```

produceert, zouden we er snel achter komen, dat het model op Latijnse teksten getraind is.

Classificatie/Toetsing

We veronderstellen dat we bij een classificatie taak in de patroonherkenning voor elke klasse van patronen een Markov model gebouwd hebben dat de elementen van de klasse goed beschrijft.

Classificatie principe: Voor een gegeven rij waarnemingen wordt voor iedere klasse van patronen berekend met welke kans de waarneming door het Markov model van deze klasse voortgebracht wordt. Het patroon wordt aan de klasse toegewezen, waarvoor deze kans maximaal is.

Dit principe laat zich als volgt onderbouwen:

Bij een Markov proces van orde 0 met N mogelijke uitkomsten $S_1, \ldots, S_N$ kunnen we de kans op de rij $x_1 x_2 \ldots x_n$ van uitkomsten eenvoudig berekenen door

$$p(x_1 x_2 \ldots x_n) = p(x_1) \cdot p(x_2) \cdot \ldots \cdot p(x_n).$$

We bekijken nu de stochast X van een Markov proces met *echte* kansen $p_i = p(X = S_i)$ en beschrijven deze door een Markov model met (geschatte) kansen $q_i = p'(q_t = S_i)$.

Als n groot is, is het aantal van uitkomsten x_i in de rij ongeveer gelijk aan $n \cdot p_i$. Dan krijgen we

$$p(x_1 x_2 \ldots x_n) = \prod_{i=1}^{N} p(X = S_i)^{n \cdot p_i} = \prod_{i=1}^{N} p_i^{n \cdot p_i}$$

voor de juiste kansen en

$$p'(x_1 x_2 \ldots x_n) = \prod_{i=1}^{N} p'(q_t = S_i)^{n \cdot p_i} = \prod_{i=1}^{N} q_i^{n \cdot p_i}$$

voor de kansen volgens het model.

Om rijen van verschillende lengtes te kunnen vergelijken moeten we hieruit nog de n-de machtswortel trekken, dit geeft

$$p(x_1 x_2 \ldots x_n)^{\frac{1}{n}} = \prod_{i=1}^{N} p_i^{p_i} \quad \text{tegenover} \quad p'(x_1 x_2 \ldots x_n)^{\frac{1}{n}} = \prod_{i=1}^{N} q_i^{p_i}.$$

Als we van deze vergelijkingen de logaritme (met basis 2) nemen, krijgen we een verband met een oude bekende uit de laatste les, namelijk de entropie:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{N} p_i \ {}^2\log(p_i) = -\frac{1}{n} \ {}^2\log(p(x_1x_2\ldots x_n)) \\ &\leq -\sum_{i=1}^{N} p_i \ {}^2\log(q_i) = -\frac{1}{n} \ {}^2\log(p'(x_1x_2\ldots x_n)) =: H. \end{aligned}$$

In de limiet $n \to \infty$ geeft dus $H := -\frac{1}{n} \ {}^2\log(p'(x_1x_2\ldots x_n))$ een schatting voor de entropie $H(X)$ van de kansverdeling van de stochast X en deze schatting is beter naarmate H een lagere waarde heeft, want we weten dat het minimum bereikt wordt als Q de juiste kansverdeling van X is.

Wat we net hebben gezien, laat zich op algemene Markov processen veralgemenen, voor de entropie van een stochast X geldt:

$$H(X) = \lim_{n\to\infty} -\frac{1}{n} \ {}^2\log(p(x_1x_2\ldots x_n))$$

waarbij $p(x_1x_2\ldots x_n)$ de juiste kansverdeling voor de stochast X aangeeft. Als we nu dezelfde kans met de kansen uit een Markov model berekenen, wordt deze kans hoger als het Markov model de stochast beter beschrijft, want voor een hogere kans $p'(x_1x_2\ldots x_n)$ is $-\frac{1}{n} \ {}^2\log(p'(x_1x_2\ldots x_n))$ kleiner en ligt dus dichter bij $H(X)$.

Een andere manier om tot dezelfde conclusie te komen berust op de interpretatie van $2^{H(X)}$ als het gemiddelde aantal alternatieven dat men voor de stochast X verwacht:

Uit de vorige les weten we dat een stochast X met entropie $H(X)$ net zo moeilijk is als een uniforme verdeling met $2^{H(X)}$ alternatieven. Maar we weten dat voor $H = -\frac{1}{n} \ {}^2\log(p'(x_1x_2\ldots x_n))$ steeds geldt dat $H \geq H(X)$ en dus ook $2^H \geq 2^{H(X)}$. We kunnen dus zeggen, dat de beschrijving van de stochast X door het Markov model met kansverdeling Q net zo moeilijk is als een uniforme verdeling met 2^H alternatieven, en natuurlijk is degene beschrijving het beste waarvoor 2^H minimaal is.

We passen dit idee nu op korte testteksten toe, waarvoor we de taal willen bepalen. We nemen aan dat we een Markov model van orde 1 hebben met overgangskansen a_{ij} van state S_i naar state S_j en met kans $b_i = \sum_{j=1}^{N} a_{ij}$ voor state S_i. Met zo'n model berekenen we de kans van een rij $x_1x_2\ldots x_n$ van letters door

$$\begin{aligned} &p(x_1x_2\ldots x_n) \\ &\qquad = p(q_1 = S_{i_1}) \cdot p(q_2 = S_{i_2} \mid q_1 = S_{i_1}) \cdot \ldots \cdot p(q_n = S_{i_n} \mid q_{n-1} = S_{i_{n-1}}) \\ &\qquad = b_{i_1} \cdot \prod_{j=1}^{n-1} a_{i_j i_{j+1}} \end{aligned}$$

waarbij S_{i_j} de state van de letter x_j is.

Voorbeeld: We onderzoeken verschillende stukken tekst in de talen Nederlands (NL), Engels (EN), Duits (DU) en Fins (FI) met Markov modellen voor deze talen en berekenen voor elke combinatie van tekst en Markov model de waarde

$$2^H \text{ voor } H = -\frac{1}{n}\ {}^2\log(p(x_1x_2\ldots x_n))$$

waarbij de kansen zo als net aangegeven berekend worden.

De testteksten zijn:

T_1 : `SINTERKLAAS KOMT NAAR ONS HUIS`

T_2 : `SANTA CLAUS COMES TO OUR HOUSE`

T_3 : `NIKOLAUS KOMMT IN UNSER HAUS`

T_4 : `HANNU MANNINEN`

Als resultaat krijgen we de volgende tabel met de waarden van 2^H:

	NL	EN	DU	FI
T_1	14.1	28.3	16.2	19.0
T_2	18.2	12.4	28.9	18.0
T_3	14.4	23.5	9.8	16.5
T_4	19.2	25.0	16.8	14.0

Het is duidelijk dat we in elk geval de juiste taal kunnen achterhalen. Hoe typisch de testteksten voor de enkele talen zijn, kunnen we zien als we de boven gevonden waarden met de waarden op de teksten vergelijken waarop de Markov modellen getraind zijn, dus met de entropieën van de Markov modellen zelfs. De waarden van $2^{H(X)}$ voor de verschillende talen zijn:

NL: 9.2 EN: 9.6 DU: 9.3 FI: 9.7.

De classificatie met behulp van Markov modellen voor letter strings in de verschillende talen is de manier hoe in tekstverwerkingsprogramma's als Word (Office) automatisch de spellchecker naar een andere taal omgeschakeld wordt, als er bijvoorbeeld in een Nederlandstalige tekst een citaat in het Engels ingebouwd wordt.

12.5 Markov modellen met verborgen states

Tot nu toe hebben we steeds naar systemen gekeken, waarvoor we de states direct konden waarnemen. We hebben daarom ook geen onderscheiding gemaakt tussen states, uitkomsten en waarnemingen. We krijgen echter een grotere vrijheid in de Markov modellen, als we de states los van de gebeurtenissen en waarnemingen bekijken. Het idee is, dat de states de mogelijke uitkomsten wel veroorzaken, maar dat verschillende states dezelfde uitkomst kunnen produceren en dat niet (noodzakelijk) bekend is, welke state een bepaalde uitkomst heeft veroorzaakt. Om deze reden noemen we de states ook *verborgen* en een

Markov model met verborgen states heet een *Hidden Markov model*, of in het kort een *HMM*.

We geven twee opzetten die het idee van de Hidden Markov modellen illustreren:

- **Het munt model**

  Achter een gordijn zit iemand die met een aantal mogelijk geladen (dus niet noodzakelijk eerlijke) munten een muntworp experiment uitvoert, maar alleen maar de rij uitkomsten (kop/munt) aan de waarnemer doorgeeft. De keuze van de munten voor de enkele worpen volgt een stochastisch proces die door overgangskansen tussen de munten bepaald wordt.

- **Het vaas model**

  Er zijn N vazen met telkens ballen van M verschillende kleuren, waarbij de aantallen van ballen met een zekere kleur per vaas mogen verschillen en ook het totale aantal ballen per vaas niet hetzelfde hoeft te zijn. Iemand trekt (met terugleggen) een bal uit een van de vazen en geeft de kleur van de bal aan de waarnemer door. Vervolgens wordt volgens een toevalskeuze, die afhankelijk van de laatst gekozen vaas is, een nieuwe vaas gekozen.

De algemene ingrediënten van een HMM (van orde 1) zijn als volgt:

(1) Mogelijke uitkomsten $x_1, \ldots, x_M$. De waargenomen uitkomst op tijdstip t word met o_t aangegeven (de letter o staat voor het Engelse *observation*).

(2) Een aantal states $S_1, \ldots, S_N$, waarbij de state op tijdstip t met q_t aangegeven wordt.

(3) De overgangskansen $a_{ij} := p(q_t = S_j \mid q_{t-1} = S_i)$ voor de overgang van state S_i naar state S_j.

(3) Voor elke state S_i een kansverdeling b_i voor de *emissiekansen*, d.w.z. $b_i(x_k) = p(o_t = x_k \mid q_t = S_i)$ is de kans dat in state S_i de uitkomst x_k geproduceerd wordt. Er wordt veronderstelt dat deze kansen onafhankelijk van het tijdstip t zijn.

(4) Een beginverdeling π die de kansen $\pi(i) := p(q_0 = S_i)$ aangeeft dat het systeem op tijdstip $t = 0$ in state S_i is.

Ook een gewoon Markov model laat zich (op een iets kunstmatige manier) als HMM opvatten: Hiervoor worden de states S_i identiek met de uitkomsten x_i gekozen en de emissiekansen b_i worden gedefinieerd door

$$b_i(x_i) = 1 \text{ en } b_i(x_k) = 0 \text{ voor } k \neq i.$$

Voorbeeld van een HMM

We bekijken een munt model met drie munten als states, waarvan de eerste eerlijk is, dus kansen $\frac{1}{2}$ voor *kop* en *munt* heeft, de tweede oneerlijk met kans $\frac{3}{4}$ op

kop en de derde oneerlijk met kans $\frac{1}{4}$ op *kop*. Als we K voor de uitkomst *kop* en M voor de uitkomst *munt* schrijven, hebben we de emissiekansen $b_1(\mathtt{K}) = b_1(\mathtt{M}) = \frac{1}{2}$, $b_2(\mathtt{K}) = b_3(\mathtt{M}) = \frac{3}{4}$, $b_3(\mathtt{K}) = b_2(\mathtt{M}) = \frac{1}{4}$, die door de volgend tabel weergegeven worden:

	$b_i(\mathtt{K})$	$b_i(\mathtt{M})$
S_1	0.5	0.5
S_2	0.75	0.25
S_3	0.25	0.75

We veronderstellen verder dat de beginverdeling uniform is, d.w.z. de kans dat het systeem in het begin in state S_i is, is voor elke state $\pi(i) = \frac{1}{3}$.

Stel we nemen de rij $O = \mathtt{KMKMK}$ waar.

In een eerste opzet nemen we aan dat alle overgangskansen hetzelfde zijn, dus alle $a_{ij} = \frac{1}{3}$.

Omdat de hoogste kans op de uitkomst K in state S_2 zit, de hoogste kans op de uitkomst M in S_3 en de overgangskansen alle hetzelfde zijn, kunnen we makkelijk zien dat de rij $q = S_2S_3S_2S_3S_2$ de rij van states is, waarvoor de kans op de waarneming O maximaal is. In dit geval is deze kans namelijk $p(O, q) = (\frac{1}{3})^5 \cdot (\frac{3}{4})^5 = (\frac{1}{4})^5 \approx 9.77 \cdot 10^{-4}$.

In tegenstelling hiermee is de kans dat deze waarneming door de rij $q' = S_1S_1S_1S_1S_1$ voortgebracht is, slechts $p(O, q') = (\frac{1}{3})^5 \cdot (\frac{1}{2})^5 = (\frac{1}{6})^5 \approx 1.29 \cdot 10^{-4}$. Deze kans is om een factor $(\frac{3}{2})^5 \approx 7.6$ kleiner dan voor de eerdere rij q van states.

Het probleem wordt iets ingewikkelder als de overgangskansen niet meer alle hetzelfde zijn. Stel we hebben de volgende matrix $A = (a_{ij})$ van overgangskansen a_{ij} tussen de states:

$$A = (a_{ij}) := \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.45 & 0.1 & 0.45 \\ 0.45 & 0.45 & 0.1 \end{pmatrix}$$

dan is de kans $p(O, q \mid A)$ (we geven hier voor de duidelijkheid de matrix van overgangskansen mee aan) voor dezelfde rijen q en q' van states als boven gegeven door

$$p(O, q \mid A) = \frac{1}{3} \cdot 0.45^4 \cdot (\frac{3}{4})^5 \approx 3.24 \cdot 10^{-3},$$
$$p(O, q' \mid A) = \frac{1}{3} \cdot 0.9^4 \cdot (\frac{1}{2})^5 \approx 6.83 \cdot 10^{-3},$$

dus is deze keer $p(O, q' \mid A)$ om een factor $2^4(\frac{2}{3})^5 \approx 2.1$ *groter* dan $p(O, q \mid A)$.

We zien dus dat in het tweede geval de hypothese dat het systeem door de rij q' van states gelopen is, een hogere kans voor de waarneming geeft dan de rij q van states.

Het is nu natuurlijk een voor de hand liggende vraag, of er een verdere rij q'' van states is, die een nog hogere kans voor de rij O van waarnemingen oplevert.

Voor korte rijen kunnen we dit met brute kracht nog wel achterhalen (voor het voorbeeld met 5 waarnemingen en 3 states zijn er $3^5 = 243$ mogelijkheden voor de rij q van states), maar voor langere rijen is dit ondoenlijk.

In het speciaal geval van het voorbeeld is de rij q' inderdaad optimaal, omdat de overgangskans $a_{11} = 0.9$ minstens twee keer groter is dan alle andere overgangskansen en de emissiekansen $b_1(\mathtt{K}) = b_1(\mathtt{M}) = \frac{1}{2}$ zijn. Maar zo'n soort redenering zal in de praktijk natuurlijk nooit werken, omdat de modellen veel ingewikkelder en onoverzichtelijker zijn.

We zitten dus met de vraag hoe we bij een rij waarnemingen de rij states vinden, die de hoogste kans aan de waarnemingen geeft. Dit is één van drie fundamentele problemen in het kader van Hidden Markov modellen die we in de volgende les gaan bespreken.

Belangrijke begrippen in deze les

- Markov processen
- overgangsmatrix
- state diagram
- stochastische automaat
- Markov model
- Hidden Markov model (HMM)

Opgaven

94. In een communicatie systeem worden bits als 0 of 1 over een aantal stappen doorgegeven, waarbij in iedere stap een bit met kans 0.8 correct blijft.

    (i) Beschrijf het communicatie systeem als een Markov proces en geef het state diagram van het proces aan.

    (ii) Bepaal de kans dat een bit met de waarde 0 na vier stappen als 0 ontvangen wordt.

95. De oogst van appels in Tasmanië wordt als *geweldig*, *middelmatig* of *slecht* geclassificeerd. Na een geweldig jaar zijn de kansen voor het volgende jaar 0.5, 0.3, 0.2 voor een geweldige, middelmatige of slechte oogst. Na een middelmatig jaar zijn de kansen voor het volgende jaar 0.2, 0.5, 0.3 en na een slecht jaar zijn de kansen 0.2, 0.2, 0.6 voor een geweldige, middelmatige of slechte oogst.

    (i) Beschrijf de ontwikkeling van de appel oogst door een Markov proces en geef het state diagram van het proces aan.

    (ii) Stel de kansen om met een geweldig, middelmatig of slecht jaar te beginnen zijn 0.2, 0.5 en 0.3. Wat zijn de kansverdelingen voor de kwaliteit van de oogst na 1 jaar, 3 jaren en 5 jaren?

(iii) Kan je de kansverdeling voor de kwaliteit van de oogst bepalen, die op lange termijn bereikt wordt?

96. Een Markov proces heet *irreducibel* als elke state in eindig veel stappen vanuit elke andere state bereikbaar is. Laat zien dat de Markov processen met overgangsmatrices

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{en} \quad B = \begin{pmatrix} 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

irreducibel zijn.

97. We bekijken de emotionele robot uit sectie 12.2 en bepalen de kansverdeling voor zijn toestand na twee inputs.

(i) Veronderstel dat de robot in het begin gelukkig is en bereken de kansverderdeling voor elk van de vier mogelijke inputs XX, XY, YX en YY.

(ii) Bereken de kansverdelingen voor de verschillende inputs ook voor de gevallen dat de robot in het begin bedroefd of mal was.

98. De states S_1, S_2, S_3 van een Hidden Markov model zijn (net als in het voorbeeld) drie munten die de emissiekansen $\frac{1}{2}, \frac{3}{4}, \frac{1}{4}$ op kop (`K`) en de emissiekansen $\frac{1}{2}, \frac{1}{4}, \frac{3}{4}$ op munt (`M`) hebben. De beginverdeling van de states is uniform, dus $\pi(1) = \pi(2) = \pi(3) = \frac{1}{3}$. We bekijken de drie rijen waarnemingen $O_1 = \texttt{KKKK}$, $O_2 = \texttt{KKKM}$, $O_3 = \texttt{KKMM}$.

(i) Veronderstel dat alle overgangskansen hetzelfde zijn, dus gelijk aan $\frac{1}{3}$. Bepaal de rijen q_1, q_2, q_3 van states, waarvoor de kans dat zij de waarnemingen O_1, O_2, O_3 geproduceerd hebben maximaal is. Bereken voor de gevonden rijen van states de kansen $p(O_1, q_1)$, $p(O_2, q_2)$, $p(O_3, q_3)$.

(ii) Vergelijk de kansen uit (i) met de kansen $p(O_i, q)$ die men krijgt, als men aanneemt dat altijd de eerlijke munt geworpen wordt, dus als $q = S_1S_1S_1S_1$ is.

(iii) Veronderstel nu dat de overgangskansen niet uniform zijn, maar gegeven door de matrix

$$A = (a_{ij}) := \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.2 \end{pmatrix}.$$

Bereken de kansen $p(O_i, q_i \mid A)$ voor de rijen van states uit deel (i) en de kansen $p(O_i, q \mid A)$ voor de rij q van states uit deel (ii) met betrekking tot deze overgangskansen.

(iv) Probeer in deel (iii) de rijen q'_1, q'_2, q'_3 van states te vinden, zo dat $p(O_i, q'_i \mid A)$ maximaal wordt.

Les 13 Hidden Markov modellen

In deze les zullen we nader op Hidden Markov modellen ingaan, in het bijzonder op de technieken en algoritmen die bij het omgaan met dit soort modellen belangrijk zijn. Om de notaties helder te hebben, spreken we nu af dat we een Hidden Markov model als volgt beschrijven:

Een Hidden Markov model (vanaf nu afgekort als HMM) λ is gegeven door $\lambda = \lambda(\mathcal{S}, \mathcal{X}, A, B, \pi)$, waarbij de parameters de volgende betekenis hebben:

- $\mathcal{S} = \{S_1, \ldots, S_N\}$ is een verzameling van states;
- $\mathcal{X} = \{x_1, \ldots, x_M\}$ is een verzameling van uitkomsten, die door de states geproduceerd worden;
- $A = (a_{ij})$ is de matrix van overgangskansen tussen de states, d.w.z. $a_{ij} = p(q_t = S_j \mid q_{t-1} = S_i)$ is de kans voor de overgang van state S_i naar state S_j (onafhankelijk van het tijdstip t);
- $B = b_i(k)$ is de matrix van emissiekansen voor de gebeurtenissen vanuit de states, d.w.z. $b_i(k) = p(o_t = x_k \mid q_t = S_i)$ is de kans dat de state S_i de uitkomst x_k produceert (onafhankelijk van t);
- $\pi = (\pi(1), \ldots, \pi(N))$ is de beginverdeling van de states.

Vaak behoren de states en de gebeurtenissen tot de algemene opzet van een probleem, in dit geval staan alleen maar de verschillende kansverdelingen ter discussie. In zo'n geval wordt een HMM iets korter door $\lambda = \lambda(A, B, \pi)$ beschreven.

Er zijn in feite drie fundamentele vragen, waarmee we ons moeten bemoeien:

(1) Gegeven een rij $O = o_1 o_2 \ldots o_T$ van waarnemingen en een HMM $\lambda = \lambda(A, B, \pi)$, hoe vinden we de kans $p(O \mid \lambda)$ op deze waarnemingen, gegeven het model λ? Deze kans kan men ook interpreteren als maat, hoe goed het model bij de waarnemingen past.

(2) Gegeven een rij $O = o_1 o_2 \ldots o_T$ van waarnemingen en een HMM $\lambda = \lambda(A, B, \pi)$, hoe vinden we de rij $q = q_1 q_2 \ldots q_T$ van states die de rij waarnemingen het beste kan verklaren?

(3) Hoe kunnen we de parameters van het HMM $\lambda = (A, B, \pi)$ zo aanpassen dat $p(O \mid \lambda)$ voor een (vaste) rij O van waarnemingen maximaal wordt?

De eerste vraag gaat over het evalueren van een gegeven model op een rij waarnemingen, de tweede over het onthullen van de verborgen states en de derde over het vinden van de parameters van een HMM, zo dat het model goed bij een gegeven rij waarnemingen past. Het laatste noemt men ook het *training* van een HMM. We zullen deze vragen nu apart bekijken.

13.1 Evalueren met behulp van een HMM

Stel we hebben een rij waarnemingen $O = o_1 o_2 \dots o_T$ en een HMM $\lambda = \lambda(A, B, \pi)$ en we willen de kans $p(O \mid \lambda)$ op de rij waarnemingen, gegeven het model, berekenen. Een typische situatie waar men dit probleem tegen komt is de classificatie van de waarneming O. Stel dat verschillende klassen $C_1, \dots, C_r$ door verschillende HMM's $\lambda_1, \dots, \lambda_r$ gekarakteriseerd zijn, dan is het een voor de hand liggende idee de waarneming O aan degene klasse C_k toe te wijzen, waarvoor $p(O \mid \lambda_k)$ maximaal is. Deze aanpak noemt men ook de *maximum likelihood* methode.

Om de kans $p(O \mid \lambda)$ te berekenen moeten we in principe voor elke rij $q = q_1 q_2 \dots q_T \in \mathcal{S}^T$ van states de kans $p(O, q \mid \lambda)$ berekenen en deze kansen voor alle mogelijke rijen q van states bij elkaar optellen. Volgens de definitie van de voorwaardelijke kans geldt

$$p(O, q \mid \lambda) = p(O \mid q, \lambda) \cdot p(q \mid \lambda)$$

en dus

$$p(O \mid \lambda) = \sum_{q \in \mathcal{S}^T} p(O, q \mid \lambda) = \sum_{q \in \mathcal{S}^T} p(O \mid q, \lambda)\, p(q \mid \lambda).$$

Met behulp van de laatste uitdrukking kunnen we de kans $p(O \mid \lambda)$ inderdaad uitrekenen: Aan de ene kant is $p(q \mid \lambda)$ juist het product van de kansen voor de overgangen tussen de states in de rij $q = q_1 q_2 \dots q_T$, dus

$$p(q \mid \lambda) = \pi(q_1) \cdot \prod_{t=1}^{T-1} a_{q_t q_{t+1}}.$$

Aan de andere kant is voor een gegeven rij van states de kans $p(O \mid q, \lambda)$ het product van de emissiekansen van de enkele states, dus

$$p(O \mid q, \lambda) = \prod_{t=1}^{T} b_{q_t}(o_t).$$

Bij elkaar genomen krijgen we zo:

$$\begin{aligned} p(O \mid \lambda) &= \sum_{q = q_1 \dots q_T} \pi(q_1) b_{q_1}(o_1) \prod_{t=1}^{T-1} a_{q_t q_{t+1}} b_{q_{t+1}}(o_{t+1}) \\ &= \sum_{q = q_1 \dots q_T} \pi(q_1) b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b_{q_T}(o_T). \end{aligned}$$

Het probleem hierbij is, dat we voor een rij van lengte T over N^T mogelijke rijen van states moeten lopen, en dit is al voor kleine waarden van T (bijvoorbeeld $T = 100$) ondoenlijk.

Gelukkig kunnen we het vermijden over alle mogelijke rijen van states te lopen. Bij de brute kracht methode zouden we erg veel dingen herhaaldelijk uitrekenen, namelijk de beginstukken van de rijen waarvoor de eerste t states hetzelfde zijn. Het idee is, de kansen voor de beginstukken stapsgewijs te

berekenen en deze te recyclen. Als we namelijk de kans voor het beginstuk $o_1 o_2 \ldots o_t$ al kennen, zijn er maar N mogelijkheden voor de state waarin het systeem op tijdstip t zit, en voor de voortzetting naar o_{t+1} hoeven we alleen maar de overgangen van deze N mogelijkheden naar de N mogelijke states op tijdstip $t+1$ te berekenen. Zo krijgen we slechts $T \cdot N^2$ waarden, die we moeten berekenen. De procedure die we zo net hebben geschetst is zo belangrijk dat ze een eigen naam heeft (ook al is die niet erg karakteristiek), ze heet *forward algoritme*.

Forward algoritme

We willen voor $O = o_1 o_2 \ldots o_T$ de kans $p(O \mid \lambda)$ berekenen. Hiervoor definiëren we de *vooruitkans*

$$\alpha_t(i) := p(o_1 o_2 \ldots o_t, q_t = S_i \mid \lambda),$$

die de kans aangeeft dat het systeem op tijdstip t in state S_i is en tot dit tijdstip de waarnemingen $o_1, \ldots, o_t$ heeft geproduceerd.

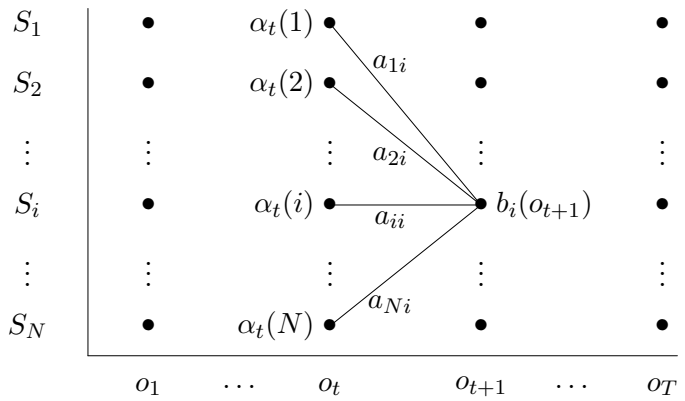
Voor $t = 1$ laten zich de vooruitkansen $\alpha_1(i)$ heel eenvoudig berekenen, er geldt

$$\alpha_1(i) = \pi(i) b_i(o_1).$$

Als we nu van tijdstip t naar tijdstip $t+1$ willen, moeten we over alle N states waarin het systeem op tijdstip t kan zijn lopen en de kans op de overgang naar de verschillende states op tijdstip $t+1$ en de emissie van waarneming o_{t+1} berekenen. Dit geeft de recursie formule:

$$\alpha_{t+1}(i) = \left(\sum_{k=1}^{N} \alpha_t(k) a_{ki} \right) b_i(o_{t+1}).$$

In Figuur III.5 is de berekening van $\alpha_{t+1}(i)$ in een schema aangegeven: De kansen $\alpha_t(1), \ldots, \alpha_t(N)$ van de voorafgaande stap worden met de overgangskansen $a_{1i}, \ldots, a_{Ni}$ en de emissiekans $b_i(o_{t+1})$ gecombineerd tot de kans $\alpha_{t+1}(i)$.



Figuur III.5: Berekening van $\alpha_{t+1}(i)$ in het *forward algoritme*.

Als we de vooruitkansen $\alpha_t(i)$ voor $t = 1, 2, \ldots, T$ berekenen, hoeven we in de laatste stap alleen maar nog de kansen voor de N verschillende states op

tijdstip $t = T$ op te tellen want omdat het systeem in een van de states moet zijn, geeft dit juist de kans op de volledige rij waarnemingen aan. Op deze manier krijgen we

$$p(o_1 o_2 \ldots o_T \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Backward algoritme

Het zou geen verrassing zijn dat er behalve van een forward algoritme ook een *backward algoritme* bestaat, waarbij de kansen op een deel van de waarnemingen van het einde af berekend worden. Men definieert de *achteruitkans* $\beta_t(i)$ als de voorwaardelijke kans

$$\beta_t(i) := p(o_{t+1} \ldots o_T \mid q_t = i, \lambda)$$

op de laatste $T - t$ waarnemingen $o_{t+1}, \ldots, o_T$, gegeven het feit dat het systeem op tijdstip t in state S_i was.

In dit geval heeft men de initialisering $\beta_T(i) = 1$ want we veronderstellen dat het systeem op tijdstip T in state S_i is.

Om van het tijdstip $t + 1$ naar t terug te komen, moeten we over alle states lopen waarin het systeem op tijdstip $t + 1$ kan zijn en de overgangen en de emissie van de waarneming o_{t+1} vanuit deze states berekenen. Dit geeft de recursie

$$\beta_t(i) = \sum_{k=1}^{N} a_{ik} b_k(o_{t+1}) \beta_{t+1}(k).$$

Door deze recursie voor $t = T - 1, \ldots, 2, 1$ te doorlopen, krijgen we uiteindelijk de kans $p(O \mid \lambda)$ door

$$p(O \mid \lambda) = \sum_{i=1}^{N} \pi(i) b_i(o_1) \beta_1(i).$$

We zullen de vooruitkansen $\alpha_t(i)$ en de achteruitkansen $\beta_t(i)$ later in deze les nog eens tegenkomen. Door een slimme combinatie van de $\alpha_t(i)$ en $\beta_t(i)$ laten zich namelijk de parameters van een HMM zo verbeteren dat het systeem een hogere kans voor een gegeven rij waarnemingen oplevert. Op deze manier wordt het HMM beter aan de waarnemingen aangepast, dus getraind.

> De combinatie van vooruit- en achteruitkansen speelt ook bij problemen een rol, waar snel een kandidaat voor een rij states met hoge kans gevonden moet worden. Het idee is, tegelijkertijd aan het begin en aan het eind te beginnen, tot dat $\alpha_t(i)$ en $\beta_t(i)$ in het midden op elkaar stoten. Daarbij worden alleen maar de meestbelovende trajecten meegenomen, d.w.z. de states op tijdstip t die de hoogste vooruit- en achteruitkansen hebben. Deze manier om de *zoekruimte* snel tot de interessante states te beperken staat bekend onder de naam *beam-search*.

13.2 States onthullen

Vaak is het niet genoeg de kans voor een rij waarnemingen, gegeven een HMM, te bepalen, men wil ook een rij states bepalen die bij de waarnemingen past. Maar omdat er verschillende rijen states zijn, die een rij waarnemingen kunnen produceren, moet men hier een criterium hebben, welke states het beste passen. Voor dat we erover na kunnen denken hoe we een optimale rij states kunnen vinden, moeten we dus eerst definiëren, wat we met de *optimale rij states* bij een rij waarnemingen überhaupt bedoelen,

Helaas is er geen *juiste* manier, om een optimaliteitscriterium te definiëren, en afhankelijk van het probleem worden ook verschillende criteria gehanteerd.

Een mogelijkheid is bijvoorbeeld, op elke tijdstip t de state $q_t = S_i$ te kiezen die op dit tijdstip optimaal is. Dat wil zeggen we kiezen q_t zo dat $p(O, q_t = S_i \mid \lambda)$ maximaal wordt. Merk op dat we dit met behulp van de vooruit- en achteruitkansen keurig kunnen formuleren, er geldt namelijk dat

$$p(O, q_t = S_i \mid \lambda) = \alpha_t(i)\beta_t(i)$$

en we hoeven dus voor q_t alleen maar de state S_i te kiezen waarvoor $\alpha_t(i)\beta_t(i)$ maximaal wordt.

Soms willen (of kunnen) we voor de state q_t op tijdstip t alleen maar de waarnemingen $o_1 \dots o_t$ tot op dit tijdstip gebruiken, bijvoorbeeld in een *real-time* systeem. In dit geval zouden we de state $q_t = S_i$ zo kunnen kiezen, dat $p(o_1 o_2 \dots o_t, q_t = S_i \mid \lambda)$ maximaal wordt. Maar dit betekent, dat we voor q_t de state S_i kiezen, waarvoor $\alpha_t(i)$ maximaal is, want dit is precies de definitie van de vooruitkans.

Een probleem bij de genoemde criteria is, dat de overgangen tussen de states enigszins buiten beschouwing blijven, en we zo misschien zelfs een rij van states krijgen die een *verboden* overgang bevat, dus een overgang met kans 0.

Het meest gebruikte criterium dat dit probleem voorkomt is, de optimale rij q_{opt} van states te definiëren als de rij waarvoor de gemeenschappelijke kans over de hele rij states en waarnemingen maximaal is.

Criterium: We noemen een rij $q_{opt} = q_1 q_2 \dots q_T \in \mathcal{S}^T$ van states *optimaal* voor de waarneming $O = o_1 o_2 \dots o_T$ als

$$p(O, q_{opt} \mid \lambda) \geq p(O, q \mid \lambda) \text{ voor alle } q \in \mathcal{S}^T.$$

We staan nu weer voor het probleem dat we in principe de kans $p(O, q \mid \lambda)$ voor alle rijen q van states moeten berekenen. Anders als bij het berekenen van de kans voor de waarneming mogen we nu niet alle mogelijkheden om tot een tussenpunt te komen bij elkaar optellen, dus helpen de vooruitkansen $\alpha_t(i)$ hier niet verder.

Maar een kleine variatie van het forward algoritme geeft ook hier een oplossing, waarbij we niet alle N^T mogelijke rijen moeten bekijken. Het idee wat hier achter zit komt uit het *dynamische programmeren* en is een bijna vanzelfsprekende opmerking, maar is wel zo fundamenteel, dat het de naam *Bellman's principe* draagt.

Bellman's principe

We bekijken een iets algemenere situatie die uit het dynamische programmeren ontleend is: Stel we hebben een rooster met punten (i, j) voor $0 \leq i \leq N$, $0 \leq j \leq M$, en we zijn op zoek naar en pad van $(0, 0)$ naar (N, M). Met elke overgang van een punt naar een andere zijn kosten verbonden, die we als *afstanden* tussen de punten zien, daarbij noteren we de kosten voor de overgang van (i', j') naar (i, j) met $d((i', j'), (i, j))$. Sommige van de kosten kunnen oneindig zijn, om uit te drukken dat deze overgang onmogelijk is.

Voor elk punt (i, j) noemt men de punten (i', j') waarvoor de overgang van (i', j') naar (i, j) mogelijk is (d.w.z. eindige kosten heeft) de *mogelijke voorgangers* een het stelsel van mogelijke voorgangers noemt men de *lokale beperkingen.* In sommige toepassingen kan men bijvoorbeeld alleen maar van $(i - 1, j - 1)$, $(i - 1, j)$ of $(i, j - 1)$ naar (i, j) komen, in andere gevallen zijn alle punten $(i - 1, j')$ mogelijke voorgangers van (i, j). Dit is bijvoorbeeld het geval als de eerste coördinaat tijdstippen en de tweede states aangeeft en we veronderstellen dat we van elke state naar elke andere state kunnen komen.

Het optimale pad van $(0, 0)$ naar (N, M) is natuurlijk het pad waarvoor de som van de kosten van de overgangen minimaal is. Bellman's principe zegt nu het volgende:

Bellman's principe: *Als het optimale pad van* $(0, 0)$ *naar* (N, M) *door het punt* (i, j) *loopt, dan is ook het deelpad van* $(0, 0)$ *tot* (i, j) *een optimaal pad tussen deze twee punten, net als het deelpad van* (i, j) *naar* (N, M) *een optimaal pad tussen deze twee punten is.*

Hier zit alleen maar de vanzelfsprekende opmerking achter dat we de kosten voor het pad van $(0, 0)$ via (i, j) naar (N, M) nog kunnen reduceren, als we de kosten voor een van de deelpaden tussen $(0, 0)$ en (i, j) of tussen (i, j) en (N, M) kunnen reduceren.

Maar als gevolg van Bellman's principe krijgen we een efficiënte manier om het optimale pad te vinden. We moeten (afhankelijk van de lokale beperkingen) stapsgewijs de optimale paden voor de punten (i, j) bepalen, door voor elke mogelijke voorganger (i', j') van (i, j) de kosten voor het optimale pad naar (i', j') bij de kosten voor de overgang van (i', j') naar (i, j) op te tellen en het minimum van deze kosten te kiezen.

Viterbi algoritme

Als we Bellman's principe op het probleem van de optimale rij van states van een HMM toepassen, krijgen we het *Viterbi algoritme.* Bellman's principe zegt in dit geval dat voor de optimale rij $q = q_1 q_2 \ldots q_T$ van states voor de waarneming $O = o_1 o_2 \ldots o_T$ ook de deelrijen tot en vanaf tijdstip t optimaal zijn, dus $p(o_1 \ldots o_t, q_1 \ldots q_t \mid \lambda)$ is maximaal en $p(o_t \ldots o_T, q_t \ldots q_T \mid \lambda)$ is maximaal.

In de opzet van het dynamische programmeren hebben we als roosterpunten de paren (t, i) die aangeven dat $q_t = S_i$ is. Hierbij beginnen we met het (formele) punt $(0, 0)$ en eindigen in een punt (T, i), waarbij we geen beperking op i opleggen. De mogelijke voorgangers van (t, i) zijn $(t - 1, k)$ voor alle

$1 \leq k \leq N$. In plaats van kosten praten we nu over kansen, en natuurlijk willen we voor de kansen niet het minimum maar het maximum vinden. De kans die bij de overgang van $(t-1,k)$ naar (t,i) hoort, is de overgangskans a_{ki} van state S_k naar state S_i en de kans $b_i(o_t)$ om in state S_i op tijdstip t de waarneming o_t te produceren. De totale kans voor de overgang $(t-1,k) \to (t,i)$ is dus

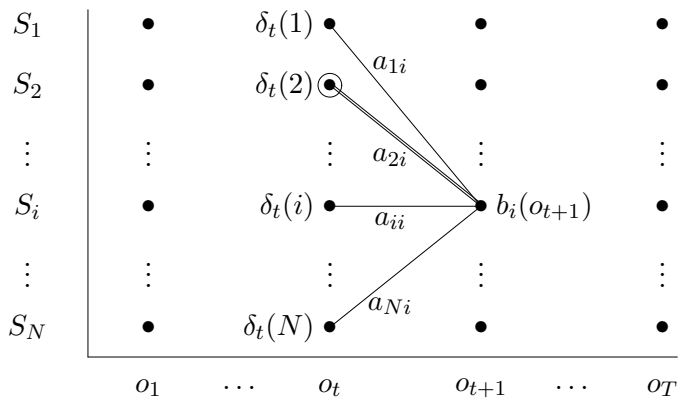
$$p((t-1,k) \to (t,i)) = a_{ki} \cdot b_i(o_t).$$

We definiëren nu $\delta_t(i)$ als de kans van de optimale rij van states voor de deelwaarneming $o_1 o_2 \ldots o_t$, die op tijdstip t in state S_i is.

We krijgen zo de recursie

$$\delta_1(i) = \pi(i) b_i(o_1) \quad \text{en} \quad \delta_{t+1}(i) = \left(\max_{1 \leq k \leq N} \delta_t(k) a_{ki} \right) b_i(o_{t+1})$$

die sterk op de recursie bij het forward algoritme lijkt. Het enige verschil is, dat in plaats van de som over de alle voorgangers nu het maximum over de voorgangers genomen wordt. Maar het schema van het Viterbi algoritme is zo als in Figuur III.6 te zien precies hetzelfde als bij het forward algoritme. Aanvullend moeten we wel bij elke punt (t,i) nog opslaan, vanuit welke voorganger $(t-1,k)$ het maximum bereikt werd, om uiteindelijk het optimale pad terug te kunnen vinden. Dit wordt meestal door een geschakelde lijst geïmplementeerd, in Figuur III.6 is als voorbeeld de overgang $(t,2) \to (t+1,i)$ benadrukt.



Figuur III.6: Berekening van $\delta_{t+1}(i)$ in het Viterbi algoritme.

Om meer efficiëntie bij het evalueren van een rij waarnemingen met verschillende HMM's te bereiken, wordt soms de evaluatie van de kans met behulp van vooruit- of achteruitkansen vervangen door de zogeheten *Viterbi benadering*. Hierbij wordt in plaats van de som over de kansen voor *alle* paden alleen maar de kans voor het beste pad bepaald (en dit natuurlijk met behulp van het Viterbi algoritme). Het idee hierachter is dat bij het evalueren uiteindelijk toch maar heel weinig paden een substantiële bijdrage aan de totale kans leveren en dat de som over de kansen voor hetgeen HMM maximaal wordt waarvoor het optimale pad de hoogste kans heeft.

Er valt nog iets over de implementatie van het Viterbi algoritme (en andere algoritmen in het kader van probabilistische modellen) op te merken:

Omdat er steeds kansen met elkaar vermenigvuldigd worden en deze soms zelf al redelijk klein zijn, worden de waarden van de $\delta_t(i)$ snel erg klein en dalen al gauw onder de rekennauwkeurigheid van een computer. Voor dit probleem bestaat er een heel simpele oplossing: Men rekent met de logaritmen van de kansen.

Merk op: Omdat de logaritme een monotone functie is, is $\delta_t(i)$ maximaal voor de i waarvoor $\tilde{\delta}_t(i) := -\log(\delta_t(i))$ minimaal is.

Als we het Viterbi algoritme op de logaritmen $\tilde{\delta}_t(i) = -\log(\delta_t(i))$ transformeren, krijgen we:

$$\begin{aligned}\tilde{\delta}_1(i) &= -\log(\pi(i)) - \log(b_i(o_1));\\ \tilde{\delta}_{t+1}(i) &= \min_{1\le k\le N}\Big(\tilde{\delta}_t(k) - \log(a_{ki})\Big) - \log(b_i(o_{t+1})).\end{aligned}$$

Natuurlijk worden de logaritmen van de a_{ij} en $b_i(k)$ niet steeds opnieuw berekend, maar ze worden bij het HMM opgeslaan.

Een soortgelijke opmerking geldt natuurlijk ook voor het forward algoritme. Daarbij is er echter het probleem, dat de kansen voor de verschillende paden bij elkaar opgeteld moeten worden. Dit lost men soms met behulp van de formule $\log(p+q) = \log(p(1+\frac{q}{p})) = \log(p)+\log(1+\frac{q}{p}) = \log(p)+\log(1+e^{\log(q)-\log(p)})$ op, maar vaak wordt hier inderdaad met kansen gerekend die op een geschikte manier geschaald worden.

Ook dit probleem wordt vermeden, als men bij het evalueren het forward algoritme door de Viterbi benadering vervangt.

Toepassing van het Viterbi algoritme

We kijken nu naar de toepassing van het Viterbi algoritme op een HMM met de drie munten, waarvan maar één eerlijk is. De drie munten zijn de drie states S_1, S_2, S_3 en de mogelijke uitkomsten zijn $x_1 = \texttt{K}$ voor *kop* en $x_2 = \texttt{M}$ voor *munt*. Het HMM $\lambda = \lambda(A, B, \pi)$ is gegeven door

$$A = (a_{ij}) := \begin{pmatrix} 0.6 & 0.2 & 0.2\\ 0.4 & 0.2 & 0.4\\ 0.4 & 0.4 & 0.2\end{pmatrix},\quad B = (b_i(k)) := \begin{pmatrix} 0.5 & 0.5\\ 0.75 & 0.25\\ 0.25 & 0.75\end{pmatrix},\quad \pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}).$$

We bekijken de waarneming $O = \texttt{KMKMM}$.

Voor de initialisering hebben we:

$$\begin{aligned}\delta_1(1) &= \pi(1)b_1(1) = 0.33\cdot 0.5 = 0.167,\\ \delta_1(2) &= \pi(2)b_2(1) = 0.33\cdot 0.75 = 0.25,\\ \delta_1(3) &= \pi(3)b_3(1) = 0.33\cdot 0.25 = 0.083.\end{aligned}$$

Voor de volgende stap berekenen we nu

$$\begin{aligned}
i = 1 : \delta_1(1)a_{11}b_1(2) &= 0.167 \cdot 0.6 \cdot 0.5 = 0.05, \leftarrow \max \\
\delta_1(2)a_{21}b_1(2) &= 0.25 \cdot 0.4 \cdot 0.5 = 0.05, \\
\delta_1(3)a_{31}b_1(2) &= 0.083 \cdot 0.4 \cdot 0.5 = 0.0167, \\
i = 2 : \delta_1(1)a_{12}b_2(2) &= 0.167 \cdot 0.2 \cdot 0.25 = 0.0083, \\
\delta_1(2)a_{22}b_2(2) &= 0.25 \cdot 0.2 \cdot 0.25 = 0.0125, \leftarrow \max \\
\delta_1(3)a_{32}b_2(2) &= 0.083 \cdot 0.4 \cdot 0.25 = 0.0083, \\
i = 3 : \delta_1(1)a_{13}b_3(2) &= 0.167 \cdot 0.2 \cdot 0.75 = 0.025, \\
\delta_1(2)a_{23}b_3(2) &= 0.25 \cdot 0.4 \cdot 0.75 = 0.075, \leftarrow \max \\
\delta_1(3)a_{33}b_3(2) &= 0.083 \cdot 0.2 \cdot 0.75 = 0.0125.
\end{aligned}$$

Dit geeft voor de $\delta_2(i)$ het volgende:

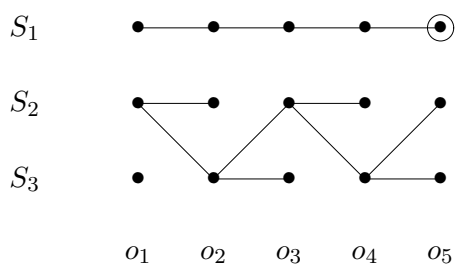
$$\begin{aligned}
&\delta_2(1) = 0.05 \text{ met } k = 1 \text{ (of } k = 2) \text{ als voorganger,} \\
&\delta_2(2) = 0.0125 \text{ met } k = 2 \text{ als voorganger,} \\
&\delta_2(3) = 0.075 \text{ met } k = 2 \text{ als voorganger.}
\end{aligned}$$

Als we zo doorgaan krijgen we voor $\delta_t(i)$ met de voorgangers k:

$$\begin{aligned}
&\delta_3(1) = 0.015, k = 1, \quad \delta_3(2) = 0.0225, k = 3, \quad \delta_3(3) = 0.00375, k = 3, \\
&\delta_4(1) = 0.0045, k = 1, \ \delta_4(2) = 0.001125, k = 2, \ \delta_4(3) = 0.00675, k = 2, \\
&\delta_5(1) = 0.00135, k = 1, \delta_5(2) = 0.000675, k = 3, \delta_5(3) = 0.0010125, k = 3.
\end{aligned}$$

We zien dat $\delta_5(1)$ het maximum van de $\delta_5(i)$ is, daarom eindigt de optimale rij van states in state S_1. Omdat in alle stappen de state S_1 voorganger S_1 heeft, is dus $S_1S_1S_1S_1S_1$ de optimale rij van states. Merk op dat tot $t = 4$ de rij $S_2S_3S_2S_3$ optimaal was geweest.

Als we de punten (t, i) als punten van een tralie (of rooster) bekijken en het punt (t, i) met degene voorganger $(t - 1, k)$ verbinden die de maximale waarde van $\delta_t(i)$ oplevert, kunnen we hieruit de optimale rij van states makkelijk achterhalen. In Figuur III.7 is dit tralie voor het net besproken voorbeeld te zien, waarbij de optimale eindstate door een extra cirkel benadrukt is.



Figuur III.7: Tralie voor het Viterbi algoritme.

13.3 Training van een HMM

Tot nu toe zijn we ervan uit gegaan dat we de parameters van het HMM al kennen. De vraag is nu, hoe we de parameters $A = (a_{ij})$, $B = (b_i(k))$ en $\pi = (\pi(1), \ldots, \pi(N))$ zo kunnen bepalen, dat het model een gegeven rij $O = o_1 o_2 \ldots o_T$ van waarnemingen zo goed mogelijk beschrijft, dus zo dat de kans $p(O \mid \lambda(A, B, \pi))$ maximaal wordt. Omdat bij deze aanpak de kans gemaximaliseerd wordt, noemt men dit ook de *maximum likelihood schatting* van de parameters.

In Wiskunde 1 hebben we in het kader van de kansrekening naar een soortgelijk, maar veel eenvoudiger probleem gekeken. We wilden toen de parameters van een kansverdeling, bijvoorbeeld een normale verdeling, zo bepalen, dat met deze parameters de kans voor een rij gebeurtenissen maximaal werd. Het idee was toen, de (logaritme van de) kans op de gebeurtenissen als functie van de parameters te interpreteren en een maximum van deze functie te bepalen door de partiële afgeleiden naar de parameters gelijk aan 0 te zetten en deze vergelijkingen op te lossen. Bij de normale verdeling hebben we zo bijvoorbeeld geconcludeerd, dat de beste keuze voor de verwachtingswaarde μ van de normale verdeling het gemiddelde van de gebeurtenissen is – een niet echt verrassend resultaat.

In principe zouden we bij de HMM's een analoge aanpak kunnen kiezen: We schrijven $p(O \mid \lambda)$ als functie van de parameters a_{ij}, $b_i(k)$ en $\pi(i)$, zo als we dat in het begin van deze les al hebben gedaan, dus als

$$p(O \mid \lambda) = \sum_{q=q_1 \ldots q_T} \pi(q_1) b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots a_{q_{T-1} q_T} b_{q_T}(o_T).$$

Vervolgens bepalen we de partiële afgeleiden naar de parameters en proberen de vergelijkingen

$$\frac{\partial}{\partial a_{ij}} p(O \mid \lambda) = 0, \quad \frac{\partial}{\partial b_i(k)} p(O \mid \lambda) = 0, \quad \frac{\partial}{\partial \pi(i)} p(O \mid \lambda) = 0$$

simultaan op te lossen. Dat we eigenlijk nog moeten eisen dat de rijen van de matrices A en B de som 1 hebben, omdat we het over kansverdelingen hebben, vergeten we hierbij even.

Het probleem is dat in alle praktische gevallen het stelsel vergelijkingen dat men zo krijgt niet analytisch oplosbaar is. Maar dit probleem doet zich ook al in veel eenvoudigere vraagstukken voor, want ook bij gewone functies van één veranderlijke kunnen we vaak de nulpunten niet expliciet bepalen. De gebruikelijke manier, om in deze situatie verder te komen, is een *numerieke benaderingsmethode* toe te passen.

Het idee in het kader van HMM's is, startwaarden voor de parameters A, B en π te gokken en vervolgens de parameters stapsgewijs zo aan te passen, dat in elke stap de *likelihood* $p(O \mid \lambda(A, B, \pi))$ toeneemt.

In het algemeen levert zo'n benaderingsmethode alleen maar een lokaal maximum van de functie $p(O \mid \lambda)$ op, en omdat deze functie zo ingewikkeld

is, is er ook geen goede manier om een globaal maximum te vinden. In de praktijk probeert men een paar verschillende stelsels van startwaarden en kiest vervolgens het beste van de gevonden lokale maxima.

Baum-Welch algoritme

We zullen nu een speciale benaderingsmethode bekijken, die de parameters van een HMM stapsgewijs verbetert, namelijk het *Baum-Welch algoritme*. Deze gebruikt de vooruit- en achteruitkansen $\alpha_t(i)$ en $\beta_t(i)$ die we al bij de evaluatie van het HMM tegen gekomen zijn.

Om de methode goed te kunnen formuleren, hebben we eerst nog twee nieuwe uitdrukkingen nodig, die zekere kansen beschrijven:

De voorwaardelijke kans dat het systeem op tijdstip t in state S_i is, gegeven de volledige rij waarnemingen $O = o_1 o_2 \ldots o_T$, noemen we $\gamma_t(i)$. Volgens de relatie $p(A \mid B) = \frac{p(A,B)}{p(B)}$ geldt dan:

$$\gamma_t(i) := p(q_t = S_i \mid O, \lambda) = \frac{p(O, q_t = S_i \mid \lambda)}{p(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}.$$

Verder definiëren we als $\xi_t(i,j)$ de voorwaardelijke kans dat het systeem tussen de tijdstippen t en $t+1$ van state S_i naar state S_j gaat, gegeven de rij O van waarnemingen. Er geldt

$$\begin{aligned}\xi_t(i,j) := p(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) &= \frac{p(O, q_t = S_i, q_{t+1} = S_j \mid \lambda)}{p(O \mid \lambda)} \\ &= \frac{\alpha_t(i)\ a_{ij}\ b_j(o_{t+1})\ \beta_{t+1}(j)}{p(O \mid \lambda)}.\end{aligned}$$

Tussen de kansen $\xi_t(i,j)$ en $\gamma_t(i)$ bestaat een eenvoudige relatie, want de kans om op tijdstip t in state S_i te zijn is de som over alle j van de kansen, tussen de tijdstippen t en $t+1$ van state S_i naar S_j te gaan. Er geldt dus

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j).$$

Als we nu de kansen $\gamma_t(i)$ over de tijdstippen $t = 1, \ldots, T$ optellen, krijgen we het verwachte aantal van waarnemingen die door de state S_i geproduceerd zijn. Net zo kunnen we de kansen $\xi_t(i,j)$ over de tijdstippen $t = 1, \ldots, T-1$ optellen, dan krijgen we het verwachte aantal overgangen van state S_i naar state S_j. We hebben dus

$$\begin{aligned}\sum_{t=1}^{T} \gamma_t(i) &= \text{ verwacht aantal emissies vanuit state } S_i; \\ \sum_{t=1}^{T-1} \xi_t(i,j) &= \text{ verwacht aantal overgangen tussen states } S_i \text{ en } S_j.\end{aligned}$$

Maar aan de hand van deze gegevens kunnen we nieuwe parameters A', B' en π' als relatieve frequenties schatten, namelijk door:

$$\begin{aligned}
\pi'(i) &= \text{ verwachte kans op state } S_i \text{ op tijdstip 1} \\
&= \gamma_1(i) = \frac{\alpha_1(i)\ \beta_1(i)}{\sum_{i=1}^{N} \alpha_t(i)\ \beta_t(i)} = \frac{\alpha_1(i)\ \beta_1(i)}{\sum_{i=1}^{N} \alpha_T(i)} \\
a'_{ij} &= \frac{\text{verwacht aantal overgangen van state } S_i \text{ naar state } S_j}{\text{verwacht aantal overgangen vanuit state } S_i} \\
&= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\ a_{ij}\ b_j(o_{t+1})\ \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \\
b'_i(k) &= \frac{\text{verwacht aantal emissies vanuit state } S_i \text{ met waarneming } x_k}{\text{verwacht aantal emissies vanuit state } S_i} \\
&= \frac{\sum_{t=1,o_t=x_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} = \frac{\sum_{t=1,o_t=x_k}^{T} \alpha_t(i)\ \beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i)\ \beta_t(i)}
\end{aligned}$$

Merk op hoe de waarneming $O = o_1 o_2 \ldots o_T$ bij de berekening van a'_{ij} en $b'_i(k)$ betrokken is, dit is uiteindelijk de reden dat de parameters aan de waarneming aangepast worden.

De grap is nu, dat we met de nieuwe parameters $A' = (a'_{ij})$, $B' = (b'_i(k))$ en $\pi' = (\pi'(1), \ldots \pi'(N))$ steeds een *beter* model voor de beschrijving van O krijgen dan met de oude parameters A, B en π, er laat zich namelijk aantonen dat geldt:

$$\lambda' = \lambda(A', B', \pi') \Rightarrow p(O \mid \lambda') \geq p(O \mid \lambda).$$

We kunnen nu de herschatting van de parameters itereren door het nieuwe model $\lambda(A', B', \pi')$ te gebruiken om de vooruit- en achteruitkansen $\alpha_t(i)$ en $\beta_t(i)$ en de kansen $\gamma_t(i)$ en $\xi_t(i,j)$ opnieuw te bepalen en hieruit een verder verbeterd stelsel parameters te verkrijgen. Deze procedure wordt herhaald tot dat de likelihood $p(O \mid \lambda)$ niet meer veranderd of een maximaal aantal iteratie stappen bereikt is.

13.4 Toegift: Levenshtein afstand

Als toegift behandelen we de toepassing van Bellman's principe op een ander belangrijk probleem in de patroonherkenning, namelijk de afstand tussen strings. Dit heeft toepassingen in de verwerking en herkenning van teksten en taal, maar ook in de beeldherkenning.

Een string is hierbij algemeen een keten van symbolen en men wil een afstand tussen twee ketens kunnen berekenen. Bij teksten zijn de symbolen gewoon letters, in de spraakherkenning zijn de symbolen vaak woorden, maar kunnen ook grammaticale etiketten zijn. In de beeldherkenning wordt vaak de omtrek van een element als keten van zekere elementaire symbolen beschreven, lijnstukken, hoeken etc.

Een mogelijke definitie van de afstand tussen twee strings is de *Edit afstand* die naar een van de uitvinders nu meestal *Levenshtein afstand* heet. Het

idee hierbij is, door *elementaire edit operaties* de ene string in de andere te transformeren, waarbij elementaire operaties de volgende zijn:

- vervangen (substitution) van een symbool, bijvoorbeeld *ki***j***ker* → *ki***k***ker*;
- invoegen (insertion) van een symbool, bijvoorbeeld *bouwer* → *b***r***ouwer*.
- weglaten (deletion) van een symbool, bijvoorbeeld *koe***k** → *koe*;

Natuurlijk zijn er verschillende manieren, om van een string door een combinatie van vervangen, invoegen en weglaten naar een andere string te komen, maar het is voor de hand liggend het minimale aantal stappen als edit afstand tussen de strings te definiëren:

Definitie: De *Levenshtein afstand* tussen twee strings is gedefinieerd als het *minimale aantal* van elementaire edit operaties waarmee de eerste string in de tweede string getransformeerd kan worden.

De vraag is nu hoe men het minimale aantal operaties vindt. Dit gebeurt analoog met het Viterbi algoritme door de methode van het dynamische programmeren.

Het idee is dat men voor twee strings $X = x_1x_2 \dots x_N$ en $Y = y_1y_2 \dots y_M$ stapsgewijs kijkt hoe men beginstukken van de twee strings in elkaar kan transformeren. Volgens Bellman's principe hoeft men hierbij alleen maar het minimale aantal operaties op te slaan om van het beginstuk $x_1 \dots x_i$ van lengte i van X naar het beginstuk $y_1 \dots y_j$ van lengte j van Y te komen. Men krijgt zo een rooster van punten (i, j) voor $0 \leq i \leq N$ en $0 \leq j \leq M$ waarbij we het aantal edit operaties als kosten voor de overgang tussen twee punten interpreteren. In dit geval hebben we (tegenover het Viterbi algoritme) sterke lokale beperkingen, want het punt (i, j) heeft slechts drie mogelijke voorgangers:

(1) het punt $(i-1, j-1)$: In het geval $x_i = y_j$ heeft de overgang van $(i-1, j-1)$ naar (i, j) kosten 0, anders kosten 1. Als $x_i \neq y_j$ is deze overgang namelijk het vervangen van x_i door y_j.

(2) het punt $(i, j-1)$: Deze overgang is het invoegen van het symbool y_j en heeft de kosten 1.

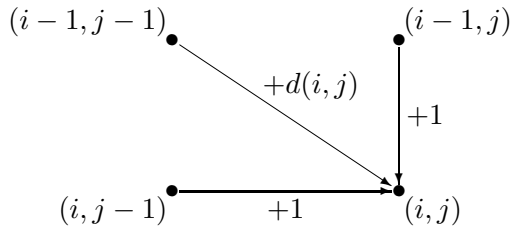
(3) het punt $(i-1, j)$: Deze overgang is het weglaten van het symbool x_i en heeft de kosten 1.

In Figuur III.8 zijn deze overgangen schematisch te zien, waarbij we met $d(i, j)$ de kosten voor het vervangen van x_i door y_j aangeven, hiervoor geldt

$$d(i,j) := \begin{cases} 0 & \text{als} \quad x_i = y_j \\ 1 & \text{als} \quad x_i \neq y_j. \end{cases}$$

Volgens Bellman's principe vinden we de minimale kosten $D(i, j)$ voor de transformatie van het beginstuk $x_1 \dots x_i$ van X naar het beginstuk $y_1 \dots y_j$ van Y als volgt:

We initialiseren $D(i, 0) := i$ voor $0 \leq i \leq N$ (dit is het weglaten van de eerste i symbolen van X) en $D(0, j) := j$ voor $0 \leq j \leq M$ (dit is het invoegen



Figuur III.8: Mogelijke voorgangers van (i,j).

van de eerste j symbolen van Y) en berekenen vervolgens voor $i = 1, 2, \ldots N$ en voor $j = 1, 2, \ldots M$:

$$D(i,j) := \min\{D(i-1,j-1) + d(i,j),\ D(i,j-1) + 1,\ D(i-1,j) + 1\}.$$

Merk op dat op het moment dat we $D(i,j)$ willen berekenen de waarden van $D(i-1,j-1)$, $D(i,j-1)$ en $D(i-1,j)$ al berekend zijn, omdat we i stapsgewijs van 1 t/m N verhogen en voor een vaste i ook met j stapsgewijs van 1 t/m M lopen.

Als we ons de waarden van $D(i,j)$ als elementen van een $N \times M$-matrix voorstellen, vullen we deze matrix rijsgewijs van boven naar beneden en de rijen van links naar rechts. Uiteindelijk zijn we geïnteresseerd in de waarde $D(N,M)$ rechts onder, die de Levenshtein afstand tussen X en Y aangeeft.

Het schema hieronder laat voor het voorbeeld $X =$ `KUNSTMATIGE` en $Y =$ `INTELLIGENTIE` de waarden $D(i,j)$ en een optimaal pad (aangeduid door de hokjes) zien.

		I	N	T	E	L	L	I	G	E	N	T	I	E
	$\boxed{0}$	1	2	3	4	5	6	7	8	9	10	11	12	13
K	$\boxed{1}$	1	2	3	4	5	6	7	8	9	10	11	12	13
U	2	$\boxed{2}$	3	3	4	5	6	7	8	9	10	11	12	13
N	3	3	$\boxed{2}$	3	4	5	6	7	8	9	9	10	11	12
S	4	4	$\boxed{3}$	3	4	5	6	7	8	9	10	10	11	12
T	5	5	4	$\boxed{3}$	4	5	6	7	8	9	10	10	11	12
M	6	6	5	4	$\boxed{4}$	5	6	7	8	9	10	11	11	12
A	7	7	6	5	5	$\boxed{5}$	6	7	8	9	10	11	12	12
T	8	8	7	6	6	6	$\boxed{6}$	7	8	9	10	10	11	12
I	9	8	8	7	7	7	7	$\boxed{6}$	7	8	9	10	10	11
G	10	9	9	8	8	8	8	7	$\boxed{6}$	$\boxed{7}$	$\boxed{8}$	$\boxed{9}$	$\boxed{10}$	11
E	11	10	10	9	8	9	9	8	7	6	7	8	9	$\boxed{10}$

Merk op dat er verschillende mogelijkheden voor het optimale pad zijn die mogelijk ook verschillende aantallen vervangingen, invoegingen en weglatingen kunnen hebben, maar de *som* van de aantallen vervangingen, invoegingen en weglatingen is bij alle optimale paden natuurlijk hetzelfde. In het voorbeeld is de Levenshtein afstand tussen de twee strings dus 10, het aangegeven pad heeft 4 vervangingen, 4 invoegingen en 2 weglatingen.

Net als bij het Viterbi algoritme moeten we ook hier opslaan vanuit welke voorganger we bij $D(i,j)$ het minimum bereiken om het optimale pad terug te kunnen vinden.

Een iets algemenere versie van de Levenshtein afstand krijgt men, door gewichten aan de verschillende edit operaties te geven, want in sommige toepassingen kan een invoeging erger zijn dan een vervanging. Als we de kosten van een vervanging met k_s, de kosten van een invoeging met k_i en de kosten van een weglating met k_d noteren, berekenen we in dit geval de kosten $D(i,j)$ voor het optimale pad door het punt (i,j) als

$$D(i,j) := \min\{D(i-1,j-1) + d(i,j)\,k_s,\ D(i,j-1) + k_i,\ D(i-1,j) + k_d\},$$

waarbij de initialiseringen $D(i,0) = i\,k_d$ en $D(0,j) = j\,k_i$ zijn.

> In de eerste fase van de spraakherkenning is een soortgelijke techniek ook op spraaksignalen toegepast, er werden namelijk de geluidssignalen in een keten van symbolen omgezet en deze werden door een variatie van de tijdschaal met opgeslagen patronen vergeleken. Deze methode noemt men *dynamic time warping*.

Belangrijke begrippen in deze les

- forward algoritme, backward algoritme
- vooruitkansen, achteruitkansen
- optimale rij van states
- Bellman's principe
- Viterbi algoritme
- training van een HMM, Baum-Welch algoritme
- Levenshtein afstand

Opgaven

99. We beschrijven twee mogelijke uitkomsten `K` en `M` door twee HMM's λ_1, λ_2 met (telkens) twee states. De beginverdelingen voor de states zijn bij beide modellen uniform, dus $\pi = (0.5, 0.5)$. Het model λ_1 heeft de overgangskansen A_1 en emissiekansen B_1, het model λ_2 de overgangskansen A_2 en emissiekansen B_2 gegeven door:

$$A_1 := \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},\ B_1 := \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix};\quad A_2 := \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix},\ B_2 := \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix}.$$

(i) Bepaal voor beide modellen de kansen $p(O \mid \lambda)$ voor de waarnemingen $O_1 =$ `KKK` en $O_2 =$ `MKM`.

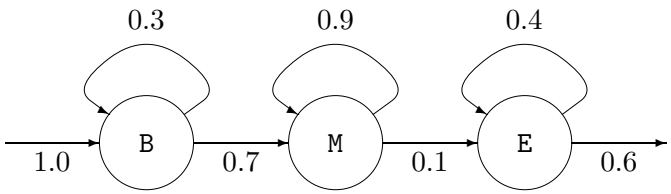
(ii) Bepaal voor beide modellen de optimale rij q van states voor de waarnemingen uit deel (i) en bereken de kansen $p(O, q \mid \lambda)$ voor de combinatie van waarnemingen en states.

100. We kijken nog eens naar het inmiddels bekende HMM met drie munten en parameters:

$$A = (a_{ij}) := \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.2 \end{pmatrix}, \quad B = (b_i(k)) := \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}, \quad \pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}).$$

Door een meting weten we, dat bij de eerste en laatste waarneming de eerlijke (eerste) munt geworpen werd. Wat is nu de optimale rij van states die de waarneming $O =$ `KMKMK` voortbrengt?

101. In de spraakherkenning worden fonemen (de kleinste onderscheidbare klanken in een taal) vaak door HMM's met drie states (begin (`B`), midden (`M`), eind (`E`)) gerepresenteerd. Stel de overgangskansen tussen de states zijn door het volgende diagram gegeven:



Het HMM heeft 7 mogelijke uitkomsten $x_1, \ldots, x_7$ en de emissiekansen voor deze uitkomsten zijn gegeven door

state	x_1	x_2	x_3	x_4	x_5	x_6	x_7
`B`	0.5	0.2	0.3	0	0	0	0
`M`	0	0	0.2	0.7	0.1	0	0
`E`	0	0	0	0	0.1	0.5	0.4

Bepaal voor de rij $x_1x_2x_3x_4x_5x_6x_7$ van uitkomsten de optimale rij van states en geef de kans op deze waarneming voor de optimale rij van states aan.

102. Bepaal de Levenshtein afstand tussen de volgende paren van strings (waarbij ook de spatie een symbool is) en geef de edit operaties aan:

(i) $X =$ `ABABAA` en $Y =$ `ABBAA`;

(ii) $X =$ `IK WEET NIETS` en $Y =$ `WEET IK WAT`;

(iii) $X =$ `SINTERKLAAS` en $Y =$ `KERSTMAN`;

(iv) $X =$ `C3POR2D2` en $Y =$ `HAL2001`.