



Arnt Inge Vistnes

Svingninger og bølger

Eget forlag
Versjon: Vår2013

Denne boka er blitt til med støtte fra Norsk faglitterær forfatter- og oversetterforening (frikjøp av undervisning i to perioider mens jeg jobbet mye med boka).

Boka kom ut i en redusert "prøveutgave" på Tapir Forlag i 2011, men på grunn av fargetrykk og lite opplag, ble boka uforholdsmessig dyr. Avtalen med Tapir Forlag ble da sagt opp.

Boka er nå gratis tilgjengelig i elektronisk format (.pdf-fil) fra www.duo.uio.no. Denne versjonen av boka ble brukt i kurset FYS2130 ved Universitetet i Oslo våren 2013. Kapitlene ble da frigitt etter som de ble revidert. Samlingen av alle kapitler innen en og samme "perm" ble foretatt sept/okt samme år.

Etter planen vil det komme en revidert versjon for våren 2014. Spesielt er det tanke å revidere betydelig kapittel 5 om fouriertransformasjon.

Svingninger og bølger

Arnt Inge Vistnes

Versjon: Vår 2013

© Copyright 2013 Arnt Inge Vistnes
på tekst og figurer/bilder (der ikke annet eksplisitt er gitt).

ISBN 978-82-999439-0-1

Innhold

1 Innledning	iii
1.0.1 Numeriske metoder	iv
1.0.2 Bakgrunn for boka	v
1.0.3 Organisering av stoffet	vi
1.0.4 Format - rettigheter	vii
1.0.5 Støttelitteratur	vii
2 Fri og dempede svingninger	1
2.1 Kinematikk	2
2.2 Dynamisk beskrivelse av et mekanisk system	3
2.3 Dempede svingninger	6
2.4 Superposisjon og ikke-lineære ligninger *	9
2.5 Elektriske svingninger	11
2.6 Energibetraktninger	13
2.7 Læringsmål	15
2.8 Oppgaver	16
3 Tvungne svingninger og resonans	19
3.1 Tvungne svingninger	20
3.2 Resonans	22
3.2.1 Fasorbeskrivelse	23
3.3 Kvalitetsfaktoren Q	26
3.4 Tidsbegrenset tvungen svingning	29
3.5 Frekvensrespons ved tidsbegrenset tvungen svingning	32
3.6 Eksempel: Hørsel *	33
3.7 Læringsmål	35
3.8 Oppgaver	36
4 Numeriske løsningsmetoder	39
4.1 Innledning	40
4.2 Grunnleggende idé bak numeriske metoder	40
4.3 Eulers metode og varianter av denne	41
4.4 Runge-Kuttas metode	43
4.4.1 Beskrivelse av metoden	43
4.5 Partielle differensialligninger	45
4.6 “Dimensjonsløs” differensialligning	47
4.7 Eksempel på numerisk løsning: Enkel pendel	49
4.8 Test av implementering	50
4.9 Krav til reproducertbarhet	51
4.10 Arbeidsgang ved numeriske metoder	52
4.11 Diverse tillegg	54
4.11.1 Oppsummering, kapittel 3	54

4.11.2	Pseudokode for Runge-Kuttas metode	55
4.11.3	Python-kode for Runge-Kuttas metode	56
4.11.4	Matlab-kode for Runge-Kuttas metode	56
4.11.5	Funksjonen som inneholder differensiallikningen	57
4.11.6	Eksempel: Matlabprogram som bruker Runge-Kutta	58
4.11.7	Bruk av Matlab's innebygde Runge-Kutta	60
4.11.8	Noen "kjøreregler" fra Hans Petter Langtangen	61
4.11.9	Forslag til videre lesing	61
4.11.10	Et annet lite tips....	62
4.12	Læringsmål	62
4.13	Oppgaver	63
4.13.1	En mer sammensatt regneoppgave	65
5	Fourieranalyse	71
5.1	Innledning	72
5.2	Fouriertransformasjon (FT)	72
5.3	Hva sier $F(\omega)$ oss?	73
5.3.1	Fouriertransformasjon av en ren sinusfunksjon	74
5.3.2	Fouriertransformasjon av en ren cosinusfunksjon	75
5.3.3	Fouriertransformasjon av en mer sammensatt funksjon	75
5.4	Fourierrekker	76
5.5	Diskret fouriertransformasjon	78
5.5.1	Diskret fouriertransform i mer fysiske termer	80
5.6	Et konkret eksempel	82
5.6.1	Fouriertransformasjonen	83
5.6.2	Tidspopløsningen, noen kommentarer	84
5.6.3	Speiling / folding	86
5.6.4	Samplingsteoremet	88
5.7	* En finurlighet	89
5.7.1	Fouriertransformasjon av mer kompliserte signaler	91
5.8	Tidsbegrenset signal	93
5.9	Til ettertanke	95
5.10	Fouriertransformasjon, eksempel på et dataprogram	97
5.11	Læringsmål	98
5.12	Oppgaver	99
6	Bølger	103
6.1	Innledning	104
6.2	Plan bølge	106
6.2.1	Bølgens hastighet	107
6.2.2	Løsning av bølgeligningen?	108
6.2.3	Hvilken vei?	109
6.2.4	Andre bølgeformer	110
6.2.5	Sum av bølger	111
6.2.6	Bølge beskrevet på kompleks form	111
6.3	Transversell og longitudinal	112
6.4	Utledning av bølgeligningen	113
6.4.1	Bølger på en streng	113
6.4.2	Bølger i luft/væske	116
6.4.3	Konkrete eksempler	118
6.4.4	Trykkbølger	120

6.5 Læringsmål	121
6.6 Opgaver	121
7 Lyd	125
7.1 Refleksjon av bølger	126
7.2 Akustisk impedans	128
7.2.1 Ultralydbilder	129
7.3 Stående bølger	130
7.4 Musikkinstrumenter og frekvensspekter	132
7.4.1 Toneintervaller	135
7.5 Svevelyd	137
7.6 Lydintensitet	138
7.6.1 Lydintensitet vs avstand og tid	140
7.7 Desibel-skalaen	141
7.8 Doppler-effekt	145
7.8.1 Doppler for elektromagnetiske bølger	147
7.9 Sjokkbølger	148
7.9.1 Eksempel: Helikoptere	149
7.10 Læringsmål	150
7.11 Opgaver	150
8 Vannbølger og dispersjon	155
8.1 Innledning	156
8.2 Bølgebeskrivelser	157
8.2.1 Enkel bølgebeskrivelse	159
8.2.2 Fasehastigheten til vannbølger	160
8.3 Fase- og gruppehastighet	163
8.3.1 Aller enkleste tilnærming	164
8.3.2 Program-listing	166
8.3.3 Kommentarer til den enkle beskrivelsen	167
8.3.4 Normal dispersjon og anomal dispersjon	167
8.4 Bølger i vann	171
8.4.1 Skipsbølger, et eksempel	174
8.5 Lyspuls gjennom et medium	175
8.6 Numerisk beregning av tidsutvikling for en bølge	178
8.6.1 Et bølge-eksempel	181
8.7 Læringsmål	183
8.8 Referanser	184
8.9 Opgaver	184
9 Maxwells ligninger og elektromagnetiske bølger	189
9.1 Innledning	190
9.2 Maxwells ligninger på integralform	191
9.3 Over til differensialform	194
9.4 Utledning av bølgeligningen	196
9.5 Én løsning av bølgeligningen	199
9.6 Det elektromagnetiske spekteret	201
9.7 Energitransport	202
9.7.1 Poynting vektor	204
9.8 Strålingstrykk	205
9.9 Feiloppfatninger	206
9.9.1 Nærfelt og fjernfelt	206

9.9.2 Fotonbegrepet	209
9.9.3 Plan og plan fru Blom	211
9.10 Hjelpestoff	213
9.10.1 Nyttige matematiske relasjoner	213
9.10.2 Nyttige relasjoner og størrelser fra elektromagnetismen	214
9.11 Læringsmål	215
9.12 Opgaver	216
10 Refleksjon og transmisjon, polarisasjon, dobbeltbrytning	221
10.1 Innledning	222
10.2 Elektromagnetisk bølge vinkelrett inn mot et grensesjikt mellom to medier	222
10.3 Refleksjon og transmisjon når en bølge kommer på skrå inn mot grenseflaten	226
10.3.1 Brewster-vinkel-fenomenet i praksis	232
10.3.2 Fresnels ligninger	232
10.4 Polarisasjon	234
10.4.1 Dobbeltbrytning	235
10.4.2 Lysets vekselvirkning med materie	237
10.4.3 Polarisasjonsfiltre	239
10.4.4 Polariometri	242
10.5 Polarisasjon i astronomien	243
10.6 Snels brytningslov	244
10.6.1 Totalrefleksjon	245
10.7 Flyktige bølger (Evanescent waves)	246
10.8 Orienteringsstoff: 3D Stereoskopi	248
10.9 Læringsmål	251
10.10 Opgaver	252
11 Lysmåling, dispersjon av lys, farger	257
11.1 Lysmåling	258
11.1.1 Lumen vs watt	261
11.2 Dispersjon	262
11.3 "Farge" hva er det?	263
11.3.1 Fargemetri	266
11.3.2 Farger på en mobiltelefon- eller dataskjerm	269
11.3.3 Additiv versus subtraktiv fargeblanding	269
11.3.4 Andre kommentarer	271
11.4 Spekter fra et prisme	272
11.4.1 En digresjon: Goethes fargelære	274
11.5 Referanser	275
11.5.1 Takk!	275
11.6 Læringsmål	276
11.7 Opgaver	276
12 Geometrisk optikk	281
12.1 Lysstråler	282
12.2 Lys gjennom en krum grenseflate	284
12.3 Linsemakerformelen	286
12.4 Lysstråleoptikk	289
12.4.1 Fortegnsregler for linseformelen	292
12.5 Optiske instrumenter	292
12.5.1 Lupen	293
12.5.2 Teleskopet	295

12.5.3 Speilteleskop	297
12.5.4 Mikroskopet	299
12.5.5 Bildekvalitet	300
12.5.6 Synsfelt	303
12.5.7 Lysstyrke, blendertall	304
12.5.8 Øyets optikk	307
12.6 Oppsummering	311
12.7 Læringsmål	313
12.8 Opgaver	314
13 Interferens - Diffraksjon	319
13.1 Superposisjon og linearitet	320
13.2 Huygens prinsipp	321
13.3 Interferens fra en dobbeltpsalt	322
13.3.1 Interferensfiltre, interferens fra en tynn film	326
13.4 Mange, parallele spalter (optisk gitter)	327
13.5 Diffraksjon fra én spalt	330
13.6 Kombinert effekt	333
13.7 Diffraksjon i en videre ramme	334
13.8 Numerisk beregning av diffraksjon	335
13.8.1 Den grunnleggende modellen	336
13.8.2 Ulike løsninger	338
13.9 Diffraksjon fra et rundt hull	340
13.9.1 Bildet av stjerner i et teleskop	341
13.9.2 Divergens i en lysstråle	344
13.9.3 Andre eksempler	345
13.9.4 Diffraksjon ved to og tre dimensjoner	346
13.10 Babinet's prinsipp	347
13.11 Læringsmål	349
13.12 Opgaver	350
14 Koherens, dipolstråling og laser	355
14.1 Koherens, en kvalitativt tilnærming	356
14.2 Sum av to reelle bølger	358
14.3 Sum av flere bølger	359
14.4 Kohrente bølger	360
14.5 Koherenstidsbestemmelse	361
14.6 Anskueliggjøring av kohärens	364
14.7 Måling av koherenslengde for lys	365
14.8 Stråling fra en elektrisk ladning	368
14.8.1 Dipolstråling	370
14.9 Lasere	373
14.9.1 Populasjonsinvertering	376
14.10 Litteratur	378
14.11 Læringsmål	379
14.12 Opgaver	380
15 Wavelettransformasjon	383
15.1 Hva slags info kan wavelettransformasjon gi oss?	384
15.2 Kort historikk	386
15.3 Kort om matematikken bak	386
15.3.1 Oppfrisking av fouriertransformasjon	386

15.3.2 Formalisme ved wavelettransformasjon	387
15.3.3 “Diskret kontinuerlig” wavelettransformasjon	390
15.4 Praktisk gjennomføring	391
15.4.1 Eksempel på råmetoden for wavelettransformasjon	391
15.4.2 En mye mer effektiv algoritme	394
15.5 Viktige detaljer	397
15.5.1 Faseinformasjon og skalering av utslaget	397
15.5.2 Frekvensoppløsning vs tidsoppløsning	398
15.5.3 Randproblem	400
15.6 Optimalisering	402
15.7 Et realistisk eksempel	405
15.8 To ytterligere eksempler	408
15.9 Wavelet-ressurser på nett	412
15.10 Læringsmål	412
15.11 Oppgaver	412
16 Skinndybde og bølgeledere	417
16.1 Husker du	418
16.2 Skinndybde	418
16.2.1 Elektromagnetiske bølger inn mot en metallflate	419
16.3 Bølgeledere	422
16.3.1 Bølgemønsteret i en rektangulær bølgeleder	424
16.4 Enkeltmode optisk fiber	426
16.5 Læringsmål	428
16.6 Oppgaver	429

Kapittel 1

Innledning

Fenomener knyttet til svingninger og bølger omfatter noe av det vakreste vi kan oppleve innen fysikk. Tenk deg en verden uten lys og lyd, så fornemmer du kanskje hvor fundamentale svingninger og bølger er for vår tilværelse og for vår sivilisasjon! Svingninger og bølger har derfor vært en sentral del av enhver utdanning i fysikk, men det er ingen ensartet måte å presentere dette stoffet på.

“*Matematikk er fysikkens språk!*”, er det mange som hevder. Selv er jeg enig et stykke på vei. Fysiske lover blir formulert som matematiske ligninger, og vi bruker disse formlene når vi skal beregne forventet utfall av eksperimenter. Skal vi imidlertid kunne sammenligne tallene fra beregningene med faktiske observasjoner, må det mer til enn ren matematikk. Fysikk er like mye en rekke begreper, og begrepene er knyttet opp til vår erfaringsverden såvel som til matematikken. Fysikk uten kontakt med dagliglivets språk, begreper og erfaringer, ville slett ikke være fysikk. Da hadde vi bare ren matematikk! Det greske ordet *φυσις* (“fysis”) betyr jo *naturen*, og fysikk hører med til *naturvitenskapene*.

Mennesker er forskjellige. Min erfaring er at noen først og fremst fascineres av matematikken og lovene i fysikken, mens andre begeistres av fenomenene i seg selv. Det er sjeldent studenter har like stor interesse for begge disse sidene av fysikken. I denne boka vil jeg forsøke å presentere både formalisme og fenomener, for som nevnt er det først og fremst kombinasjonen som er fysikk! En god fysiker bør ha nær kontakt med både fenomenene og formalismen. Av praktiske og volummessige grunner har jeg valgt å legge mye vekt på matematikken for enkelte deler av fenomenene som presenteres, mens andre deler nesten er fri for matematikk.

Matematikken kommer inn på to forskjellige måter. Bevegelsen til f.eks. en gitarstreng kan beskrives matematisk som en funksjon av posisjon og tid. Funksjonen er en løsning av differensielligninger. En slik beskrivelse er grei nok, men har en ad hoc funksjon. Vet vi utslaget ved en viss tid, kan vi finne utslaget ved en senere tid. En slik beskrivelse er en nødvendighet for videre analyser, men har egentlig liten interesse ut over dette. I mekanikken kalles dette en *kinematisk* beskrivelse.

Det sies ofte at vi *i fysikken forsøker å forstå hvordan naturen fungerer*. Vi er altså ikke fornøyd med bare å kunne gi en matematisk beskrivelse av gitarstrengens bevegelse. Vi ønsker å gå et nivå bakenfor denne beskrivelsen. Hvordan kan vi “forklare” at en tynn stålstreng som strekkes så og så mye faktisk gir tonen C når vi klimprer på den? Det fascinerende er at vi ved hjelp av relativt få og enkle fysiske lover er i stand å forklare mange og tilsynelatende helt forskjellige fenomener. Det gir en ekstra tilfredsstillelse. Vi vil kalle dette en *mekanistisk* eller *dynamisk* beskrivelse.

Matematikken har tradisjonelt sett fått stor plass i forhold til utfordringen med å forstå mekanismer synes jeg. Dette tror jeg til dels skyldes at vi hittil stort sett har anvendt analytiske matematiske metoder for å løse differensielligningene som fremkommer. Når vi bruker analytiske metoder må vi riktignok innom mekanismene for å sette opp ligningene



Figur 1.1: *Svingninger og bølger inngår i et vell av fenomener vi opplever hver eneste dag, så som lyd og lys. Ut fra temmelig generelle prinsipper kan vi forklare hvorfor den vanligste regnbuen nettopp har en radius på 40-42 grader og er rød ytterst og at himmelen like utenfor regnbuen er noe mørkere enn himmelen like innenfor. De samme prinsippene gir oss også egenskapene ved den ytre regnbuen når det er to av dem.* Foto: Bjørn Lybekk.

som beskriver fenomenene. Fokus skyves imidlertid raskt over til utfordringene ved å løse differensialligningen og å drøfte den analytiske løsningen vi kommer fram til.

Denne fremgangsmåten har flere begrensninger. For det første forsvinner fokus fra de bakenforliggende ligningene som forteller om viktige mekanismer for at en bølge skal kunne oppstå. For det andre er det bare noen ganske få forenklede problemstillinger vi da er i stand til å takle, ellers blir ligningene for vanskelige å løse analytisk. Vi må da ofte nøye oss med å betrakte løsninger med forenklede randbetingelser og/eller løsninger som først gjelder etter at transiente forløp har dødd ut.

Det betyr at fysikere gjennom mange generasjoner sitter igjen med forenklede bilder av svingninger og bølger og tror at disse gjelder generelt. For eksempel er min erfaring at det er urovekkende mange fysikere som tror at elektromagnetiske bølger generelt er synonymt med plane elektromagnetiske bølger. De antar at denne forenklede løsningen er en generell løsning som kan anvendes overalt. Ved å fokusere på numeriske løsningsmetoder vil vi forklare hvorfor dette er feil.

1.0.1 Numeriske metoder

Det skjer for tiden en dramatisk omlegging av fysikkundervisning i verden. Studenter er nå vant til å bruke datamaskiner og omtrent alle har sin egen, eller har lett adgang til en datamaskin. Dataprogrammer og programmeringsverktøy er blitt mye bedre enn de var for få tiår siden, og det er utviklet og systematisert numeriske metoder vi kan benytte oss av. Det betyr at bachelorstudenter tidlig i studiet kan ta i bruk like avanserte metoder som tidligere bare ble brukt innen snevre forskningsområder på master- og PhD-nivå. Det betyr at de kan arbeide med fysikken på en annen og mer spennende måte enn tidligere.

Riktignok må vi også nå sette opp differensialligninger og løse dem, men numeriske løsningsmetoder forenkler arbeidet betydelig. Følgen er at vi kan leke oss med å beskrive ulike mekanismer på forskjellig vis og studere hvordan løsningene avhenger av modelleringen vi starter ut med. Videre åpner numeriske løsningsmetoder opp for mye mer virkelighetsnære problemstillinger enn tidligere, fordi en ”stygg” differensialligning ikke er særlig vanskeligere å løse numerisk enn en enkel. Vi kan for eksempel legge inn en ikke-lineær beskrivelse av friksjon og få ut resultatene omrent like enkelt som uten friksjon,

mens vi overhodet ikke kunne løst problemet rent analytisk.

Det betyr at vi nå kan legge mindre vekt på ulike løsningsstrategier for differensialligninger, og heller bruke tiden vi sparer på dette til å ta tak i mer virkelighetsnære problemstillinger. Jeg selv tilhører en generasjon som lærte å finne kvadratroten av et tall ved direkte utregning. Etter at kalkulatoren kom på markedet, har jeg ikke hatt behov for denne kunnskapen. Vi er nå i en lignende fase innen fysikk og matematikk. Bruker vi f.eks. dataprogrammene Maple eller Mathematica, får vi ut analytiske uttrykk for et vell av differensialligninger, og dersom en differensialligning ikke har en grei analytisk løsning, kan problemet løses numerisk. Noen ferdigheter fra tidligere år har derfor mindre verdi i dag, mens andre ferdigheter har fått større verdi.

Denne boka er skrevet i omveltingstiden hvor vi skifter fra å bruke bare analytiske metoder i bachelorkurs, til en situasjon der datamaskiner inngår som et naturlig hjelpemiddel både pedagogisk og faglig. Vi kommer til å dra direkte nytte av dette, ikke bare for å bygge opp en kompetanse som hver enkelt vil ha glede av i senere yrker, men også som et pedagogisk hjelpemiddel for å forstå stoffet bedre. Ved numeriske beregninger kan vi lettere fokusere på selve algoritmene, basisligningene, enn ved analytiske metoder. Dessuten kan vi ta fatt i et vell av interessante problemstillinger vi ikke kunne studere bare ved analytiske metoder, noe som bidrar til økt forståelse. Numeriske metoder gir oss dessuten verktøy til å analysere funksjoner/signaler på en elegant måte, slik at vi nå kan få ut mye mer relevant informasjon enn vi kunne med de metodene som var tilgjengelig tidligere.

Bruk av numeriske metoder er interessant også fordi vi lettere kan gi “forskningsbasert undervisning”. Studentene vil være i stand til å gjøre beregninger tett opp til det som faktisk gjøres i forskning i dag. Det er nok av temaer å ta tak i, for det skjer en enorm utvikling innen ulike bølgebaserete fenomener for tiden. Eksempelvis kan det nevnes at vi bruker mange transdusere som ligger i en rekke (“array”) i ultralyddiagnostikk, oljeleting, ekkolodd og radarteknologi. I alle disse eksemplene brukes det veldefinerte faseforskjeller for å få fram romlige variasjoner på elegante måter. Videre kan vi ved såkalte fotoniske krystaller og andre hi-tech strukturer på nanonivå oppnå bedre oppløsning i målinger enn tidligere, til og med bedre enn teoretiske grenseverdier vi trodde på for få år siden. Videre utnytter vi i dag ikke-lineære prosesser som ikke var kjent for få tiår siden. Det er utrolig mye spennende som skjer i fysikken nå, og mange vil møte problemstillinger og metoder som tas opp i denne boka også etter endt studium.

1.0.2 Bakgrunn for boka

Denne boka ble skrevet for bruk i kurset “FYS2130 Svingninger og bølger” ved Fysisk institutt, Universitetet i Oslo. Kurset tas av fjerde semester bachelor-studenter som har vært gjennom klassisk mekanikk og elektromagnetisme på forhånd. Disse kursene har dog vært strippet for mange svinge- og bølgfenomener.

Det finnes mange bøker innen svingninger og bølger på markedet, men ingen passet godt for de læringsmål som var satt opp for kurset. Siden Universitetet i Oslo er blant de første som innførte bruk av numeriske metoder som en integrert del i nesten alle kurs i matematikkunge realfag, var det svært få egnede lærebøker. Når vi som en del av samme strategi valgte å gi grundigere og mer praktisk innføring i noen numeriske metoder enn det som har vært vanlig tidligere, fant vi ingen lærebøker på markedet som kunne dekke hele stoffet som skulle undervises.

Vi levde noen år med en situasjon der vi brukte et standardverk i fysikk som basis, og lagde ekstra kompendier for de delene av pensum som læreboka ikke dekket. Det ble en uheldig blanding, og jeg valgte derfor å lage en mer helhetlig løsning for kurset vårt. Etter hvert har jeg blitt mer og mer overbevist om at kombinasjonen med å bruke både

analytisk matematikk, numeriske metoder, fokus på hverdagsfenomener, og noen state of the art eksempler, kan være av interesse godt ut over et enkelt kurs ved Universitetet i Oslo. Jeg håper derfor at denne boka kan være interessant for langt flere enn våre egne studenter.

Jeg vil benytte anledningen til å takke alle som har bidratt til denne boka. Jeg vil spesielt nevne Borys Jagielski, Knut Kvaal, Jan Henrik Wold, Karl A. Maaseide og kolleger ved Fysisk institutt, og i særdeleshet Anders Johnsson for svært nytte tips og kommentarer. Morten Hjorth-Jensen takkes for generell støtte og interesse for undervisningsspørsmål over mange år. Jeg vil også takke tidligere lærere så som Svenn Lilledal Andersen og Kristoffer Gjøtterud for å skape et miljø hvor min fysikkforståelse fikk vokse og utvikle seg, og til Gunnar Handal som utfordret meg på en fin måte innen universitetspedagogikk. Støtte fra Faglitterære Forfatter og Oversetterforening gjorde det mulig å kjøpe meg fri fra undervisning to høstsemestre for å frigi tid til å skrive boka og lage de fleste illustrasjonene. Aller mest takker jeg for stor forståelse og tålmodighet fra min kjære Kirsten og barna våre i perioder da jeg bidro lite til familielivet.

1.0.3 Organisering av stoffet

I boka har jeg forsøkt å gi standard beskrivelser for en god del av stoffet, men har forsøkt å legge mer vekt enn vanlig på å vise begrensinger i de beskrivelsene vi da kommer fram til. Jeg trekker analogier på tvers av flere ulike fenomener der jeg synes det er interessant. Jeg bruker erfaringer fra bruk av numeriske metoder for å få en noe dypere forståelse av enkelte fenomener enn det som er vanlig på dette nivået. Jeg innser at vi ennå ikke klarer å utnytte bruk av numeriske metoder og algoritmisk tenkning i så stor grad som vi burde. Dette er en modningsprosess som tar lang tid. Etter som vi som fysikklærere prøver å la studentene utnytte dagens teknologi, vil læringsmiljø og -metoder endre seg til dels betydelig de neste 10-20 årene. Det viktigste er at vi åpner opp for studentenes kreativitet, så vil utviklingen tvinge seg fram av seg selv.

yBoka er skrevet i LaTeX for å forenkle skriving av matematiske uttrykk. Dessverre medfører det svært begrensede layoutmuligheter. Viktige deler av kjernestoffet er markert med fargefelt. Eksempelstoffet er ofte skrevet med noe mindre skrift og viser hvordan kjernestoffet kan anvendes i ulike sammenhenger. Læringsmål peker på de viktigste deler av hvert kapittel, og oppgaver gis slik at hver enkelt kan teste sin forståelse av stoffet som er presentert.

Det finnes tre typer oppgaver i boka. De fleste er forståelses-/diskusjonsspørsmål og regneoppgaver. Det er viktig å prøve seg på begge disse typer oppgaver. Vi ønsker å stimulere hver enkelt til å lære seg å *argumentere* for hvordan en oppgave kan analyseres og hvilke lovmessigheter som ønskes anvendt. Et “riktig svar” uten tilstrekkelige begrunnelser, er egentlig lite verdt. Det kan legges til at mange oppgaver gis uten at alle størrelser som inngår er gitt. Meningen er at vi da må søke f.eks. på internett for å finne de størrelsene vi trenger. Dette er en naturlig del av det å arbeide med fysikk i dag.

En tredje type oppgave kalles AKBD-oppgaver. Disse er karakterisert ved at de er meget upresise, men peker på en problemstilling der det ligger gjemt et problem som kan løses. Oppgaven blir da å analysere (A) den ulne problemstillingen, konkretisere (K) hva man faktisk ønsker å ta tak i, og så foreta beregningene (B) og diskutere resultatene (D). Hensikten med disse oppgavene er å trenere studentene i å ta tak i problemstillinger på egen hånd, uten at noen har defineret og konkretisert hvilke konkrete ting som må gjøres.

Noen få oppgaver er store, sammensatte og gjerne litt “åpne” oppgaver som egner seg som prosjektoppgaver. De er viktige for at studenter skal lære seg å arbeide med problemer på en måte som kreves etter endt utdanning.

1.0.4 Format - rettigheter

I 2011 kom denne boka ut på Tapir forlag, men prisen ble høy siden opplaget var lite og vi ønsket fargetrykk. Nå tilbys boka gratis til alle, både innenfor og utenfor Universitetet i Oslo, i form av pdf-filer. Brukes tekst og figurer fra boka i andre sammenhenger, må kilden til stoffet/figurene oppgis. Videre er jeg frimodig nok til å be om et lite økonomisk bidrag dersom du finner deler av boka nyttig. Bokprosjektet har hittil ført til endel utgifter. Jeg avsto fra royalties for å få boka så billig som mulig og støtten fra Faglitterære Forfatter og Oversetterforening gikk til Fysisk institutt for å dekke vikar. Dersom du føler for det, ville jeg derfor sette pris på 50 - 100 kroner til min bankkonto 0532 923 0491. Føler du ikke for å betale, er det helt greit det også.

1.0.5 Støttelitteratur

Det er skrevet mange bøker om svingninger og bølger, men ingen tidligere med samme kombinasjonen av emner som denne. Det er ofte nyttig å lese hvordan andre har beskrevet et emne, og av den grunn anbefaler vi at du samtidig som du leser denne boka leser i andre bøker og sjekker f.eks. Wikipedia og andre seriøse beskrivelser på web. Her er noen bøker som kan være av interesse:

- Jonas Persson: "Vågrörelselslära, akustik och optik". Studentlitteratur, 2007.
- H.J.Pain: "The Physics of Vibrations and Waves". 6th edition. Wiley, 2005.
- A.P.French: "Vibrations and Waves". W.W.Norton & Company, 1971.
- Dudley H. Towne: "Wave Phenomena". Dover, 1967.
- John R. Pierce: "Almost all about Waves". Dover, 1974.
- Daniel Fleisch: "A Student's Guide to Maxwell's Equations". Cambridge University Press, 2008.
- Tor Halmrast: "Klangen" (bok om lyd, musikk og akustikk), Eget forlag, kontaktes via torhalm@online.no, 2013.
- Sir James Jeans: "Science & Music". Dover, 1968 (oppriinnelig fra 1937!).
- Eugene Hecht: "Optics", 4th edition. Addison Wesley, 2002.
- Geoffrey Brooker: "Modern Classical Optics". Oxford University Press, 2003.
- Grant R. Fowles: "Introduction to Modern Optics". 2nd editon. Dover Publications, 1975.
- Ian Kenyon: "The Light Fantastic". 2nd edition. Oxford University Press, 2010.
- Karl Dieter Möller: "Optics. Learning by Computing, with Model Examples Using MathCad, Matlab, Mathematica, and Maple". 2nd editon. Springer 2007.
- Peter Coles: "From Cosmos to Chaos". Oxford University Press, 2010.
- Helmut Ormestad: "Svingninger og bølger". Universitetsforlaget 1964 (også interessant fra en historisk synsvinkel.)

Lykke til!

Jeg håper at når dere jobber med denne boka, vil dere oppleve at svingninger og bølger er en morsom del av fysikken, og sitte igjen med en dypere forståelse enn dere hadde før dere startet.

Blindern, januar 2013

Arnt Inge Vistnes

Kapittel 2

Fri og dempede svingninger



Foucault-pendelen i Fysikkbygget på Blindern.

Svingninger er en mer sentral del av fysikk enn folk ofte tenker over. Pendelbevegelse er det mest kjente eksemplet på svingninger. Svingninger inngår imidlertid også i alle bølgefенomener. Vårt syn, vår hørsel, ja til og med nerveledning i kroppen, har nært tilknytning til svingninger, for å ikke snakke om nesten all kommunikasjon via teknologiske hjelpe midler. I kapittel 1 skal vi se på de enkleste matematiske beskrivelsene av svingninger. De er enkle, men ikke undervurder dem! Det finnes små detaljer innimellom som kan synes ubetydelige i første omgang, men som er viktige for å forstå mer kompliserte fenomener vi kommer til senere i boka.

“Svingninger” er av praktiske grunner delt opp i to kapitler. Første kapittel omhandler “passive” svingninger (systemer som ikke blir utsatt for en periodisk påvirkning). Andre kapittel tar for seg “tvungne svingninger”. Nøkkelord for kapittel 1 er: krefter, virkningsmekanismer/fysiske lover, amplituder, frekvenser, faser, tidsutvikling, svingeligningen, annen ordens differentialligning, lineær og ikke-lineære ledde.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 04012013

I mekanikken skiller vi mellom kinematikk og dynamikk, og skillet er relevant også når vi betrakter svingninger. Innen kinematikken er fokus først og fremst å *beskrive* en bevegelse. Beskrivelsen er gjerne selve *løsningen* av differensialligninger eller eksperimentelle målinger. De bakenforliggende fysiske lovene trekkes ikke inn.

I dynamikken, derimot, setter vi opp differensialligningene for bevegelsene ut fra kjente fysiske lover. Ligningene løses enten ved analytiske eller numeriske metoder, og vi sammenholder løsningene med modelleringen av fysikken vi startet ut med. Ettertrakter vi fysisk forståelse, er det dynamikken som er mest interessant, men kinematikken kan også være nyttig for å bli vant med relevant matematisk beskrivelse og størrelsene som inngår.

2.1 Kinematikk

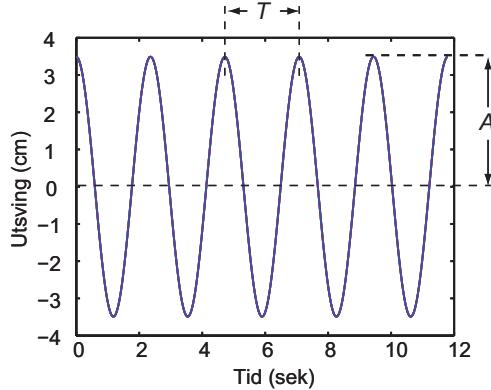
Hvordan beskriver vi en svingning? La oss ta et eksempel: Et lodd henger i en fjær og svinger vertikalt opp og ned.

Den kinematiske beskrivelsen kan da være som denne: Loddet svinger omkring et likevektspunkt. Maksimalt utslag A relativt til likevektspunktet kalles svingningens *amplitude*. Tiden loddet bruker på hver fulle svingning kalles *periodetiden* T . *Svingefrekvensen* f er den inverse av periodetiden, dvs. $f \equiv 1/T$ og måles i inverse sekunder eller hertz (Hz).

For et passe tungt lodd og amplitude for fjæra, vil utsvinget (posisjon som funksjon av tid) $Z(t)$ tilnærmet følge en matematisk sinus/cosinus-bevegelse:

$$Z(t) = A \cos(2\pi t/T)$$

såfremt utslaget Z er maksimalt ved det tidspunktet vi velger som nullpunkt for t . Med "passe tungt lodd" menes at utslaget er slik at fjæra alltid er noe strukket, men aldri så mye at den deformeres. Vi antar at Hookes lov gjelder for hele bevegelsen.



Figur 2.1: En harmonisk svingning karakteriseres ved amplitude, frekvens og fase, se teksten.

Tas det hensyn til at det ikke finnes et absolutt nullpunkt for posisjon i rommet, og heller ikke noe absolutt nullpunkt for tid, burde vi kanskje heller skrive:

$$Z(t) - \bar{Z} = A \cos(2\pi(t - t_0)/T)$$

hvor \bar{Z} er middelverdien for posisjon, og t_0 er et tidspunkt der Z har sin maksimale verdi. For å gjøre det matematiske uttrykket så kort som mulig, *velges* gjerne middelverdien for svingningen som referansepunkt i rommet ved å skrive $z(t) = Z(t) - \bar{Z}$, og det innføres et

faseledd $\phi = -2\pi t_0/T$ i stedet for å referere til tiden t_0 da utslaget var størst. Resultatet blir:

$$z(t) = A \cos(2\pi t/T + \phi)$$

Størrelsen $2\pi/T$ går igjen i mange beskrivelser av svingbevegelser, og vi forenkler skrivingen mye ved å definere en vinkelfrekvens ω som følger:

$$\omega \equiv 2\pi/T = 2\pi f$$

Den enkle "harmoniske" svingbevegelsen kan da beskrives på flere ekvivalente måter:

$$z(t) = A \cos(\omega t + \phi) \quad (2.1)$$

$$\begin{aligned} z(t) &= A \cos(\omega t) \cos(\phi) - \\ &\quad A \sin(\omega t) \sin(\phi) \end{aligned} \quad (2.2)$$

$$z(t) = B \sin(\omega t) + C \cos(\omega t) \quad (2.3)$$

$$z(t) = \Re \{ A e^{i(\omega t + \phi)} \} \quad (2.4)$$

$$z(t) = \Re \{ \mathcal{D} e^{i\omega t} \} \quad (2.5)$$

hvor $\Re \{ \}$ betyr at realdelen av det komplekse uttrykket i parantesen, og \mathcal{D} er et komplekst tall. Eulers formel for eksponentialfunksjonen (kompleks form) er brukt i de siste to uttrykkene. Eulers formel sier:

$$e^{i\alpha} = \cos(\alpha) + i \sin(\alpha)$$

Denne formelen danner forresten grunnlaget for en grafisk representasjon av en harmonisk bevegelse: Tenk deg først at vi tegner en vektor med lengde 1 i et plan. Startpunktet til vektoren legges i origo, og vektoren danner en vinkel α med x-aksen. Vektoren kan da skrives på følgende måte:

$$\cos(\alpha)\hat{x} + \sin(\alpha)\hat{y}$$

hvor \hat{x} er enhetsvektor i x -retning, og tilsvarende for y . Likheten med det forrige uttrykket er slående, forutsatt at realdelen av uttrykket oppfattes som komponenten i x -retning og imaginærdelen som komponenten i y -retning.

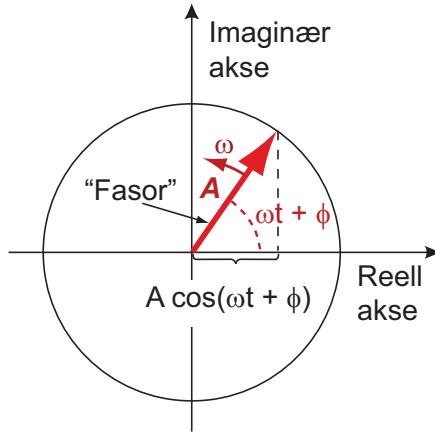
Dette grafiske vektorbildet kan utvides til også å representere en harmonisk svingning. Vi bruker da en vektor med en lengde som svarer til amplituden for den harmoniske bevegelsen. Vektoren roterer med en fast vinkelfrekvens ω om origo. Vinkelen mellom vektoren og x -aksen er til enhver tid $\omega t + \phi$. Da angir x -komponenten til vektoren til enhver tid det momentane utslaget til den harmoniske svingningen. En slik grafisk beskrivelse er illustrert i figur 2.2 og kalles en *fasorbeskrivelse* av bevegelsen.

Fasorer er svært nyttige når flere bidrag til en bevegelse eller et signal med samme frekvens skal summeres. Summen av alle bidragene finnes ved vektoraddisjon. Spesielt i vekselstrømsteknikk, når spenninger over ulike kretskomponenter summeres, er fasorer et utmerket hjelpemiddel. Vi kommer tilbake til dette senere. Også i andre sammenhenger er fasorer nyttige, men først og fremst når alle bidrag i en summasjon har samme vinkelfrekvens.

Det er viktig å lære seg alle de matematiske uttrykkene (2.1) - (2.5) for enkel svingbevegelse slik at uttrykkene straks kan gjenkjennes når de dukker opp. Det er også viktig raskt å kunne konvertere mellom de ulike formene. Boka er full av disse uttrykkene!

2.2 Dynamisk beskrivelse av et mekanisk system

En fjær følger ofte Hookes lov: Utslaget fra likevektspunktet er proporsjonalt med kraften fjæra trekkes med.

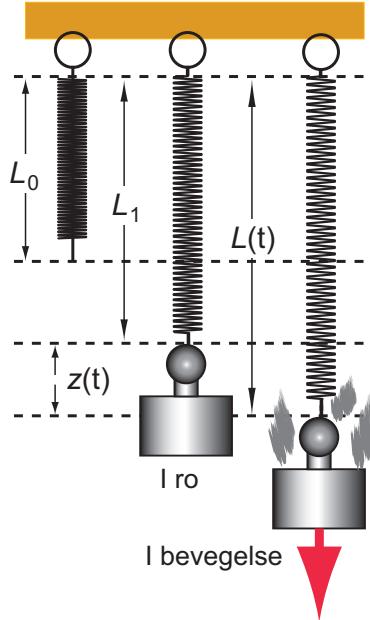


Figur 2.2: En fasor er en vektor med en gitt lengde. Fasoren roterer med en gitt vinkel-frekvens og fase. Figuren viser fasorens posisjon ved ett tidspunkt. Se teksten.

Anta at fjæra henger loddrett uten lodd i enden. Den har da en lengde L_0 . Henges et lodd med masse m i fjæra, og vi venter til systemet har kommet til ro, vil fjæra ha en ny lengde L_1 som tilfredsstiller ligningen

$$k(L_1 - L_0) = mg$$

hvor den eksperimentelt bestemte størrelsen k kalles fjærkonstanten, og g er tyngdens akselerasjon og regnes her som en konstant (ser bort fra variasjon med høyden).



Figur 2.3: Definisjon av ulike lengder på fjæra med og uten lodd, se teksten.

Trekkes nå loddet litt ned og slippes, vil kraften som virker på loddet til enhver tid være

$$F(t) = k(L(t) - L_0) - mg$$

hvor $L(t)$ er den momentane lengden av fjæra. Kombineres de siste to ligningene, følger

$$\begin{aligned} F(t) &= k(L(t) - L_1) + k(L_1 - L_0) - mg \\ &= k(L(t) - L_1) \end{aligned}$$

Utslaget relativt til likevektpunktet, dvs $L(t) - L_1$ døpes om til $-z(t)$. Kraften som virker på loddet blir da

$$F(t) = -kz(t)$$

Det negative fortegnet skyldes at kraft og posisjon relativt til likevektpunktet til enhver tid har motsatt retning.

Ifølge Newtons 2. lov er summen av krefter som virker på loddet lik massen til loddet multiplisert med den momentane akselerasjonen:

$$F(t) = m\ddot{z}(t) = -kz(t)$$

I dette uttrykket er både fjærkraft og gravitasjonskraft tatt med. \ddot{z} er den dobbelt-deriverte av z med hensyn på tid, altså akselerasjonen i vertikal retning oppover):

$$\ddot{z} \equiv \frac{d^2 z}{dt^2}$$

Ligningen kan nå skrives:

$$\ddot{z}(t) = -\frac{k}{m}z(t) \quad (2.6)$$

Dette er en annen-ordens homogen differensiell ligning med konstante koeffisienter. Fra matematikken vet vi at denne har en generell løsning

$$z(t) = B \sin \left(\sqrt{\frac{k}{m}} t \right) + C \cos \left(\sqrt{\frac{k}{m}} t \right)$$

hvor B og C er to konstanter (med dimensjon lengde). Vi gjenkjenner denne løsningen som ligning (2.3) ovenfor såfremt vi setter vinkelfrekvensen ω til

$$\omega = \sqrt{\frac{k}{m}}$$

Konstantene B og C bestemmes ut fra initialbetingelsene. Både amplitude og fase til svingbevegelsen blir da fastlagt.

Vinkelfrekvensen ω er en praktisk størrelse å operere med i de matematiske uttrykkene. Ved observasjoner av et svingende system er det imidlertid mest praktisk å benytte seg av frekvens f og periodetid T . Sammenhengen er:

$$f = \frac{\omega}{2\pi}$$

$$T = \frac{1}{f} = \frac{2\pi}{\omega}$$

For den mekaniske fjærpendelen følger da:

$$f = \frac{1}{2\pi} \sqrt{\frac{k}{m}}$$

$$T = 2\pi \sqrt{\frac{m}{k}}$$

Hva har vi lært i dette underkapitlet? Jo, vi har sett at et lodd som henger i en fjær og påvirkes av fjærkraft og tyngdekraften vil svinge pent og pyntelig i en enkel harmonisk bevegelse, opp og ned, med en viss amplitud og periodetid. Vi har derved "forklart" svingbevegelsen ved å ta utgangspunkt i Hookes lov og Newtons annen lov.

Den kinematiske beskrivelsen vi hadde i delkapittel 1.1, er identisk med *løsningen* av den dynamiske ligningen vi satte opp i dette delkapittelet ut fra Newtons 2. lov.

2.3 Dempede svingninger

Ingen makroskopiske svingninger varer ved i det uendelige uten at det tilføres energi. Grunnen er at det alltid vil være krefter som forsøker å motsette seg bevegelsen. Vi kaller disse for friksjonskrefter.

Friksjonskrefter er ofte ganske vanskelige å forholde seg til, for de representerer komplisert fysikk i grenseland mellom atomær og makroskopisk beskrivelse. En grunnleggende forståelse av friksjon har vi først *begynt* å få de siste tiår, fordi denne delen av fysikken er nesten helt avhengig av omfattende modellering ved hjelp av datamaskiner.

Friksjon i luft er kompleks og vi bør ha med minst to ledd for å beskrive friksjonen:

$$F_f = -bv - Dv^2$$

hvor v er hastigheten (med retning), og b og D er positive konstanter, begge er en slags friksjonskoeffisienter.

Et uttrykk som også angir riktig fortegn og retning, er:

$$\vec{F}_f = -b\vec{v} - Dv^2 \frac{\vec{v}}{v} \quad (2.7)$$

Friksjonskraften \vec{F}_f virker med andre ord i motsatt retning av hastigheten \vec{v} .

Starter vi med et system med harmonisk bevegelse uten friksjon, og legger til friksjon som gitt i ligning (2.7), er det ikke mulig å finne en generell løsning ved hjelp av analytisk matematikk alene. Hvis problemet forenkles ved å sette friksjonskraften kun lik $-bv$, er det likevel mulig å bruke analytiske metoder. Løsningen er brukbar for langsomme bevegelser i luft. For små hastigheter vil nemlig ledet Dv^2 være mindre enn ledet bv i ligning (2.7), slik at v^2 -leddet kan neglisjeres.

[♠ ⇒ Kommentar: $-Dv^2$ er et ikke-lineært ledd som ofte har sammenheng med turbulens, et av de vanskelige områdene innen fysikk, ofte knyttet til kaotiske systemer. Friksjon av denne typen avhenger av en rekke parametre som delvis kan trekkes sammen i et såkalt “Reynolds tall”. I enkelte beregninger må størrelsen D erstattes av en funksjon $D(v)$ dersom ligning (2.7) skal anvendes. Alternativt kan Navier-Stokes ligning brukes som et utgangspunkt. Rimelig nøyaktige beregninger av friksjonen til en ball, fly eller rakett kan bare gjennomføres ved bruk av numeriske metoder. (Interesserte kan finne mer stoff på Wikipedia under søkeordene “Reynolds number” og “Navier-Stokes equation”).] ← ♠]

Siden det ikke er noen kunst å løse den aktuelle *forenkled*e differensialligningen, tar vi den utfordringen! Løsningsmetoden kan være nyttig å kjenne til fordi vi vil bruke komplekse eksponenter og får vist formalismens eleganse. Dessuten er dette standard, klassisk lærebokfysikk, og resultatene er attpåtil nyttige i mange sammenhenger. Selve den matematiske fremgangsmåten finner vi også igjen i mange andre deler av fysikk.

Utgangspunktet er som før Newtons annen lov, og vi anvender den for et lodd som svinger sakte opp og ned i enden av en fjær i luft. Ligningene kan nå skrives:

$$\begin{aligned}\Sigma F &= ma \equiv m\ddot{z} \\ -kz(t) - b\dot{z}(t) &= m\ddot{z}(t) \\ \ddot{z}(t) + \frac{b}{m}\dot{z}(t) + \frac{k}{m}z(t) &= 0\end{aligned}\tag{2.8}$$

Dette er en homogen annenordens differensiell ligning, og vi forsøker oss med en løsning av typen:

$$z(t) = Ae^{\alpha t}\tag{2.9}$$

MERK: Her antas såvel A som α å være komplekse tall.

Derivering av eksponentialfunksjonen (2.9), innsetting i (2.8) og til slutt forkortning med eksponentialleddene og faktoren A gir

$$\alpha^2 + \frac{b}{m}\alpha + \frac{k}{m} = 0$$

Vi omdørper brøkene for å få et enklere sluttuttrykk:

$$\frac{b}{m} \equiv 2\gamma\tag{2.10}$$

$$\frac{k}{m} \equiv \omega^2\tag{2.11}$$

Ligningen blir da:

$$\alpha^2 + 2\gamma\alpha + \omega^2 = 0$$

Dette er en vanlig kvadratisk ligning som har følgende løsning (når faktoren 2 i løsningsuttrykket er forkortet bort):

$$\alpha = -\gamma \pm \sqrt{\gamma^2 - \omega^2}\tag{2.12}$$

Det skiller nå mellom tre ulike former for løsninger etter fortegn under rottegnet:

- $\gamma > \omega$: Overkritisk demping.

Dersom friksjonskraften blir for stor i forhold til km , får vi overkritisk demping. Kriteriet for overkritisk demping $\gamma > \omega$ er matematisk ekvivalent med: $b > 2\sqrt{km}$.

I dette tilfellet blir såvel A som α i ligning (2.9) reelle tall, og en generell løsning blir:

$$z(t) = A_1 e^{(-\gamma + \sqrt{\gamma^2 - \omega^2})t}\tag{2.13}$$

$$+ A_2 e^{(-\gamma - \sqrt{\gamma^2 - \omega^2})t}\tag{2.14}$$

hvor A_1 og A_2 bestemmes av initialbetingelsene og har betydning for amplitud og tidsforløp for bevegelsen.

Dette er en sum av to eksponentielt avtakende funksjoner der den ene går raskere mot null enn den andre. Det er ikke mye som minner om svingninger i bevegelsen.

Merk at A_1 og A_2 for visse initialbetingelser kan ha forskjellige fortegn, og at tidsforløpet derfor kan by på overraskelser!

- $\gamma = \omega$: Kritisk demping.

Friksjonskraften og den effektive fjærkraften matcher nå hverandre på en slik måte at bevegelsen blir spesielt enkel. Med utgangspunkt i ligning (2.9) og (2.12), finner vi én løsning: Den kan beskrives som en enkel eksponentialsfunksjon:

$$z(t) = Ae^{-\gamma t}$$

Fra matematikken er det kjent at det må finnes *to* valgfrie konstanter i en generell løsning av en annenordens differensielligning for å kunne tilfredsstille initialbetingelsene. Vi har derfor ikke funnet den fulle løsningen ennå. Den finnes ved å benytte en metode kalt “reduksjon av orden”. Vi bruker da en prøveløsning av typen:

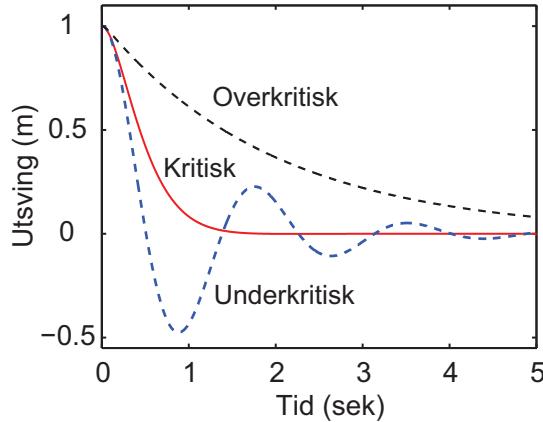
$$z(t) = f(t)e^{\alpha t}$$

Settes denne prøveløsningen inn i vår differensielligning (2.8) når $\gamma = \omega$, finner vi nokså raskt at \ddot{f} må være lik 0. Etter to ganger integrering mhp t finner vi til slutt et egnert uttrykk for $f(t)$.

Den endelige generelle løsningen av differensielligning (2.8) for kritisk demping er da:

$$z(t) = Ae^{-\gamma t} + Bte^{-\gamma t} \quad (2.15)$$

Kritisk demping svarer i mange tilfeller til den raskeste dempingen for et system, og er den som etterstrebtes for f.eks. bilfjærer.



Figur 2.4: Eksempler på overkritisk, kritisk og underkritisk demping av en svingning som ville vært harmonisk uten friskjon. Friksjonen er økt med en faktor fire mellom hver av beregningene: Underkritisk, kritisk og overkritisk demping.

- $\gamma < \omega$: Underkritisk demping.

I dette tilfellet blir α i ligning (2.12) kompleks, og det medfører at løsningen vil inneholde såvel en eksponentielt avtakende faktor som et sinus-cosinus-ledd. Fra ligning (2.12) får vi da:

$$\alpha = -\gamma \pm \sqrt{\gamma^2 - \omega^2} \quad (2.16)$$

$$= -\gamma \pm i\omega' \quad (2.17)$$

hvor $\omega' \equiv \sqrt{\omega^2 - \gamma^2}$ er et reelt tall. Den generelle løsningen blir da:

$$z(t) = e^{-\gamma t}(\mathcal{A}e^{i\omega't} + \mathcal{B}e^{-i\omega't})$$

hvor \mathcal{A} og \mathcal{B} er komplekse tall, og \Re betyr at vi tar realdelen av uttrykket.

Løsningen for underkritisk demping kan skrives på en enklere måte slik:

$$z(t) = e^{-\gamma t} A \cos(\omega' t + \phi) \quad (2.18)$$

Her må konstantene A og ϕ bestemmes for at den generelle løsningen skal bli en konkret løsning for et gitt fysisk system. Utslaget vil oscillere på begge sider av likevektpunktet mens amplituden avtar mot null. Svingefrekvensen er lavere enn om vi ikke hadde demping (noe som er naturlig siden friksjonen forsøker å bremse på all bevegelse).

Det er vanlig i lærebøker å ta med en figur som viser typisk tidsforløp for en dempet bevegelse, og vi følger opp tradisjonen med figur 2.4. Det bør imidlertid bemerkes at slike figurer kan være svært misvisende, for de tar ofte utgangspunkt i at hastigheten er lik null ved tiden $t = 0$ (som i vår figur). I en oppgave sist i dette kapitlet ber vi deg om å undersøke hvordan en overkritisk bevegelse ser ut for noen andre initialbetingelser. Løser du den oppgaven, vil du se at løsningsmengden er mer mangfoldig enn de tradisjonelle figurene tilslører!

2.4 Superposisjon og ikke-lineære ligninger *

Da vi forsøkte å finne ut hvordan en dempet svingning utvikler seg med tiden, tok vi utgangspunkt i differensialligningen:

$$\ddot{z}(t) + \frac{b}{m}\dot{z}(t) + \frac{k}{m}z(t) = 0$$

og vi fant en generell løsning som bestod av to ledd. For overkritisk demping så løsningen slik ut:

$$z(t) = A_1 e^{(-\gamma + \sqrt{\gamma^2 - \omega^2})t} + A_2 e^{(-\gamma - \sqrt{\gamma^2 - \omega^2})t}$$

hvor γ og ω er definert i ligning (2.10) og (2.11) ovenfor. For enkelhets skyld setter vi:

$$f_1(t) = e^{(-\gamma + \sqrt{\gamma^2 - \omega^2})t}$$

og

$$f_2(t) = e^{(-\gamma - \sqrt{\gamma^2 - \omega^2})t}$$

Da kan vi skrive den generelle løsningen slik:

$$z(t) = A_1 f_1(t) + A_2 f_2(t)$$

Dersom vi ikke tar hensyn til initialbetingelsene, vil f_1 og f_2 være to uavhengige løsninger av den opprinnelige differensialligningen (2.8). En annenordens differensialligning har den egenskapen at løsningsrommet blir utspent av to uavhengige løsninger. Løsningene har med andre ord bare to frihetsgrader.

Vi kan betrakte differensialligningen på en litt annerledes måte, nemlig ved å innføre såkalte "operatorer". I dette tilfellet vil vi kunne skrive differensialligningen på følgende måte:

$$\left(\frac{d^2}{dt^2} + \frac{b}{m} \frac{d}{dt} + \frac{k}{m} \right) z(t) = 0$$

Innholdet i den første parantesen kan betraktes som en “oppskrift” for hvordan vi skal behandle uttrykket $z(t)$. Denne oppskriften er det vi kaller en operator. Operatoren kan angis slik:

$$\hat{F} \equiv \frac{d^2}{dt^2} + \frac{b}{m} \frac{d}{dt} + \frac{k}{m}$$

Differensialligningen angitt som en operatorligning vil da se slik ut:

$$\hat{F}z(t) = 0$$

Vi har sett at denne ligningen har to uavhengige løsninger, slik at:

$$\hat{F}f_1(t) = 0$$

og

$$\hat{F}f_2(t) = 0$$

Dersom f_1 multipliseres med en konstant, vil også resultatet bli en løsning av differensialligningen

$$\hat{F}(A_1 f_1(t)) = 0$$

Tilsvarende:

$$\hat{F}(A_2 f_2(t)) = 0$$

Alt dette synes trivielt, men nå kommer vi straks til noe mer interessant.

Vi ser at:

$$\hat{F}(A_1 f_1(t) + A_2 f_2(t)) = \hat{F}(A_1 f_1(t)) + \hat{F}(A_2 f_2(t)) \quad (2.19)$$

Matematisk sier vi at \hat{F} er en lineær operator.

Ligning (2.19) viser at dersom vi har én løsning av differensialligningen, og en annen løsning av samme ligning, så vil også summen av løsningene (og alle mulige lineære kombinasjoner av de to) være løsninger av differensialligningen.

Dette kalles “superposisjonsprinsippet”. Dette prinsippet går igjen i store deler av fysikken (ikke minst i kvantefysikken).

Tidligere anså mange superposisjonsprinsippet som en fundamental egenskap ved naturen, men slik er det ikke. Grunnen til misforståelsen er kanskje at fysikere flest på den tiden bare jobbet med lineære systemer hvor superposisjonsprinsippet fungerer ok. I dag kan vi takket være datamaskiner og numeriske metoder ta fatt på fysiske systemer som tidligere var utilgjengelige. Det betyr at det har skjedd en “eksplosjon” innen fysikk de siste få tiår, og utviklingen er langt fra over.

La oss se hva som blir forskjellig når ikke-lineære beskrivelser benyttes. Med ikke-lineær beskrivelse mener vi f.eks. at krefter som beskriver et system ikke er lineært avhengig av posisjon eller hastighet. Da vi f.eks. fant løsningene for dempede svingninger, så vi at friksjon ofte må modelleres med minst to ledd:

$$F = -bv - Dv^2$$

Det siste leddet her er et ikke-lineært bidrag til kraften.

For å finne en analytisk løsning gjorde vi forenklingen å sette $D = 0$. Hvis vi hadde beholdt D , ville operatoren for den tilsvarende differensialligningen ha sett slik ut:

$$\hat{F} \equiv \frac{d^2}{dt^2} + \frac{b}{m} \frac{d}{dt} + \frac{D}{m} \left(\frac{d}{dt} \cdot \right)^2 + \frac{k}{m}$$

forutsatt at:

$$\frac{D}{m} \left(\frac{d}{dt} \cdot \right)^2 f(t) \equiv \frac{D}{m} \left(\frac{df(t)}{dt} \right)^2$$

Hvordan ville det da gå med superposisjonsprinsippet? Inntil nå har det vært slik at alle ledd i \hat{F} har fungert på enkleste måte når de er anvendt på en sum, for eksempel er

$$\frac{d^2}{dt^2} (f_1 + f_2) = \frac{d^2 f_1}{dt^2} + \frac{d^2 f_2}{dt^2}$$

Men når neste ledd ved friksjonsbeskrivelsen tas med, ser vi at:

$$\begin{aligned} \left(\frac{d}{dt} \cdot \right)^2 (f_1 + f_2) &= \left(\frac{df_1}{dt} + \frac{df_2}{dt} \right)^2 \\ &= \left(\frac{df_1}{dt} \right)^2 + 2 \frac{df_1}{dt} \frac{df_2}{dt} + \left(\frac{df_2}{dt} \right)^2 \\ &\neq \left(\frac{d}{dt} \right)^2 f_1 + \left(\frac{d}{dt} \right)^2 f_2 \end{aligned}$$

Med andre ord, når vi inkluderer et annen ordens ledd for å komplettere friksjonsbeskrivelsen, ser vi at superposisjonsprinsippet ikke lenger gjelder! Selv om vi finner en mulig løsning for et slikt svingesystem, og deretter en annen løsning, så er *ikke* summen av disse enkeltløsningene noen løsning av ligningsystemet.

Leddet Dv^2 er et “ikke-lineært” ledd, og når fysikken er slik at ikke-lineære ledd spiller en viss rolle, gjelder ikke superposisjonsprinsippet.

Ta en titt på “List of nonlinear partial differential equations” på den engelske Wikipedia så får du et levende inntrykk av hvor viktig ikke-lineære prosesser er blitt innen f.eks. fysikk i dag. Oversikten viser indirekte hvor mange flere problemstillinger vi kan studere i dag sammenlignet med hva som var mulig før få tiår siden. Til tross for dette, har vi fortsatt en lei tendens til å bruke en formalisme og tolke fenomener både innen vanlig klassisk fysikk og i kvantefysikk, som om verden var strengt lineær. Når det går noen tiår til, tror jeg fysikerne vil ha et så rikt erfaringsmateriale å bygge på, at tankegangen vil endre seg. Time will show!

2.5 Elektriske svingninger

Før vi går videre med tvungne svingninger skal vi utlede svingeligningen for en elektrisk svingekrets. Hensikten er å vise at matematikken blir helt analog til den vi brukte i det mekaniske systemet.

I elektromagnetismen inngår først og fremst tre karakteristiske kretselementer: Resistanser, induktanser (spoler) og kapasitanser (kondensatorer). Deres lovemessigheter i en elektrisk krets er gitt ved følgende relasjoner (hvor Q er ladning, $I = dQ/dt$ er elektrisk strøm, V er spenning, R er resistans, L induktans og C kapasitans):

$$V_R = RI \quad (2.20)$$

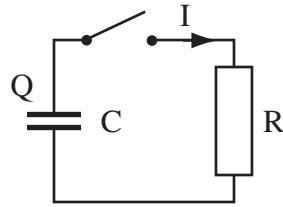
$$V_C = Q/C \quad (2.21)$$

$$\begin{aligned} V_L &= L dI/dt \\ &= L d^2Q/dt^2 \end{aligned} \quad (2.22)$$

Kobles kretselementer sammen i en lukket sløyfe, vil den totale spenningsendringen være null når vi følger sløyfa fra et vilkårlig punkt rundt til samme punkt (Kirchhoffs lov). Kobler vi f.eks. en (oppladet) kondensator til en resistans (ved at vi lukker bryteren i figur 2.5), vil spenningen over kondensatoren hele tiden være motsatt lik spenningen over resistansen. Herav følger:

$$RI = -Q/C$$

$$\frac{dQ}{dt} = -\frac{1}{RC}Q$$



Figur 2.5: Spenningen over en oppladet kondensator vil synke eksponentielt mot null etter at kondensatoren er koblet mot en resistans.

Dersom ladningen på kondensatoren var Q_0 ved tiden $t = 0$, er løsningen av denne differensiellligningen:

$$Q = Q_0 e^{-t/RC}$$

Kondensatoren får altså en eksponentielt avtakende ladning som går mot null. (Dette regner vi som kjent fra før.)

I “svingninger og bølger”-sammenheng, skal vi for elektriske kretser koncentrere oss om svingekretser. En elektrisk svingekrets består vanligvis av minst en kondensator og en induktans (spole). Kobles de to i serie slik at de danner en lukket sløyfe, gir Kirchhoffs lov:

$$\frac{Q}{C} = -L \frac{dI}{dt} = -L \frac{d^2Q}{dt^2}$$

$$\frac{d^2Q}{dt^2} = -\frac{1}{LC}Q$$

Vi kan på samme måte som for det mekaniske systemet skrive dette på følgende form:

$$\ddot{Q}(t) = -\frac{1}{LC}Q(t) \quad (2.23)$$

Sammenligner vi ligning (2.23) med ligning (2.6), ser vi at de er helt analoge. Konstantleddet heter $\frac{k}{m}$ for det mekaniske systemet, og $\frac{1}{LC}$ i det elektriske, men begge er uansett konstanter.

Dette er svingelingningen, på ny, og vi vet at den generelle løsningen er:

$$Q = Q_0 \cos(\omega t + \phi)$$

hvor $\omega = \frac{1}{\sqrt{LC}}$. Q_0 og ϕ er to variabler som er bestemt ut fra initialbetingelsene for systemet ved tiden $t = 0$.

[♣ ⇒ Det kan være verdt å reflektere over hvorfor det må to initialbetingelser til for å angi en konkret løsning for LC-kretsen sammenlignet med RC-kretsen. I RC-kretsen er strømmen entydig gitt dersom ladningen er gitt. Vi kan da ved hjelp av ett øyeblikksbilde, enten av ladning eller spenning, bestemme hvordan tidsforløpet vil arte seg videre (forutsatt at vi kjenner R og C). For LC-kretsen er det ikke tilfelle. Der må vi kjenne f.eks. både ladning og strøm ved ett tidspunkt, eller ladning ved to nærliggende

tidspunkt, for å bestemme det videre forsløp. Grunnen er at vi ikke kan dedusere strøm ut fra én ladning (eller spenning) alene. $\Leftarrow \spadesuit$

En elektrisk svingekrets inneholder i praksis en eller annen form for tap/resistans. La oss ta for oss det enkleste eksemplet, nemlig at tapet skyldes en konstant serieresistans R i den lukkede sløyfa. Brukes Kirchhoff lov på nytt, får vi følgende differensialligning:

$$\frac{Q}{C} = -RI - L\frac{dI}{dt} = -R\frac{dQ}{dt} - L\frac{d^2Q}{dt^2}$$

eller

$$\frac{d^2Q}{dt^2} + \frac{R}{L}\frac{dQ}{dt} + \frac{1}{LC}Q = 0 \quad (2.24)$$

Dette er en homogen annenordens differensialligning som kan løses ved å bruke det karakteristiske polynom:

$$a^2 + \frac{R}{L}a + \frac{1}{LC} = 0$$

som har løsningen:

$$a = -\frac{R}{2L} \pm \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}}$$

Differensialligningen har da følgende generelle løsning:

$$Q = Q_{0,1}e^{-\frac{R}{2L}t + \left(\sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}}\right)t} + Q_{0,2}e^{-\frac{R}{2L}t - \left(\sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}}\right)t} \quad (2.25)$$

Vi merker oss at for $R = 0$ faller vi tilbake til ligning (2.23), og løsningen blir:

$$\begin{aligned} Q &= Q_{0,1}e^{(\sqrt{-1/LC})t} + Q_{0,2}e^{-(\sqrt{-1/LC})t} \\ &= Q_{0,1}e^{i(\sqrt{1/LC})t} + Q_{0,2}e^{-i(\sqrt{1/LC})t} \\ &= Q_0 \cos(\omega t + \phi) \end{aligned}$$

hvor $\omega = 1/\sqrt{LC}$. Igjen ser vi at det er to konstanter som bestemmer initialbetingelsene.

Når $R \neq 0$, får vi et eksponentielt avtakende ledd $e^{-(R/2L)t}$ multiplisert med enten et oscillatorende ledd eller et nytt eksponentielt avtakende ledd, avhengig av hvorvidt $(R/2L)^2$ er mindre eller større enn $1/LC$. Når $(R/2L)^2 = 1/LC$ blir innholdet under rottegnet i ligning (2.25) lik null, og det samsvarer med hva vi har sett tidligere med to sammenfallende røtter. I det tilfellet får den generelle løsningen en tilsvarende form som ligning (2.15). Igjen er det naturlig å snakke om underkritisk, kritisk og overkritisk demping, på tilsvarende måte som for en mekanisk pendel.

Vi ser at vi ender opp med helt analoge ligninger som ved en mekanisk pendel. Andre fysiske fenomener fremviser tilsvarende svingende systemer.

Felles for alle systemene vi har sett på til nå er *svingeligningen*. I sin enkleste form er den gitt som

$$\frac{d^2f}{dt^2} + c_1\frac{df}{dt} + c_2f = 0$$

hvor c_1 og c_2 er positive konstanter.

2.6 Energibetrakninger

La oss beregne hvor mye energi elektriske svingekretser inneholder og hvilken tidsutvikling av energien de fremviser. Vi begrenser oss til et elektrisk svingesystem uten tap, det vil si at resistansen $R = 0$. Løsningen av differensialligningen er da:

$$Q = Q_0 \cos(\omega t + \phi)$$

hvor $\omega = \frac{1}{\sqrt{LC}}$. Q_0 og ϕ er to variable som bestemmes av initialbetingelsene for systemet ved tiden $t = 0$.

Energien som til enhver tid er lagret i kondensatoren, er gitt ved:

$$E_C = \frac{1}{2} QV = \frac{1}{2} \frac{Q^2}{C}$$

Den momentane energien er altså:

$$E_C(t) = \frac{1}{2} \frac{(Q_0 \cos(\omega t + \phi))^2}{C}$$

$$E_C(t) = \frac{1}{2} \frac{Q_0^2}{C} \cos^2(\omega t + \phi)$$

Fra elektromagnetismen er det kjent at energien som er lagret i en induktans er gitt ved:

$$E_L = \frac{1}{2} LI^2 = \frac{1}{2} L \left(\frac{dQ}{dt} \right)^2$$

Settes den generelle løsningen inn her, følger det at den momentane energien i induktansen er

$$E_L(t) = \frac{1}{2} L \left(\frac{d(Q_0 \cos(\omega t + \phi))}{dt} \right)^2$$

$$E_L(t) = \frac{1}{2} L Q_0^2 \omega^2 \sin^2(\omega t + \phi)$$

Siden $\omega = \frac{1}{\sqrt{LC}}$, kan uttrykket også skrives på formen:

$$E_L(t) = \frac{1}{2} \frac{Q_0^2}{C} \sin^2(\omega t + \phi)$$

Summeres de to energibidragene, følger:

$$E_{tot}(t) = E_C(t) + E_L(t)$$

$$E_{tot}(t) = \frac{1}{2} \frac{Q_0^2}{C} (\cos^2(\omega t + \phi) + \sin^2(\omega t + \phi))$$

$$E_{tot}(t) = \frac{1}{2} \frac{Q_0^2}{C}$$

Vi merker oss at totalenergien er konstant, det vil si tidsuavhengig. Selv om energien i kondensatoren og induktansen varierer fra null til en maksimal verdi og tilbake i et oscillerende tidsforløp, er disse variasjonene tidsforskjøvet med en kvart periode slik at summen blir uavhengig av tiden. Engerien "skvulper" fram og tilbake mellom kondensatoren og induktansen. En tidsforskyvning mellom to energiformer synes å være et karakteristisk trekk ved alle svingninger. Enkle svingninger er ofte løsninger av en annen ordens svingeligning, men svingninger kan også ha opphav i fenomener som må beskrives matematisk på annet vis.

For det mekaniske systemet er det potensiell energi (fra den konservative fjærkraften) og kinetisk energi som er de to energiformene. Du anbefales å gjennomføre en liknende beregning som vi har gjort i dette underkapitlet også for det mekaniske systemet for å se at resultatet blir analogt med det vi fant for det elektriske systemet. (Dette er tema for en regneoppgave sist i kapitlet.)

Energiberegningene vi nettopp har gjennomført er basert på at det ikke er noe tap i systemet. Dersom tap på grunn av resistans eller friksjon tas med, vil energien selvfølgelig avta med tiden. Litt avhengig av hvorvidt vi har overkritisk, kritisk eller underkritisk demping, vil energitapet per tidsenhet få litt ulikt tidsforløp, men i hovedtrekk vil energitapet følge et eksponentielt avtakende forløp.

2.7 Læringsmål

Bokas tittel er “Svingninger og bølger”, men omtrent all grunnleggende teori om *svingninger* er presentert allerede i kapittel 1 og 2. Likevel vil grunnleggende tanker fra disse to kapitlene dukke opp på ny mange ganger når vi omtaler bølger. Vi tror derfor at det vil lønne seg å bruke en god del tid på kapittel 1 og 2 for at tilegning av stoff i de påfølgende kapitlene skal gå så glatt som mulig.

Etter å ha jobbet deg gjennom kapittel 1 bør du

- - kjenne til at en harmonisk svingbevegelse kan angis matematisk på en rekke måter, både med sinus- og/eller cosinusledd, eller på kompleks form (vha Eulers formel). Det er et mål å gjenkjenne de ulike formene og å kunne gå matematisk fra en hvilken som helst av disse representasjonene til en annen.
- - kjenne til at svingninger kan forekomme når et system påvirkes av en “kraft” som forsøker å føre systemet tilbake til likevekt. Matematisk kan dette i enkle tilfeller skrives:

$$\ddot{x} = -qx$$

hvor x er utslaget og q er et reelt, positivt tall.

- - vite at enhver svingeligning må inneholde de to leddene gitt i ligningen i forrige punkt, men at også andre ledd kan inngå.
- - kjenne til hvordan fysiske lover/relasjoner kombineres ved utledning av svingeligningen både for et mekanisk og elektrisk system.
- - kjenne til at for å løse en svingeligning entydig, må to uavhengige initialbetingelser være kjent, og foreslå minst et par ulike valg av initialbetingelser.
- - kunne utlede og løse svingeligningen både for fri og dempet svingning med linært dempeledd. Det innebærer at du må kunne skille mellom overkritisk, kritisk og underkritisk demping, og kunne skissere grafisk typiske trekk for ulike initialbetingelser.
- - kunne utlede svingeligningen også for ikke-lineært dempeledd, og finne løsningen numerisk (etter å ha vært gjennom kapittel 3).
- - kunne forklare hvorfor superposisjonsprinsippet ikke gjelder når ikke-lineære ledd tas med i svingeligningen.

2.8 Oppgaver

MERK: For alle oppgaver gjelder en generell regel at riktig svar alene ikke regnes som en fullgod løsning. Full uttelling oppnås bare om det i tillegg til riktig svar er gitt begrunnelser og forutsetninger og tilnærmingar som gjøres o.l. Denne generelle regelen må brukes med skjønn siden oppgaver kan være ganske forskjellige i utgangspunktet. Regelen gjelder selvfølgelig også ved vurdering av eksamensoppgaver, så det er lurt å øve inn gode vaner på dette området så snart som mulig i tilfelle du ikke er helt vant med dette fra tidligere.

Forståelses- / diskusjonsspørsmål

1. Lag en skisse lignende figur 1.1 som viser et tidsforløp for én svingning, men tegn også inn tidsforløpet for en annen svingning med samme amplitud og faseledd, men forskjellig frekvens sammenlignet med den første. Gjenta det samme for tilfellet der amplitudene er forskjellige, mens fase og frekvens er de samme. Lag til slutt den tredje varianten av slike skisser. (Finn selv ut hva som menes med dette.)
2. Hvilke krav må vi stille til en kraft for at den skal kunne danne grunnlaget for svingninger?
3. Dersom en fjær kuttes på midten, hvilken fjærkonstant får da hver av delene sammenlignet med konstanten for den opprinnelige fjæra? Hvor stor blir svingtiden for et lodd i enden av den halve fjæra sammenlignet med svingtiden for loddet i den opprinnelige fjæra?
4. Anta at vi har et lodd i en fjær som svinger opp og ned med en bestemt periodetid her på Jorda. Anta at vi tok med oss fjær og lodd til Månen. Vil periodetiden endres? Forklar.
5. Anta at vi gjør som i forrige oppgave, men har en pendel i stedet for et lodd i en fjær. Vil periodetiden endres?
6. En god sprettboll kan hoppe opp og ned mange ganger mot en hard horisontal flate. Er dette en harmonisk bevegelse slik vi har brukt ordet?
7. I teksten er det brukt en vag formulering om tilpasning mellom fjær, masse og utslag for å få en tilnærmet harmonisk svingbevegelse. Kan du gi eksempler på hvilke forhold som kan ødelegge for en harmonisk bevegelse?

Regneoppgaver

8. Vis matematisk at totalenergien for et svingende lodd i en fjær er konstant i tiden såfremt det ikke er noe friksjon til stede.(Se bort fra endringer i potensiell energi i gravitasjonsfeltet i denne omgang.)
9. En fjær henger vertikalt i et stativ. Uten noe lodd er fjæra 30 cm lang. Henger vi et 100 g lodd i enden, strekkes fjæra, og blir 48 cm lang når loddet har kommet til ro. Vi trekker så loddet 8.0 cm loddrett nedover, holder loddet i ro, og så slipper det. Finn svingtiden for loddets bevegelse. Angi et matematisk uttrykk som kan beskrive svingbevegelsen. Finn maksimal og minimal kraft som virker mellom loddet og fjæra.
10. Et svingende lodd i en fjær beveger seg med en frekvens 0.4 Hz. Ved tiden $t = 2.0$ s har loddet posisjonen + 2.4 cm over likevektsposisjonen, og hastigheten til loddet er - 16 cm/s. Finn akselerasjonen til loddet ved tiden $t = 2.0$ s. Finn en matematisk beskrivelse som passer til bevegelsen.

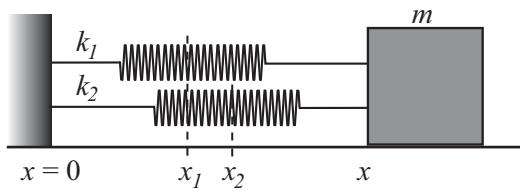
11. En lodd med masse m henger i en masseløs fjær med fjærstivhet k . Amplituden er A . Hvor stort er utslaget fra likevektstillingen når den kinetiske energien er lik halvparten av den potensielle energien?
12. Det er iblant en fordel å beskrive en bevegelse ved å plotte hastighet vs posisjon i stedet for posisjon versus tid slik vi hittil har gjort. Lag et slikt plot for et lodd som svinger opp og ned i enden av en fjær (et plot i faserommet). Hvilken form får plottet i vårt tilfelle?
13. Lag et plot i faserommet (se forrige oppgave) for bevegelsen til en sprettboll som spretter vertikalt opp og ned på et hardt underlag (praktisk talt uten tap). Hvilken form får plottet? Kommentér likheter/forskjeller mellom plottet i denne og den forgående oppgaven.
14. En svingebevegelse kan beskrives ved ligningen $z(t) = A \cos(\omega t + \phi)$ hvor $A = 1.2$ m, frekvensen $f = \omega/(2\pi) = 3.0$ Hz, og $\phi = 30$ grader. Finn ut hvordan denne svingebevegelsen kan angis formelt når vi a) Ikke bruker faseledd, men bare en kombinasjon av sinus og cosinus-ledd, og b) Når vi bruker en kompleks beskrivelse basert på Eulers formel.
15. En svingebevegelse beskrives ved ligningen $z(t) = A \sin(\omega t) + B \cos(\omega t)$, hvor A og ω er gitt via forrige oppgave, og $B = 0.7$ m. Angi denne svingebevegelsen ved å bruke bare cosinus-ledd pluss et faseledd. Angi også et uttrykk for svingebevegelsen basert på komplekse tall (Eulers formel).
16. En annen svingebevegelse er gitt ved $y(t) = \Re((-5.8 + 2.2i)e^{i\omega t})$. Omform ligningen til den kommer på samme form som ligning (2.1) og omform den videre til den kommer på formen i ligning (2.3).
17. Et lodd som veier 1.00 N henges i enden av en lett fjær med kraftkonstant 1.50 N/m. Lar vi loddet svinge opp og ned, er periodetiden T . Dersom vi i stedet lar loddet komme til ro, og trekker det litt ut til siden og slipper det, vil pendelbevegelsen ha en periodetid $2T$ (utslaget i pendelbevegelsen er svært lite). Hvilken lengde har fjæra uten lodd?
18. Anta at vi har to fjærer med ulik fjærkonstant. Fjærene kan kobles sammen enten i parallel eller i serie. Finn et uttrykk for svingtiden til et lodd med masse m som henger i enden av enten de parallelkoblede eller seriekoblede fjærene. Loddet kan også spennes fast mellom de to fjærene (som da begge er strukket en viss lengde). Kan du finne et uttrykk for svingtiden til loddet i dette tilfellet også?
19. Vis at energitapet for en dempet pendel der friksjonskraften er $F_f = -b \cdot v$ er gitt ved $\frac{dE}{dt} = -b \cdot v^2$. Her er b et positivt tall (friksjonskoeffisienten) og v er hastigheten. (Ta utgangspunkt i den mekaniske energien for systemet, $E = E_{potensial} + E_{kinetisk}$.)
20. En masse $m = 2.0$ kg henger i en fjær med fjærstivhet $k = 50$ N/m. Vi ser bort fra fjærens masse. Systemet settes i svingninger og er dempet. Når massens hastighet er 0.5 m/s er den dempende kraften 8.0 N.
 - a) Hva er systemets naturlige svingefrekvens f (dvs. hvis demping ikke var til stede)?
 - b) Bestem frekvensen for de dempede svingningene.
 - c) Hvor lang tid tar det før amplituden er redusert til 1 % av den opprinnelige verdi?

Eksempler på tidligere eksamensoppgaver

21. a) En fjær henger vertikalt. En masse $m_1 = 0.100$ kg festes i den nedre enden av fjæren. Svingeperioden til denne harmoniske oscillatoren er $T_1 = 2.0$ s. Vi fester en ukjent masse m_2 til m_1 slik at totalmassen blir $m_1 + m_2$. Svingeperioden blir nå 2.4 s. Finn den ukjente massen m_2 . Fjærkonstanten k er ukjent.

- b) To fjærer med fjærkonstanter $k_1 = 40$ N/cm og $k_2 = 20$ N/cm er koblet i parallelle til en masse m som kan bevege seg uten friksjon på et underlag. Figuren 2.6 viser massen ved en posisjon x .

$x_1 = 4.0$ cm og $x_2 = 6.0$ cm er posisjonene til fjærendene når de ikke er strukket og når de ikke er koblet til massen. m settes i svingninger. Hva er x i likevektsposisjonen? Fjærene regnes som masseløse.



Figur 2.6: Se oppgaven for forklaring.

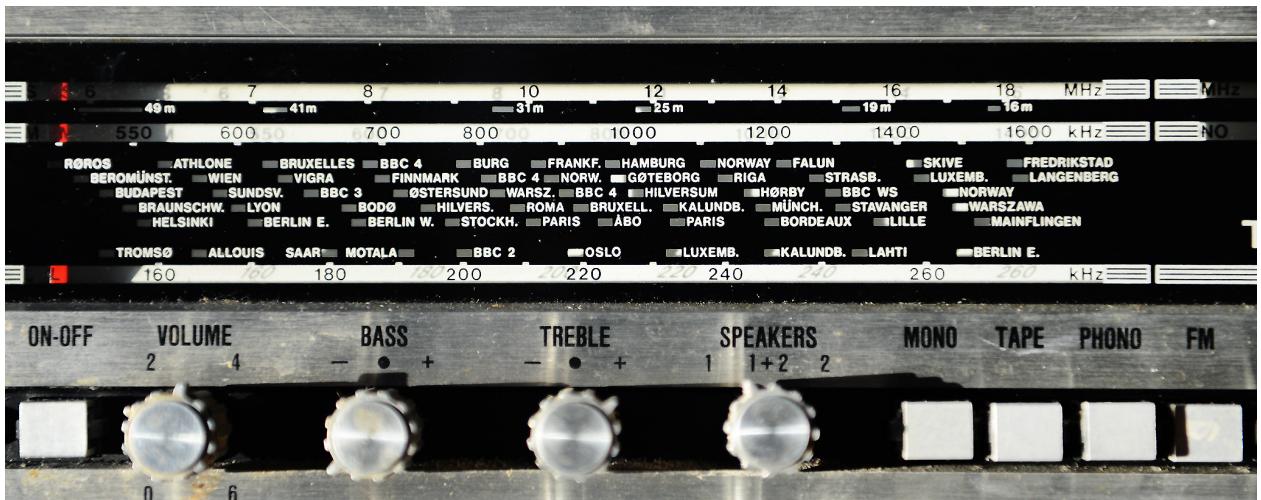
22. En harmonisk svingning kan beskrives matematisk på flere måter. Her er tre eksempler:

$$\begin{aligned}y(t) &= A \sin(\omega t) + B \cos(\omega t) \\y(t) &= \Re \{ C e^{i\omega t} \} \\y(t) &= \Re \{ D e^{i\omega t} \}\end{aligned}$$

hvor A , B og C er reelle tall, og D er et komplekst tall. Hvilke av disse beskrivelsene må anses som identiske? Hva ligger forskjellen i for den beskrivelsen som er litt forskjellig fra de to andre?

Kapittel 3

Tvungne svingninger og resonans



Litt av fronten på Norges stolthet for mange år siden: En Tandberg Sølvsuper radio!.

Ordet resonans kommer kanskje fra det engelske uttrykket "resound". Synger vi med riktig tonehøyde, kan vi få et hulrom til å synge med, - og på en måte forsterke lyden vi puttet inn. I dag brukes ordet i mange ulike sammenhenger, men alltid har det noe med at en impuls vinner gjenklang i et eller annet medium. Når vi stiller radioen på å motta signaler fra en radiosender, sørger vi for at meget svake signaler som antennen fanger opp, blir plukket ut fra alle andre ikke-interessante signaler som kommer samtidig. Det kan synes som ren magi.

Fysikken bak slike fenomener er på en måte enkel når vi begrenser oss til de aller enkleste tilfellene. Går vi litt mer i dybden kommer det fram detaljer som gjør det hele mye mer utfordrende og spennende. En god fysiker bør unne seg å trenge litt bak den enkleste teoretiske fasaden. Da får fenomenet litt mer liv, og man er i stand til å utnytte effektene på en nydelig måte.

God fornøyelse!

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 10012013

3.1 Tvungne svingninger

Foucault-pendelen i vestbylen på Fysikk-bygget på Blindern svinger med samme amplitud i år etter år, selv om den blir påvirket av luftfriksjon som i prinsippet skulle ha gitt en dempet bevegelse. Grunnen er at pendelen får et lite elektromagnetisk dytt hver gang den passerer det laveste punktet. Når det skjer, lyser en liten rød lysdiode. Dyttet kommer akkurat på det tidspunktet pendelen er på vei bort fra likevektpunktet. På den måten blir svingetiden praktisk talt fullstendig bestemt av pendelens egen naturlige svingetid (gitt av pendellengde og jordens gravitasjon).

I andre sammenhenger kommer "dyttene" i en annen takt enn systemet selv kunne ønsket å bevege seg i. Elektroner i en antenn, membranen i en høyttaler, vuggingen av en båt når bølger passerer, er alle eksempler på systemer som blir påtvunget en svingbevegelse fra en ytre kraft som varierer i tid uavhengig av systemets egen bevegelse. Vi snakker da om *tvungne svingninger* ("forced oscillations" eller "driven oscillations" på engelsk).

Den ytre tidsavhengige kraften kan i prinsippet variere på uendelig mange måter, men en interessant underklasse er karakterisert ved en harmonisk tidsvarierende kraft, det vil si som en sinus- eller cosinusfunksjon. I første del av kapitlet antar vi at den harmoniske kraften varer ved "lenge" (skal konkretisere hva som menes med dette senere).

Dersom vi går tilbake til den mekaniske pendelen vi studerte tidligere, holder oss til den enkle friksjonsbeskrivelsen, og nøyer oss med harmoniske ytre krefter, kan bevegelsen beskrives analytisk.

For et mekanisk system er ugangspunktet igjen Newtons annen lov (se kapittel 1): Summen av kreftene er lik massen ganger akselerasjonen:

$$F \cos(\omega_F t) - kz(t) - b\dot{z}(t) = m\ddot{z}(t)$$

hvor $F \cos(\omega_F t)$ er den ytre kraften som svinger med sin egen vinkelfrekvens ω_F . Dersom vi setter

$$\omega_0^2 = k/m$$

(vinkelfrekvensen for svingningen i et fritt svingende system), kan ligningen også skrives slik:

$$\ddot{z}(t) + (b/m)\dot{z}(t) + \omega_0^2 z(t) = (F/m) \cos(\omega_F t) \quad (3.1)$$

Dette er en inhomogen annenordens differensielligning, og den har en generell løsning av typen:

$$z(t) = z_h(t) + z_p(t)$$

hvor z_h er en generell løsning av den tilsvarende homogene differensielligningen (F satt lik null), mens z_p er en partikulær løsning av den fulle inhomogene differensielligningen.

Vi har allerede i kapittel 1 funnet den generelle løsningen av den tilsvarende homogene ligningen på en konstant faktor nær, så utfordringen blir å finne en partikulær løsning.

Vi vet at løsningen av den homogene ligningen avtar med tiden mot null. Når det er gått lang tid fra start, vil derfor bevegelsen være dominert av den ytre periodiske kraften.

Det er da naturlig å undersøke om en partikulær løsning kan ha følgende form:

$$z_p(t) = A \cos(\omega_F t + \phi) \quad (3.2)$$

hvor A er reell.

Når de to uttrykkene for $z_p(t)$ og $F(t)$ settes inn i ligning (3.1), følger med litt ordning

av leddene:

$$(\omega_0^2 - \omega_F^2) \cos(\omega_F t + \phi) - (b/m)\omega_F \sin(\omega_F t + \phi) = F/(Am) \cos(\omega_F t)$$

Setter vi inn uttrykkene for sinus og cosinus til en sum (se Rottmann), følger:

$$\begin{aligned} (\omega_0^2 - \omega_F^2) \{\cos(\omega_F t) \cos \phi - \sin(\omega_F t) \sin \phi\} - (b/m)\omega_F \{\sin(\omega_F t) \cos \phi + \cos(\omega_F t) \sin \phi\} \\ = F/(Am) \cos(\omega_F t) \end{aligned}$$

Leddene med $\sin(\omega_F t)$ og leddene med $\cos(\omega_F t)$ danner tilsammen to ligninger som hver for seg må tilfredsstilles. Tar vi for oss leddene som inneholder $\sin(\omega_F t)$ -ledd (og forkorter med dette sin-uttrykket), følger:

$$(\omega_F^2 - \omega_0^2) \sin \phi = (b\omega_F/m) \cos \phi$$

Faseforskjellen mellom utslag og påtrykt kraft er da gitt ved følgende uttrykk:

$$\tan \phi = \frac{b\omega_F/m}{\omega_F^2 - \omega_0^2} \quad (3.3)$$

Bruker vi leddene som inneholder $\cos(\omega_F t)$ (og forkorter med dette cos-uttrykket), får vi:

$$(\omega_0^2 - \omega_F^2) \cos \phi - (b\omega_F/m) \sin \phi = F/(Am)$$

Vi bruker så uttrykket $\sin x = \pm \tan x / \sqrt{1 + \tan^2 x}$ fra Rottmann (og et tilsvarende uttrykk for cos) sammen med ligning (3.3).

Med litt mellomregning får vi da følgende uttrykk for amplituden i de tvungne svingningene:

$$A = \frac{F/m}{\sqrt{(\omega_F^2 - \omega_0^2)^2 + (b\omega_F/m)^2}} \quad (3.4)$$

Det er nå på tide med en oppsummering av hva vi har gjort:

For en tvungen svingning med en harmonisk kraft som varer lenge, har vi vist at en partikulær løsning (som gjelder lenge etter at kraften er koblet til) faktisk er en harmonisk svingning som er faseforskjøvet i forhold til den opprinnelige kraften, som gitt i ligning (3.2).

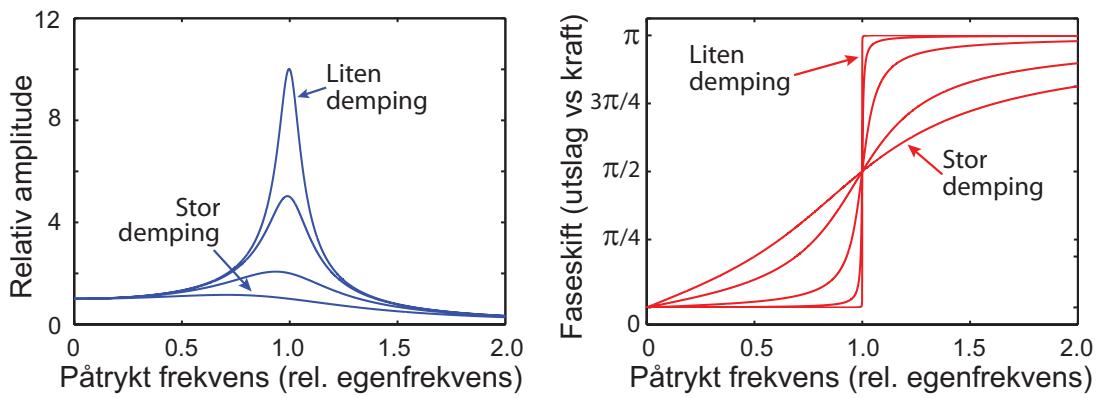
Amplituden i svingningene er da gitt av ligning (3.4) og faseforskjellen mellom utslaget og kraften er gitt av ligning (3.3). Figur 3.1 viser skjematiske hvordan amplituden og fasen varierer med frekvensen til den påtrykte kraften. Frekvensen til kraften er gitt relativt til frekvensen til svingningene i samme system dersom det ikke var noe påtrykt kraft og heller ingen friksjon/demping.

Vi ser at amplituden er størst når frekvensen til kraften er nær den samme som svingningene i samme system uten påtrykt kraft og demping. Vi kaller dette fenomenet for *resonans*, og detaljer vil bli diskutert i neste delkapittel.

Fasen er omtrent lik $\pi/2$ ved resonans, det vil si kraften er omtrent $\pi/2$ faseforskjøvet foran utslaget. For fjærpendelen betyr det at kraften er størst i retning oppover når pendelen passerer likevektpunktet på vei oppover.

Utenfor resonans er faseforskjellen mindre enn $\pi/2$ når den påtrykte frekvensen er lavere enn den "naturlige", og større enn $\pi/2$ når påtrykt frekvens er høyere enn den naturlige. Disse relasjonene kan oppsummeres slik at pendelen forsøker å bevege seg raskere enn optimalt når den påtrykte kraften endrer seg for sakte i forhold til resonans. Pendelen forsøker å bevege seg langsmere enn optimalt når kraften endrer seg for raskt i forhold til resonans.

Faseforskjellen er et viktig karakteristisk trekk ved tvungne svingninger.



Figur 3.1: Amplituden i en tvungen svingning (venstre) og faseforskjellen mellom utslag og den påtrykte kraften (høyre) som funksjon av frekvensen til den påtrykte kraften.

3.2 Resonans

Ut fra ligning (3.4) er det klart at amplituden i de tvungne svingningene varierer med frekvensen til den påtrykte kraften. Når frekvensen er slik at amplituden er størst, sier vi at vi har *resonans*.

Det kan være nyttig å reflektere litt om hva som skal til for å få størst mulig utslag, noe som svarer til høyest mulig svingeenergi for systemet.

La oss ta utgangspunkt i den mekaniske fjærpendelen igjen. Vi har da en mekanisk kraft som virker på et system i bevegelse. Vi husker fra mekanikken at arbeidet kraften gjør er lik kraftens størrelse multiplisert med hvor langt systemet beveger seg mens kraften virker. For en konstant kraft er effekten kraften leverer lik kraften multiplisert med hastigheten til systemet kraften virker på. Kraft og hastighet er vektorielle størrelser, og det er prikkproduktet som teller.

I vårt tilfelle vil kraften leve til størst mulig effekt til systemet dersom kraften har størst verdi samtidig som pendeloddet har størst mulig hastighet. Kraft og hastighet må virke i samme retning. Dette vil skje dersom kraften f.eks. oppover er størst samtidig som pendeloddet passerer likevektposisjonen på vei oppover. Dette svarer til at posisjonen er faseforskjøvet $\pi/2$ etter kraften. For å oppnå en slik tilstand, må den ytre kraften svinge med *resonansfrekvensen*.

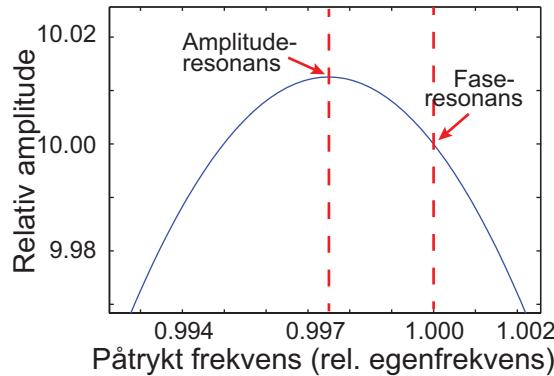
Hittil har vi vært litt upresise når vi har omtalt resonans. Dette skyldes at vi strengt tatt må skille mellom to ulike definisjoner på resonans, nemlig *faseresonans* og *amplituderesonans*. Forskjellen mellom disse er ofte i praksis så liten at vi ikke behøver å bry oss om den.

Faseresonans er karakterisert ved at kraft og utslag er faseforskjøvet $\pi/2$ i forhold til hverandre. Det skjer når frekvensen til påtrykt kraft er identisk med systemets egenresonans (uten demping).

Ser vi imidlertid nøyne på øvre del av kurven i venstre del av figur 3.1, får vi fram at amplituden faktisk er størst ved en litt lavere frekvens (se figur 3.2). Den lille, men signifikante forskjellen, skyldes en detalj vi nevnte da vi drøftet dempet svingning. I det tilfellet er svingefrekvensen litt lavere enn ved ingen demping. Størst utslag i svingningen forekommer derfor ved en litt lavere frekvens enn systemets svingefrekvens uten demping. Frekvensen der amplituden er størst angir *amplituderesonansen* til systemet. De to resonansfrekvensene ligger ofte, som allerede nevnt, temmelig nær hverandre.

La oss nå finne matematiske uttrykk for de to resonansfrekvensene.

Amplituderesonansfrekvensen kan vi finne ved å derivere uttrykket for amplituden til



Figur 3.2: Detalj i amplituden i en tvungen svingning som funksjon av frekvensen til den påtrykte kraften.

svingningen i ligning (3.4) (det vil si en vanlig prosedyre for å finne ekstremalverdier). Beregner vi vinkelfrekvensen ω_F som tilfredsstiller:

$$\frac{dA}{d\omega_F} = 0$$

finner vi at

$$\omega_F = \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}$$

Angir vi frekvenser i stedet for vinkelfrekvenser, har vi da følgende uttrykk for resonansfrekvensene:

Ampliteresonansfrekvensen er:

$$f_{amp.res.} = \frac{1}{2\pi} \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}$$

hvor $\omega_0 = \sqrt{k/m}$.

Faseresonansfrekvensen er:

$$f_{fase.res.} = \frac{1}{2\pi} \omega_0$$

Vi ser at de to resonansfrekvensene sammenfaller kun dersom dempingen $b = 0$.

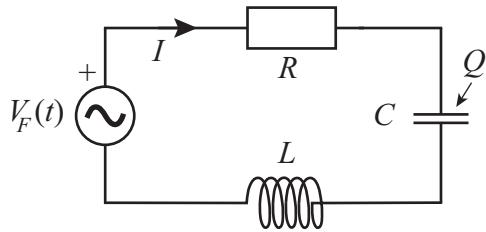
3.2.1 Fasorbeskrivelse

Vi vil nå betrakte tvungne svingninger i en elektrisk svingekrets. Først vil vi gå fram om lag på samme måte som for det mekaniske systemet vi nettopp har vært gjennom, men til slutt vil vi gi en alternativ beskrivelse av tvungne svingninger basert på fasorer. Systemet er en serie-RCL-krets hvor det er koblet inn en harmonisk varierende spenningskilde $V_0 \cos(\omega_F t)$, som vist i figur 3.3. Differensialligningen for systemet blir da (sammenlign med ligning (1.24)):

$$L \frac{d^2Q}{dt^2} + R \frac{dQ}{dt} + \frac{1}{C} Q = V_0 \cos(\omega_F t) \quad (3.5)$$

Denne ligningen er ikke-homogen, og løsningen finnes på samme måte som for mekanisk tvungne svingninger som vi nylig betraktet. Løsningen består av en sum av en partikulær løsning og en løsning av den homogene ligningen (når $V_0 = 0$). Løsningen av den homogene ligningen er allerede kjent, nå gjenstår det bare å finne en partikulær løsning. Vi forsøker følgende løsning:

$$Q_p(t) = A e^{i\omega_F t} \quad (3.6)$$



Figur 3.3: En serie RCL-krets påtrykt en ytre harmonisk varierende spenning. Markeringene +, I og Q angir fortegnsvalgene våre.

hvor A kan være et komplekst tall.

Samtidig velges en eksponensiell form for beskrivelsen av den ytre påtrykte spenningen:

$$V(t) = V_0 \cos(\omega_F t) \rightarrow V_0 e^{i\omega_F t} \quad (3.7)$$

Her er det underforstått at bare den reelle delen av uttrykket benyttes.

De to uttrykkene for $Q_p(t)$ og $V(t)$ settes inn i ligning (3.5), og etter å ha forkortet bort den felles faktoren $e^{(i\omega_F t)}$ får vi:

$$-L\omega_F^2 A + iR\omega_F A + \frac{1}{C} A = V_0$$

Løser ligningen med hensyn på A :

$$A(-L\omega_F^2 + iR\omega_F + \frac{1}{C}) = V_0$$

$$A = \frac{V_0}{\frac{1}{C} - L\omega_F^2 + iR\omega_F}$$

A blir på ny et komplekst tall (bortsett fra når $R = 0$).

Den momentane strømmen i RCL-kretsen finnes ved å anvende Ohms lov på resistansen:

$$I_{krets} = \frac{V_R}{R} = \frac{dQ}{dt}$$

Hvis vi venter lenge nok slik at løsningen av den homogene ligningen har dødd ut og bare den partikulære løsningen eksisterer, er:

$$I_{krets} = \frac{dQ_p}{dt} = A i \omega_F e^{i\omega_F t}$$

Herav følger (med få mellomregninger):

$$I_{krets}(t) = \frac{V_0 \omega_F}{R \omega_F + i(L\omega_F^2 - \frac{1}{C})} e^{i\omega_F t} \quad (3.8)$$

Dette uttrykket bør sammenholdes med påtrykt spenning V_F på kretsen, som på kompleks form er gitt ved:

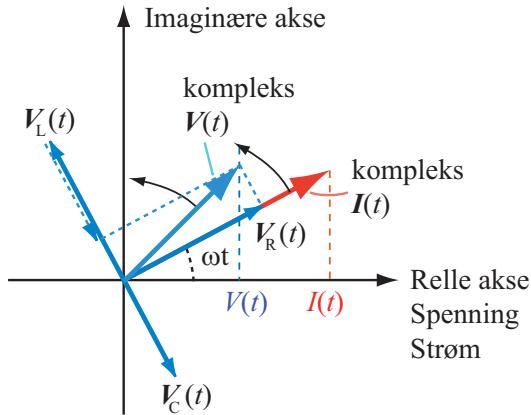
$$V_F(t) = V_0 e^{i\omega_F t}$$

Dette er komplekse talluttrykk, og den virkelige strømmen og spenningen er realverdiene av uttrykkene.

Av ligning (3.8) fremgår det at dersom $R = 0$ vil strømmen være faseforskjøvet 90 grader i forhold til påtrykt spenning. Dersom i tillegg $L = 0$, vil strømmen komme 90 grader før spenningen. Men dersom $L \neq 0$ og C er meget stor (C "kortsluttet"), vil strømmen være forskjøvet 90 grader etter spenningen. (I en regneoppgave sist i kapitlet bes du å vise dette.)

Dersom $R \neq 0$, men $L\omega_F^2 - \frac{1}{C} = 0$, vil strøm og spenning være i fase, og $I = V_0/R$. Dette svarer til $\omega_F = \frac{1}{\sqrt{LC}}$ som vi kalte faseresonans ovenfor.

Sammenhengen mellom R , C , L , strøm og fase kan anskueliggjøres på en elegant måte ved hjelp av fasorer. Vi har allerede omtalt fasorer, men nå utvider vi bildet ved å trekke inn flere roterende vektorer samtidig. Figur 3.4 viser et eksempel.



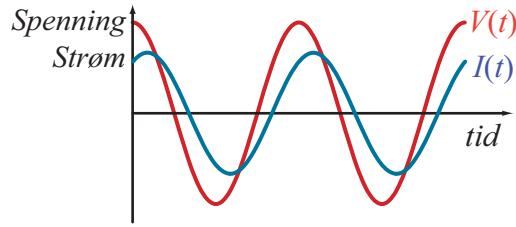
Figur 3.4: Eksempel på fasorbeskrivelse av en RCL-krets påtrykt en harmonisk varierende spenning. Strømmen er til enhver tid (hvor som helst i kretsen) x -komponentene av $I(t)$ -vektoren, mens spenningen over de ulike kretskomponentene er gitt ved x -komponenten av vektorene $V_R(t)$, $V_C(t)$ og $V_L(t)$, og summen av dem er $V(t)$. Se teksten for detaljer.

Både strøm og spenninger er tegnet inn i samme diagram. Vi starter med en vektor som representerer strømmen som den påtrykte spenningen fører til. Deretter tegner vi inn vektorer som representerer spenningen over henholdsvis resistans, kapasitans og induktans ut fra strømmen som går. Vektoren som viser spenningen over kondensatoren, vil da ligge 90 grader etter vektoren som viser strømmen, vektoren for spenning over resistansen vil ha samme retning som strømmen, og vektoren for spenningen over induktansen er 90 grader foran strømmen. Den samlede spenningen over seriekoblingen av R , C og L skal da være like stor som påtrykt spenning. Vi ser at faseforskjellen mellom strøm og spenning vil ligge mellom +90 og -90 grader.

Fasordiagrammer kan også baseres på andre størrelser enn de vi har valgt her. En variant er å bruke komplekse impedanser som adderes vektorielt. Styrken med fasordiagrammer er at vi på en lettfattelig måte kan forstå f.eks. hvordan faseforskjellene endrer seg med frekvensen. Bildet i figur 3.4 gjelder bare for en bestemt påtrykt vinkelfrekvens ω_F . Dersom vinkelfrekvensen øker, vil spenningen over kondensatoren avta, mens spenningen over induktansen vil øke. Faseresonans inntreffer når de to spenningsvektorene er nøyaktig like store (men motsatt rettet) slik at summen av dem er null.

Figur 3.5 viser tidsutviklingen av spenning og strøm i et tidsdiagram. Strømmen i kretsen er forskjøvet litt foran den påtrykte vekselspenningen. For en serie RCL-krets med påtrykt vekselspenning, vil det si at påtrykt frekvens er lavere enn kretsens egen resonansfrekvens.

Merk at fasorer stort sett bare kan brukes når det transiente innsvingningsforløpet er over, og vi har fått en stabil svingning svarende til den partikulære løsningen av differensielligningen.



Figur 3.5: Et tidsdiagram som viser at strøm er forskjøvet litt foran spenningen.

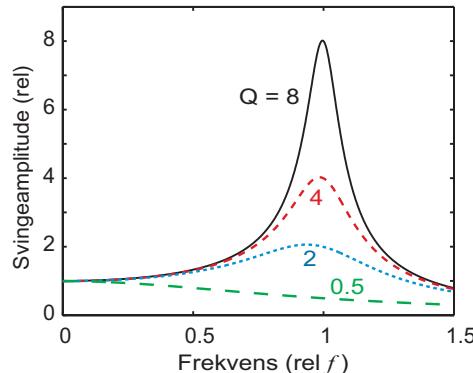
3.3 Kvalitetsfaktoren Q

For tvungne svingninger er det vanlig å karakterisere systemet med en *Q-faktor*. (Må ikke sammenblandes med ladningen *Q* i en elektrisk krets!) *Q* står for “quality”, så faktoren kalles også kvalitetsfaktoren. Faktoren sier oss noe om hvor lett det er å få systemet til å svinge, eller hvor lenge systemet vil fortsette å svinge etter at drivkraften har sluttet å virke. Dette er mer eller mindre ensbetydende med hvor lite tap/friksjon det er i systemet.

Kvalitetsfaktoren for en svingende fjær-pendel er gitt som:

$$Q = \frac{m\omega}{b} = \sqrt{\frac{mk}{b^2}} \quad (3.9)$$

Vi ser av formelen at jo mindre *b* er, desto større blir kvalitetsfaktoren *Q*.



Figur 3.6: Når frekvensen til den påtrykte kraften endrer seg relativt til systemets egen svingefrekvens, vil amplituden bli størst når de to frekvensene er omtrent like store. Jo større kvalitetsfaktor *Q* (dvs jo mindre tap), desto høyere blir amplituden ved resonans.

Figur 3.6 viser hvordan svingeamplittuden varierer med den påtrykte kraftens frekvens for fire ulike kvalitetsfaktorer. En *Q*-verdi på 0.5 svarer i dette tilfellet til kritisk demping, og vi ser ingen antydning til noe resonans for så stor demping.

Det finnes to vanlige måter å definere *Q* på. Den første er:

$$Q \equiv 2\pi \frac{\text{Lagret energi}}{\text{Tap av energi per periode}} = 2\pi \frac{E}{P_{\text{tap,periode}}} \quad (3.10)$$

Denne definisjonen impliserer en spesiell detalj få kjenner til, men som er særdeles viktig for tvungne svingninger i mange sammenhenger. Vi ser av ligning (3.10) at et svingesystem med høy *Q*-verdi bare mister en liten del av den totale energien per periode. Når vi har oppnådd en stabil tilstand (når kraften har virket lenge og fortsatt virker) vil tap av energi erstattes av arbeidet den påtrykte kraften tilfører systemet.

Anta nå at vi kutter ut den påtrykte kraften. Da vil energien i systemet etter hvert forsvinne. Det vil ta i størrelsesordenen $Q/(2\pi)$ perioder før energien er brukt opp og svingningen slutter. Vi kommer tilbake til dette.

Tap av energi per periode er en litt uvant størrelse. La oss omforme uttrykket til "tap per sekund" som da vil få enheten watt, som vi betegner P_{tap} . Vi vet:

$$P_{tap} = -\frac{dE}{dt} \quad (3.11)$$

Tid kan måles enten i sekunder eller i en enhet lik periodetiden T . Kaller vi tid målt i antall periodetider for t' , følger det at $t = t'T$. Følgelig:

$$\begin{aligned} P_{tap} &= -\frac{dE}{dt'} \frac{dt'}{dt} \\ &= P_{tap,periode} \frac{1}{T} \end{aligned}$$

Benytter vi oss nå av definisjonen i ligning (3.10), får vi:

$$P_{tap} = \frac{2\pi}{TQ} E$$

Kombinerer vi dette med definisjonen av P_{tap} i ligning (3.11) og sammenhengen mellom vinkelfrekvens og periodetid, får vi en differensialligning som viser tidsutviklingen av lagret energi etter at drivkraften for den tvungne svingningen opphører. Ligningen blir:

$$P_{tap} = -\frac{dE}{dt} = \frac{\omega}{Q} E$$

Løsningen er:

$$E(t) = E_0 e^{-\omega t/Q}$$

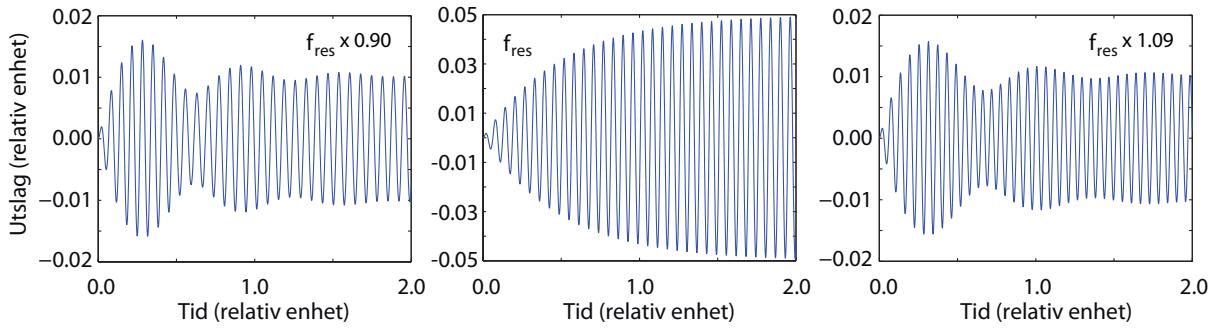
Energien synker til $1/e$ av opprinnelig energi etter en tid

$$\Delta t = \frac{Q}{\omega} \quad (3.12)$$

Vi har sett at amplituden i en svingning avtar på en pen eksponentiell måte etter at en påtrykt oscillerende kraft har opphørt. Tidskonstanten er gitt i ligning (3.12).

Det kan vises at samme tidskonstant gjelder også etter at vi starter en svingning ved hjelp av en påtrykt oscillerende kraft. Riktignok er ikke tidsforløpet like enkelt, fordi det blant annet avhenger av hvorvidt frekvensen til den påtrykte kraften er lik resonansfrekvensen for kretsen eller ikke (se figur 3.7). Likevel er det slik at dersom det tar i størrelsesordenen 10 ms for en oscillasjon å dø ut etter at en påtrykt kraft er fjernet, vil det også ta i størrelsesordenen 10 ms å bygge opp en stabil amplitude etter at vi starter å anvende den påtrykte kraften.

Figur 3.7 viser tre eksempler på svingningsforløpet for en RCL-krets like etter at en ytre oscillerende spenning ble påtrykt. I alle tilfellene var initialbetingelsen ingen svingninger i kretsen idet spenningen ble koblet til. Øverst har den påtrykte spenningen 10 % lavere frekvens enn resonansfrekvensen (faseresonans), i midten har den samme frekvens, og nederst har den en frekvens 9 % høyere enn resonansfrekvensen. Legg merke til at det er mulig å angi en tidskonstant for den eksponentielt økende oscillasjonen (ca 1.5 relative tidsenheter), selv når forstyrrende oscillasjoner er til stede. Legg også merke til den maksimale amplituden i oscillasjonene etter at grenseverdien er oppnådd. Amplituden blir størst ved resonansfrekvensen!



Figur 3.7: Tre eksempler på innsvingningsforløp i en RCL resonanskrets etter at en påtrykt oscillatorende spenning ble koblet over kretsen. Frekvensen til den påtrykte spenningen er litt lavere, lik og litt høyere enn kretsens egen resonansfrekvens. Kretsens Q-verdi er 25.

Kurvene i figur 3.7 viser at etter at vi har startet å anvende en påtrykt kraft, øker amplituden i svingningene, men ikke i det uendelige. Før eller siden blir tapet like stort som den effekten som tilføres gjennom den oscillatorende kraften. Etter at likevekt er oppnådd, vil amplituden for svingningene holde seg konstant så lenge den påtrykte kraften har konstant amplitud.

[♣ ⇒ Løsningen av den inhomogene differensiell ligningen for en RCL-krets som er påtrykt en oscillatorende spenning med gitte initialbetingelser, blir et ganske komplisert uttrykk. Vi har brukt dataprogrammet Maple for å finne dette uttrykket. (Maple kan løse mange differensiell ligninger på analytisk form.) Det analytiske uttrykket fra Maple ble brukt i et lite Matlab-program for å generere de tre forløpene vist i figur 3.7. Programsnuttet er gitt i en oppgave sist i dette kapitlet. ⇐ ♠]

I eksperimentell sammenheng benyttes ofte en annen definisjon av Q-verdi enn den i ligning (3.10). Lager vi et plot som viser energi (NB: Ikke amplitude) i det svingende systemet som funksjon av frekvens (som i figur 3.8), er Q-verdien definert som:

$$Q = \frac{f_0}{\Delta f} \quad (3.13)$$

hvor halvverdibredden Δf , vist i figuren, sammenholdes med resonansfrekvensen f_0 .

Denne relasjonen kan vises å være i overensstemmelse med relasjonen gitt i ligning (3.10), i alle fall for høye Q-verdier.

Definisjonene gitt både i ligning (3.10) og (3.13) gjelder for alle fysiske svingesystemer, ikke bare de mekaniske.

[♣ ⇒ Det er nå mulig å gjøre en interessant observasjon: En resonanskrets responderer betydelig for frekvenser innen et frekvensbånd om lag

$$\Delta f = \frac{f_0}{Q}$$

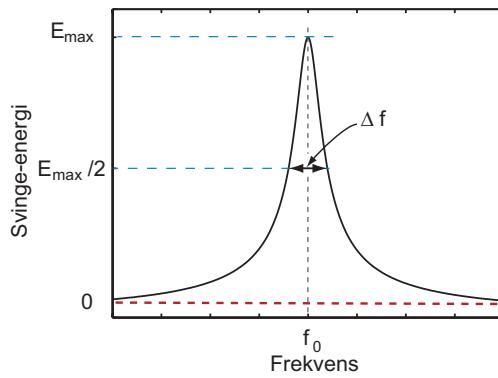
bredd. Kretsen trenger imidlertid en viss tid

$$\Delta t = \frac{Q}{\omega}$$

å bygge opp responsen dersom vi starter fra null. Det tar om lag samme tid også for at en respons som allerede er bygget skal kunne dø ut.

Produktet av Δf og Δt blir:

$$\Delta t \Delta f = \frac{Q}{\omega} \frac{f_0}{Q}$$



Figur 3.8: *Q-verdi kan også defineres ut fra en grafisk fremstilling av energi lagret i svingesystemet som funksjon av frekvens. Q-verdien er da gitt som resonansfrekvensen f_0 dividert på halverdibredden Δf .*

$$\Delta t \Delta f = \frac{1}{2\pi} \quad (3.14)$$

Multipliseres dette uttrykket med Plancks konstant h , og anvendes et av kvantefysikken postulater at energien til et foton er lik $E = hf$, får vi:

$$\Delta t \Delta E = \frac{h}{2\pi} \quad (3.15)$$

Dette uttrykket er nesten identisk med det vi kjenner som Heisenbergs uskaphetsrelasjon for energi og tid. Det savnes en faktor $1/2$ foran ledet etter likhetstegnet, men en slik faktor vil avhenge av hvordan vi velger å definere bredder i frekvens- og tidsforløp.

Det er visse paralleller mellom et makroskopisk svingende system og relasjonene vi kjenner fra kvantefysikken. I kvantefysikken tolkes Heisenbergs uskaphetsrelasjon som en "usikkerhet" i tid og energi: Vi kan ikke "måle" tiden for en hendelse nøyaktigere enn det som er gitt i relasjonen

$$\Delta t = \frac{h}{2\pi \Delta E}$$

forutsatt at vi ikke endrer energien til et system med mer enn ΔE .

Vår makroskopiske variant gjelder uavhengig av hvorvidt vi gjør målinger eller ikke, men målinger vil selvfølgelig reflektere den relasjonen som finnes. Vi kommer tilbake til denne relasjonen senere i boka, men da i form av ligning (3.14) i stedet for (3.15).

"Treheten" i en svingekrets er viktig for hva vi kan gjøre av målinger. For en høykvalitets svingekavitet for mikrobølgeområdet (kalles en "kavitet") kan vi nokså lett oppnå Q-verdier på 10 000 eller mer. Dersom en slik kavitet benyttes i pulset mikrobølgespektroskopi, vil det ta i størrelsesorden 60 000 perioder å endre energien i kavitten betydelig. Dersom mikrobølgefrekvensen er 10 GHz (10^{10} Hz), vil tidskonstanten for energiendringer være i størrelsesorden 6 mikrosekunder. Dersom vi studerer nokså langsomme atomære prosesser, kan dette være akseptabelt, og sensitiviteten til systemet er da gjerne proporsjonal med kvalitetsfaktoren. Dersom vi imidlertid ønsker å studere tidsforløp som er bare noen få ganger periodetiden til svingningene som observeres, må vi bruke kaviteter med langt lavere Q-verdi. Det er mer om dette i neste underkapittel. ← ♠]

3.4 Tidsbegrenset tvungen svingning

Hittil har vi betraktet et system som påvirkes av en oscillerende kraft som varer ved "uendelig lenge" eller som har vart lenge og slutter brått. I en slik situasjon kan vi definere en kvalitetsfaktor Q eksperimentelt ut fra frekvensresponsen til systemet som vist i figur 3.8 og ligning (3.13). Relativ svingeenergi (relativ amplitude kvadrert) må bestemmes

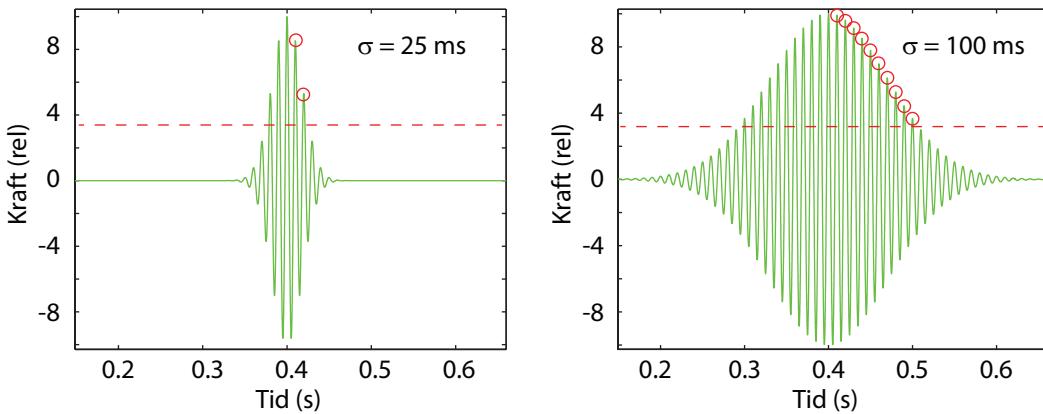
etter at systemet har kommet i en stasjonær tilstand, dvs. at amplituden ikke lenger endrer seg når tiden går.

Hvordan vil et slik system oppføre seg dersom den oscillerende kraften bare varer ved en kort tid? Vi skal nå undersøke dette nærmere.

Når vi innfører en tidsbegrenset kraft, må vi velge hvordan kraften skal begynne, vare ved og hvordan den skal avsluttes. Vi ønsker av flere grunner å unngå brå endringer, og har derfor valgt en kraft der amplituden følger en gaussisk funksjon, men at tidsutviklingen forøvrig matematisk sett er gitt med en eneste cosinusfunksjon (én frekvens). Matematisk er en slik kraft gitt ved:

$$F(t) = F_0 \cos(\omega(t - t_0)) e^{-(t-t_0)/\sigma^2} \quad (3.16)$$

hvor σ angir varigheten på kraften (fra amplituden har hatt sin maksimale verdi til amplituden har sunket til $1/e$ av max). ω er vinkelfrekvensen til den underliggende cosinusfunksjonen, og t_0 er tiden der kraften har maksimal amplitud (toppen av kraftpulsen forekommer ved tiden t_0). Det svingende systemet antas å være i ro før kraften settes inn.



Figur 3.9: Kraften $F(t)$ for senterfrekvens 100 Hz og σ lik 0.025 s og 0.10 s. Se teksten for ytterligere forklaringer.

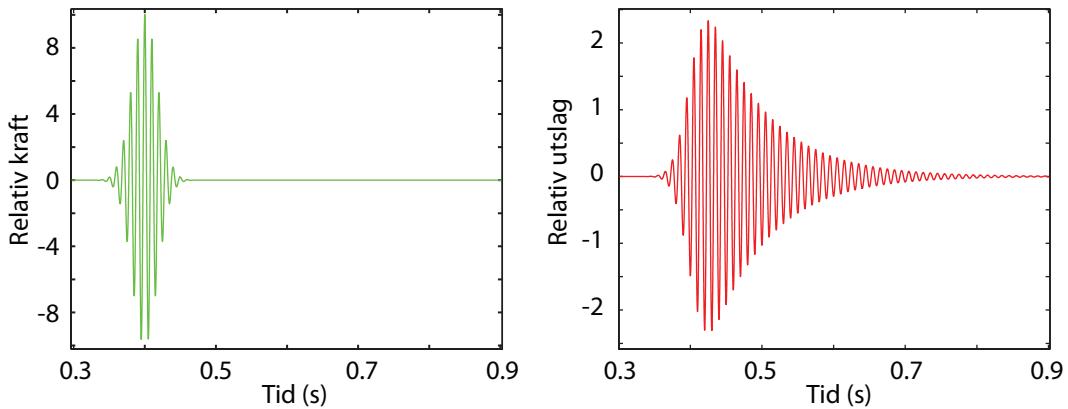
Figur 3.9 viser to eksempler på kraft med ulik varighet. Her har kraften en frekvens lik 100 Hz (periodetid 10 ms). I figuren til venstre er σ lik 25 ms, dvs $2.5 \times$ periodetiden. I figuren har vi merket av antall toppunkt etter max inntil amplituden har sunket til $1/e$ for å få fram hva σ faktisk innebefører. I figuren til høyre er σ 100 ms, dvs $10 \times$ periodetiden. Igjen gir markeringene en indikasjon på sammenhengen mellom omega (eller rettere sagt frekvensen eller periodetiden) og σ .

Vi ønsker nå å studere hvordan et svingesystem vil oppføre seg når det utsettes for en slik tidsbegrenset kraft. Ut fra figur 3.7 vil vi forvente at responsen kan bli temmelig komplisert. Det er ikke enkelt å gjøre beregningene analytisk, så vi har valgt numeriske beregninger i stedet.

Figur 3.7 viser tidsforløpet for ett valg av kraft sammen med systemets respons på kraften. Vi har for enkelhets skyld valgt at frekvensen til kraften er lik resonansfrekvensen til systemet, og ifølge de valgte initialbetingelsene er systemet i ro før kraften startet opp.

Figur 3.7 viser interessante trekk. Systemet forsøker å følge med etter som den påtvungne kraften vokser, men er alltid litt på etterskudd. Det ser vi av at toppen på responsen (utslaget) forekommer litt senere enn tidspunktet da kraften hadde sin maksimale verdi.

Kraften tilfører systemet en del energi. Når kraften avtar så raskt som den gjør i dette tilfellet, klarer ikke systemet å kvitte seg med den tilførte energien like raskt som kraften avtar. Systemet går derfor inn i en periode med dempede svingninger og følger da de karakteristiske trekkene som gjelder da. Det kan nevnes at σ her er 25 ms og at



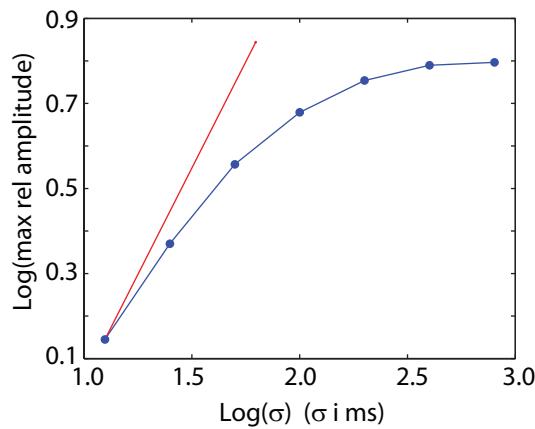
Figur 3.10: *Tidsforløpet for utslaget til systemet (til høyre) som følge av den påtrykte kraften vist i venstre del av figuren.*

Q-verdien for svingekretsen er valgt til å være 25, hvilket svarer til en decaytid for de tvungne svingningene på 40 ms.

Her kan det være nyttig å peke på en del sammenhenger mellom parametere:

- Hvor mye energi som kan puttes inn i systemet innen en gitt tid avhenger av styrken på kraften (proporsjonalitet?).
- Hvor mye energi som kan puttes inn for en viss maksimal styrke av kraften, vil avhenge av hvor lang tid kraften virker.
- Tapet av energi er uavhengig av styrken på kraften etter at kraften har forsvunnet.
- Tapet av energi er proporsjonalt med amplituden til det svingende systemet.

Som nevnt forventer vi at amplituden vil øke når kraften varer lengre og lengre tid, men det er ikke selvvinnlysende hvordan denne sammenhengen er. I figur 3.11 er det vist beregnede resultater for den maksimale amplituden vårt system oppnår for ulike σ -verdier. Omega svarer hele tiden til resonansfrekvensen til systemet. Figuren har logaritmiske akser for å få med et stort variasjonsområde for σ . Den rette linjen representerer det tilfellet at amplituden øker lineært med σ (varigheten av kraften).



Figur 3.11: *Maksimal utslag for systemet for ulik varighet (σ) for kraften. Merk at det er logaritmiske akser.*

Vi ser at for små σ (kraften varer ved bare noen få periodetider av svingefrekvensen), øker den maksimale amplituden omrent proporsjonalt med varigheten av kraften. Når

kraften varer lenger, gjelder ikke dette lenger, og når vi passerer en viss grense, får vi ikke større amplitude på svingningen uansett hvor mye lenger kraften varer.

Dette skyldes at ved det utslaget vi da har, er tapet like stort som energien som tilføres via kraften.

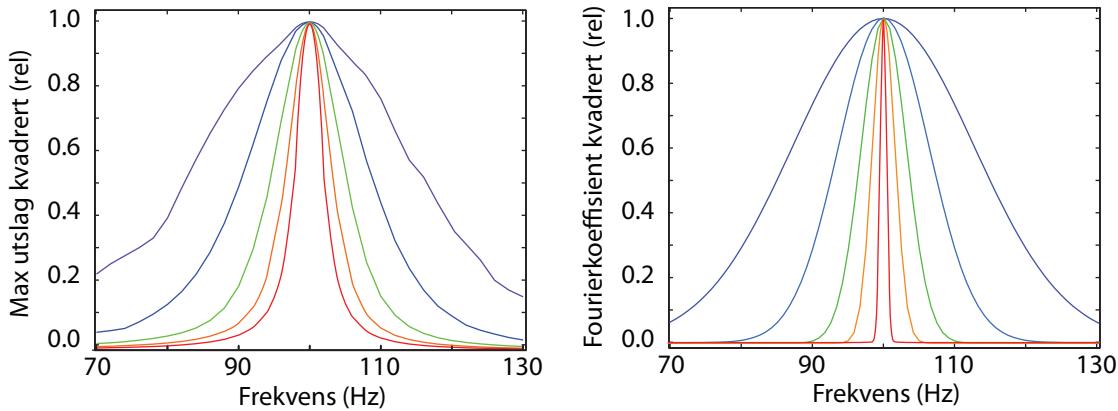
Dersom amplituden på kraften gjøres større, vil svingningene også få større amplitude, men da øker også tapet. Det viser seg derfor at varigheten av kraften som trengs for å få max amplitude er tilnærmet uavhengig av amplituden på kraften.

3.5 Frekvensrespons ved tidsbegrenset tvungen svingning

Det finnes en uventet konsekvens av å bruke kortvarige "kraftpulser". Vi skal ta opp dette temaet allerede nå, men kommer tilbake til dette temaet flere andre ganger i løpet av boka. Full forståelse av det aktuelle fenomenet vil du ikke få i dette delkapitlet. Forståelsen vil forhåpentligvis komme ettersom du har fått belyst fenomenet også på annet vis senere.

I figur 3.8 viste vi hvor stor svingeenergi (proporsjonalt med amplitude kvadrert) et system får dersom det utsettes for en harmonisk kraft som varer "uendelig lenge". Svingeenergien som oppnås ble plottet som funksjon av frekvensen til den påtrykte kraften. Et diagram som dette kalles gjerne for "frekvensresponsen" til systemet, og kurven kan brukes for å bestemme svingesystemets kvalitetsfaktor (Q-verdi) ut fra ligning (3.13). Jo smalere frekvensrespons, desto høyere Q-faktor.

Det er naturlig å bestemme frekvensresponsen også for det tilfellet at kraften bare varer en kort tid. Maksimal energi systemet oppnår som følge av kraften plottes som funksjon av senterfrekvensen til kraften. Relativ energi er proporsjonal med kvadratet av amplituden for svingningene.



Figur 3.12: Frekvensresponsen til det svingende systemet for ulike varigheter (σ) for kraften (venstre del). De σ -verdiene som er brukt er hhv. 25, 50, 100, 200, 400 og 800 ms (for blåfiolett/bredeste kurvene til rød/smalleste). I høyre del av figuren er det vist tilsvarende frekvensanalyser av selve kraften. Se teksten for ytterligere forklaringer.

Det viser seg da (venstre del av figur 3.12) at frekvensresponsen til systemet blir annerledes ved kortvarige "kraftpulser" enn ved uendelig lang varighet på den harmoniske kraften. Frekvensresponsen blir bredere og bredere (spreter seg over stadig større frekvensområde på begge sider av resonansfrekvensen) etter som kraftpulsen varer kortere og kortere tid.

Dersom vi derimot øker varigheten på "kraftpulsene" mer og mer, vil frekvensresponsen til systemet nå en grenseverdi. Det er en nedre grense for bredden i kurven, og dermed en maksimal grense for beregnet Q-faktor. Generelt sett brukes begrepet Q-faktor egentlig

bare for denne grenseverdien. For kortere kraftpulser angis frekvensresponsen heller enn å angi Q-verdi.

Det er imidlertid mulig å foreta en “frekvensanalyse” av selve *tidsforløpet til kraften*. Vi skal gå nærmere inn på hvordan dette gjøres i kapittel 4 når vi omtaler fourieranalyse. For å angi svært omtrentlig allerede her hva en frekvensanalyse går ut på, kan vi si at den forteller oss “hvilke frekvenser et signal består av” eller “hvilke frekvenser som må til for å lage det aktuelle signalet”.

Høyre del av figur 3.12 viser frekvensanalysen av “kraften som funksjon av tid” for de samme σ -verdiene som i venstre del av figuren. Figuren viser faktisk en klassisk analogi til Heisenbergs uskarphetsrelasjon. Dette var vi innom allerede i ligning (3.14), og vi kommer tilbake til dette på ny i kapittel 4.

Det viser seg at når σ avtar (kortere og kortere varighet på kraften), vil bredden på frekvensanalysen av kraften øke omvendt proporsjonalt med σ . For korte kraftpulser (så små σ -verdier at kraften bare svinger noen få ganger mens kraften har en betydelig verdi) er bredden på frekvensanalysen av kraften omtrent identisk med bredden av frekvensresponsen til systemet.

Ut fra disse observasjonene kan vi si at:

- Kvalitetsfaktoren er en parameter/størrelse som karakteriserer *det svingende systemet*. Jo mindre tap i systemet, desto høyere Q-faktor og smalere frekvensrespons, vel og merke for harmoniske krefter som varer lenge.
- Når kraften varer kort tid (få svingninger) er frekvensen for kraften dårlig definert. Når et svingende system utsettes for en slik kraft, domineres frekvensresponsen av *frekvenskarakteristikken til selve kraften*, og i mindre grad systemet selv.

3.6 Eksempel: Hørsel *

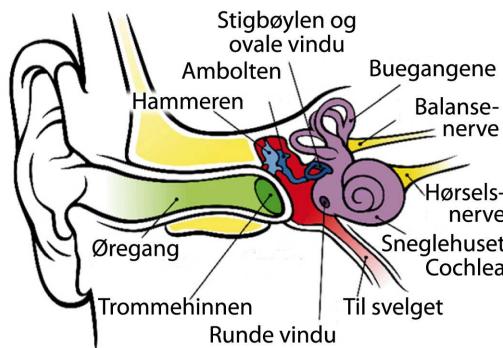
Til slutt i dette kapitlet vil vi si litt om vår hørsel og mekanismene som ligger bak. Tvangne svingninger står i sentrum i det følgende, mens andre sider av hørselen blir tema for tekst og oppgaver i et senere kapittel.

I øret vårt (se figur 3.13 og figur 3.14) vil lydbølger i luften føre til svingninger i øregang, trommehinne, de tre små ørekneklene og et system med tredelt væskerom i sneglehuset, og basillarmembranen (se f.eks. Wikipedia under “Basilar membrane”). Ulike deler av membranen vil svinge ved ulike frekvenser. Amplituden på svingningene blir “plukket opp” av følehår, og informasjonen sendt videre via nerver til hjernen (via ulike signalbehandlingsentre underveis).

[♠ ⇒ Det var biofysikeren Georg von Békésy fra Budapest som fant ut hvordan basillarmembranen fungerer som et “posisjon-frekvens-kart”. Han fikk Nobelprisen i fysiologi og medisin for dette arbeidet i 1961. ⇐ ♠]

Tidligere i dette kapitlet har vi omtalt tvangne svingninger. Analysen derfra kan forsøkes anvendt på svingninger i basillarmembranen. Basillarmembranen strekker seg dialektromatisk over det koniske hulrommet i sneglehuset i det indre øret (se figur 3.14).

Membranen kan vibrere, omtrent som klangbunnen i en fiolin, i takt med trykkvariasjonene i lyden. Membranen endrer imidlertid karakter fra ytterst i sneglehuset til de indre delene. Ytterst er den tynn og smal, mens lengre inne i sneglehuset blir den tykkere og bredere. Det fører til at dersom vi hører en mørk lyd (lav frekvens), vil bare den indre delen av basillarmembranen vibrere. Dersom vi hører en lys lyd (høy frekvens), vil bare den ytre delen vibrere. Dette er et fabelaktig design som gir oss mulighet til å kunne høre mange



Figur 3.13: Anatomiske strukturer i menneskeøret. Skissen er laget med utgangspunkt i en figur fra Wikipedia under oppslagsordet "Ear". (en.wikipedia.org/wiki/File:Anatomy_of_the_Human_Ear.svg)

ulike frekvenser samtidig som separate lydinntrykk. Vi kan høre både en basslyd og en diskantlyd samtidig, fordi de to lydstimuliene eksiterer ulike deler av basillarmembranen. Følehårene og nerveledningen plukker opp vibrasjoner fra ulike deler av membranen parallelt.

Basillarmembranen er et mekanisk svingesystem som oppfører seg på liknende måte som fjær-pendelen og RCL-kretsen når disse utsettes for tvungne svingninger. Ulike deler av membranen vil ha egenskaper som passer for det frekvensområdet delen skal respondere på. Vi kan tilordne ulike Q-verdier for ulike deler av basillarmembranen.

Ut fra det vi har lært i kapittel 2 må vi forvente at selv om vi hører en lyd som gir en harmonisk kraft på trommehinnen og bare en skarp frekvens, vil basillarmembranen svinge ikke bare ett sted, men over et område. Siden vi har "parallelprosessering" av signalene fra hårcellene, vil hjernen likevel kunne "beregne seg til" en ganske veldefinert senterfrekvens.

Dersom vi imidlertid lytter til kortere og kortere lydpulser, må vi forvente at bredere og bredere deler av basillarmembranen blir eksitert. Dette vil gjøre det vanskeligere for hjernen å avgjøre hvilken senterfrekvens lydpulsen hadde. Det betyr at det er vanskeligere å bestemme tonehøyden for en lyd når lyden varer meget kort tid.

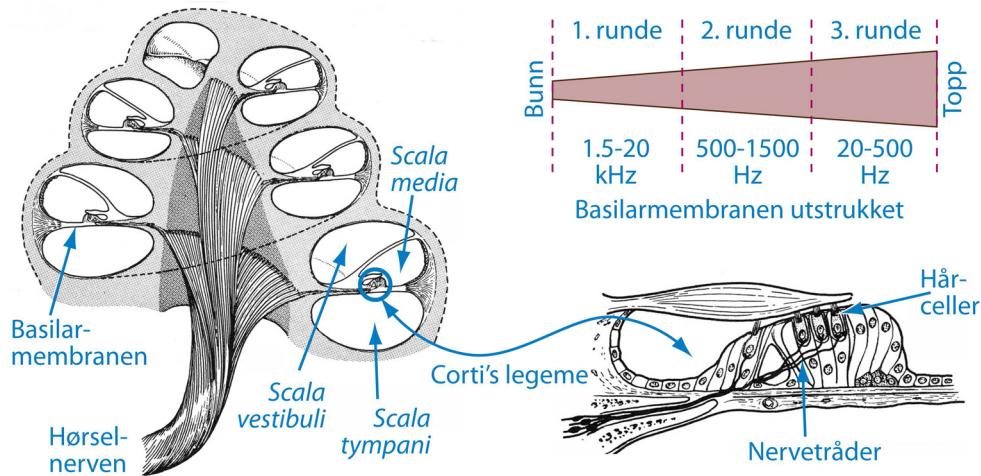
Når musikere spiller raske passasjerer på f.eks. en fiolin, kan de bomme *litt* på tonehøyden uten at vi merker det. Dersom de bommet like mye på toner som varer lenge, ville vi mye lettere kunne oppfange feilen.

Når lydpulsen bare varer én periode (og denne perioden f.eks. svarer til 1000 Hz), hører vi bare et "knepp". Det er umulig å fortelle hvilken frekvens som ble brukt for å lage selve lydbildet.

På den annen side er det lettere å oppfatte retningen til lydkilden til et knepp enn lydkilden til en vedvarende lyd. Evnen til bestemme tidspunktet nokså presist når en lyd opptrer, sammen med det faktum at vi har to ører, er meget viktig for å kunne bestemme retningen lyden kommer fra. (Bør likevel minne om at det finnes også andre mekanismer for å bestemme hvor en lyd kommer fra.)

Øret vårt er ifølge Darwins utviklingsteori et resultat av årtuseners utvelgelse av de mutasjoner som ga oss størst overlevelsesevne. Øret er derfor et system der det er et optimalt forhold mellom evne til å skille mellom ulike frekvenser og evne til å følge temmelig raske endringer med tiden. Resonans, tidsrespons og frekvensrespons er svært viktige detaljer for å forstå vår hørsel.

[♣] En interessant detalj mhp hørsel går på fase-sensitivitet. Nerveimpulser (de er digitale!) kan ikke sendes over nervefibre med en repetisjonsfrekvens stort høyere enn ca 1000 Hz. Det er derfor umulig for øret å sende signaler til hjernen med en bedre tidsoppløsning enn ca 1 ms. Det betyr at øret i prinsippet



Figur 3.14: Sneglehuset i det indre øret har en tredelt kanal som snor seg nesten tre runder fra bunn til topp. Det ovale vindu er knyttet til scala vestibuli og det runde til scala tympani. Når basillarmembranen vibrerer et sted, vil hårceller generere nervesignaler, og vi hører lyd. Figuren er omarbeidet fra McNaught og Callander: *Illustrated Physiology*, 2. utgave, Churchill Livingstone, 1972.

ikke kan gi opplysninger om fasen til en lydsvingning for frekvenser høyere enn noen få hundre hertz. (Noen er uenige og hevder at vi kan følge faser opp til 2000 Hz). Mest vanlig er å si at lydinntrykkene blir de samme uansett fasen på de ulike frekvenskomponentene i et lydsignal. ← ♠]

3.7 Læringsmål

Etter å ha jobbet deg gjennom kapittel 2 bør du ...

- kunne sette opp differensialligningen for et system som utsettes for tvungne harmoniske svingninger, og finne analytisk løsningen for denne når friksjonsleddet er lineært.
- kunne finne numerisk løsning av den nevnte differensialligningen også for ikke-lineært friksjonsledd og for ikke-harmonisk kraft (etter å ha vært gjennom kapittel 3).
- kunne finne uttrykk for resonansfrekvens, faseskift og kvalitetsfaktor for et enkelt mekanisk svingesystem.
- kunne sette opp et fasordiagram for å forklare typiske trekk for en RCL-krets for ulike frekvenser til en påtrykt spenning.
- kjenne til tidsforløpet for svingningene i en svingekrets idet en påtvunget kraft begynner og når den slutter.
- kjenne til hvordan responsen til et svingesystem endres når kraften bare varer ved en begrenset tidsperiode.

[♠ ⇒ Både for det mekaniske og elektriske svingesystemet vi har sett på til nå, ender vi opp med en svingeligning der en annen-derivert av en størrelse sammen med størrelsen i seg selv inngår. Det kan lede til en oppfatning at alle svingninger må beskrives ved en annengrads differensialligning.]

Det finnes imidlertid også svingninger som kan beskrives ved en første ordens differensiell ligning. Forutsetningen er at det er en markant tidsforsinkelse mellom "kraften" og "responsen". I biologien er slike sammenhenger ikke uvanlige. ← ♠]

3.8 Oppgaver

Forståelses- / diskusjonsspørsmål

1. For et mekanisk system ble faseforskyvningen på $\pi/2$ mellom utslag og påtrykt kraft forklart ved at denne konstellasjonen svarte til at kraften tilførte mest mulig effekt til systemet (maksimal kraft anvendt over lengst mulig vei). Forsøk å forklare faseforskyvningen på en liknende måte også for den elektriske RCL-kretsen med påtrykt harmonisk varierende spenning. Hvordan er forresten faseforskyvningen i en serie-RCL-krets?
2. Forsøk å forklare faseforskyvningen for serie-RCL-svingekretsen med påtrykt spenning i tilfellet når frekvensen er langt mindre og langt større enn resonansfrekvensen til kretsen alene. Ta utgangspunkt i hvordan impedansen til en kondensator og impedansen til en induktans endrer seg med frekvensen.

Regneoppgaver

3. En serie-RCL-krets består av en resistans R på 5.0 ohm, en kondensator C på $1.0 \mu\text{F}$, og en induktans L på $2.5 \mu\text{H}$.
 - a) Beregn resonansfrekvensene (både for fase- og amplituderesonans) for kretsen.
 - b) Beregn Q-verdien for kretsen.
 - c) Hvor stor faseforskjell er det mellom påtrykt spenning og strøm i kretsen ved faseresonans og ved en frekvens der strømamplituden er halvparten av det den var ved faseresonans?
 - d) Hvor stor bredde er det på frekvensresponsen til kretsen når den påtrykte spenningen varer "lenge"?
 - e) Hvor "lenge" må den påtrykte spenningen faktisk være for å oppnå en stasjonær tilstand (at amplituden ikke lenger endrer seg med tiden)?
 - f) Anta at kretsen blir påvirket av en "kraftpuls" med senterfrekvens lik resonansfrekvensen og at kraftpulsen har en gaussisk amplitude-omhyllingsfunksjon (ligning (3.16)) der σ har en verdi lik tre ganger periodetiden for senterfrekvensen for kretsen. Estimér bredden på frekvensresponsen til kretsen ved denne påvirkningen.
4. Ved gammeldags radiomottaking i mellombølgemrådet brukte vi svingekretser bestående av en induktans (spole) og en kapasitans (kondensator) for å skille en radiostasjon fra en annen. Radiostasjonene tok opp 9 kHz på frekvensbåndet, og to radiostasjoner kunne ligge så tett som 9 kHz. For at vi skulle kunne skille en radiostasjon fra en annen, måtte da mottakeren ha en variabel resonanskrets som passet til én radiostasjon, men ikke til en annen. Frekvensen på Stavanger-senderen var 1313 kHz. Hvilken Q-faktor må radiomottakerens resonanskrets ha? [Disse betraktningene er fortsatt gjeldende i vår moderne tid, selv om digitalteknikken gir visse endringer.]
5. a) Vi har en elektrisk krets som består av en seriekoppling av en resistans R , en kapasitans C , en induktans L og en spenningskilde $V(t) = V_m \cos(\omega t)$. Tidsmidlet av effekten \bar{P} , avsatt i motstanden, kan skrives som:

$$\bar{P} = \frac{\frac{1}{2}RV_m^2}{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

Bruk dette uttrykket til å vise at resonansfrekvensen er $f_0 = \frac{1}{2\pi\sqrt{LC}}$. Hva menes med halvverdibredden Δf . Illustrer gjerne svaret med en figur. Vis at Q-verien til kretsen er $\frac{\omega_0 L}{R}$

b) Vi har en radiomottaker som består av en RCL -krets som i a). $L = 0.1 \mu\text{H}$. To nærliggende radiostasjoner, stasjon 1 og stasjon 2, sender på henholdsvis $f_1 = 100.0 \text{ MHz}$ og $f_2 = 100.4 \text{ MHz}$. Vi antar at signalstyrken fra de to stasjonene er like. Vi ønsker å tilpasse komponentene i vår RCL -krets slik at når vi stiller den inn på stasjon 1, vil signalet fra stasjon 2 være om lag 1 % av signalet fra stasjon 1. På den måten vil stasjon 1 ikke forstyrre oss når vi lytter på stasjon 2. Bestem kapasitansen C , motstanden R og kvalitetsfaktoren Q . Det er oppgitt at $Q = \frac{f_0}{\Delta f} = \frac{\omega_0 L}{R}$.

6. Lydpulsene flaggermus bruker for å orientere seg (og finne et bytte) har en frekvens på 40-100 kHz. Anta at Q-verdien for hørselen ved disse frekvensene er omtrent 100.
 - a) Finn minste avstand mellom flaggermusa og f.eks. en vegg for at flaggermusa skal klare å oppfatte et ekko etter en kort lydpuls. Lydhastigheten i luft er om lag 340 m/s. (Hint: Bruk ligning (3.12)).
 - b) Hvor stor ville den minste avstanden vært dersom flaggermusa brukte lydpulser ved om lag 1000 Hz (anta fortsatt at $Q \approx 100$)?
7. Søk på web og finn fram til minst ti ulike former for resonans innen fysikk. Angi en web-adresse hvor vi kan lese litt om hver av disse formene for resonans.
8. Utled uttrykkene gitt i ligning (3.9) fra ligning (3.10) og andre uttrykk for en svингende fjær-pendel.

Eksempler på tidligere eksamensoppgaver

9. a) Skriv opp en generell svingeligning med og uten demping og med og uten en påtrykt harmoinsk tidsvariabel kraft. (Velg selv om ligningen skal beskrive et lodd i en fjær, en pendel, en torosjonspendel, en elektrisk svingekrets eller andre systemer.)
 b) Skissér hovedlinjene i hvordan vi kan gå fram for å finne en løsning av den mest generelle av disse ligningene.
 c) Følg hovedlinjene i utledningen noen få trinn, men nok til at du kan angi hva som fører til underkritisitk, kritisitk og overkritisitk demping.
 d) Skissér hvordan disse løsningene ser ut når vi ikke har en påtrykt harmonisk tidsvarierende kraft. Spesifiser hvilke initialbetingelser du tar utgangspunkt i for dine skisser.
 e) Forklar kort hvordan løsningen vil se ut når også den påtrykte harmoniske tidsvarierende kraften er til stede. Også her er det fint om du sier noe om initialbetingelsene du velger for din forklaring.
10. a) Vi har en elektrisk krets som består av en seriekobling av en resistans, R , en kapasitans, C , og en induktans, L , og en spenningskilde, $V(t) = V_m \cos(\omega t)$. (Den samme som vist i figur 3.3). Vis at differensialligningen som beskriver strømmen $I(t)$ i kretsen kan skrives som

$$a \frac{d^2I}{dt^2} + b \frac{dI}{dt} + cI = U(t)$$

Uttrykk a , b , c og U ved R , C , L og $V(t)$.

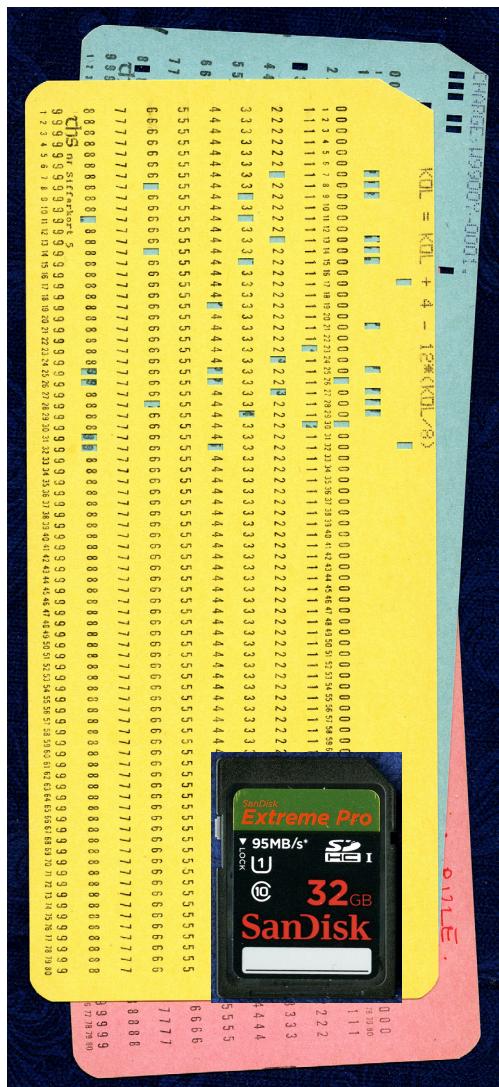
Vi fjerner nå spenningskilden $V(t)$ og erstatter den med en kortslutning i resten av

oppgaven.

- b) Vis at strømmen $I(t) = A \cdot e^{\gamma \cdot t} \cos(\omega t + \psi)$ kan være en løsning av differensialligningen i a) (når altså $V(t) = 0$). A , γ , ω og ψ er konstanter. Hvilke av disse er bestemt av differensialligningen? Bestem i så fall disse uttrykt ved R , L og C .
- c) Ladningen på kondensatoren ved tiden t kan skrives som $q(t) = Q_m e^{-\frac{R}{2L} \cdot t} \cos(\omega t + \psi)$. Anta at strømmen i kretsen ved $t = 0$ er null, dvs. $I(t = 0) = 0$, men at det er en ladning Q på kondensatoren. Bestem $q(t)$ for $t > 0$ uttrykt ved I_0 , R , L og C .
- e) Bestem R for gitt L og C slik at kretsen blir kritisk dempet.

Kapittel 4

Numeriske løsningsmetoder



I min studietid (1969-1974) hadde Norges største data-maskin 250 kb hukommelsete og fylte et helt rom. Vi laget programmer ved å punche, hver linje - ett hullkort. Kortbunken ble båret forsiktig til Abels hus (det var en katastrofe å miste den!). Noen timer opp til et døgn senere kunne vi hente resultatet i form av en utskrift på sideperforerte ark. En punsje feil førte til at et kort måtte punches på nytt, og riktig kort i kortbunken byttes. Så var det en ny innlevering og ny venting. Gjett om utesting av programmer tok lang tid!

I dag er situasjonen totalt forskjellig. Datamaskinen er allemans eie. Programutvikling er utrolig effektiv sammenlignet med tidligere tider. Og numeriske metoder har blitt et like naturlig verktøy som analytisk matematikk.

Men verktøy er som verktøy flest: Det trengs opplæring i hvordan de brukes. I dette kapitlet skal vi først og fremst se hvordan svingeligningen og senere bølgeligningen kan løses på en god måte. Det er ikke nok å lese seg til hvordan ting kan gjøres. Praktisk trening må til for å få den nødvendige ferdigheten og rutinen.

Eksempler på datakort (65% av naturlig størrelse), sammen med en moderne brikke (i naturlig størrelse) med lagerkapasitet svarende til 400 millioner hullkort (som hadde veid 950 tonn!).

¹Deler av teksten er skrevet av David Skålid Amundsen som en sommerjobb for CSE 2008. Amundsens tekst er siden bearbeidet og utvidet av Arnt Inge Vistnes. Copyright 2013 for tekst og figurer: David Skålid Amundsen og Arnt Inge Vistnes. Versjon 14012013.

4.1 Innledning

Når vi i “gamle dager” (det vil si for mer enn 15 år siden) skulle beregne bevegelsen til en matematisk eller fysisk pendel i et laveregrads fysikkurs, måtte vi nøye oss med “små utslag”. Vi hadde den gang bare analytisk matematikk i verktøykassen, og da var det bare små utslag der bevegelsen er tilnærmet en ren harmonisk svingning vi kunne håndtere. Større utslag er det langt vanskeligere å håndtere rent analytisk, og dersom vi eventuelt legger inn kompliserende friksjon, finnes det rett og slett ikke noe analytisk løsning på problemet.

Når vi har lært oss å bruke numeriske løsningsmetoder, er det ofte omrent like enkelt å bruke en realistisk, ikke tilnærmet beskrivelse av en bevegelse som en idealisert forenklet beskrivelse.

Denne boka er basert på at leseren allerede kjenner litt til hvordan vi løser f.eks. differensialligninger ved hjelp av numeriske metoder. Likevel tar vi med en rask repetisjon av noen av de enkleste løsningsmetodene slik at de som ikke har tidligere erfaring med numeriske løsningsmetoder, tross alt skal kunne henge med. Etter den raske gjennomgangen av de enkle metodene, bruker vi litt mer tid på en mer robust metode. I tillegg sier vi litt om hvordan disse metodene kan generaliseres til å løse partielle differensialligninger.

Det bør allerede her nevnes at de enkleste numeriske løsningsmetodene ofte er gode nok for f.eks. å beregne en kastbevegelse, også når friksjon er til stede. De enkleste metodene summerer imidlertid ofte opp feil og gir ganske dårlige resultat for svingbevegelser. Det er med andre ord ofte nødvendig å bruke litt avanserte numeriske metoder når vi gjør beregninger på svingninger og bølger.

Dette kapitlet er bygd opp på følgende måte:

Først gis en rask gjennomgang av de enkleste numeriske metodene som brukes ved løsning av differensialligninger. Dernest beskrives Runge-Kuttas metode av fjerde orden. Denne første delen av kapitlet er temmelig matematisk.

Deretter gis det et praktisk eksempel, og til slutt tar vi med eksempler på programkoder som kan anvendes i oppgaveløsning videre i boka.

4.2 Grunnleggende idé bak numeriske metoder

I mange deler av fysikken møter vi annen ordens ordinære differensialligninger:

$$\ddot{x} \equiv \frac{d^2x}{dt^2} = f(x(t), \dot{x}(t), t). \quad (4.1)$$

I mekaniske system har differensialligningen ofte opphav i Newtons 2. lov. I elektriske svingekretser er det gjerne Kirchhoffs lov sammen med generalisert Ohms lov og komplekse impedanser som er opphav til differensialligningene.

Uttrykket $f(x(t), \dot{x}(t), t)$ sier at f (i tilfelle x er en posisjonsvariabel og t tiden) er en funksjon av såvel tid, posisjon og hastighet.

Når vi løser annen ordens differensialligninger numerisk, betrakter vi ofte ligningen som en kombinasjon av to koblede første ordens differensialligninger:

$$\frac{d\dot{x}}{dt} = f(x(t), \dot{x}(t), t)$$

$$\frac{dx}{dt} = g(x(t), t)$$

hvor g er en funksjon som angir den tidsderiverte av x . Vi vil straks se noen enkle eksempler på dette i praksis.

4.3 Eulers metode og varianter av denne

Vi kan løse en differensialligning ved å angi en startverdi for den løsningen vi er interessert i, og bruke vår kunnskap om funksjonens deriverte for å beregne løsningen en kort tid Δt etterpå. Vi lar så den nye verdien fungere som ny startverdi for å beregne verdien som følger Δt etter dette igjen. Slik fortsetter vi helt til vi har beskrevet løsningen i så mange punkter n som vi er interessert i.

Utfordringen er å finne hvordan vi kan bestemme neste verdi ut fra den vi allerede kjenner. Det kan gjøres på en mer eller mindre raffinert metode. Den enkleste metoden er kanskje Eulers metode. Den tar utgangspunkt i den velkjente definisjonen av den deriverte:

$$\dot{x}(t) = \lim_{\Delta t \rightarrow 0} \frac{x(t + \Delta t) - x(t)}{\Delta t}$$

Dersom Δt er tilstrekkelig liten, kan vi omskrive dette uttrykket og få:

$$x(t + \Delta t) \approx x(t) + \Delta t \dot{x}(t)$$

Anta at startverdiene er gitt ved: (x_n, \dot{x}_n, t_n) . Da følger for en første ordens differensialligning:

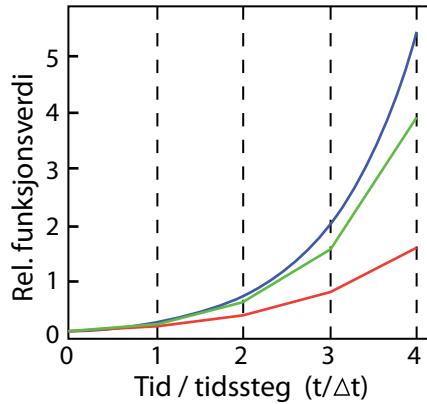
$$x_{n+1} = x_n + \dot{x}_n \Delta t$$

Ved å bruke en slik oppdateringsligning på både $x(t)$ og $\dot{x}(t)$, får vi den velkjente Eulers metode (i vår sammenheng for løsning av annen ordens differensiallikning):

$$\dot{x}_{n+1} = \dot{x}_n + \ddot{x}_n \Delta t$$

$$x_{n+1} = x_n + \dot{x}_n \Delta t$$

Vi har altså to koblede differensialligninger.



Figur 4.1: *Eulers enkle metode for å beregne en funksjon numerisk. Den øverste (blå) kurven er den korrekte. Den nederste (røde) kurven er beregnet ved hjelp av Eulers enkle metode, mens den midtre er beregnet med midtpunktsmetoden. Tidssteget er det samme i begge tilfeller og er valgt meget stort for å få fram prinsippet i metodene.*

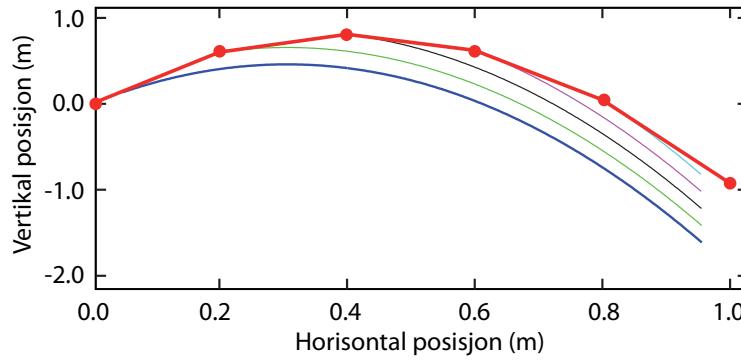
Figur 4.1 skisserer hvordan metoden virker. Dette er den mest vanlige måten å lage en slik illustrasjon på, men den gir etter min mening bare en overfladisk forståelse. For hva skjer når vi får større og større avvik mellom den korrekte løsningen og den beregnede løsningen? Her er det noen detaljer vi bør kjenne til.

I figur 4.2 er det vist en figur lignende den i figur 4.1, men med en rekke nye kurver. Den mellomtykke blå kurven (nederst) viser hvordan et skrått kast vil forløpe for en initiell hastighet på 1.0 m/s i horisontal retning og 3.0 m/s i vertikal oppover retning. Beregningen er basert på en analytisk løsning av dette enkle problemet.

I figuren er det også tegnet inn løsningen når Eulers metode er benyttet med meget store tidssteg (0.2 s). Allerede etter ett tidssteg er den beregnede nye posisjonen temmelig langt fra det den skulle vært.

Etter dette tidssteget er det beregnet nye verdier for posisjon og hastighet i både horisontal og vertikal retning. Det er disse verdiene som nå plugges inn i differensialligningen. Hadde vi beregnet banen for akkurat disse verdiene, ville vi fått løsningen gitt ved en grønn kurve (nest nederst). *Dette er en annen løsning av differensialligningen enn den vi opprinnelig startet ut med!*

Heller ikke nå klarer vi å følge denne nye løsningen på en god måte siden tidssteget er så stort, og når vi bruker Eulers metode enda en gang, får vi en posisjon (og hastigheter) som ligger temmelig langt også fra den andre løsningen av differensialligningen vi var innom.



Figur 4.2: Beregning av et skrått kast ved hjelp av Eulers metode med stort tidssteg. Hvert nytt punkt som beregnes er utgangspunkt for en ny løsning av den opprinnelige differensialligningen. Se teksten for detaljer.

Slik fortsetter det hele. For hvert nytt tidssteg er vi innom en ny løsning av differensialligningen, og i vårt tilfelle er feilen som gjøres systematisk og gir større og større feil etter hvert tidssteg.

Det kan vises at dersom vi reduserer tidssteget betydelig (!) i forhold til hva som er brukt i figur 4.2, blir løsningen langt bedre enn i figuren. Likevel er det ikke alltid nok å bare redusere tidssteget.

For det første kan vi ikke redusere tidssteget så mye at vi får problemer med hvor nøyne tall kan angis på en datamaskin (uten at vi må anvende ekstremt tidkrevende teknikker). Når vi beregner $x_{n+1} = x_n + \dot{x}_n \Delta t$, må ikke bidraget $\dot{x}_n \Delta t$ bestandig være så lite at det bare kan bidra i det minst signifikante sifferet til x_{n+1} .

En annen begrensning ligger i den numeriske metoden i seg selv. Gjør vi systematiske feil som adderer seg til å bli værre og værre for hvert tidssteg, uansett hvor små tidsstegene er, får vi også problemer. Da må vi bruke andre numeriske metoder enn denne aller enkleste Euler.

En forbedret versjon av Eulers metode blir kalt *Euler-Cromers metode*. Anta at startverdiene er (x_n, \dot{x}_n, t_n) . Første trinn er identisk med Eulers enkle metode:

$$\dot{x}_{n+1} = \dot{x}_n + \ddot{x}_n \Delta t.$$

Andre trinn i Euler-Cromers metode adskiller seg imidlertid fra den enkle Euler: For å finne x_{n+1} , bruker vi \dot{x}_{n+1} og ikke \dot{x}_n som vi gjør i Eulers metode. Det gir følgende oppdateringslikning for x :

$$x_{n+1} = x_n + \dot{x}_{n+1} \Delta t.$$

Årsaken til at Euler-Cromers metode fungerer, og at den ofte (men ikke alltid) fungerer bedre enn Eulers metode er ikke triviell, og vi skal ikke gå nærmere inn på det her. Eulers metode fører ofte til at energien i det modellerte systemet ikke er bevart, men øker sakte men sikkert etterhvert. Dette problemet blir dramatisk redusert med Euler-Cromers metode, som gjør at den i de fleste tilfeller fungerer bedre.

En annen forbedring av Eulers metode, som er enda bedre enn Euler-Cromers metode, er *Eulers midtpunktsmetode*. Istedentfor å bruke stigningstallet i begynnelsen av intervallet, og bruke dette i hele intervallet, bruker vi stigningstallet i midten av intervallet. Ved å bruke stigningstallet i midten av intervallet vil vi normalt få et mer nøyaktig resultat enn ved å bruke stigningstallet i begynnelsen av intervallet når vi er ute etter å finne den gjennomsnittlige vekstraten.

I Eulers midtpunktsmetode bruker vi først stigningstallet i begynnelsen av intervallet, men istedenfor å bruke dette stigningstallet for hele intervallet, bruker vi det for halve intervallet. Så beregner vi stigningstallet i midten av intervallet og bruker dette for hele intervallet. Matematisk blir dette, ved å bruke samme notasjon som tidligere:

$$\begin{aligned}\dot{x}_{n+1/2} &= \dot{x}_n + f(x_n, \dot{x}_n, t_n)\Delta t/2, \\ x_{n+1/2} &= x_n + \dot{x}_n\Delta t/2.\end{aligned}$$

Her er altså $\dot{x}_{n+1/2}$ og $x_{n+1/2}$ verdien til den ukjente funksjonen og verdien til den deriverte av denne funksjonen midt i intervallet. Oppdateringsligningen for hele intervallet blir så:

$$\begin{aligned}\dot{x}_{n+1} &= \dot{x}_n + f(x_{n+1/2}, \dot{x}_{n+1/2}, t_{n+1/2})\Delta t, \\ x_{n+1} &= x_n + \dot{x}_{n+1/2}\Delta t.\end{aligned}$$

4.4 Runge-Kuttas metode

I Eulers metode fant vi neste verdi ved å bruke stigningstallet i starten av det steget vi skal ta. I Eulers midtpunktsmetode brukte vi stigningstallet i midten av steget vi skal ta. I begge tilfeller er det nokså lett å tenke seg at for noen funksjoner vil vi kunne få en systematisk feil som vil summeres opp til betydelige totale feil etter hvert som det gjennomføres mange etterfølgende beregninger. Det kan vises at feilen vi gjør blir vesentlig mindre dersom vi går over til å bruke enda mer raffinerte metoder for å finne neste verdi. En av de mest populære metodene kalles Runge-Kuttas metode av fjerde orden. Hele fire forskjellige estimater for stigningstallet, ett i begynnelsen, to i midten og ett i slutten blir da brukt for å beregne det gjennomsnittlige stigningstallet i intervallet. Dette gjør Runge-Kutta metoden mye bedre enn Eulers midtpunktsmetode, og siden den ikke er stort vanskeligere å programmere, er det denne som ofte blir brukt i praksis.

La oss se hvordan Runge-Kuttas metode av 4. orden fungerer, og hvordan den kan brukes for å løse en annen ordens differensialligning. (Helt til slutt i kapitlet følger pseudokode og full kode for et program som bruker Runge-Kuttas metode av fjerde orden.)

4.4.1 Beskrivelse av metoden

Runge-Kuttas metode er faktisk ikke så vanskelig å forstå, men la oss si litt om notasjonen først. Ta utgangspunkt i differensialligningen gitt ved

$$\ddot{x}(t) = f(x(t), \dot{x}(t), t). \quad (4.2)$$

Anta at vi befinner oss i punktet (x_n, \dot{x}_n, t_n) og at skritt lengden for tidsøkningen er Δt . I det videre kommer vi til å finne estimater for x_n , \dot{x}_n og \ddot{x}_n , og hvilket nummer estimatet

har i rekken blir angitt med en ny indeks. Det første estimatet til \ddot{x}_n blir angitt som $a_{1,n}$, det første til \dot{x}_n blir angitt som $v_{1,n}$ og det første til x_n blir angitt som $x_{1,n}$. Det andre estimatet til \ddot{x}_n blir da angitt som $a_{2,n}$, og tilsvarende for resten av estimatene.

Vi kan finne første estimatet til \ddot{x}_n ved å bruke ligning (4.2):

$$a_{1,n} = f(x_n, \dot{x}_n, t_n).$$

Samtidig er den førstederiverte kjent i starten av intervallet:

$$v_{1,n} = \dot{x}_n.$$

Neste skritt på veien er å bruke Eulers metode for å finne $\dot{x}(t)$ og $x(t)$ i midten av intervallet:

$$\begin{aligned} x_{2,n} &= x_{1,n} + v_{1,n} \frac{\Delta t}{2}, \\ v_{2,n} &= v_{1,n} + a_{1,n} \frac{\Delta t}{2}, \end{aligned}$$

Videre kan vi finne et estimat for den andrederiverte i midten av intervallet ved å bruke $v_{2,n}$, $x_{2,n}$ og ligning (4.2):

$$a_{2,n} = f(x_{2,n}, v_{2,n}, t_n + \Delta t/2).$$

Neste skritt på veien i Runge-Kuttas metode er å bruke den nye verdien for den andrederiverte i midten av intervallet til å finne et nytt estimat for $x(t)$ og $\dot{x}(t)$ i midten av intervallet ved hjelp av Eulers metode:

$$\begin{aligned} x_{3,n} &= x_{1,n} + v_{2,n} \frac{\Delta t}{2}, \\ v_{3,n} &= v_{1,n} + a_{2,n} \frac{\Delta t}{2}. \end{aligned}$$

Med det nye estimatet for $x(t)$ og $\dot{x}(t)$ i midten av intervallet kan vi finne et nytt estimat for den andrederiverte i midten av intervallet:

$$a_{3,n} = f(x_{3,n}, v_{3,n}, t_n + \Delta t/2).$$

Ved hjelp av det nye estimatet for den andrederiverte i tillegg til estimatet for den førstederiverte i midten av intervallet kan vi nå bruke Eulers metode for å finne et estimat for $x(t)$ og $\dot{x}(t)$ i slutten av intervallet. Det gjøres da slik:

$$x_{4,n} = x_{1,n} + v_{3,n} \Delta t,$$

$$v_{4,n} = v_{1,n} + a_{3,n} \Delta t.$$

Til slutt kan vi på tilsvarende måte som før finne et estimat for $\ddot{x}(t)$ i slutten av intervallet ved hjelp av disse nye verdiene:

$$a_{4,n} = f(x_{4,n}, v_{4,n}, t_n + \Delta t)$$

Vi kan nå regne ut et vektet gjennomsnitt av estimatene, og vi får da et rimelig godt estimat for den gjennomsnittlige verdien til den andrederiverte og førstederiverte i intervallet:

$$\begin{aligned} a_n &= \frac{1}{6} (a_{1,n} + 2a_{2,n} + 2a_{3,n} + a_{4,n}) \\ v_n &= \frac{1}{6} (v_{1,n} + 2v_{2,n} + 2v_{3,n} + v_{4,n}) \end{aligned}$$

Ved hjelp av disse gjennomsnittene, som er ganske gode tilnærminger til middelverdiene av stigningstall over hele intervallet, kan vi bruke Eulers metode for å finne et godt estimat for $x(t)$ og $\dot{x}(t)$ i slutten av intervallet:

$$x_{n+1} = x_n + v_n \Delta t \quad (4.3)$$

$$v_{n+1} = v_n + a_n \Delta t \quad (4.4)$$

$$t_{n+1} = t_n + \Delta t \quad (4.5)$$

Dette er da ekvivalent med startverdiene for neste intervall.

I Runge-Kuttas metode henter vi ut mye mer informasjon fra differensialligningen enn i Eulers metode. Dette gjør Runge-Kuttas metode betydelig mer stabil enn Eulers metode, Euler-Cromers metode og Eulers midtpunktsmetode. Runge-Kuttas metode krever ikke så enormt mye mer ressurser for en datamaskin å utføre, samtidig er den forholdvis enkel å programmere. Runge-Kuttas metode, i en eller annen variant, er derfor ofte den metoden vi først tyr til når vi ønsker å løse ordinære differensialligninger numerisk.

Programmeringen av den grunnleggende delen av Runge-Kutta gjøres nærmest en gang for alle. Det er bare to små filer som endres fra ett problem til et annet. Disse to filene angir nøyaktig differensialligningene som skal benyttes i akkurat de beregningene som da skal gjennomføres.

For mer stoff om Runge-Kuttas metode se Tom Lindstrøms “Kalkulus” eller Wikipedia.

4.5 Partielle differensialligninger

Mange fysiske problemer er beskrevet ved hjelp av partielle differensialligninger, kanskje de mest kjente er Maxwells ligninger, Schrödingerligningen og bølgeligningen. Med partielle differensialligninger menes at den ukjente funksjonen består av flere variable og at derivasjon med hensyn på disse inngår i differensialligningen.

Det finnes flere forskjellige løsningsmetoder for partielle differensialligninger, men den enkleste og letteste å forstå er nok den såkalte “endelig-differanse-metoden” (finite difference method). Den går ut på å erstatte differentialene i differensialligningen med endelige differanser. Betrakt den enkle differensialligningen

$$\frac{\partial y}{\partial x} = K \frac{\partial y}{\partial t}. \quad (4.6)$$

Den enkleste måten å gjøre om differentialene i denne ligningen til diskrete differensialer, er å bruke definisjonen av den deriverte, som vi har gjort tidligere. Ligningen over blir da

$$\begin{aligned} \frac{y(x + \Delta x, t) - y(x, t)}{\Delta x} &= \\ K \frac{y(x, t + \Delta t) - y(x, t)}{\Delta t}. \end{aligned}$$

Denne ligningen kan så løses med hensyn på $y(x, t + \Delta t)$, som gir

$$\begin{aligned} y(x, t + \Delta t) &= \\ y(x, t) + \frac{\Delta t}{K \Delta x} (y(x + \Delta x, t) - y(x, t)). \end{aligned}$$

Dersom $y(x, t)$ er kjent ved et tidspunkt $t = t_0$, slik at $y(x, t_0) = f(x)$, gir ligningen over funksjonsverdien ved neste tidspunkt $y(x, t_0 + \Delta t)$. Men, legg merke til at vi også bruker funksjonsverdien ved en annen x enn den vi gjør beregning på i uttrykket over. Det betyr

at vi får et problem når vi skal beregne funksjonsverdier for x nær yttergrensen for det området vi gjør beregninger for. Av ligningen over ser vi at vi må vite hva funksjonen var ved neste x -koordinat ved forrige tidspunkt, og i det ytterste x -punktet går ikke dette.

Dette medfører at vi må kjenne *randbetingelsene*^a, det vil si tilstanden til systemet i grensen av området vi ser på, for at problemet skal ha en entydig løsning. Disse må avklares før beregningene i det hele tatt kan begynne. Randbetingelser finnes også for ordinære differensialligninger, men i partielle differensialligninger er de mye mer fremtredende og kommer i tillegg til initialbetingelsene.

^aInitialbetingelser og randbetingelser er to forskjellige ting, og må ikke blandes sammen. Initialbetingelser sier noe om hvilken tilstand systemet befinner seg i helt i begynnelsen av beregningene, og må også brukes her. Randbetingelsene sier noe om systemets tilstand ved endepunktene av beregningene ved *alle* tidspunkt.

De endelige differensialene vi brukte over blir derimot sjeldent brukt, da de kan ertattes av noe som er bedre og ikke stort mer vanskelig å forstå. Istedetfor å bruke Eulers metode i differensialene som over, brukes Eulers midtpunktsmetode, som reduserer feilen i beregningene betraktelig. Dersom vi gjør dette, blir diskretiseringen av ligning (4.6) slik:

$$\frac{y(x + \Delta x, t) - y(x - \Delta x, t)}{2\Delta x} = K \frac{y(x, t + \Delta t) - y(x, t - \Delta t)}{2\Delta t}.$$

Det er ikke så vanskelig å forstå at resultatet nå vil bli bedre, for istedetfor beregne den gjennomsnittlige veksten gjennom det aktuelle punktet og neste punkt, brukes den gjennomsnittlige veksten gjennom forrige og neste punkt. På samme måte som isted kan denne ligningen løses med hensyn på $y(x, t + \Delta t)$, og resultatet blir:

$$y(x, t + \Delta t) = y(x, t - \Delta t) + \frac{\Delta t}{K\Delta x} [y(x + \Delta x, t) - y(x - \Delta x, t)]$$

Vi ser at vi får samme problem med randbetingelser som over, faktisk krever den en randbetingelse til, også i begynnelsen av x -gridet. Siden dette er et problem som omhandler en romlig dimensjon, må vi angi to randbetingelser for at løsningen skal være entydig (det er to randpunkter). For å bruke den første oppdateringsligningen må vi altså ta hensyn til den andre randbetingelsen også.

På samme måte som vi erstattet førstederiverte med et endelig differensial, kan n -tederiverte tilnærmes på samme måte. Et eksempel er den andrederiverte som kan tilnærmes med følgende differensial:

$$f''(x) \approx \frac{f(x + \Delta x) - 2f(x) + f(x - \Delta x)}{\Delta x^2}$$

Bevis: Taylorutviklingen til funksjonen $f(x)$ om $x = a$ blir:

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2 + \dots$$

En tilnærming til $f(a + \Delta x)$ er gitt ved

$$f(a + \Delta x) \approx f(a) + f'(a)(a + \Delta x - a) + \frac{f''(a)}{2}(a + \Delta x - a)^2,$$

som gir

$$f''(a) \approx \frac{2}{\Delta x^2} (f(a + \Delta x) - f(a) - f'(a)\Delta x).$$

Ved å sette inn for $f''(a)$ ved å bruke ligningen over for den førstederiverte, gir det

$$f''(a) \approx \frac{2}{\Delta x^2} (f(a + \Delta x) - f(a)) - \frac{2}{\Delta x^2} \left(\frac{f(a + \Delta x) - f(a - \Delta x)}{2\Delta x} \Delta x \right).$$

Ved å rydde får vi formelen

$$f''(a) \approx \frac{f(a + \Delta x) - 2f(a) + f(a - \Delta x)}{\Delta x^2}. \quad (4.7)$$

Vi kan også bevise samme uttrykk ved å ta utgangspunkt i den deriverte av den deriverte, og anvende den enkleste definisjonen av den deriverte. Metoden har den fordelen at det fremgår tydeligere hvorfor vi må kjenne funksjonsverdien i (minst) tre punkt for at vi overhodet skal kunne beregne en annenderivert.

Som med de ordinære differensialligningene kan vi gå videre og bruke metoder som gir et enda bedre resultat.

Det finnes en rekke metoder tilgjengelig for ulike deler av fysikken. Interesserte henvises til spesielle kurs/bøker i numeriske beregninger.

4.6 “Dimensjonsløs” differensialligning

Når vi løser fysiske problemer numerisk, brukes ofte såkalte “dimensjonsløse variable”. Grunnen er at vi lett kan miste numerisk presisjon om vi ikke gjør dette. Et reelt tall i en datamaskin kan bare være innenfor et bestemt intervall, og dersom vi i beregningene kommer nær grensene for hva som kan representeres med full presisjon, går det ut over nøyaktigheten i beregningene. Selv om et tall i seg selv kan være langt fra grensene, vil vi kunne støte på høyere potenser av tallet i en eller annen mellomregning, og da kan skaden være gjort.

Av den grunn forsøker vi å holde alle variable på et nivå som ikke er alt for langt fra 1. Når vi f.eks. regner på en atomkjerne og har dimensjoner langt ned i 10^{-15} m, eller gjør beregninger innen kosmologi og har avstander på over en milliard lysår, dvs over $9.46 \cdot 10^{24}$ m, er bruk av såkalte dimensjonsløse variable ofte gunstig.

En annen grunn til å bruke dimensjonsløse variable, er at vi kan løse et problem for en hel klasse problemer i en og samme beregning, og at vi kan finne løsning av alle varianter av et problem ved bare å skalere løsningen som er oppnådd ved hjelp av dimensjonsløse variable.

I vår sammenheng er det ikke mye å vinne på å bruke dimensjonsløse variable. Grunnen er at vi oftest arbeider med fenomener fra dagligdagse fenomener der SI-enhetene meter og sekund er meget velegnet. Vi skal heller ikke studere mange varianter av eksakt samme type problem, og får derfor ikke noe særlig rasjonaliseringsevinst ved å velge dimensjonsløse variable. Likevel nevner vi hovedprinsippet som benyttes slik at du kan velge selv om du vil bruke teknikken eller ikke.

Som et eksempel velger vi å ta utgangspunkt i differensialligningen som beskriver en dempet svingning av en fjærpendel (ligning (1.8) i kapittel 1). Den kan skrives som:

$$\frac{d^2z}{dt^2} + 2\gamma \frac{dz}{dt} + \omega^2 z = 0 \quad (4.8)$$

Her er $2\gamma = b/m$, det vil si friksjonskoeffisienten dividert på massen, mens $\omega^2 = k/m$ hvor k er fjærstivheten.

Alle ledd i summen skal ha samme enhet, nemlig m/s². Du bør verifisere dette selv.

Vi vil nå skalere alle lengde- og tidsangivelser, og skriver:

$$z = z_0 \hat{z}$$

$$t = t_0 \hat{t}$$

Her er z_0 en referanselengde, angitt med riktig benevning/enhet. Skal vi studere en fjærpendel med amplitud f.eks. 50 cm, kunne vi latt z_0 være 0.5 m. Men siden dette er en størrelse så nær opp til SI-enheten 1 m, og så lenge det ikke er andre faktorer i ligningen som ville bli forenklet av et annet valg, vil det i dette tilfellet være naturlig å la z_0 være lik 1 m. I så fall vil \hat{z} være ubenevnt, nemlig lik måletallet for en lengde (posisjon rel. likevektpunktet) når måleenheten er 1 m.

For tiden er det litt annerledes. Her kan vi velge SI-enheten 1 sek, eller vi kan velge en tidsenhet som er typisk for dempingen (dvs som bestemmes ut fra gamma, dvs fra friksjonsleddet), eller vi kan velge en tidsenhet som er typisk for selve svingtiden for pendelen dersom det ikke var friksjon. Vi velger siste variant, og for å få enklest mulig uttrykk velger vi helt konkret å la

$$t_0 = \frac{1}{\omega}$$

Igjen vil \hat{t} være et dimensjonsløst måletall for tidsangivelser gitt i tidsenhet lik $1/\omega$.

Da er det på tide å sette inn for disse uttrykkene i den opprinnelige differensiellligningen, og vi starter med å finne et uttrykk for hastighet med de nye variablene.

$$\frac{dz}{dt} = \frac{d(z_0 \hat{z})}{d\hat{t}} \frac{d\hat{t}}{dt}$$

z_0 er en konstant og kan settes utenfor. Det siste ledet finnes lett når vi ser at $\hat{t} = t/t_0$ hvor også t_0 er en konstant. Da følger:

$$\frac{dz}{dt} = \frac{z_0}{t_0} \frac{d\hat{z}}{d\hat{t}}$$

På lignende måte kan vi finne et uttrykk for akselerasjonen, og resultatet blir:

$$\frac{d^2z}{dt^2} = \frac{z_0}{t_0^2} \frac{d^2\hat{z}}{d\hat{t}^2}$$

Vi setter nå inn disse uttrykkene i den opprinnelige differensiellligningen, og setter også inn at $t_0 = \frac{1}{\omega}$. Resultatet blir da:

$$z_0 \omega^2 \frac{d^2\hat{z}}{d\hat{t}^2} + 2\gamma z_0 \omega \frac{d\hat{z}}{d\hat{t}} + z_0 \omega^2 \hat{z} = 0$$

Vi "forkorter" med faktoren $z_0 \omega^2$ og får:

$$\frac{d^2\hat{z}}{d\hat{t}^2} + 2\frac{\gamma}{\omega} \frac{d\hat{z}}{d\hat{t}} + \hat{z} = 0 \quad (4.9)$$

Sammenligner vi nå den opprinnelige ligningen (4.8) med den dimensjonsløse ligningen (4.9), ser vi at vi har fått en viss forenkling. Ved å løse ligning (4.9), kan vi behandle alle dempede svingninger hvor γ/ω er identiske, ved å bare skalere lengder og tider i tråd med hva vi har gjort i utledningen. I dette tilfellet er det imidlertid praktisk talt ingen rasjonaliseringsgevinst å hente ut fra dette.

For å være konkret må vi bruke $\hat{t} = \omega t$ i beregningene, og når resultatet er fremkommet, må vi transformere tilbake til SI-enheter ved å bruke relasjonen $t = \hat{t}/\omega$. For posisjoner/lengder kan det i utgangspunktet se ut for at uansett hvilken z_0 vi hadde valgt, ville løsningen være identisk. Men dersom vi husker at vi må spesifisere en startposisjon (rel. likevektpunktet) for å starte beregningen, innser vi at valget av z_0 betyr noe for hvilke tall beregningene ender opp med likevel. Velges en z_0 forskjellig fra 1 m, må initialverdien skaleres tilsvarende, og alle posisjoner må tilbakeskaleres etterpå. Initialbetingelsene må

nemlig også angis på dimensjonsløs form når vi bruker ligning (4.9). Det betyr også at dersom vi angir initialbetingelsene i form av posisjon z_1 og hastighet v_1 ved tiden $t = 0$, må initialbetingelsene for ligningen (4.9) henholdsvis være $\hat{z} = z_1/z_0$ og

$$\hat{v}_1 \equiv \left(\frac{d\hat{z}}{dt} \right)_1 = \frac{t_0}{z_0} v_1$$

4.7 Eksempel på numerisk løsning: Enkel pendel

La oss ta et konkret eksempel, nemlig en tilnærmet matematisk pendel som kan svinge med vanlig stort utslag (opp til $\pm\pi$) uten å klappe sammen (dvs. snora er “stiv”). Vi regner at all masse ligger i en liten kule i enden av snora.

Mekanikken sier oss at kraften som trekker pendelen langs banen mot likevektpunktet er

$$F_\theta = -mg \sin(\theta)$$

når vinkelutslaget er θ . Kraftmomentet omkring opphengspunktet for pendelen med lengde L er da:

$$\tau = -mgL \sin(\theta)$$

Spinsatsen anvendt omkring opphengspunktet gir:

$$\tau = I\alpha = I\ddot{\theta}$$

Her er $\alpha = \ddot{\theta}$ vinkelakselerasjonen og I er treghetsmoment om aksen. Vi velger den enkleste beskrivelse:

$$I = mL^2$$

og ender opp med den endelige differensiellligningen:

$$mL^2\ddot{\theta} = -mgL \sin(\theta)$$

$$y\ddot{\theta} = -\frac{g}{L} \sin(\theta)$$

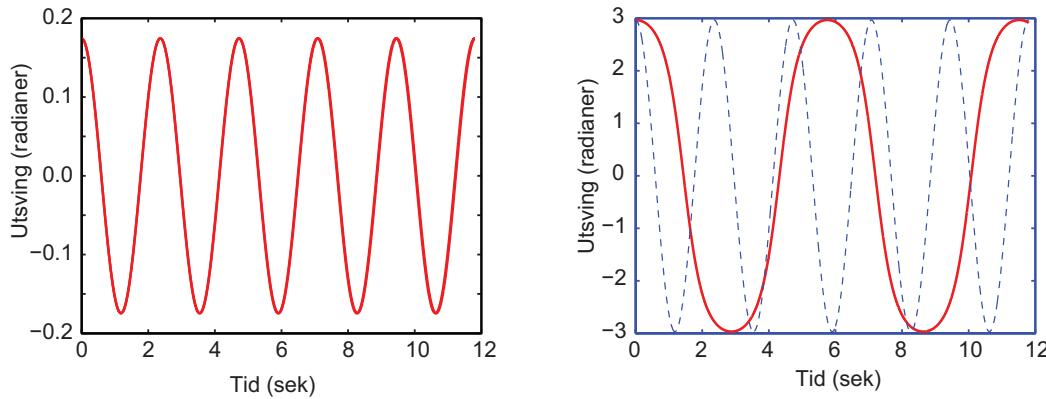
I mekanikken ble denne ligningen løst ved å anta at vinkelen θ er så liten at $\sin(\theta) \approx \theta$. Løsningen ble en enkel harmonisk bevegelse med svingfrekvens (vinkelfrekvens) gitt ved:

$$\omega = \sqrt{\frac{g}{L}}$$

Tilnærmingen $\sin(\theta) \approx \theta$ ble gjort for å kunne bruke analytiske metoder. Denne tilnærmingen var ikke helt nødvendig i akkurat dette spesielle tilfellet, fordi vi *kan* løse den opprinnelige differensiellligningen analytisk også for store vinkler ved å benytte oss av rekkeutviklingen til sinusfunksjonen. Det er imidlertid enklere å bruke numeriske metoder.

Resultatet av numeriske beregninger hvor vi bruker fjerde ordens Runge-Kutta, er vist i figur 4.3. Vi ser at bevegelsen er nær harmonisk for små vinkelutslag, men svært forskjellig fra en sinus for et stort utslag. Dessuten har periodetiden endret seg mye. Merk at vi i høyre del av figuren har valgt en bevegelse der pendelen *nesten* når retningen “rett opp” både på “framover” og “bakover” (utsving nær $+\pi$ og $-\pi$).

Dersom vi ønsket å inkludere friksjon ved beskrivelsen av pendelbevegelsen, ville det representere et mer komplisert uttrykk for den effektive kraften enn vi hadde i vårt tilfelle. For ikke-lineær beskrivelse av friksjon finnes det ingen analytiske løsning.



Figur 4.3: En pendel svinger harmonisk når utslaget er lite, men svingeforløpet endres mye når svingevinkelen øker. Også svingetiden endres. Forøvrig: Se teksten.

Siden hovedstrukturen i en numerisk løsning ville være den samme, uansett hvilken beskrivelse vi har av effektiv kraft som virker på systemet, kan de mer kompliserte fysiske forholdene ofte håndteres forbausende lett med numeriske løsningsmetoder.

Det er også en ekstra bonus ved numeriske løsningsmetoder: Kraften som virker blir mer sentral i vårt arbeid med å finne løsningen, og derved selve fysikken i problemet! Hvilken kraft gir hvilket resultat? Numerikken blir den samme, og vi behøver ikke pønske ut ulike til dels vrine analytiske løsningsmetoder og triks som er forskjellige for hvert uttrykk vi har for kraften. Fokus blir der den bør være: på utgangspunktet som er effektiv kraft og differensialligningen som gjelder, hvilke initialbetingelser vi har, og så hvilket resultat vi får.

4.8 Test av implementering

Det er fort gjort å gjøre en feil, enten i den analytiske matematikken, eller når vi lager dataprogrammet med numerisk løsningsmetoder. Vi har meget tragiske eksempler på hvor galt det kan gå i slike sammenhenger!

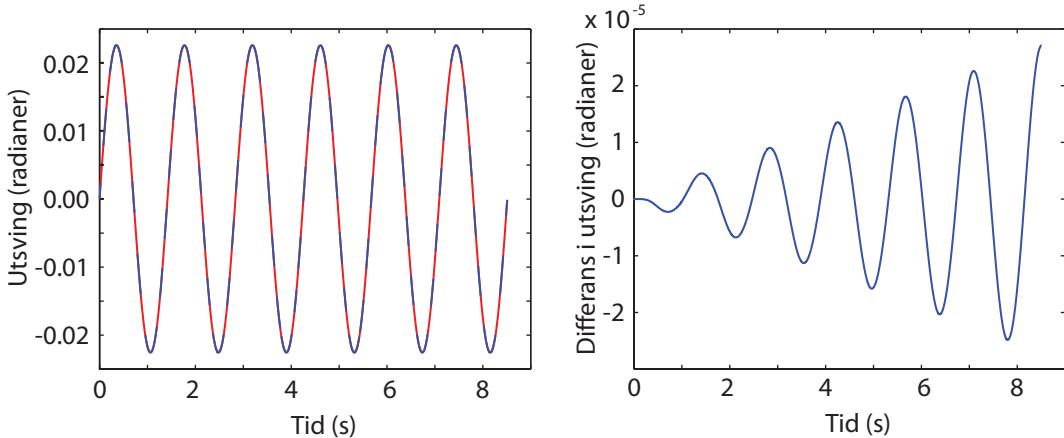
Det er derfor meget viktig å teste den numeriske implementeringen for å luke ut så mange feil vi bare kan. Det er ofte lettere sagt enn gjort! Vi bruker jo ofte numeriske metoder fordi vi ikke har noen analytiske metoder å falle tilbake på.

I vårt tilfelle med den mekaniske pendelen, er det likevel et triks vi kan gjøre. Det finnes en analytisk løsning som er tilnærmet riktig for *små* utslag. For *det* spesialtilfellet, kan vi teste om den numeriske løsningen blir omrent den samme som den analytiske. Dersom det er uoverenskomst mellom disse to løsningene, er det opplagt en feil et eller annet sted.

Dersom den numeriske løsningen er lik den analytiske i dette spesialtilfellet, er det dessverre ikke et bevis på at programmet er feilfritt! Implementeringen av det spesielle som gjelder ut over spesialtilfellet, kan likevel være feil. Her er det nødvendig å vurdere de fysiske prediksjonene: virker de rimelige eller urimelige? Det er ofte umulig å være helt sikker på at et dataprogram er fullstendig korrekt. Innen informatikk er det spesielle teknikker som kan brukes i en del tilfeller. Vi kan ikke gå inn på disse. Hovedpoenget er at vi må være ydmyke og åpne for at feil kan finnes, og at vi forsøker å teste implementeringen av numeriske metoder hver gang vi utvikler et dataprogram.

Som et eksempel skal vi nå forsøke å sjekke programmet vi brukte i beregningene som førte til figur 4.3. Det er bare tilfellet der utslaget er lite vi kan bruke i testen vår.

I figur 4.4 er det til venstre vist resultatet av de numeriske beregningene (rød kurve) sammen med analytisk løsning (stiplet blå kurve) i et tilfelle når pendelutsvinget er lite



Figur 4.4: Sammenligning mellom analytisk og numerisk løsning av en pendelbevegelse. For forklaringer: Se teksten.

(maksimalt ± 0.023 radianer). Det er ikke mulig å se forskjeller mellom de to kurvene.

Å plotte analytisk og numerisk løsning i samme diagram er en vanlig måte å sjekke at to løsninger er lik hverandre. Det er imidlertid en svært grov test, for det er begrenset opplösning i en grafisk framstilling. I høyre del av figuren har vi valgt en bedre test. Her er *differansen* mellom analytisk og numerisk resultat angitt, og vi ser at det sannelig var litt forskjeller selv om vi ikke så dette i venstre del.

Vi kan nå se at forskjellen øker på en systematisk måte. Etter seks perioder har forskjellen økt til $2.7 \cdot 10^{-5}$ radianer. Er dette en indikasjon på at dataprogrammet vårt er feil?

Vi vet imidlertid at den analytiske løsningen egentlig bare var tilnærmet rett, og tilnærmingen bør bli bedre desto mindre utslaget er. Vi kan da redusere utslaget og se hva som skjer. Beregninger viser at dersom utslaget blir redusert til $1/10$ av det vi har i figuren, reduseres maksimal forskjell etter seks perioder til $1/1000$ av det vi hadde i stad. Reduserer vi utslaget til $1/100$ av det opprinnelige, reduseres maksimal forskjell til 10^{-6} av den opprinnelige forskjellen. Vi ser at numerisk og analytisk løsning blir mer og mer lik hverandre, og på en slik måte som vi måtte forvente. Dersom vi at påtar tar en titt på rekkeutviklingen for sinus-funksjonen, gir det oss enda et holdepunkt på at resultatene våre er slik vi måtte forvente.

Vi kan da føle oss rimelig sikre på at programmet oppfører seg som det bør for små vinkler, og at det synes å håndtere økte vinkler som det bør, i alle fall så lenge de er små.

[♣ ⇒ Det er også en annen test vi ofte må gjøre i forbindelse med numeriske beregninger. Vi valgte å bruke 1000 tidssteg innen hver periode i beregningene som ligger bak figur 4.3 og 4.4. For beregninger som strekker seg over svært mange perioder, kan vi ikke bruke så små tidssteg. Går vi ned til f.eks. 100 beregninger per periode, vil resultatet vanligvis fortsatt være ok (avhengig av hvilke krav vi setter), men går vi f.eks. ned til 10 tidssteg per periode, vil resultatet nesten garantert avhenge mye av valget av tidssteg. Vi må ofte gjøre et sett beregninger for å forsikre oss om at "opplosningen" i beregningene er akseptabel (verken for høy eller for lav). ← ♣]

4.9 Krav til reproduserbarhet

I dag er det lekende lett å endre på et program fra en kjøring til den neste. Det gir ekstra utfordringer som må tas alvorlig. Når vi gjør beregninger som skal brukes i en vitenskapelig artikkkel, en masteroppgave, en prosjektoppgave, ja nærmest i hvilken som helst

sammenheng hvor vårt program brukes, må vi kjenne eksakt program og parametre som er brukt dersom resultatene skal ha full verdi. I eksperimentell fysikk vet vi at det er viktig å angi i labjournalen alle detaljer om hvordan eksperimentene er gjennomført. Hensikten er at det skal være mulig å etterprøve resultatene vi får. Dette er helt essensielt for å få reproducert barhet og for at vi skal kunne oppnå såkalt “intersubjektivitet” (at resultatet skal være uavhengig av hvilken person som faktisk gjennomfører eksperimentet), noe som er ekstremt viktig innen vitenskap og utvikling.

I eksperimentell virksomhet faller man iblant for fristelsen til å ikke notere alle relevante detaljer mens eksperimentet foregår. Vi er interessert i resultatet og tenker gjerne at når vi har kommet litt lenger og fått enda bedre resultater, *da* skal vi skrive ned alle detaljer. En slik praksis fører ofte med seg en del frustrasjon på et senere tidspunkt, for plutselig oppdager vi at noen viktige opplysninger faktisk aldri ble notert. I verste fall kan resultatet bli at vi må gjøre eksperimenter om igjen og lete oss fram til betingelser som var slik de var i et tidligere eksperiment hvor resultatene viste seg å være spesielt interessante.

Moderne bruk av numeriske metoder kan på flere måter sammenlignes med eksperimentelt arbeid i laboratoriet. Vi tester ut hvordan ulike parametre i beregningen innvirker på resultatene, og vi bruker ulike numeriske metoder på tilsvarende måte som vi bruker ulike måleinstrumenter og protokoller i eksperimenter. Det betyr at det stilles like strenge krav til dokumentasjon for den som driver med numeriske metoder som for eksperimentlisten.

For å etterkomme dette kravet bør vi innarbeide gode vaner i programmeringen. En måte vi kan etterkomme krav om reproducert barhet, er å gjøre følgende:

- Angi i programkoden et “versjonsnummer” for programmet ditt.
- I resultatfilen du genererer må versjonsnummeret legges inn automatisk.
- Hver gang du endrer programmet forut for en beregning du vil ta vare på, må versjonsnummer oppdateres.
- Hver versjon av programmet (som faktisk brukes i praksis) må lagres på disk slik at det alltid er mulig å kjøre om igjen et program med gitt versjonsnummer.
- Parametre som brukes og som varierer fra kjøring til kjøring innenfor samme versjon av programmet, må skrives ut til en fil sammen med resultatet av kjøringen.

Lever vi opp til disse reglene, vil vi alltid kunne gå tilbake og reproducere de resultatene vi fikk. Det er her antatt at resultatene er uavhengig av hvilken datamaskin beregningene foregår ved. Dersom vi har mistanke om at en kompilator eller bakenforliggende program eller operativsystem kan ha svakheter, kan det være aktuelt å angi også tilleggsopplysninger om dette sammen med resultatene (i en resultatfil).

I programeksemplene i denne boka er det de fleste steder *ikke* tatt med i koden de linjene som trengs for dokumentasjon av parametre og versjonsnummer. Grunnen er at programsnuttene som er angitt først og fremst er ment å vise hvordan selve beregningene kan gjennomføres. I eksempelprogrammet “Eksempelprogram i Matlab hvor Runge-Kutta brukes i praksis” er det imidlertid vist et eksempel på hvordan dokumentasjon om versjon og parametre kan håndteres.

4.10 Arbeidsgang ved numeriske metoder

Mange av problemstillingene vi møter i denne boka er knyttet til oppintegrering av differensielligninger som beskriver de prosessene vi er interessert i. Det er mange mulige måter

å foreta denne oppintegreringen på, som vi allerede har sett. Tiden er nå moden for å se på kodingen av de numeriske metodene.

En generell arbeidsgang ved løsning av numeriske oppgaver kan forsøksvis se slik ut:

- Før du setter deg til datamaskinen, må du ha etablert differensialligningen som beskriver prosessen du er interessert i. Tenk gjennom hvordan du kan finne/bestemme alle størrelser som inngår i differentalligningen. Bestem deg for initialverdier.
- Gå så til datamaskinen ...
- Skriv inn programmet: Definer alle variable du trenger og gi dem deres verdier, blant annet oppløsningen (N) du skal bruke i oppintegreringen.
- Nullstill arrays, eller gi arrays verdier.
- Slå sammen alle uttrykk av faste konstanter som skal brukes i hovedløkken (se nedenfor), slik at du ikke får flere regneoperasjoner enn nødvendig i løkken.
- Angi initialbetingelser.
- Gå inn i en løkke der du foretar oppintegreringen av differensialligningen, f.eks. ved hjelp av fjerde ordens Runge-Kutta.
- Bearbeid dataene videre dersom det er ønskelig.
- Plot resultatene eller presenter dem på annet vis. Lagre data på fil dersom det er ønskelig.
- Sjekk at programmet gir korrekt resultat for en forenklet versjon av problemet hvor det også eksisterer en analytisk løsning. Dette er viktig!
- Gjenta beregningene med ulik oppløsning (ofte gitt ved Δt) for å se hvor mange punkter som trengs for å få god overensstemmelse med det analytiske svaret, eller at resultatet i liten grad avhenger av moderate endringer i oppløsning.
- Tenk minst to ganger gjennom hva du har gjort i beregningene og hvilke resultater du har fått, og forsøk å oppdag faktorer som kan ha ødelagt for kvaliteten i beregningene.
- Tenk også gjennom om den presentasjonsformen du har valgt er tilstrekkelig for det som ønskes studert i beregningene. Ofte er enkle plot ok, men vi kan sjeldent lese ut nøyaktige detaljer fra et plot, i alle fall ikke uten at vi har valgt helt spesielle plot som egner seg akkurat for det vi vil vise.(Eksempel: Enkelte ganger er valg av lineære eller logaritmiske akser i plot avgjørende for om du oppdager interessante sammenhenger eller ikke.)
- Når du mener at programmet fungerer slik det skal, kan du endelig tenke på å gjøre de beregningene som skal inn i det prosjektet du arbeider med. Da må krav om reproducertbarhet etterleves ved at programmet nå får et høytidelig versjonsnummer, og det legges inn utskriftsrutiner som sørger for at alle parametre som er brukt dokumenteres gjennom programkoden (som må lagres og ikke endres uten at det blir nytt versjonsnummer) og/eller i resultatfilen for kjøringene som skal gjøres.

- Gjennomfør så beregningene som skal brukes videre i ditt arbeid.
- Filer som dokumenterer kjøringene for ettertiden må tas vare på på tilsvarende måte som en laboratoriejournal.

Mens du driver med programutviklingen og testingen er det viktig å lagre programmet flere ganger underveis, og helst skifte navn noen ganger i tilfelle noe katastrofalt skjer. Da slipper du å måtte starte helt fra nytt av om du mister alt i en fil. Det er også lurt å teste ut *deler* av programmet underveis når det lar seg gjøre.

Får du ikke programmet til å virke slik det skal, må du forsøke å teste ut bit for bit av programmet i den rekkefølgen beregningene gjennomføres. Forsøke å teste mellomverdier (f.eks. skrive dem ut til skjerm). Lag gjerne forenklinger i noen av uttrykkene for å se hvor ting skjærer seg. Det er møysommelig og ofte fryktelig tidkrevende å finne feil. Av den grunn er tipset ovenfor uhyre viktig, nemlig å teste ut bit for bit underveis mens du programmerer. Det dummeste du kan gjøre er å forsøke å skrive hele programmet på en gang, uten noe som helst testing underveis!

4.11 Diverse tillegg

4.11.1 Oppsummering, kapittel 3

La oss forsøke oss med en oppsummering av viktige punkter i vårt kapittel:

- En annen ordens differensielligning kan anses som ekvivalent med to koblede første ordens differensielligninger.
- I en enkel differensielligning kan vi tilnærmet erstatte den deriverte df/dt med differentialet $\Delta f/\Delta t$. Ved å ta utgangspunkt i denne tilnærmede ligningen og initialbetingelsene, kan vi suksessivt regne oss fram til alle senere verdier av $f(t)$. Denne metoden kalles Eulers metode. Metoden gir ofte store feil, spesielt når vi har med svingninger å gjøre!
- Det finnes bedre metoder for å estimere gjennomsnittlig stigningstall for funksjonen i intervallet Δt enn å bare bruke den deriverte i begynnelsen av intervallet slik vi gjør ved Eulers metode. En av de mest praktiske og robuste metodene kalles Runge-Kuttas metode av fjerde orden. I denne metoden benyttes et veit gjennomsnitt av fire ulike beregnede stigningstall i intervallet Δt som utgangspunkt for beregningene. Metoden gir ofte god overensstemmelse med analytiske løsninger der disse finnes, også for svingefenomener. Vi må imidlertid være klar over at også denne metoden har feil, og for enkelte systemer vil den ikke fungere tilfredsstillende.
- For annen ordens ordinære differensielligninger så som svingeligningen, kan vi finne løsningen såfremt vi kjenner differensielligningen og initialbetingelsene. For annen ordens partielle differensielligninger, for eksempel en bølgeligning, må vi *i tillegg* kjenne de såkalte randbetingelsene, ikke bare ved starten, men også hele tiden underveis i beregningene. Dette gjør at det ofte er langt vanskeligere å løse partielle differensielligninger enn ordinære annen ordens diffligninger.

- Det er verdifullt å sammenligne numeriske beregninger og analytiske beregninger (der disse finnes) for å oppdage feil i programeringen vår. Men selv om overensstemelsen er god i slike spesialtilfeller, er det ingen garanti for at løsningene er korrekte også for andre parametere (der analytiske løsninger ikke finnes).
- I blant skrives en differensiell ligning om slik at bare "dimensjonsløse variable" benyttes ved den numeriske løsningen. En viktig grunn er å redusere faren for tap av numerisk presisjon. I vår sammenheng er dette ikke så viktig, blant annet fordi vi arbeider med fenomener der lengdeskalaen ikke er svært langt fra en meter, og tidsskalaen ikke er svært forskjellig fra et sekund.
- Siden vi lett kan endre på programmer og parametere, er det en stor utfordring å holde rede på hvordan dataprogrammet så ut og hvilke parametere vi brukte da vi foretok beregninger og kom fram til resultater vi vil anvende. En eller annen systematisk form for dokumentasjon er tvingende nødvendig, der program, inputparametere og resultater kan kobles mot hverandre på en entydig måte.

4.11.2 Pseudokode for Runge-Kuttas metode

Funksjonen tar inn $x[n-1]$, $v[n-1]$ og $t[n-1]$ og returnerer $x[n]$ og $v[n]$.

1. Bruk innparameterene til å finne akselerasjonen, a_1 , i begynnelsen av intervallet. Farten i starten av intervallet, v_1 , er gitt som innparameter.

$x_1 = x[n-1]$

$v_1 = v[n-1]$

$a_1 = \dots$

2. Bruk denne akselerasjonen og farten til å finne et estimat for farten (v_2) og posisjonen i midten av intervallet.

$x_2 = \dots$

$v_2 = \dots$

3. Bruk den nye posisjonen og farten for å finne et estimat for akselerasjonen, a_2 , i midten av intervallet.

$a_2 = \dots$

4. Bruk nå den nye akselerasjonen og farten (a_2 og v_2) til å finne et nytt estimat for posisjonen og farten (v_3) i midten av intervallet.

$x_3 = \dots$

$v_3 = \dots$

5. Bruk deretter den nye posisjonen, farten og tiden i midten av intervallet til å finne et nytt estimat for akselerasjonen, a_3 , i midten av intervallet.

$a_3 = \dots$

6. Benytt så det siste estimatet for akselerasjonen og farten i midten av intervallet for å finne et estimat for posisjonen og farten (v_4) i slutten av intervallet.

$x_4 = \dots$

$v_4 = \dots$

7. Bruk så det siste estimatet for posisjonen og farten for å finne et estimat for akselerasjonen i slutten av intervallet, a_4 .

$a_4 = \dots$

8. En middelverdi for farten og akselerasjonen i intervallet beregnes så ved hjelp av et vektet gjennomsnitt:

$v_{\text{Middle}} = 1.0/6.0 * (v_1 + 2*v_2 + 2*v_3 + v_4)$

```

aMmiddle = 1.0/6.0 * (a1 + 2*a2 + 2*a3 + a4)

9. Til slutt brukes denne midlede verdien for farten og akselerasjonen i intervallet
til å beregne posisjonen og farten i slutten av intervallet. Funksjonen returnerer
denne farten og posisjonen.
x[n] = ...
v[n] = ...
return x[n], v[n]

```

4.11.3 Python-kode for Runge-Kuttas metode

```

# Først en egen selvstendig funksjon som på basis av kun ETT punkt (x,v,t) kan
# angi hvordan differensialligningen ser ut akkurat i dette punktet.
# Svaret returneres og brukes av den generelle Runge-Kutta funksjonen.
# MERK: Det er KUN denne lille funksjonen som endres når vi går
# fra ett system til et annet. Den generelle RK-funksjonen er identisk
# for alle annen ordens diff ligninger vi skal bruke.
def diffEq(xNow,vNow,tNow):
    aNow = f(xNow,vNow,tNow)
    # Her må vi erstatte linjen foran
    # med den differensialligningen
    # vi skal løse numerisk.
    return aNow

# Funksjon som tar inn startpunktet og bruker Runge-Kuttas metode for å finne
# neste punkt. Kaller på funksjonen diffEq over.
def rk(xStart,vStart,tStart):
    a1 = diffEq(xStart,vStart,tStart)
    v1 = vStart

    xHalf1 = xStart + v1 * dt/2.0
    vHalf1 = vStart + a1 * dt/2.0

    a2 = diffEq(xHalf1,vHalf1,tStart+dt/2.0)
    v2 = vHalf1

    xHalf2 = xStart + v2 * dt/2.0
    vHalf2 = vStart + a2 * dt/2.0

    a3 = diffEq(xHalf2,vHalf2,tStart+dt/2.0)
    v3 = vHalf2

    xEnd = xStart + v3 * dt
    vEnd = vStart + a3 * dt

    a4 = diffEq(xEnd,vEnd,tStart + dt)
    v4 = vEnd

    aMiddle = 1.0/6.0 * (a1 + 2*a2 + 2*a3 + a4)
    vMiddle = 1.0/6.0 * (v1 + 2*v2 + 2*v3 + v4)

    xEnd = xStart + vMiddle * dt
    vEnd = vStart + aMiddle * dt

    return xEnd, vEnd

```

4.11.4 Matlab-kode for Runge-Kuttas metode

```
unction [xp, vp, tp] = rk4r(xn, vn, tn, delta_t, param)
```

```
% Runge-Kutta integrator (4. orden) Versjon 13012013.
%*****
% Denne versjonen av fjerde ordens Runge-Kutta rutine for Matlab
% er skrevet av Arnt Inge Vistnes for bruk i FYS2130 våren 2013.
% Programmet ligger tett opp mot beskrivelsen av metoden i kapittel 3.4.1.

% Rutinen passer for det tilfellet at vi har to koblede diffligninger
%   dv/dt = ffa(x,v,t,param)
%   dx/dt = v
% der x,v,t kan være hhv posisjon, hastighet og tid. param er ulike
% parametre som inngår i diffligningene. delta_t er steplengden i tid.
% Vår rutine kaller på funksjonen ffa(x,v,t,param).
% Dette er en funksjon som brukeren selv må sette opp.

% Input argumenter (n: "nå")
%   [xn,vn,tn] = nåværende verdier for x, v og t.
% Output argumenter (p : "n plus 1")
%   [xp, vp, tp] = nye verdier for x, v og t etter et step i delta_t.
%*****
```

halv_delta_t = 0.5*delta_t;
t_p_halv = tn + halv_delta_t;

x1 = xn;
v1 = vn;
a1 = ffa(x1,v1,tn,param);

x2 = x1 + v1*halv_delta_t;
v2 = v1 + a1*halv_delta_t;
a2 = ffa(x2,v2,t_p_halv,param);

x3 = x1 + v2*halv_delta_t;
v3 = v1 + a2*halv_delta_t;
a3 = ffa(x3,v3,t_p_halv,param);

tp = tn + delta_t;
x4 = x1 + v3*delta_t;
v4 = v1 + a3*delta_t;
a4 = ffa(x4,v4,tp,param);

% Returnerer (tilnærmet) (x,v,t) i slutten av intervallet.
delta_t6 = delta_t/6.0;
xp = xn + delta_t6*(v1 + 2.0*(v2+v3) + v4);
vp = vn + delta_t6*(a1 + 2.0*(a2+a3) + a4);

return;

4.11.5 Funksjonen som inneholder differensiallikningen

```
function dvdt = ffa(xn,vn,tn,param)

%*****
% Funksjon for bruk i FYS2130 våren 2013. Versjon 13012013
% Returnerer venstresiden i en diffligning for dv/dt
% Benyttes av Runge-Kutta 4 numerisk løsning av to koblede
% diffligninger, f.eks. for svingebevegelse for ulike varianter forhold.
% Input argumenter ("n" indikerer "nå")
%   xn = posisjon
%   vn = hastighet
%   tn = tid
```

```
% Output argument
% dvdt = Venstresiden av første ordens diff ligningen for v
% Kommenterer ut og inn ulike beskrivelser etter behov
%*****
%
% Enkel harmonisk svingning
%dvdt = -param.kn*xc;

% Dempet svingning
%dvdt = -param.kn*xc - param.bn*vc;

% Dempet og påtrykt svingning
%dvdt = -param.kn*xc - param.bn*vc + param.Fmaxn*sin(param.omegad*tc);

% Pendelsvingning med demping
dvdt = -param.lg*sin(xc) - param.bml*vc; % DENNE er i bruk nå!

% Dempet og påtrykt svingning
%dvdt = -param.kn*xc - param.kn*xc*xc - param.bn*vc + ...
%     param.Fmaxn*sin(param.omegad*tc);
return;
```

4.11.6 Eksempel: Matlabprogram som bruker Runge-Kutta

Nedenfor er det gitt et program for beregning av tvungne mekaniske svingninger (fjærpendel). Det viser hvordan Runge-Kutta brukes i praksis dersom vi programmerer Runge-Kutta rutinen selv. I programmet er det lagt inn et eksempel på kode som vil kunne dokumentere hvilket program og hvilke parametre som ligger bak en konkret beregning. En slik dokumentasjon er helt nødvendig for å tilfredsstille dagens krav til reproducibilitet.

```
% Eksempelprogram for å undersøke ulike former for svingninger
% i kurset FYS2130 ved UiO våren 2013. Skrevet av Arnt Inge Vistnes.
% Programmet kaller på funksjonene rk4r og ffa.

function svingning12(startpos, startfart)
% Inputparametre:
% startposisjon (rel likevektpos) i meter, og startfart i m/s

versjon = '13 januar 2013';

global param;

% Her velges parametrene for beregningene (bortsett fra
% startpos og startfart somgis gjennom kall til funksjonen).
%*****
AntallPerioder=80;      % Antall perioder beregningene bør gå over (ca 40)
n_pr_T=1000;             % Velger 1000 punkt pr periode (ca)
k=1.16;                  % "Fjærstivheten" (default 1.16)
m=85.6e-3;               % Loddets masse (default 0.0856)
b=0.25*0.025;            % Friksjons-faktor (default 0.02)
Fmax=0.4e-00;             % Amplituden for påtrykt kraft (default 2)
omega0=sqrt(k/m);        % Grunn-vinkelrekvensen beregnes og skrives ut
omegad=1.2*omega0;        % Velger (relativ) frekvens for påtrykt kraft
filnavn='sving11c002.txt'; % Navn på fil hvor parametre og resultater
                           % skal lagres for senere dokumentasjon
% MERK: Også ffa bør dokumenteres, f.eks. slik:
fvLigning1 = 'dvdt = -param.kn*xc - param.bn*vc + ';
fvLigning2 = 'param.Fmaxn*sin(param.omegad*tc);';
%*****
```

```
% Parametre nedenfor denne linjen skal ikke endres uten at programmet får
% nytt versjonsnummer (eller at endringen dokumenteres på annet vis).

delta = b/(2.0*m)          % Parameter for å bedømme om man har kritisk demping
                            % eller ikke ( i så fall er delta=omega0). Skrives ut.
param.kn=k/m;              % Normerte størrelser inn i en struktur
param.bn=b/m;              % "param"
param.Fmaxn=Fmax/m;
param.omegad=omegaD;
T0=2*pi/omega0;            % Svingetid for enkel harmonisk bevegelse

N=n_pr_T*AntallPerioder;   % Antall punkter alt i alt i beregningen
delta_t=double(T0/n_pr_T); % Delta_t valgt ut fra det ovenforstående

% Initierer/allokerer arrayene vi trenger
x=zeros(N,1);
t=zeros(N,1);
v=zeros(N,1);

% Setter initialbetingelsene
x(1)=double(startpos);
v(1)=double(startfart);
t(1)=0;

% Dernest kjøres løkken for å følge videre utvikling
for j=1:N-1
    [x(j+1), v(j+1), t(j+1)]=rk4r(x(j),v(j),t(j),delta_t,param);
end

% Diverse plot, først for posisjon vs tid, sammenlikner med analytisk
% x_analyt = startpos*cos(2*pi*t/T0); % Ikke generelt!
plot(t,x,'-r');
%axis([0.0 5.0 -0.55 1.1]) % Bør kuttes ut vanligvis!
%plot(t,x,'-r',t,x_analyt,'-b');
%legend('numerisk løsning','analytisk løsning');
title('Posisjon vs tid'); % Skift ut tallet etter hva som beregnes
xlabel('Tid (sek)');
ylabel('Utsving (meter)');

figure; % Et nytt plott, for faserom-diagram
plot(x,v);
xlabel('Posisjon (m)');
ylabel('Hastighet (m/s)');
title('Faserom-presentasjon av svingbevegelsen');

% Dokumentasjon på kjøring (for lagring)
fileID = fopen(filnavn, 'w');
fprintf(fileID,'Kjøring %s av programmet "svingning", , date);
fprintf(fileID, ' versjon: "%s " \r\n\r\n',versjon);
fprintf(fileID,'AntallPerioder %d \r\n', AntallPerioder);
fprintf(fileID,'n_pr_T %d \r\n', n_pr_T);
fprintf(fileID,'Fjærstivhet k %f \r\n', k);
fprintf(fileID,'Massen m %f \r\n', m);
fprintf(fileID,'Friksjonsparameter b %f \r\n', b);
fprintf(fileID,'Påtrykt kraft Fmax %f \r\n', Fmax);
fprintf(fileID,'omega0 %f \r\n', omega0);
fprintf(fileID,'omegaD %f \r\n', omegaD);
fprintf(fileID,'Diff ligninger: \r\n fv: %s \r\n', fxLigning);
fprintf(fileID,' fv: %s %s \r\n', fvLigning1, fvLigning2);
fprintf(fileID,'Beregnde data t(i), x(i), v(i) \r\n');
for i=1:N
```

```

fprintf(fileID,'%f %f %f \r\n', t(i), x(i), v(i));
end;
fclose(fileID);

```

4.11.7 Bruk av Matlab's innebygde Runge-Kutta

Her følger til slutt et eksempelprogram for beregning av dempede svingninger dersom vi bruker Matlab's innebygde løser av ordinær diffligninger (ode) ved hjelp av 4. ordens Runge-Kutta. Først angir vi hovedprogrammet som vi har kalt *dempetSvingning.m* (navnet er uvesentlig her), og dernest følger en liten programsnutt *vaarDiffLign.m* som hovedprogrammet kaller på. Matlab's ligningsløser krever nemlig en liten ekstra funksjon som angir den aktuelle differensialligningen som sådan, og det er den som gis i *vaarDiffLign.m*.

```

% Program for å simulere dempede svingninger. Laget 31. januar 2010 med utgangspunkt
% i et program laget av Filip Nicolaisen noen dager før.
% Løser de to koblede differensiallikningene
% dz/dt = v
% dv/dt = - koef1 v - koef2 z

clear all;

% Angir systemets fysiske egenskaper (i SI-enheter)
b = 3.0; % Friksjonstall
m = 7.0; % Massen
k = 73.0; % Fjærstivhet
% Retningslinje:
%    Overkritisk dempning : b > 2 sqrt(k m)
%    Kritisk demping : b = 2 sqrt(k m)
%    Underkritisk demping: b < 2 sqrt(k m)

koef1 = b/m;
koef2 = k/m;

% Initialbetingelser (i SI-enheter)
z0 = 0.40; % Posisjon rel. likevektspunkt
v0 = 2.50; % Hastighet

% Tid vi ønsker å følge systemet i [start, slutt]
TID = [0,20];

% Initialverdier
INITIAL=[z0,v0];

% Lar Matlab selv gjennomføre en full fjerde ordens Runge-Kutta oppintegrrering
% av differensialligningen. Vår konkrete diffligning er spesifisert i funksjonen
% vaarDiffLign.
% T er tiden, F er løsningene [z v], tilsvarende er t den løpende variabelen
% tid og f den løpende variablene [z(t) v(t)] som Matlab benytter ved beregningene.
% Matlab velger selv hvor tett punktene skal ligge for å gi ok nøyaktighet.
% Punktene er ikke ekvidistante i tid!

[T F] = ode45(@(t,f) varaDiffLign(t,f,koef1,koef2),TID, INITIAL);

% Plotting av resultatet, velger bare å plotte posisjon vs tid.
plot(T,F(:,1));

% length(T) % Kan ta med for å se hvor mange punkter Matlab faktisk
% brukte ved løsning av difflign over ditt valgte tidsintervall.

```

```
% Her burde det legges inn en test på at vår beregning gir identisk resultat
% med analytisk uttrykk i et tilfelle hvor analytisk uttrykk finnes.
```

Her kommer så den lille funksjonen som gir selve differensialligningen (i form av to koblede differensialligninger):

```
function df = vaarDiffLign(~,f,koef1,koef2)

% Denne funksjonen evaluerer funksjonene f, hvor f(1) = z og f(2)=v.
% Som første variabel i parametre inn har vi skrevet ~ fordi tiden ikke
% inngår eksplisitt i uttrykkene våre.

df = zeros(2,1);

%HER kommer det vesentlige: Den ene diff ligningen: dz/dt = v
df(1) = f(2);

% Den andre diff ligningen: dv/dt = -koef1 v - koef2 z
df(2) = -koef1*f(2)-koef2*f(1);
```

4.11.8 Noen “kjøreregler” fra Hans Petter Langtangen

Programmering er et eget fag forskjellig fra fysikk. Denne boka er skrevet av meg, en fysiker, som finner programmering svært nyttig for mitt fag, men som ikke har like stor stolthet som en informatiker knyttet til utformingen av de programmene jeg skriver. For meg er det viktigst å være i stand til å foreta numeriske beregninger der analytiske beregninger er umulige eller for kompliserte til å gjennomføre i praksis.

Det betyr at mine dataprogrammer sjeldent ville fått “beste karakter” av en ren informatiker. Til tross for dette har jeg svært mye glede av min programmering.

Du får selv velge hvor elegante dataprogrammer i informatikerens øyne du vil skrive.

Hans Petter Langtangen har gitt noen prinsipper for god programmering. Noen av hans poeng er følgende (per desember 2012):

- Det bør være en-til-en korrespondanse mellom matematisk beskrivelse av et problem (algoritme) og koden. Det gjelder variable, formler osv.
- En kode bør stykkes opp i logiske funksjoner. I Python kan flere funksjoner legges i samme fil. I Matlab må ulike funksjoner legges i separate filer. Dette fører ofte til at programmering i Matlab gir “flate” programmer, mye hardkoding av formler og mangel på logisk oppdeling av programmet.
- Programmér generelt der det ikke er uhensiktsmessig. For eksempel når man skal integrere et uttrykk, programmeres et generelt integral av $f(x)$ og så sender man inn sin spesielle f som argument. Dette krever hyppig bruk av funksjoner.
- Bruk energi på å konstruere testproblemer for å sjekke at implementasjonen er riktig.

4.11.9 Forslag til videre lesing

Foruten bøker skrevet av Hans Petter Langtangen og andre her ved UiO, kan følgende kilder være nyttige dersom du ønsker å gå litt dypere i dette stoffet:

- http://en.wikipedia.org/wiki/Semi-implicit_Euler_method (tilgjengelig 13.01.2013)

- http://en.wikipedia.org/wiki/Numerical_partial_differential_equations
(tilgjengelig 13.01.2013)
- “Kalkulus”, 3. utgave av Tom Lindstrøm (Universitetsforlaget).

4.11.10 Et annet lite tips....

Du vil antakelig lagre en masse plot etter dine beregninger i dette og andre kurs. Plottene legges så inn i obliger, rapporter, og eventuelt senere i masteroppgaver og denslags. Mange studenter gjør da noe ganske dumt.

Det er et krav at tall og tekst langs aksene på plottene må være godt leseelige uten bruk av lupe (!) i den endelige størrelsen figurene har i et dokument. Det betyr at tall og bokstaver bør ha en størrelse på mellom 9 og 12 pt i endelig størrelse (kan gå ned til 8 pt om sterkt ønskelig, og indekser kan være enda litt mindre).

Det kan være lurt å lagre figurer i Matlab mens figurene ikke har fylt hele skjermen (bruk default display av figurer på skjermen). Da blir fontstørrelsen ofte tilstrekkelig stor selv om figuren forminskes omtrent til samme format som er brukt i denne boka. Reduseres imidlertid figurstørrelsen for mye, vil fontstørrelsen i det endelige dokumentet bli for liten. Du kan selv velge fontstørrelle i plot som genereres av Matlab og Python.

Lær deg gode vaner så raskt som mulig, - det vil lønne seg i det lange løp!

4.12 Læringsmål

Etter å ha jobbet deg gjennom kapittel 3 bør du ...

- kjenne til at en annen ordens differensielligning kan anses som ekvivalent med to koblede første ordens differensielligning.
- kunne løse en annen ordens differensielligning numerisk ved hjelp av fjerde ordens Runge-Kuttas metode.
- kunne forklare hvorfor numeriske mye oftere enn analytiske løsningsmetoder kan behandle kompliserte fysiske lovmessigheter, f.eks. ikke-lineær friksjon.
- kunne peke på faktorer som kan føre til at numeriske beregninger fungerer dårlig.
- kunne forklare i grove trekk hvorfor fjerde ordens Runge-Kutta vanligvis fungerer bedre enn Eulers metode.
- kunne foreta en rimelig god test på at et dataprogram som bruker numeriske løsningsmetoder fungerer som det skal.
- kjenne til hvordan vi kan gå fram for å sikre dokumentasjon av program og parametre som hører sammen med beregnede verdier.
- kjenne til hvorfor det er lurt å lagre et dataprogram under nye navn rett som det er mens man driver programutviklingen.
- kjenne til noen prinsipper som bør anvendes for å unngå omfattende feilsøking “til slutt”, og kjenne noen triks som kan brukes ved feilsøking av programmer.

4.13 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Hvorfor fungerer fjerde ordens Runge-Kutta vanligvis bedre enn Eulers metode?
2. Ved numerisk løsning av en svingeligning, trenger vi å kjenne systemets karakteristiske egenskaper, og vi må kjenne initialbetingelsene. Angi karakteristiske egenskaper for en mekanisk fjær-pendel og for en RCL-svingekrets. Angi typiske initialbetingelser for de samme systemene.
3. I kapittel 1 fortalte vi litt om begrensinger i superposisjonsprinsippet. I kapittel 3 har vi foretatt beregninger hvor superposisjonsprinsippet ikke gjelder. Pek på et system som passer til denne karakteriseringen, og forklar hvorfor superposisjonsprinsippet ikke gjelder. Kan du foreslå hvordan du kan bevise dette?

Regneoppgaver

4. Lag ditt eget dataprogram i Matlab eller Python for å beregne tidsutviklingen til en serie-RCL-krets. Bruk programmet til å reproduse hovedtrekkene i figur 1.4, 2.7 og 2.10 fra kapittel 1 og 2. Velg selv parametre som kan egne seg når faseresonansfrekvensen skal være 1000 Hz, og for de siste to figurene skal Q-verdien være 25. Det kan være nyttig å bruke en relasjon gitt i oppgave 5a i kapittel 2.
5. Lag ditt eget program for å beregne tidsutviklingen til en dempet svingning ved hjelp av fjerde ordens Runge-Kutta løsningsmetode. Test at den fungerer ved å sammenligne resultatene for analytisk løsning og numerisk løsning i det tilfellet de skal være identiske. Hvor stor er feilen i posisjonen for den numeriske løsningen (relativt til max utslag)? Dersom du selv velger tidssteget Δt ber vi deg å teste ut minst to-tre ulike valg for Δt for å se hvor mye dette betyr for nøyaktigheten.
6. Vi utledet ligning (4.7) ved å ta utgangspunkt i Taylorutvikling. I teksten like etter ligningen antyder vi at uttrykket kan utledes også på en annen måte. Gjennomfør denne utledningen.
7. Gjennomfør beregninger av tvungne svingninger for en rekke forskjellige påtrykte frekvenser, og sjekk at uttrykket for kvalitetsfaktor i kapittel 1 stemmer overens med frekvenskurven og den alternative beregningen av Q basert på halvverdibredde og sentrfrekvens.
8. Studér hvor raskt amplituden vokser ved tvungne svingninger når den påtrykte frekvensen er litt forskjellig fra resonansfrekvensen. Sammenlign med tidsforløpet ved resonansfrekvensen. Initialbetingelser: Systemet starter i ro fra likevektspunktet.
9. Finn ut hvordan beregningene i de forgående oppgavene måtte modifiseres dersom vi f.eks. ønsket å innlemme et ekstra ledd $-cv^2 \frac{\ddot{v}}{v}$ for friksjonen. Kommenter gjerne din oppfatning av hvorfor numeriske metoder har en del fortrinn framfor analytiske matematiske metoder alene.
10. Denne oppgaven går ut på å sjekke om superposisjon gjelder for et svingende fjærpendelsystem med demping, først i det tilfellet at friksjonen kan beskrives kun med et ledd av typen $-bv$, dernest i det tilfellet at friksjonen må beskrives ved $-bv - sv^2$, eller rettere sagt: $-bv - s|v|v$ for å ta hensyn til retningen (se kapittel 1 hvor denne detaljen er nevnt). Rent praktisk innebærer oppgaven at du må gjøre beregninger

for én svingetilstand, dernest for en annen, og så sjekke om summen av løsninger er lik løsningen av summen av tilstander.

Fjærpendelens fysiske egenskaper er karakterisert ved $b = 2.0$, $s = 4.0$, $m = 8.0$ og $k = 73.0$, alt i SI-enheter. Gjør beregninger først med initialbetingelsene $z_0 = 0.40$ og $v_0 = 2.50$, og dernest for initialbetingelsene $z_0 = 0.40$ og $v_0 = -2.50$. Legg sammen de to løsningene. Sammenlign denne med løsningen av differensialligningen når initialbetingelsene er lik summen av initialbetingelsene vi brukte i de to første kjøringene. Husk at du skal sjekke superposisjonsprinsippet både for kjøringer hvor $-s|v|v$ -leddet er med og der det ikke er med. Kan du trekke en foreløpig konklusjon / fremsette en hypotese om gyldigheten til superposisjonsprinsippet ut fra resultatene du har kommet til?

[♣ ⇒ NB: I tilfelle du bruker Matlabs innebygde løsner, vil tidspunktene ikke stemme overens mellom de to kjøringene. Du må da ta utgangspunkt i tidsrekken svarende til den ene kjøringen og bruke interpolasjon når addisjon av resultat for den andre kjøringen skal gjennomføres. Nedenfor er det gitt et eksempel på hvordan en slik addisjon kan foretas. Spør gruppelærer dersom du ikke forstår koden godt nok til å bruke den eller noe lignende i eget program.

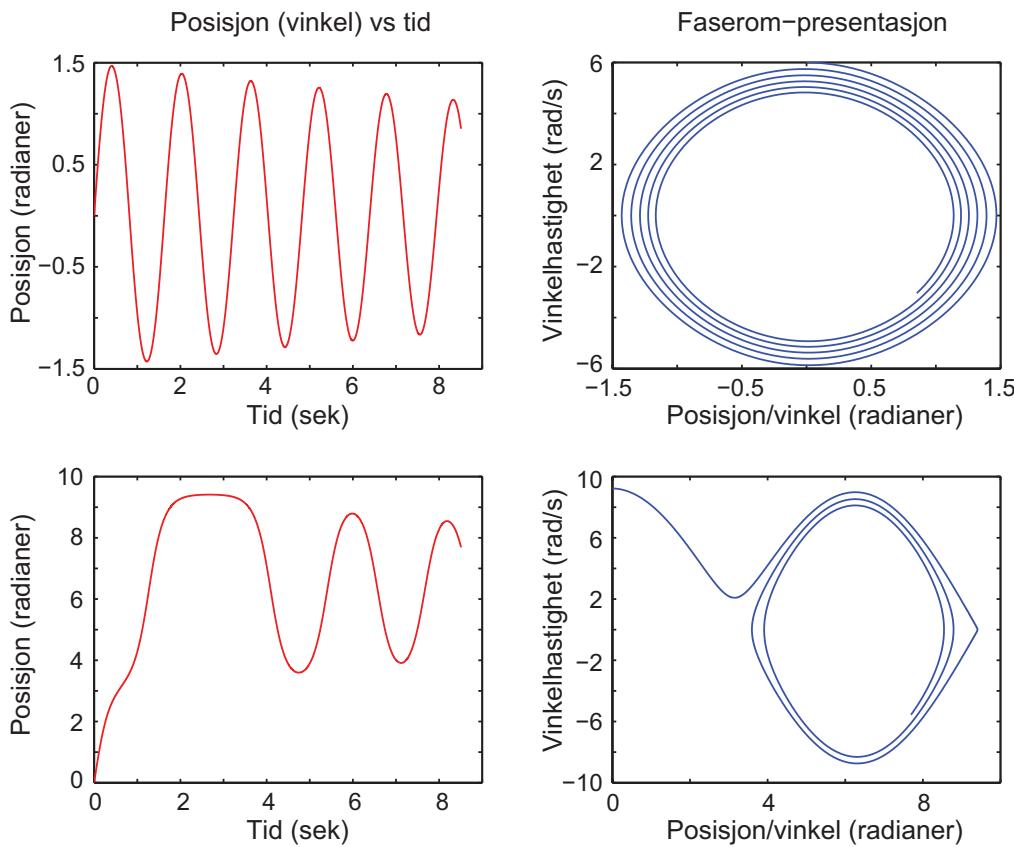
```
% Addisjon av to funksjoner Z1(t) og Z2(t'), hvor t er elementer i T1 og t' i T2.
% De to settene har samme startverdi (og sluttverdi), men ellers ulike.
% n1 = length(T1) og n2 = length(T2). Rutinen virker bare dersom n2>=n1.
% I motsatt fall må koden justeres tilsvarende.
```

```
% Tar T1 som basis for summasjonen
Z12(1)=Z1(1)+Z2(1);
for i = 2:n1
    % Finner først indeks til siste punkt i T2 mindre enn T1(i)
    j = 1;
    kL = -1;
    while kL<0
        if (T2(j)<T1(i)) j=j+1;
        else;
            kL=j-1;
        end;
    end;
    % Da har første punkt i T2 større eller lik T1(i) indeksen:
    kH = kL+1;
    % Summerer de to løsningene (lineær interpolasjon)
    Z12(i) = Z1(i)+Z2(kL) + (Z2(kH)-Z2(kL))...
        *(T1(i)-T2(kL))/(T2(kH)-T2(kL));
end;
```

⇐ ♣]

11. I figur 4.5 er resultatet av beregninger av en pendelbevegelse vist for tilfellet at det er litt friksjon til stede. Figuren viser posisjon (vinkel) som funksjon av tid (venstre del) og vinkelhastighet som funksjon av posisjon (vinkel) i høyre del (også kalt fasediagram). De to øvre figurene fremkommer ved en initialbetingelse der pendelen ved tiden $t = 0$ henger rett ned, men samtidig har en liten vinkelhastighet. De nedre figurene fremkommer ved at initialbetingelsene er som for øvre del, men at den initielle vinkelhastigheten er en god del større enn i det første tilfellet.

Forklar hva figurene sier om bevegelsene (forsøk å få med så mange interessante detaljer som mulig). Hvordan ville figuren sett ut dersom vi økte den initielle vinkelhastigheten enda litt mer enn den vi har i nedre del av figuren?



Figur 4.5: Bevegelsen til en enkel pendel. Se teksten for omtale.

4.13.1 En mer sammensatt regneoppgave

I denne oppgaven skal du gjennomgå en løsning av en annen ordens inhomogen differensiellligning ved hjelp av et spesielt dataprogram som gir analytiske uttrykk for løsningen, og dernest skal du gjøre egne beregninger ut fra numeriske metoder og sammenligne resultatet. Underveis gis det en rekke spørsmål om detaljer som tester forståelsen av hva som foregår.

12. I kapittel 1 og 2 ble det vist hvordan vi kan løse den andre ordens differensiellligningen som beskriver strømmen i en RCL-krets med eller uten ytre påtrykt vekselspenning. Det var lett å finne løsningen av den homogene differensiellligningen (uten påtrykt spenning), men litt mer styr for å finne en partikulær løsning av den inhomogene ligningen (som svarer til påtrykt ytre vekselspenning). Så lenge den påtrykte spenningen var et rent sinusignal hvor verken frekvens eller amplitude endret seg med tiden, var det likevel en overkommelig oppgave.

Så lenge vi opererer bare med generelle løsninger, inngår det et par koeffisienter i uttrykkene vi kommer fram til. Da ser løsningen rimelig grei ut. Dersom vi imidlertid skal finne en bestemt løsning ut fra et sett initialbettinger, får ofte de ellers valgfrie koeffisientene et nokså komplisert uttrykk. Dette vil du få se et eksempel på i denne oppgaven.

Det finnes dataprogrammer som kan foreta analytiske beregninger i matematikken i mange tilfeller der dette lar seg gjøre. De mest kjente programmene av denne typen er Mathematica og Maple. Du har muligens tilgang til Maple eller Mathematica, og vi har derfor valgt å vise et eksempel på hvordan et slikt program kan brukes.

Figur 4.6 og 4.7 viser en kjøring i Maple hvor vi løser den inhomogene differensiellligningen som beskriver hvordan ladningen på en kondensator varierer med tiden

i en RCL-krets når vi har en påtrykt spenning med fast vinkelfrekvens. Først får vi en generell løsning. Dernest angis initialbetingelser, og Maple gir oss den spesielle løsningen for disse initialbetingelsene. Til slutt setter vi inn et sett verdier for resistans, kapasitans og induktans, såvel som amplitude og vinkelfrekvens på den påtrykte spenningen, og ser hvordan ladningen da varierer som funksjon av tid.

De ivrigste av dere kan forsøke å starte Maple selv og teste ut de kommandoene som er brukt.

Studer figur 4.6 og forsøk å gjenkjenne kommandoer ut fra de differensialligningene vi kjenner fra kapittel 1. Merk at det vi selv gir som input til Maple er angitt med sort, mens responsen fra Maple er gitt i blått. Derivering kan angis på to måter i Maple, både som " $diff(f(t),t)$ " og som " $D(f)(t)$ " for enkeltderivert av en funksjon f (som i vårt tilfelle er ladningen q på kondensatoren).

Konkrete oppgaver (refererer ofte til figurene 4.6 og 4.7):

- Skriv ned differensialligningen som er utgangspunktet for Maple-beregningene (angi den i vår vanlige matematiske språkdrakt).
- Hvordan angis en differensialligning i Maple, og hvordan løses den?
- Kjenner du igjen løsningen av differensialligningen ut fra den generelle løsningsprosedyren angitt i kapittel 1 og 2?
- Hvilke symboler bruker Maple på de to valgfrie koeffisientene i den generelle løsnogene av diffligningene?
- Hvilke initialbetingelser har vi valgt i eksemplet vårt?
- Beskriv kort hvordan kompleksiteten i uttrykket for løsningen endret seg idet vi måtte ta hensyn til initialbetingelsene. Kan du gi en forklaring på hvorfor endringene gikk i den retning de faktisk gikk?
- Beregn systemets naturlige svingefrekvens for de valgte verdiene for R , L og C . Hvordan er den valgte påtrykte frekvensen sammenlignet med den beregnede naturlige svingefrekvensen for systemet? (Hint: Husk forskjell mellom frekvens og vinkelfrekvens).
- Forsøk å angi ut fra plottet nederst i figur 4.7 omtrentlig hvor lenge "innsvingningsforløpet" varer etter oppstart av svingningen med en ytre harmonisk påtrykt spenning. (Merk: Du klarer ikke å se hver enkelt sinussvingning i detalj i plottet fordi linjene ligger for tett. Det fremkommer derfor et indre mønster i kurvene som ikke har basis i virkelig tidsforløp, bare av den begrensede oppløsningen i plottet. Du må bare bruke omhyllingskurven når du gjør estimatet.)
- Forsøk å angi kvalitetsfaktoren for kretsen bak figur 4.6 og 4.7. (Hint: Ta utgangspunkt i omtalen av kvalitetsfaktoren Q i kapittel 1.)
- Bruk den vedlagte Matlabkoden (helt til slutt) som bruker Maple-resultatet og regner ut den spesielle løsningen vi har av den inhomogene differensialligningen med de angitte initialbetingelsene. Sjekk at du får samme resultat som angitt i figur 4.7 for de valgte verdier for R , C , L , w og V_0 . Bruk gjerne zoom for å se bedre hvordan tidsforløpet faktisk er.
- Foreta deretter en systematisk endring i w (frekvensen til påtrykt spenning) for å bestemme omtrentlig Q-verdien til denne kretsen ut fra halvverdibredden for frekvensresponsen (for de angitte verdier for R , C og L). (Hint: Det holder å lese av omtrentlige verdier fra plottene dersom du zoomer inn resultatplottene på en lur

måte. Du kan da notere tallverdier som du siden kan bruke for å vise formen til resonanskurven. Fra denne kan du så estimere kvalitetsfaktoren Q).

l. Legg til noen få linjer i Matlabprogrammet for å demonstrere faseforskjell mellom påtrykt spenning og ladningsendringen vs tid. Zoom inn på et egnet område for å få et kvalitativt bilde av faseskiftet både ved resonansfrekvensen, og litt over og litt under denne. Argumenter for hvor i tidsbildet du gjennomfører sammenligningen.

m. Bruk en numerisk løsningsmetode basert på fjerde ordens Runge-Kutta for å bestemme innsvingningsforløpet under betingelser identiske med de som ligger bak plottet nederst i figur 4.7. Sammenlign resultatet fra kjøringen med det analytiske uttrykket og resultatet fra den numeriske beregningen.

o. Hvordan ville det være å finne en analytisk og numeriske løsning dersom den påtrykte spenningen ikke var en ren sinus?

Her er til slutt Matlab-programmet for simulering av en elektrisk RCL-krets med gitte initialbetingelser:

```
% Beregner tidsforløpet i starten av en svingning i ladning i en RCL-krets
% pådyttet av en ytre vekselspenning med frekvens w og amplitude V0.
% Koden er skrevet med bakgrunn i Maple-beregninger.

R = 10;
L = 20;
C = 2.0e-6;
w = 157*1.00;
V0 = 10;
N = 4000;
t = linspace(0,20,N);

q = zeros(N,1);
rq = zeros(N,1);
partikular = zeros(N,1);

underRottegn = C^2*R^2-4*C*L;
alpha1 = -(1/2)*(C*R-sqrt(C^2*R^2-4*C*L))/(C*L);
alpha2 = -(1/2)*(C*R+sqrt(C^2*R^2-4*C*L))/(C*L);
teller1 = C*L*(-R*sqrt(C^2*R^2-4*C*L)+R^2*C-4*L)*V0;
nevner1 = 2*C^2*R^2*w^2*L^2-8*w^2*L^3*C+C^2*R^4-6*R^2*C*L+8*L^2- ...
sqrt(C^2*R^2-4*C*L)*R^3*C+4*R*sqrt(C^2*R^2-4*C*L)*L;
faktor1 = teller1/nevner1;

teller2 = C*L*(R*sqrt(C^2*R^2-4*C*L)+R^2*C-4*L)*V0;
nevner2 = (R^2*C-4*L)*(R*sqrt(C^2*R^2-4*C*L)+R^2*C-2*L+2*w^2*C*L^2);
faktor2 = teller2/nevner2;
teller3 = C*V0;
nevner3 = R^2*C^2*w^2+1-2*w^2*C*L+w^4*C^2*L^2;
faktor3 = teller3/nevner3;

partikular = cos(w.*t)-cos(w.*t)*w^2*C*L+sin(w.*t)*w*C*R;
q = exp(alpha1.*t).*faktor1 + exp(alpha2.*t).*faktor2 + partikular.*faktor3;

rq = real(q);
plot(t,rq,'-b');
```

```

> Clear(all);
                                         Clear(all)

Diffligning for ladning for en RCL - krets med en ytre spenningskilde koblet til :
> PDE2 := L*(diff(f(t),t,t)) + R*(diff(f(t),t)) + f(t)/C = V0 * cos(w*t);
                                         PDE2 = L \left( \frac{d^2}{dt^2} f(t) \right) + R \left( \frac{d}{dt} f(t) \right) + \frac{f(t)}{C} = V0 \cos(w t)

> ans2 := dsolve(PDE2);
                                         ans2 := f(t) = e^{-\frac{1}{2} \frac{(CR-\sqrt{C^2 R^2-4 CL}) t}{CL}} \frac{-\frac{1}{2} \frac{(CR+\sqrt{C^2 R^2-4 CL}) t}{CL}}{C2+e^{-\frac{1}{2} \frac{(CR+\sqrt{C^2 R^2-4 CL}) t}{CL}}} -CI + \frac{V0 C (\cos(w t)-\cos(w t) w^2 C L+\sin(w t) w C R)}{R^2 C^2 w^2+1-2 w^2 C L+w^4 C^2 L^2}

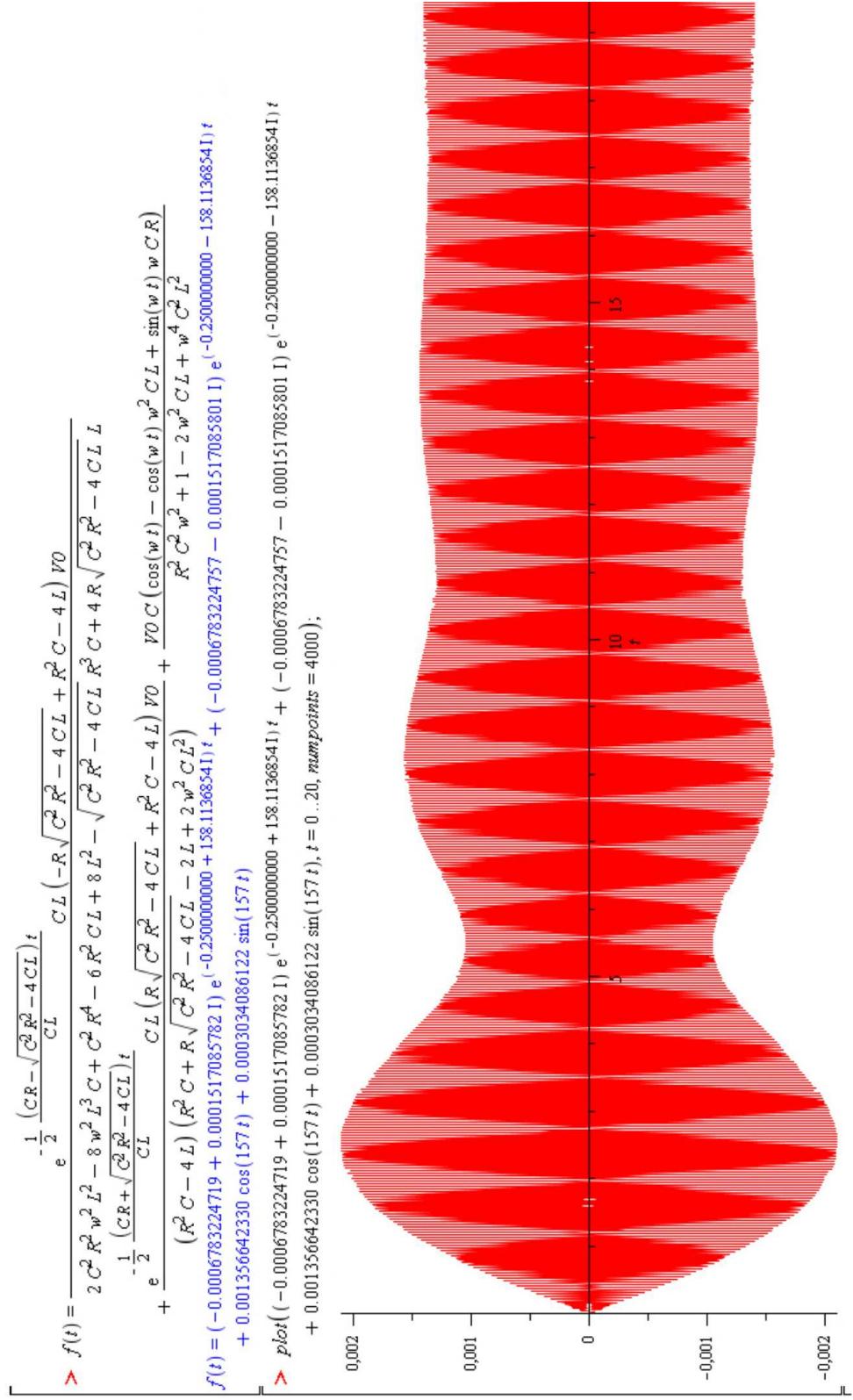
Initialbettingelser: Verken ladning eller strøm ved oppstart.
> ics2 := f(0) = 0, D(f)(0) = 0;
                                         ics2 := f(0) = 0, D(f)(0) = 0

> ans2x := dsolve({ics2,PDE2});
                                         ans2x := f(t) = \frac{e^{-\frac{1}{2} \frac{(CR-\sqrt{C^2 R^2-4 CL}) t}{CL}} C L \left(-R \sqrt{C^2 R^2-4 C L}+R^2 C-4 L\right) V0}{2 C^2 R^2 w^2 L^2-8 w^2 L^3 C+C^2 R^4-6 R^2 C L+8 L^2-\sqrt{C^2 R^2-4 C L} R^2 C+4 R \sqrt{C^2 R^2-4 C L} L}
                                         + \frac{e^{-\frac{1}{2} \frac{(CR+\sqrt{C^2 R^2-4 CL}) t}{CL}} C L \left(R \sqrt{C^2 R^2-4 C L}+R^2 C-4 L\right) V0}{(R^2 C-4 L) \left(R^2 C+C \sqrt{C^2 R^2-4 C L}-2 L+2 w^2 C L^2\right)} + \frac{V0 C (\cos(w t)-\cos(w t) w^2 C L+\sin(w t) w C R)}{R^2 C^2 w^2+1-2 w^2 C L+w^4 C^2 L^2}

> R := 10; L := 20; C := 2.0e-6; w := 157; V0 := 10;
                                         R := 10
                                         L := 20
                                         C := 0.0000020
                                         w := 157
                                         V0 := 10

```

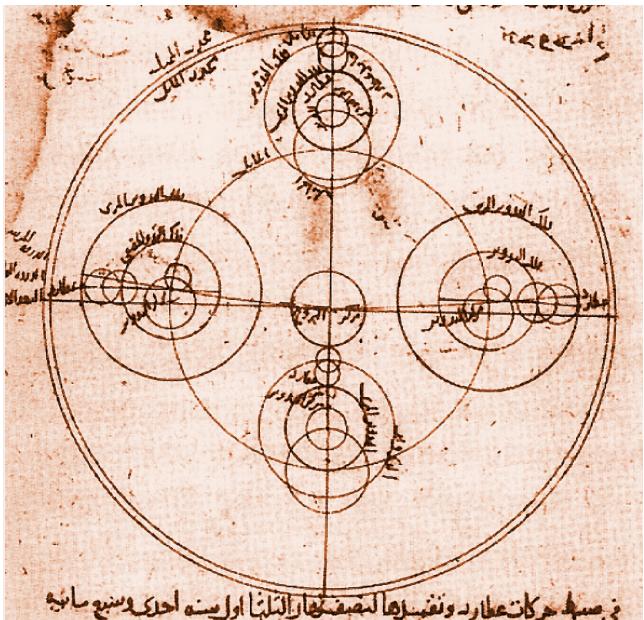
Figur 4.6: Eksempel på løsning av en differensialligning i dataprogrammet Maple. Del 1.



Figur 4.7: Eksempel på løsning av en differensialligning i dataprogrammet Maple. Del 2.

Kapittel 5

Fourieranalyse



Fouriertransformasjon og fourieranalyse har klare likhetstrekk med middelalderens bruk av episykler for å beregne hvordan planeter og sola beveget seg i forhold til hverandre. Figuren er nedlastet fra s1.hubimg.com/u/7219468_f520.jpg 30.01.2013 og lett bearbeidet.¹

Senere i boka vil vi omtale lyd fra ulike musikkinstrumenter, og til og med hvordan vi kan lage syntetisk lyd som ligner på den virkelige. Her er Fouriertransformasjon et meget nyttig verktøy. Men ethvert verktøy har sitt bruksområde der det egner seg best. En Fouriertransformasjon egner seg slett ikke for analyse av alle typer signaler. Vi trenger å være på vakt og ikke bruke Fouriertransformasjon når metoden egentlig ikke passer.

I dette kapitlet skal vi ta for oss en meget anvendelig metode for å studere periodisitet i en funksjon eller et signal. Vi kommer nesten utelukkende til å foreta Fouriertransformasjon av et signal som varierer i tid. Fouriertransformasjonen gir oss da et frekvensspekter, og ut fra frekvensspekteret kan vi foreta en Fourieranalyse eller spektralanalyse.

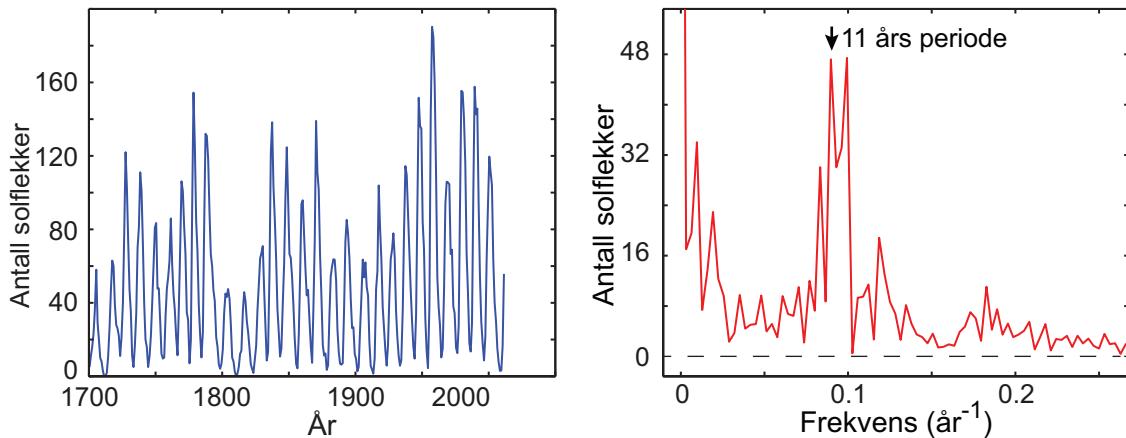
Signalet vi starter ut med kaller vi "tidsbilde". En Fouriertransformasjon gir oss da et "frekvensbilde" av det samme signalet.

Det er mange detaljer og nye ord knyttet til Fouriertransformasjon, så som grunnfrekvens og harmoniske, folding, samplingsteoremet, reell og imaginær del av Fourierkoeffisienter, fullstendig sett funksjoner og mera til. Jeg er redd du rett og slett må gjennomføre noen Fouriertransformasjoner på egen hånd for å forstå alt sammen.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 30012013.

5.1 Innledning

Mengden av solflekker varierer med en periodetid på 11 år, hører vi rett som det er. Hva er grunnlaget for en slik påstand? Vi kan plotte antall solflekker per år i løpet av en del år. Vi får da en kurve som i venstre del av figur 5.1. Vi kan se at det er omrent 11 år mellom de ulike toppene, men vi skulle gjerne hatt en metode for å kvantifisere periodetiden og dessuten se om det også synes å være andre lengre periodetider som solflekkaktiviteten følger. I slike sammenhenger bruker vi ofte fouriertransformasjon i analysen, og i høyre del av figur 5.1 er det vist et utdrag av resultatene etter en fouriertransformasjon av dataene i venstre del



Figur 5.1: Venstre del viser solflekker som dukket opp hvert år gjennom de siste tre hundre år. Høyre del viser et utdrag fra den tilsvarende fouriertransformerte funksjonene. Dataene er hentet 30.1.2012 fra <http://sidc.be/DATA/yearssn.dat>

I dette kapitlet skal vi ta for oss fouriertransformasjon. Vi starter med den klassiske matematiske beskrivelsen, går så raskt innom såkalte fourierrekker, men bruker mesteparten av tiden på diskret fouriertransformasjon. Den diskrete transformasjonen brukes svært mye i fysikk i dag. Den er godt anvendelig både ved analyse av eksperimentelle data og ved numeriske beregninger og simuleringer basert på teori.

Det er et mål for oss at kapitlet skal bidra til at du skal lære å beherske diskret fouriertransformasjon slik at du kan hente maksimalt med informasjon ut av de diagrammene fouriertransformasjonene gir. Samtidig ønsker vi at du skal unngå noen av de til dels alvorlige feilslutningene som dessverre også forekommer ved bruk av metoden.

5.2 Fouriertransformasjon (FT)

Det var den franske matematikeren og fysikeren Joseph Fourier (1768-1830) som introduserte fouriertransformasjon i analytisk matematikk.² Fouriertransformasjonen er siden videreført til også å omfavne funksjoner gitt på digital form (som en endelig streng med tall). Vi kommer tilbake til den såkalte diskrete fouriertransformasjonen om litt.

Du har muligens møtt fouriertransformasjon i matematikken tidligere. I matematikken er transformasjonen gjerne knyttet opp til indreprodukt mellom to funksjoner, og vi definerer en basis av sinus og cosinusfunksjoner og anvender Gram-Schmidt på en funksjon for å finne dens Fouriertransformerte. Vi velger en mer praktisk tilnærming i vår sammenheng.

La oss se på den opprinnelige formalismen:

²Fourier er forøvrig kjent for å ha påvist/forklart drivhuseffekten for global oppvarming i 1824.

La $f(t)$ være en integrerbar funksjon med t (gjerne tid) som parameter. I fysikk er $f(t)$ ofte en reell funksjon, men matematisk sett kan den gjerne være kompleks.

Med basis i $f(t)$ kan det beregnes en ny funksjon $F(\omega)$, på følgende måte:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt \quad (5.1)$$

Parameteren ω er vinkelfrekvens dersom t representerer tid.

Det morsomme med denne funksjonen er at vi kan ta en tilsvarende "omvendt" transformasjon:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)e^{i\omega t} d\omega \quad (5.2)$$

og ende opp med eksakt den opprinnelige funksjonen igjen. Merk fortegnskiftet i den komplekse eksponensialfunksjonen.

I ulike bøker angis faktorene foran integraltegnene gjerne på ulikt vis. Vi har valgt den varianten som gir mest symmetriske uttrykk. Produktet av disse faktorene er $1/2\pi$ i denne formen for fouriertransformasjon.

Fra ligningene (5.1) og (5.2) ser vi at selv om $f(t)$ er reell, vil $F(\omega)$ være kompleks. Det er viktig å forstå årsaken til dette, og vi vil komme tilbake til denne problemstillingen mange ganger i dette kapitlet.

La oss aller først forsøke å gjennomskue hva $F(\omega)$ står for i et enkelt eksempel.

5.3 Hva sier $F(\omega)$ oss?

Erfaringsmessig er det mange som har vanskelig for å se for seg hva som skjer i en fouriertransformasjon. Vi vil derfor starte med en matematisk argumentasjon basert på enkle prinsipper i håp om at det skal lette innføringen i metoden.

Første utgangspunkt er å merke seg at fouriertransformasjonen inneholder et komplikt eksponentialledd. Vi minner derfor om Eulers formel:

$$e^{ix} = \cos x + i \sin x$$

Brukes Eulers formel på ligning (5.1), får vi:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \{f(t) \cos(\omega t) - if(t) \sin(\omega t)\} dt \quad (5.3)$$

Andre utgangspunkt er følgende kjente formler fra Rottmann:

$$\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$$

$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$$

Fra disse formlene kan vi utlede følgende relasjoner:

$$\sin x \cos y = \frac{1}{2} (\sin(x+y) + \sin(x-y)) \quad (5.4)$$

$$\cos x \cos y = \frac{1}{2} (\cos(x+y) + \cos(x-y)) \quad (5.5)$$

$$\sin x \sin y = \frac{1}{2} (\cos(x-y) - \cos(x+y)) \quad (5.6)$$

Tredje utgangspunkt er at integralet av en sinus eller cosinus over en periode er null. Det samme gjelder for “gjennomsnittsnivået” over mange perioder. En bestemt skrivemåte fører til et spesialtilfelle av interesse for oss:

$$\lim_{T \rightarrow \infty} \left\{ \frac{1}{2T} \int_{-T}^T \sin(at) dt \right\} = 0 \quad \text{for alle reelle } a \quad (5.7)$$

$$\lim_{T \rightarrow \infty} \left\{ \frac{1}{2T} \int_{-T}^T \cos(at) dt \right\} = 0 \quad \text{for alle reelle } a \neq 0 \quad (5.8)$$

$$\lim_{T \rightarrow \infty} \left\{ \frac{1}{2T} \int_{-T}^T \cos(at) dt \right\} = 1 \quad \text{for } a = 0 \quad (5.9)$$

I denne sammenheng kan det også være nyttig å minne om at:

$$\lim_{T \rightarrow \infty} \left\{ \frac{1}{2T} \int_{-T}^T \sin^2(t) dt \right\} = \lim_{T \rightarrow \infty} \left\{ \frac{1}{2T} \int_{-T}^T \cos^2(t) dt \right\} = 1/2$$

5.3.1 Fouriertransformasjon av en ren sinusfunksjon

Vi velger nå å finne den fouriertransformerte av funksjonen

$$f(t) = \sin(\omega_0 t)$$

Fouriertransformen av denne funksjonen blir:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \{ \sin(\omega_0 t) \cos(\omega t) - i \sin(\omega_0 t) \sin(\omega t) \} dt$$

Vi bruker ligningene (5.4) og (5.6), og får:

$$\begin{aligned} F(\omega) &= \frac{1}{\sqrt{8\pi}} \int_{-\infty}^{\infty} (\sin\{(\omega_0 + \omega)t\} + \sin\{(\omega_0 - \omega)t\}) dt \\ &\quad + \frac{1}{\sqrt{8\pi}} \int_{-\infty}^{\infty} i(\cos\{(\omega_0 + \omega)t\} - \cos\{(\omega_0 - \omega)t\}) dt \end{aligned} \quad (5.10)$$

Vi har nå et problem, nemlig at integralet av sinus eller cosinus egentlig ikke er veldefinert og endelig når integrasjonsgrensene er uendelig. For den videre argumentasjonen er det derfor en fordel å heller tenke seg et integral av typen gitt i ligning (5.9). Det er slik fouriertransformasjonen skjer i praksis i de situasjonene vi faktisk vil møte i dette kurset.

Det essentielle er å innse at det resulterende integralet *bare* får bidrag dersom

$$\omega = \omega_0 \quad \text{eller} \quad \omega = -\omega_0$$

Videre ser vi at det bare er imaginærdelen av uttrykket som overlever. Nærmore bestemt får vi (når vi ikke bruker faktoren $\sqrt{2\pi}$ i definisjonen av en fouriertransformasjon pga problemet med integrerbarheten):

$$F(\omega = \omega_0) = - \lim_{T \rightarrow \infty} \left\{ \frac{1}{4T} \int_{-T}^T i \cos(\omega_0 - \omega_0) dt \right\} = -i/2$$

Og tilsvarende:

$$F(\omega = -\omega_0) = i/2$$

og

$$F(\omega) = 0 \quad \text{for alle andre } \omega.$$

Ikke la deg irritere eller forvirre for mye av tilsynelatende inkonsistenser mhp konstantfaktorene som er brukt i ulike varianter av fouriertransformasjon i dette kapitlet. Det er ulike måter å gjøre dette på, og det spiller liten rolle hvilken variant vi velger så lenge vi vet hva vi gjør. I vårt kurs vektlegger vi numerisk fouriertransformasjon, og da er problemet mindre enn for analytiske uttrykk. Vær oppmerksom på at ulike dataprogrammer velger ulike konstantfaktorer. Dette spiller ingen rolle dersom vi bare betrakter relative verdier i den transformerte funksjonen. Imidlertid, dersom absoluttverdier er viktig, må vi sjekke med en kjent funksjon for å se hvilket valg av konstantledd som er brukt i det aktuelle programmet.

5.3.2 Fouriertransformasjon av en ren cosinusfunksjon

Dersom vi i stedet velger å finne den fouriertransformerte av funksjonen

$$f(t) = \cos(\omega_0 t)$$

blir fouriertransformen:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \{\cos(\omega_0 t) \cos(\omega t) - i \cos(\omega_0 t) \sin(\omega t)\} dt$$

Vi bruker i så fall ligningene (5.4) og (5.5), og får:

$$\begin{aligned} F(\omega) &= \frac{1}{\sqrt{8\pi}} \int_{-\infty}^{\infty} (\cos\{(\omega_0 + \omega)t\} + \cos\{(\omega_0 - \omega)t\}) dt \\ &\quad - \frac{1}{\sqrt{8\pi}} \int_{-\infty}^{\infty} i(\sin\{(\omega_0 + \omega)t\} + \sin\{(\omega_0 - \omega)t\}) dt \end{aligned} \quad (5.11)$$

Også i dette tilfellet får vi bare bidrag for $\omega = \omega_0$ og $\omega = -\omega_0$, men denne gangen er bidragene reelle og har samme fortegn:

$$F(\omega = \omega_0) = F(\omega = -\omega_0) = 1/2$$

$$F(\omega) = 0 \quad \text{for alle andre } \omega.$$

Figur 5.2 illustrerer det vi hittil har vist: Den fouriertransformerte av en ren sinusfunksjon og en ren cosinusfunksjon er null overalt unntatt ved vinkelfrekvensen til funksjonen vi starter ut med og den negative av denne vinkelfrekvensen. For sinusfunksjonen blir den fouriertransformerte funksjonen rent imaginær og $F(\omega)$ skifter fortegn når vi går fra ω_0 til $-\omega_0$. For cosinusfunksjonen er den fouriertransformerte rent reell og endrer ikke fortegn ved å gå fra ω_0 til $-\omega_0$.

5.3.3 Fouriertransformasjon av en mer sammensatt funksjon

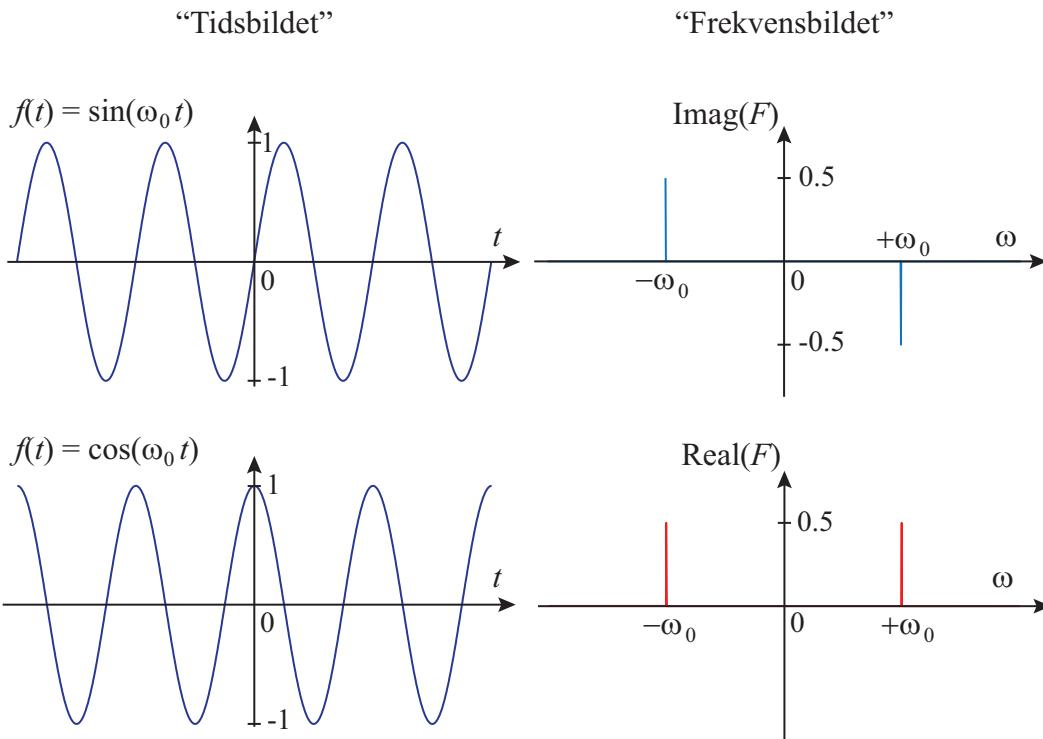
Dersom vi har funksjonen:

$$f(t) = A \cos(\omega_0 t + \phi) = a \cos(\omega_0 t) + b \sin(\omega_0 t)$$

sier resultatene ovenfor oss at den fouriertransformerte av $f(t)$ er:

$$F(\omega) = \begin{cases} a/2 - ib/2 \equiv c & \text{for } \omega = \omega_0, \\ a/2 + ib/2 \equiv c^* & \text{for } \omega = -\omega_0, \\ 0 & \text{for alle andre } \omega. \end{cases} \quad (5.12)$$

Vi innførte et komplekst tall c i ligningene ovenfor for å vise at $F(\omega_0)$ og $F(-\omega_0)$ er komplekskonjugert av hverandre.



Figur 5.2: Venstre: Tidssignalet til en harmonisk funksjon; $\sin(t)$ øverst og $\cos(t)$ nederst. Høyre: De tilsvarende fouriertransformerte funksjonene.

Vi innser forhåpentligvis nå at fouriertransformasjonen er en ganske nydelig metode for å analysere funksjoner. Dersom funksjonen er en ren harmonisk funksjon, vil vi kunne hente ut både amplitud og fase for funksjonen ved å bruke en fouriertransformasjon.

Dersom en funksjon er en sum av flere harmoniske funksjoner, vil vi kunne hente ut både amplitud og fase til hver eneste komponent i summen, siden analysen behandler funksjonen lineært.

Dersom vi har en sum av mange harmoniske funksjoner, vil sumfunksjonen kunne bli ganske uregelmessig og vanskelig å gjennomskue. I slike sammenhenger er i blant den fouriertransformerte av sumfunksjonen langt mer oversiktlig. Det er i slike sammenhenger at fouriertransformasjonen er et nyttig hjelpemiddel i fysikk.

Merk at matematikken fører til at vi får bidrag både ved positiv og negativ vinkelfrekvens i analysen, selv om funksjonen vi analyserer bare kan sies å ha én frekvens (positivt reelt tall). Vi kommer tilbake til dette siden.

Det er viktig for utbyttet av resten av dette kapittelet at du raskest mulig gjennomskuer forskjellen og sammenhengen mellom $f(t)$ og $F(\omega)$. I fysikk omtaler vi gjerne $f(t)$ som *tidsbildet* av en funksjon, mens $F(\omega)$ angis som *frekvensbildet* eller *frekvensspakteret*.

5.4 Fourierrekker

Transformasjonene i ligning (15.1) og (15.2) forutsetter at vi kjenner funksjonen vi skal transformere i et uendelig langt tidsrom. Det er greit nok i idealiserte tilfeller (ren matematikk), men i praktiske fysiske eksperimentelle situasjoner er dette uaktuelt. Dersom vi imidlertid kjenner en funksjon over et endelig tidsrom T , kan vi lage en beskrivelse som strekker seg over uendelig tid ved å anta at funksjonen er periodisk med periodelengde lik T . Dette gir en interessant forenkling.

Dersom $f(t)$ er en periodisk funksjon med periode T , kan fouriertransformasjonen

gjøres mer effektiv enn i den generelle transformasjonen i ligning (15.1). Transformasjonen kan angis ved en uendelig tallrekke, såkalte fourierkoeffisienter, $\{c_k\}$, der indeksen k er et naturlig tall mellom minus og pluss uendelig (!).

Fourierkoeffisientene beregnes slik:

$$c_k = \frac{1}{T} \int_{t_0}^{t_0+T} f(t) e^{-ik\omega_1 t} dt \quad (5.13)$$

hvor $\omega_1 = 2\pi \frac{1}{T}$, det vil si den vinkelfrekvensen som svarer til en funksjon som har nøyaktig én periode innenfor tiden T .

Siden $f(t)$ nå antas å være periodisk, kan startpunktet t_0 for integrasjonen i prinsippet velges fritt. Det antas at $f(t)$ er stykkevis glatt og kontinuerlig, og at $\int |f(t)|^2 dt < +\infty$ når integralet går over et intervall med lengde T .

Den inverse transformasjonen er da gitt ved:

$$f(t) = \sum_{k=-\infty}^{+\infty} c_k e^{ik\omega_1 t} \quad (5.14)$$

hvor igjen $\omega_1 \equiv 2\pi/T$ og svarer til frekvensen som har nøyaktig en sinusperiode innenfor intervallet T .

Dersom $f(t)$ er reell, kan det på grunn av symmetrien i ligning (5.12) enkelt vises at

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \{a_k \cos(k\omega_1 t) + b_k \sin(k\omega_1 t)\} \quad (5.15)$$

hvor

$$a_k = c_k + c_{-k} = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \cos(k\omega_1 t) dt \quad (5.16)$$

$$b_k = i(c_k - c_{-k}) = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \sin(k\omega_1 t) dt \quad (5.17)$$

Ligning (5.15) sammen med uttrykkene (5.16) og (5.17) er gull verdt! De viser at ethvert periodisk signal med periode T kan skrives som en sum av harmoniske signaler (sinussignaler med ulike amplituder og faser). I beskrivelsen inngår bare harmoniske signaler med eksakt et heltalls svingninger innenfor periodetiden T .

Harmoniske signaler med heltallige multippel av en grunnfrekvens danner et ortogonalt basis-sett av funksjoner. Fourierrekken forteller da hvor mye vi har av hver av disse basisfunksjonene.

I prinsippet har vi uendelig mange basisfunksjoner med en frekvens som går fra $\omega_1 \equiv 2\pi/T$ til uendelig. Hva vinner vi da på å bruke en fourierrekke framfor en vanlig beskrivelse i tidsdomenet? Vel, i noen tilfeller er faktisk fourierrekken en mer komplisert beskrivelse enn en tidsbeskrivelse, men i mange fysikkrelaterte situasjoner er det motsatt. Dessuten gir fourierrekken et diagram som iblant avslører lovmessigheter i svingningene det ellers ville vært vanskelig å se.

Vi kommer tilbake til disse vurderingene senere i kapitlet, for det er svært viktig å være bevisst at vi ikke må blande rent matematiske analyser med fysiske tolkninger i hytt og vær, slik det dessverre gjøres en god del i dag!

5.5 Diskret fouriertransformasjon

En generell fourierrekke gitt i ligning (5.13) er basert på en kontinuerlig funksjon angitt i et intervall med lengde T . I det tilfellet endte vi opp med uendelig mange fourierkoeffisienter.

I vår moderne tid er eksperimentelle og computergenererte data bare kvasi-kontinuerlige. Vi sampler en kontinuerlig funksjon, og ender opp med en tallrekke med et endelig antall datapunkter. Anta at dette er datapunkter tatt etter hverandre med en fast tidsdifferanse Δt . Vi kaller datapunktene x_n hvor $n = 0, \dots, N - 1$. Integralet i ligning (5.13) blir da erstattet med en sum.

Den diskrete fouriertransformasjonen er da gitt ved:

$$X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-i \frac{2\pi}{N} kn} \quad (5.18)$$

for $k = 0, \dots, N - 1$. Dersom settet x_n består av verdier gitt i tidsdomenet, vil X_k være tilsvarende sett verdier i frekvensdomenet.

Den omvendte diskrete fouriertransformasjonen ser da naturlig nok slik ut:

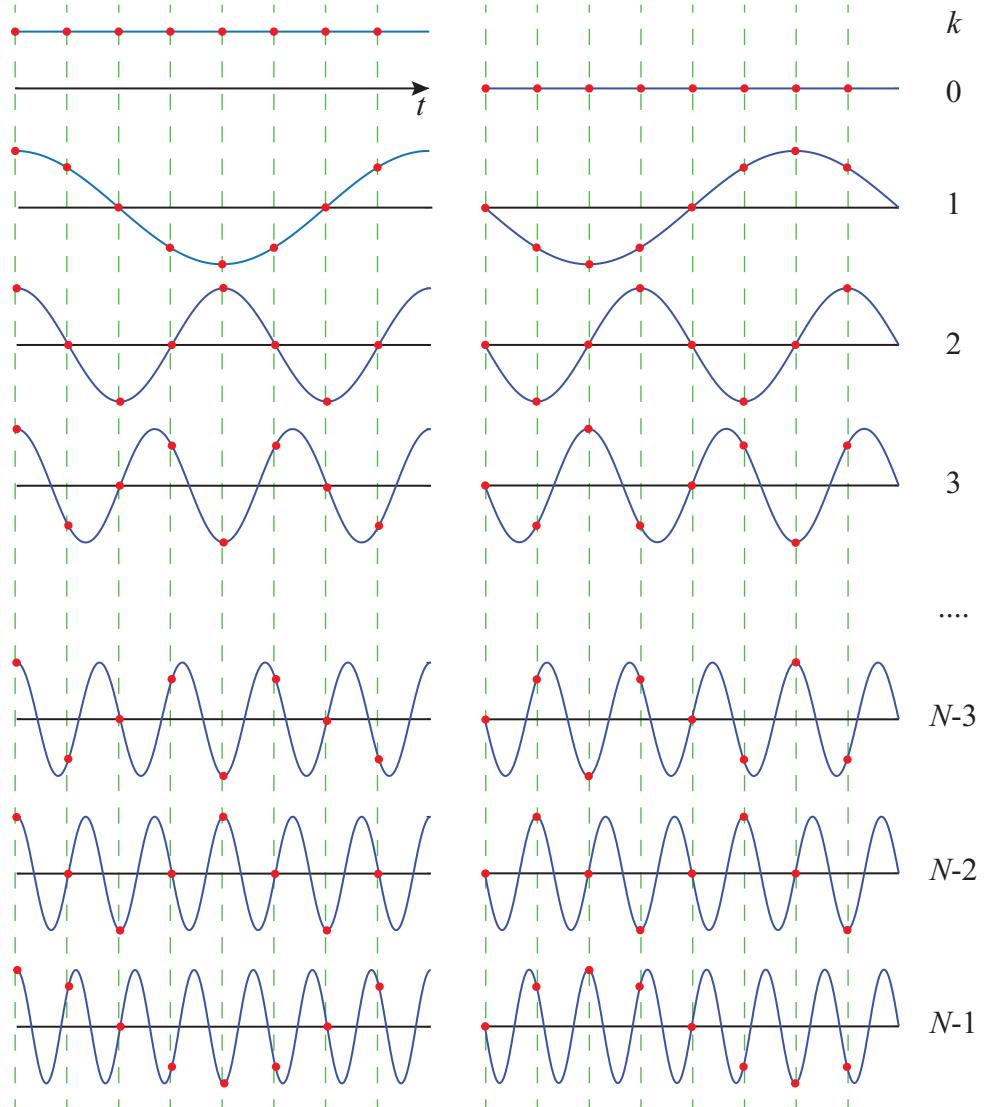
$$x_n = \sum_{k=0}^{N-1} X_k e^{i \frac{2\pi}{N} kn} \quad (5.19)$$

for $n = 0, \dots, N - 1$.

Vi så allerede i ligning (5.3) at fouriertransformasjon betyr at vi multipliserer funksjonen som skal transformeres med sinus- og cosinusfunksjoner med varierende frekvens. For den opprinnelige transformasjonen var funksjonen definert for alle tider t og vi multipliserte denne i prinsippet med sinuser og cosinuser med alle mulige frekvenser.

Når vi gjennomfører en diskret fouriertransformasjon, er funksjonen vi starter ut med bare definert i N punkter. Da kan vi bare multiplisere denne med en diskret representasjon av sinus- og cosinusfunksjoner, det vil si sinus- og cosinusfunksjoner som igjen bare er definert i N punkter.

Videre er funksjonen vår bare definert for en begrenset tid T . Da vet vi fra det vi har lært om fourierrekker, at det er tilstrekkelig å bruke sinus- og cosinusfunksjoner som har et helt antall perioder innenfor tiden T .



Figur 5.3: Funksjonene $e^{-i\frac{2\pi}{N}kn}$ som inngår i en fouriertransformasjon, cosinusfunksjoner til venstre og sinusfunksjonene til høyre. N ulike frekvenser blir brukt, og hver av frekvensene gir oss en fourierkoeffisient X_k .

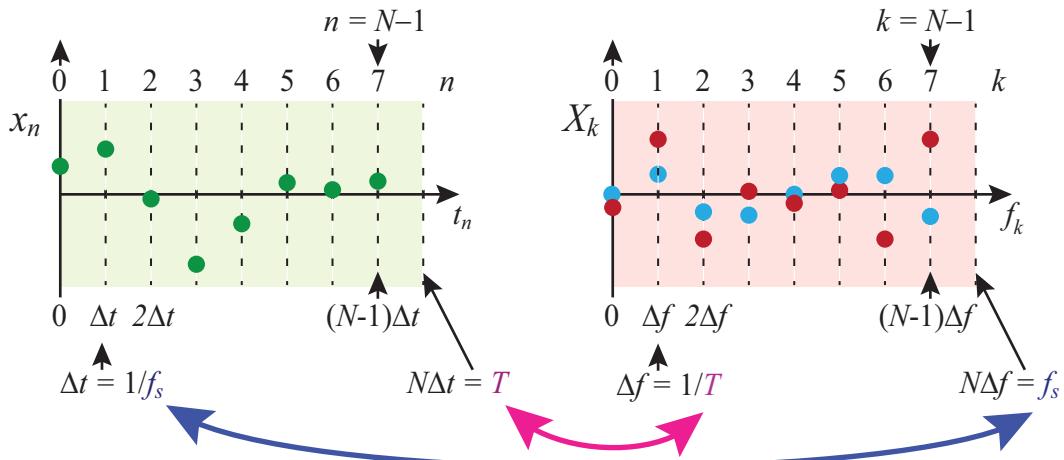
Alt dette er faktisk på plass i beskrivelsen av en diskret fouriertransformasjon slik den er gitt i ligning (5.18). Figur 5.3 viser hvilke funksjoner som ledet $e^{-i\frac{2\pi}{N}kn}$ representerer i et tilfelle der funksjonen som skal transformeres er definert i åtte punkter. Da blir sinus- og cosinusfunksjonene også bare definert i åtte punkter, og bare åtte ulike frekvenser er involvert (0, 1, 2, ..., 7 ganger den frekvensen som har nøyaktig én periode innenfor den tiden funksjonen er definert).

5.5.1 Diskret fouriertransform i mer fysiske termer

I ligning (5.18) og (5.19) er det matematisk sett bare snakk om et sett $\{x_n\}$ med N tall som kan transformeres til et nytt sett $\{X_k\}$ med N tall og tilbake igjen.

La oss nå koble matematikken til litt mer praktisk fysikk ved å se hva indeksene k , n og størrelsen N representerer.

Vi tenker at vi foretar N registreringer av en fysisk størrelse x_n over en begrenset tid T (se venstre del av figur 5.4). Dersom registreringene foretas med et mellomrom i tid lik Δt , sier vi at *samplingsfrekvensen* $f_s = 1/\Delta t$. Sammenhengen mellom størrelsene er som følger: $N = T f_s = T/\Delta t$.



Figur 5.4: En funksjon samplet i $N = 8$ tidspunkt (til venstre) sammen med fouriertransformasjonen til funksjonen (til høyre) som består av $N = 8$ komplekse tall. Realverdiene er gitt ved røde sirkelskiver og imaginærverdiene ved blå. Hvert punkt svarer til et lite tids- og frekvensintervall (hhv i venstre og høyre del). Merk sammenhengen mellom samplingsfrekvensen f_s og Δt og i særdeleshet sammenhengen mellom T og Δf . For å få en høy oppløsning i frekvensangivelsen i frekvensspekteret, må vi altså sample et signal i tilstrekkelig lang tid T .

Merk at hver sampling svarer til et helt lite tidsintervall Δt . I vår figur er det signalet i *begynnelsen* av hvert tidsintervall som registreres.

Fouriertransformasjonen i ligning (5.18) gir oss frekvensbildet (høyre del av figur 5.4). Frekvensbildet består av N komplekse tall, og disse er frekvenskomponentene fra frekvensen 0 (konstant-ledd), og frekvensene $k\Delta f$ hvor $\Delta f = 1/T$. (Denne detaljen bør du merke deg!).

Sammenhengen mellom ligning (5.18) og våre måletider $t_n = n\Delta t$ og de resulterende analysefrekvenser $f_k = k\Delta f$ kan illustreres slik (merk endringene i eksponensialfunksjonen):

$$\begin{aligned} X_k &= \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-i \frac{2\pi}{N} kn} = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-i 2\pi \frac{nT}{N} \frac{k}{T}} = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-i 2\pi (n\Delta t) (k\Delta f)} \\ X_k &= \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-i 2\pi t_n f_k} = \frac{1}{N} \sum_{n=0}^{N-1} x_n \{ \cos(\omega_k t_n) - i \sin(\omega_k t_n) \} \end{aligned} \quad (5.20)$$

for $k = 0, \dots, N-1$.

Fourierkoeffisienten X_k er altså på en måte integralet av produktet av signalet $\{x_n\}$ og en cosinus og en sinusfunksjon med frekvensen $\omega_k = k\Delta f$.

Gjør vi en lignende omskriving av den omvendte fouriertransformasjonen i ligning (5.19), får vi:

$$x_n = \sum_{k=0}^{N-1} (\Re(X_k) \cos(\omega_k t_n) - \Im(X_k) \sin(\omega_k t_n)) \quad (5.21)$$

for $n = 0, \dots, N - 1$. \Re og \Im står som før for realdelen og imaginærdelen.

Realdelen av X_k kan med andre ord betraktes som amplituden til hver enkelt cosinusfunksjon som inngår i summen, mens imaginærdelen av X_k er amplitudene til sinusfunksjonene. De komplekse kryssleddene i den opprinnelige ligning (5.19) er blitt borte på grunn av en symmetri som har rot i relasjonen vi viste i ligning (5.12), men som for diskret fouriertransform ser slik ut:

$$X_{k+1} = X^*_{N-(k+1)} \quad (5.22)$$

for $k = 0, \dots, N - 2$. Vi kommer tilbake til denne symmetrien når vi om litt skal diskutere "folding".

Siden tallrekken x_n er en vilkårlig valgt funksjon, samplet ved N tidspunkt, viser ligning (5.21) at funksjonen alltid kan beskrives som en lineær superposisjon av sinus- og cosinusfunksjoner som har eksakt et helt antall perioder innenfor den tiden T signalet er samplet. Frekvenskomponenten med lavest frekvens f_1 (ser bort fra konstantleddet f_0) har nøyaktig én periode i løpet av tiden T , mens den høyeste frekvensen som inngår i superposisjonen er lik samplingfrekvensen f_s (eller mer nøyaktig $(N - 1)/N \cdot f_s$). I tillegg til disse sinus- og cosinusfunksjonene kommer en konstantfunksjon (X_0) som er halve gjennomsnittsverdien av alle x_n .

Alternativt kan hver enkelt frekvenskomponent angis "på polar form", som en amplitude og fase. Amplituden for hver enkelt frekvenskomponent er da gitt ved:

$$A_k = |X_k| = \sqrt{\Re(X_k)^2 + \Im(X_k)^2} \quad (5.23)$$

hvor $\Re(X_k)$ og $\Im(X_k)$ er hhv realdel og imaginærdel av X_k . Fasen for denne frekvenskomponenten er gitt ved:

$$\phi_k = \arctan(\Im(X_k)/\Re(X_k)) \quad (5.24)$$

Fase er i denne sammenhengen definert ut fra uttrykket $f(t) = A_k \cos(\omega_k t + \phi_k)$.

Husk at dersom du skal regne ut fasen på en datamaskin eller kalkulator, må du få med deg alle fortegnskombinasjoner for realdel og imaginærdel av X_k . Det betyr at du må bruke funksjonen $\text{atan2}(\Im(X_k), \Re(X_k))$.

[♣ ⇒ Kommentar: Det finnes i dag en overordentlig effektiv algoritme for diskret fouriertransformasjon. Effektiviteten har medvirket sterkt til at fouriertransformasjon blir brukt mye i mange fag, ikke minst fysikk. Algoritmen ble visstnok oppdaget allerede i 1805 av Carl Friedrich Gauss, men ble glemt (den var ikke så interessant så lenge vi ikke hadde datamaskiner). Algoritmen ble i 1965 lansert av J.W.Cooley og J.Tukey som da arbeidet ved Princeton University. Deres fire siders artikkel: "An algorithm for the machine calculation of complex Fourier series." i Math.Comput. 19 (1965) 297-301, hører til de "klassiske" artiklene som forandret fysikken.]

I Matlab bruker vi Cooley og Tukey's algoritme når vi anvender *fft* ("Fast Fourier Transform") eller *ifft* ("Invers Fast Fourier Transform"). For å få full uttelling av metoden bør vi passe på at antall punkter N er eksakt et av tallene 2^n hvor n er et heltall. Det er først da vi får benyttet oss fullt ut av symmetrien i en sinus- og cosinusfunksjon.

⇐ ♣]

5.6 Et konkret eksempel

Vi vil nå gi et eksempel på fouriertransformasjon av en konkret funksjon og vil da i praksis se de lovmessighetene vi fant matematisk tidligere i kapitlet. Vi vil gå i stor detalj i den hensikt at kapitlet skal kunne brukes som en praktisk guide når vi ønsker å hente maksimalt og mest mulig presis informasjon ut av en fouriertransformasjon. Vi tror også at en meget detaljert gjennomgang av et eksempel vil kunne hjelpe på forståelsen av fouriertransformasjon, men forståelsen kommer ikke av seg selv! Vær nøyne med detaljer!

Aller først vil vi skrive om ligning (5.18) og (5.19) på en slik måte at indeksene starter fra 1 og ikke fra 0 (for å klargjøre uttrykkene for Matlab-kjøring).

Uttrykket for en diskret fouriertransformasjon blir da som følger:

$$X_k = \frac{1}{N} \sum_{n=1}^N x_n e^{-i\frac{2\pi}{N}(k-1)(n-1)} \quad (5.25)$$

for $k = 1, \dots, N$. Dersom $\{x_n\}$ er en beskrivelse i tidsbildet, vil $\{X_k\}$ være en tilsvarende beskrivelse i frekvensbildet.

Den omvendte diskrete fouriertransformasjonen er gitt ved:

$$x_n = \sum_{k=1}^N X_k e^{i\frac{2\pi}{N}(k-1)(n-1)} \quad (5.26)$$

for $n = 1, \dots, N$.

Innfører vi fysiske størrelser på lignende måte som ovenfor, vil vi med indeks fra 1 til N få følgende uttrykk (symbolene betyr det samme her som ovenfor, bl.a. i figur 5.4):

$$N = Tf_s, \quad f_k = \frac{k-1}{T} = (k-1)\Delta f, \quad \text{og} \quad t_n = \frac{n-1}{N}T = (n-1)\Delta t$$

for $f_k = 0, \dots, f_s(N-1)/N$ og $t_n = 0, \dots, T(N-1)/N$.

Med disse symbolene blir uttrykket for den diskrete fouriertransformasjonen:

$$X_k = \frac{1}{N} \sum_{n=1}^N x_n e^{-i2\pi f_k t_n} \quad (5.27)$$

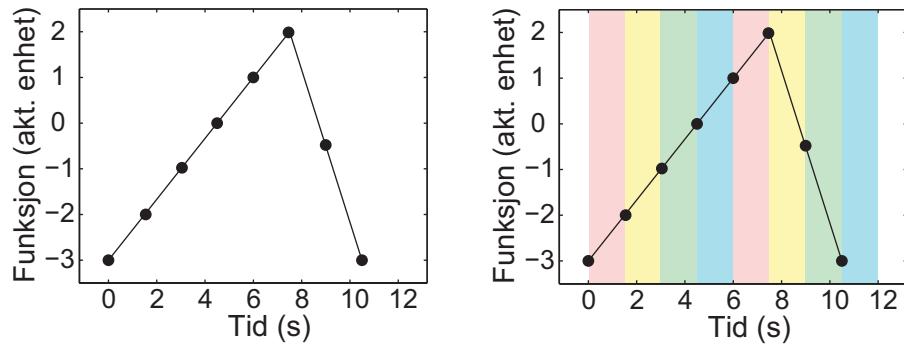
for $k = 1, \dots, N$.

og uttrykket for den inverse diskrete fouriertransformasjonen blir:

$$x_n = \sum_{k=1}^N X_k e^{i2\pi f_k t_n} \quad (5.28)$$

for $n = 1, \dots, N$.

For å gjøre eksemplet så oversiktlig og enkelt som mulig, har vi valgt å beskrive en funksjon i bare $2^3 = 8$ punkter. Funksjonen ser i først omgang ut som en skjev sagtann, og venstre del av figur 5.5 viser funksjonen. Matlab-koden som ble brukt for å definere og plottte funksjonen er vist nedenfor.



Figur 5.5: Venstre: Tidssignalet til en funksjon, definert i åtte punkter. Det er trukket en rett linje mellom hvert av punktene. Høyre: Hvert målepunkt svarer til hvert sitt lille tidsintervall. Funksjonen er derfor definert i en lengre tid enn vi ofte tenker over.

```
%*****
% Første del: Definerer funksjonen vi skal analysere
%*****
tmin = 0.0; % Tid ved begynnelsen av første tidsintervall
tmax = 12.0; % Tid ved slutten av siste tidsintervall
N = 8; % Antall punkter funksjonsverdien skal angis
t = linspace(tmin, tmax*(N-1)/N, N); % Tidspunktene der funksjonen er definert
y = [-3.0 -2.0 -1.0 0.0 1.0 2.0 -0.5 -3.0]; % Vår valgte funksjon
%y = sin(2*pi*(1/tmax)*t); % Alternativt valg av funksjon
plot(t,y,'.-k'); % Funksjonen plottes i sine N punkter
xlabel('Tid (s)');
ylabel('Funksjon (vilkårlig enhet)');
dt = 0.1*(tmax-tmin); % Detaljer for å få ønskede akser
ymin = min(y); % (uvesentlig i vår sammenheng)
ymax = max(y);
dy = 0.1*(ymax-ymin);
axis([tmin-dt tmax+dt ymin-dy ymax+dy]);
```

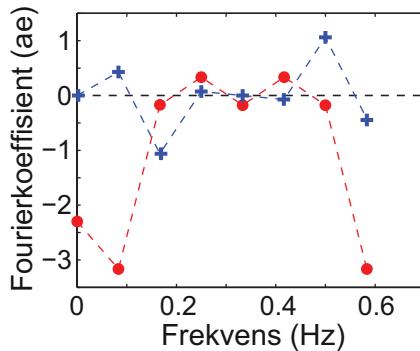
Merk at selve funksjonsverdiene er åtte reelle tall. De kan f.eks. representere en avstand målt i meter, spenning målt i volt eller hva som helst. Vi bruker bare betegnelsen "aktuell enhet" (ae), som kan stå for meter, volt eller hva det nå skulle være.

Funksjonens argument kan være så mangt. Vi har valgt at funksjonen som varierer med tiden (benevning sekunder). Måletall og enheter langs x- og y-aksen er isolert sett totalt uavhengig av hverandre i vårt oppkonstruerte eksempel. Virkelige enheter er bestemt av hva slags fysisk måling dataene $\{x_n\}$ representerer.

I høyre del av figur 5.5 markerer en viktig detalj. Dersom det er åtte målepunkter tatt med 1.5 s mellomrom, er det bare $(8-1)*1.5$ s = 10.5 s mellom første og siste tidspunkt der målingen er angitt. Det er fort gjort å tro at våre åtte målepunkter svarer til en tidsperiode på 10.5 s. Det er feil. Hver måleverdi representerer et helt lite tidsintervall. Total måleperiode svarer til det halvåpne tidsintervallet $[0,12)$ s,

5.6.1 Fouriertransformasjonen

Fouriertransformeres funksjonen, får vi resultatet gitt i figur 5.6. Matlab-koden som er brukt er gitt nedenfor. Her er det flere detaljer å kommentere.



Figur 5.6: Fouriertransformasjonen til funksjonen i forrige figur. Fouriertransformasjonen gir komplekse tall. Realverdiene er gitt ved røde sirkelskiver og imaginærverdiene ved blå kryss. Hvert punkt svarer til et lite frekvensintervall. Totalt frekvensområde strekker seg derfor lenger ut enn bare mellom punktene isolert sett.

```
%*****
% Andre del: Fouriertransformerer funksjonen og plotter denne
%*****
z = (1/sqrt(N))*fft(y); % Selve fouriertransformasjonen
figure;
T = tmax - tmin; % Beregning av hvilke frekvenser de
deltaF = 1.0/T; % forskjellige komponentene representerer
for i = 1:N
    f(i) = deltaF*(i-1);
end;
plot(f,imag(z),'.-b'); % Plotting av realdel og imaginærdel av
hold on; % komponentene
plot(f,real(z),'.-r'); % Plotter absoluttverdi også (iblant)
%plot(f,abs(z),'.-k'); % Plotter absoluttverdi også (iblant)
xlabel('Frekvens (Hz)');
ylabel('Fourierkoeffisient (vilkårlig enhet)');
```

For det første ser vi av programkoden at fouriertransformasjonen baserer seg *bare* på selve funksjonsverdiene. Benevninger er skrett bort, og tidspunktene hvor funksjonen er definert er totalt irrelevant for selve transformasjonen. Resultatet fra transformasjonen er åtte komplekse tall, og realverdi og imaginærverdi for disse åtte tallene er angitt i figuren.

I figur 5.6 er det angitt verdier og benevning langs x-aksen. Dette er verdier vi selv må bestemme ut fra hva vi vet om verdiene langs x-aksen i tidsbildet (som i figur 5.5). Sammenhengen er som følger:

Dersom funksjonen i tidsbildet er definert i N punkter med tidsforskjell Δt , tilsammen over en total tid $T = N\Delta t$, vil funksjonen i frekvensbildet være definert i N punkter med en frekvensforskjell $\Delta f = 1/T$ og represenere et totalt (halvåpent) frekvensintervall fra null til $N/T = 1/\Delta t = f_s$ der f_s er samplingsfrekvensen..

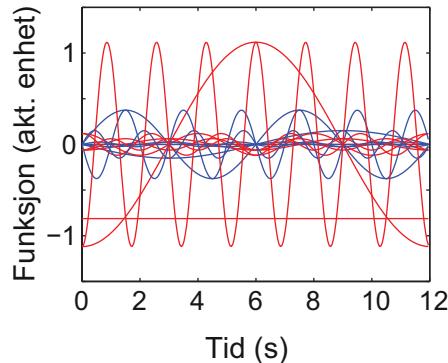
Tallene langs y-aksen i et frekvensbilde av en funksjon avhenger blant annet av hvilken faktor som er brukt i transformasjonen (vi har i ligning (5.18) valgt faktoren $1/N$). Ofte angis derfor funksjonen i frekvensbildet bare i relative verdier. Det er imidlertid ganske enkelt å angi fourierkoeffisientene som amplitudeverdier med samme enhet som den opprinnelige funksjonen $\{x_n\}$.

5.6.2 Tidspoppløsningen, noen kommentarer

Fra ligning (5.21) vet vi at den omvendte fouriertransformasjonen svarer til summasjon av cosinus- og sinusfunksjoner med et heltall perioder. Amplitudene på cosinusfunksjonene

er gitt fra realdelen av fourierkoeffisientene X_k , mens amplituden på sinusfunksjonene er gitt ved imaginærdelen av koeffisientene.

La oss teste dette ut i praksis. I figur 5.7 har vi plottet alle de harmoniske signalene vi får når realdeler og imaginærdel av alle koeffisientene X_k multipliseres med henholdsvis cosinus- og sinusfunksjoner i henhold til enkeltleddene i ligning (5.21). I vårt eksempel var realdelen av X_2 (og dermed også realdelen av X_8) størst slik at disse signalene dominerer litt over de øvrige. Vi ser at disse to svarer til henholdsvis én og syv hele perioder over tidsstrekken.



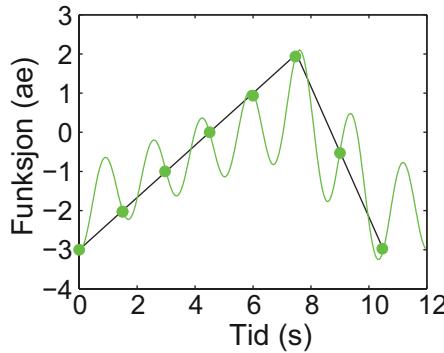
Figur 5.7: Alle harmoniske funksjoner hver for seg som inngår i summasjonen når funksjonen transformeres fra frekvensbildet til tidsbildet. Se tekst for detaljer.

Programkoden for å skrive ut de harmoniske signalene hver for seg er gitt nedenfor.

```
%*****
% Tredje del: Plotter hver enkelt komponent i Fourierspekteret
%*****
nfin = 256; % Ønsker å plotta sin(t) og cos(t)
tfin = linspace(tmin, tmax*(nfin-1)/nfin, nfin); % mellom tidspunkter som brukes
ytotI = zeros(1,nfin);
ytotR = zeros(1,nfin);
figure;
faktor = (2.0*pi/(tmax-tmin));
% Beregner nå etter tur sinus og cosinussignaler som koeffisientene
% representerer, og plotter alle sammen hver for seg (i samme diagram)
for i = 1:N
    ks = - (1.0/sqrt(N))*imag(z(i))*sin(tfin*faktor*(i-1));
    kc = (1.0/sqrt(N))*real(z(i))*cos(tfin*faktor*(i-1));
    plot(tfin,ks,'-r'); % Sinussignalene plottes rødt
    hold on;
    plot(tfin,kc,'-b'); % Cosinussignalene plottes blått
    ytotI = ytotI + ks; % Summerer opp (for fjerde del)
    ytotR = ytotR + kc;
end;
xlabel('Tid (s)');
ylabel('Funksjon (vilkårlig enhet)');
```

Summerer vi opp alle sinus- og cosinussignalene i figur 5.7, får vi den grønne kurven gitt i figur 5.8. I samme figur har vi også tegnet inn den opprinnelige funksjonen, med eksakt samme skalering både i x- og y-retning som det grønne sumsignalet.

Det ser ut for at det slett ikke er samsvar mellom den opprinnelige funksjonen $\{x_n\}$ og resultatet etter en vanlig og en invers fouriertransformasjon etter hverandre. Her er det imidlertid en viktig detalj å merke seg! Sinus og cosinusfunksjonene vi har plottet i figur 5.7 er plottet for "alle" tidspunkt, ikke bare de tidspunktene den opprinnelige funksjonen er angitt i. Dette er noe vi har lagt inn i vår egen presentasjon for å få fram et viktig poeng. Detaljer er gitt i koden ovenfor.



Figur 5.8: *Summen av alle harmoniske funksjoner som fourierspekteret svarer til når vi transformerer funksjonen tilbake til tidsbildet sammen med opprinnelig funksjon. I tidspunktene den opprinnelige funksjonen er definert, har summen av de harmoniske signalene identisk verdi med den opprinnelige funksjonen, men ikke mellom disse punktene. Se tekst for detaljer.*

Vi ser at dersom vi begrenser oss til å angi sumsignalet i bare nøyaktig de tidspunktene som den opprinnelige funksjonen ble angitt i, får vi nøyaktig samme resultat som den opprinnelige funksjonen! Utenfor disse tidspunktene er det *ikke* samsvar mellom vår heltrukne kurve mellom de opprinnelige målepunktene og sumfunksjonen av harmoniske funksjoner. Det er imidlertid nokså meningsløst å sammenligne funksjonsverdier der funksjonen faktisk ikke er definert.

Programkoden for å plotte summen av alle harmoniske funksjoner i den omvendte transformasjonen, er gitt nedenfor.

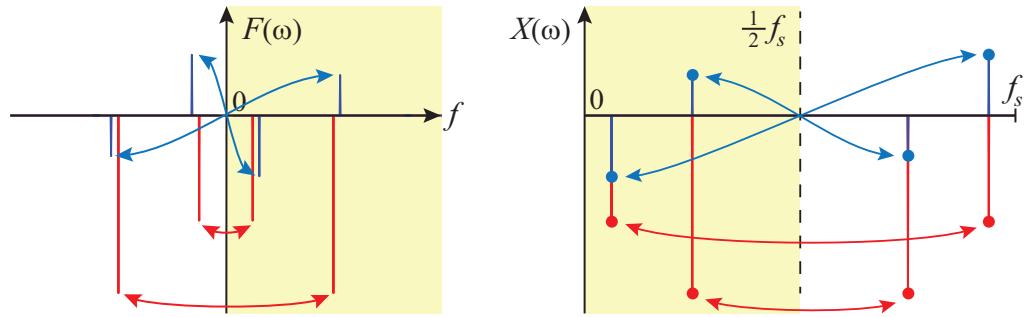
```
%*****
% Fjerde del: Plotter summen av alle komponentene i Fourierspekteret
%*****
figure;
plot(t,y,'-k'); % Plotter opprinnelig funksjon (n1 punkt)
hold on;
ytot = ytotI+ytotR; % Plotter sum av sinus og cosinussignaler
plot(tfin,ytot,'-g');
xlabel('Tid (s)');
ylabel('Funksjon (vilkårlig enhet)');
```

5.6.3 Speiling / folding

Vi viste tidligere (ligning (5.12)) i et spesialtilfelle at for en kontinuerlig fouriertransformasjon gjelder det at $F(\omega_0) = F^*(-\omega_0)$, det vil si at den fouriertransformerte ved en vinkelfrekvens er den kompleks konjugerte av den fouriertransfomerte ved den negative vinkelfrekvensen. For diskret fouriertransformasjon har vi laget indeksene slik at vi slipper å arbeide med negative frekvenser. Likevel kan vi gjenfinne oppsplittingen i det fouriertransformerte signalet også ved diskret fouriertransformasjon. Figur 5.9 forsøker å skissere dette.

Årsaken til symmetrien i figur 5.9 for diskret fouriertransformasjon, har den oppmerksomme leser kanskje allerede oppdaget. I figur 5.3 ser vi at for cosinusdelen av $e^{-i\frac{2\pi}{N}kn}$ er punktene identiske for $k = 1$ som for $k = N - 1$, og tilsvarende identiske for $k = 2$ som for $k = N - 2$. Generelt er cosinusleddene de samme for $k = m$ som for $k = N - m$ for $m = 1, \dots, N/2$. (NB: $k = 0, 1, 2, \dots, N - 1$ i denne figuren.)

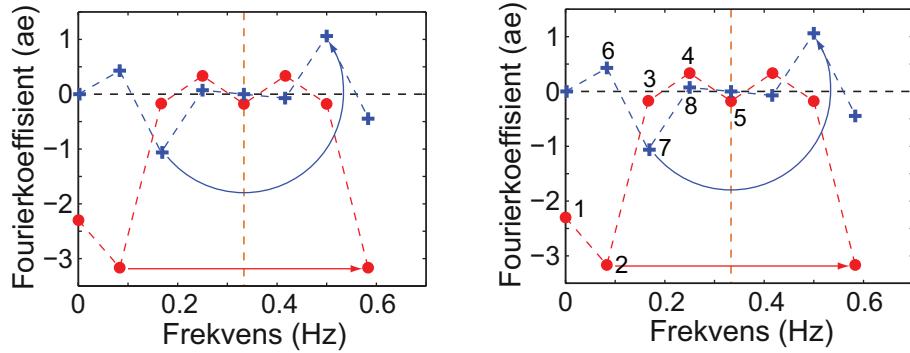
For sinusdelen av $e^{-i\frac{2\pi}{N}kn}$ er punktene identiske, men med motsatt fortegn, for $k = 1$ som for $k = N - 1$, og det samme gjelder generelt for $k = m$ sammenliknet med $k = N - m$.



Figur 5.9: Til venstre: Et fourierspekter ved kontinuerlig fouriertransformasjon av et signal som varer ved i det uendelige, inneholder alle frekvenser mellom $-\infty$ og $+\infty$, men det er en “speiling/folding” og kompleks konjugering omkring frekvensen null. Reell del av den fouriertransformerte funksjonen er markert i rødt, den imaginære i blått. Ved diskret fouriertransformasjon av et signal (høyre del) som bare er definert i et endelig antall punkter i en begrenset tid, inngår bare frekvenser mellom 0 og (nesten) samplingsfrekvensen f_s . Her får vi også en speiling/folding og kompleks konjugering, men denne gang omkring halve samplingsfrekvensen $f_s/2$. (Vi har forskjøvet de reelle i forhold til de imaginære punktene noe i venstre del for at alt skulle bli lett synlig.) Den delen av diagrammene som har en lett bakgrunnsfarge, inneholder all informasjon i det fouriertransformerte signalet siden den øvrige halvparten bare er den kompleks konjugerte av den første.

for $m = 1, \dots, N/2$. Det betyr nettopp at $e^{-i\frac{2\pi}{N}kn}$ og $e^{-i\frac{2\pi}{N}(N-k)n}$ er komplekst konjugerte av hverandre, og dermed ser vi ut fra ligning (5.18) at $X_k = X_{N-k}^*$ for $k = 1, 2, \dots, N/2$. Det betyr at vi har en klar symmetri omkring frekvensen $f_{N/2}$ (for $k = 0, 1, 2, \dots, N - 1$), det vil si halve samplingsfrekvensen.

[♣ ⇒ Merk at i en diskret fouriertransformasjon får vi like mange tall ut som tall inn. Det vil si, det kan diskuteres! Dersom tidsbeskrivelsen $\{x_n\}$ er reell, er det nok med N reelle tall for å angi tallrekken $\{x_n\}$. Det må imidlertid $2N$ reelle tall til for å angi $\{X_k\}$ fordi $\{X_k\}$ er komplekse. Det betyr at ikke alle X_k kan være uavhengige av hverandre i slike tilfeller. Halvparten av alle X_k er egentlig overflødige når $\{x_n\}$ er reelle. Vi vil se nøyere på dette ved å gå tilbake til eksemplet vårt i figur 5.6.]



Figur 5.10: Venstre: Fouriertransformasjonens komplekse tall viser en form for symmetri omkring frekvensen $f_{N/2+1}$. “ae” står for “aktuell enhet” (samme enhet som opprinnelige signal, men iblant kombinert med en skaleringsfaktor). Se teksten for detaljer. Høyre: Av de $2N$ punktene som fouriertransformasjonen fører til, er det bare N uavhengige verdier (markert med numrene 1 - 8). Dette er som forventet siden vi startet ut nettopp med N reelle verdier.

Vi ser en symmetri i frekvensbildet: $X_k = X_{N+2-k}^*$ for $k = 2, 3, \dots, N/2 + 1$ (forutsatt av vi opererer med Matlab-indekser som starter på 1, dvs. $k = 1, 2, \dots, N$). Dette kommer fram i venstre del av figur 5.10. Realdelen av tallene har samme verdi etter refleksjon (speiling) omkring linjen $f_{N/2+1}$. Imaginærdelene av

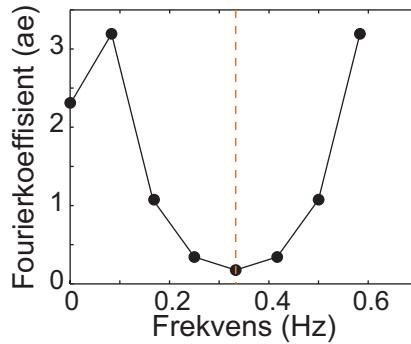
tallene har en rotasjon 180 grader rundt punktet $(f_{N/2+1}, 0)$. Denne formen for symmetri svarer nettopp til $f_k = f_{N+2-k}^*$.

Det er imidlertid et klart unntak fra symmetriregelen. Første tall i frekvensbildet er noe for seg selv. Det svarer til en frekvens lik null, med andre ord til en konstantverdi. Konstantverdien er alltid reell når den opprinnelige funksjonen er reell.

I vårt tilfelle startet vi ut med åtte reelle tall og foretok en transformasjon som på en entydig måte skal inneholde all informasjon i den opprinnelige representasjonen. Da kan det ikke være mer enn åtte frihetsgrader i det endelige resultatet.

På grunn av “speilingen” omkring punktet ved frekvensen $f_{N/2+1}$, er det klart at alle punkter over denne frekvensen følger lovmessig av verdiene til punktene under denne frekvensen. Videre vet vi at ved frekvensen null er imaginærverdien alltid null. På grunn av symmetrien vil det samme gjelde for punktet ved $f_{N/2+1}$. Vi ser da av høyre del av figur 5.10 at vi står igjen med $N/2 + 1$ reelle verdier og $N/2 - 1$ imaginære verdier som er uavhengige av hverandre, tilsammen N uavhengige verdier. Transformasjonen utspenner da et funksjonsrom som er akkurat like stort som det vi startet ut med, og det er tilfredsstillende å konstatere. Dette gjelder også i de tilfellene der vi starter ut med en kompleks funksjon i tidsbildet, men detaljene blir da til dels ganske annerledes enn i vårt tilfelle. $\Leftarrow \spadesuit$

Dersom vi bare betrakter absoluttverdien av fourierkoeffisientene X_k , bør de være symmetrisk rundt frekvensen $f_{N/2+1}$ (for $k = 1, 2, \dots, N$). Koeffisienten som representerer null frekvens, det vil si en konstant-funksjon, er ikke med i denne symmetrien. Dersom vi plotter bare absoluttverdiene av fourierkoeffisientene i figur 5.6, får vi figur 5.11. Vi ser klart symmetrien rundt $f_{N/2+1}$, og at konstantfunksjonsleddet er noe for seg selv. Symmetrien vi ser kaller vi “speiling” eller “folding”.



Figur 5.11: Absoluttverdien av fourierkomponentene viser klart speilingen eller foldingen omkring frekvensen $f_{N/2+1}$.

Ved fourieromvending av et reelt signal $\{x_n\}$ for $n = 0, \dots, (N - 1)$, er det bare de første $N/2 + 1$ koeffisientene som er av interesse. De representerer frekvenser fra null (konstantledd) opp til og med halve samplingsfrekvensen. De øvrige koeffisientene er bare kompleks konjugerte av de koeffisientene vi faktisk bruker.

5.6.4 Samplingsteoremet

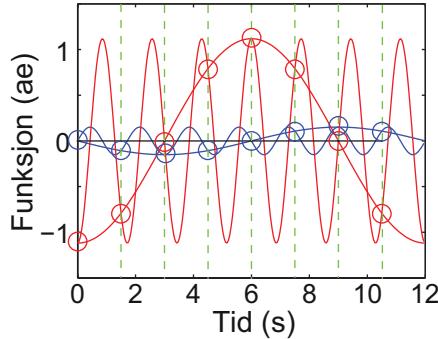
Det er nok en grunn til at vi ikke er særlig interessert i den øvre halvdelen av det symmetriske forløpet i frekvensbildet. Vi så et eksempel på dette allerede i figur 5.3. I figur 5.12 er det vist et lignende diagram, men denne gang viser vi de fire største sinus og cosinussignalene som tilsammen gir den skjeve sagtannfunksjonen vår i figur 5.5.

Funksjonene er tegnet ved “alle” tidspunkt, men tidspunktene der den opprinnelige funksjonen er definert er markert med vertikale stippled linjer. Vi ser da at funksjonene med svært forskjellig frekvens likevel har nøyaktig samme verdi i disse tidspunktene, selv

om verdiene utenfor disse tidspunktene er vidt forskjellige. Dette er i samsvar med ligning

$$e^{i \frac{2\pi}{N} kn} = e^{-i \frac{2\pi}{N} k(N-n)} \quad (5.29)$$

for k og $n = 1, \dots, N - 1$ i det tilfelle at disse indeksene generelt går fra 0 til $N - 1$.



Figur 5.12: *De harmoniske funksjonene med frekvensene f_k og f_{N-k} (her: $k = 1$) har nøyaktig samme verdi akkurat i de tidspunktene den opprinnelige funksjonen var definert (forutsatt at $k = 0, 1, \dots, N - 1$). Vi kan derfor ikke skille mellom de to ved dette valget av tidspunkt der funksjonen er definert (for den samplingsfrekvensen som ligger bak). For å kunne skille funksjoner med ulik frekvens, må samplingsfrekvensen være minst dobbelt så stor som høyeste frekvenskomponent.*

Dette er et eksempel på et generelt prinsipp:

Skal vi representere en harmonisk funksjon på en entydig måte ved et begrenset antall målinger, må måletettheten (målefrekvensen, samplingsfrekvensen) være så stor at vi får minst to målinger innen hver periode på det harmoniske signalet. “Nyquist-Shannons samplingsteorem” sier dette på en mer kompakt måte: Samplingsfrekvensen må være minst dobbelt så stor som høyeste frekvenskomponent i et signal for at et samplet signal skal gi et entydig bilde av signalet. Dersom det kan forekomme høyre frekvenser i det opprinnelige signalet, må disse filtreres bort før sampling for at resultatet skal bli entydig.

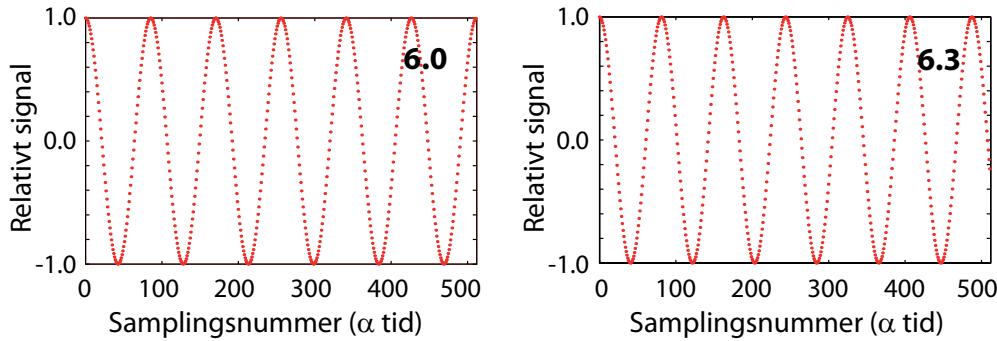
5.7 * En finurlighet

I høyre del av figur 5.2 så vi at en *kontinuerlig* fouriertransformasjon av en perfekt sinus eller cosinus gir “skarpe linjer” (“deltafunksjon”). Vi kan ledes til å tro at det alltid er slik dersom signalet er en ren sinus, men for en diskret fouriertransformasjon er dette ikke slik, generelt sett. Vi får et sammenlignbart resultat bare dersom sinus-signalet vi starter ut med har *eksakt* et helt antall perioder innenfor samplingstiden vår.

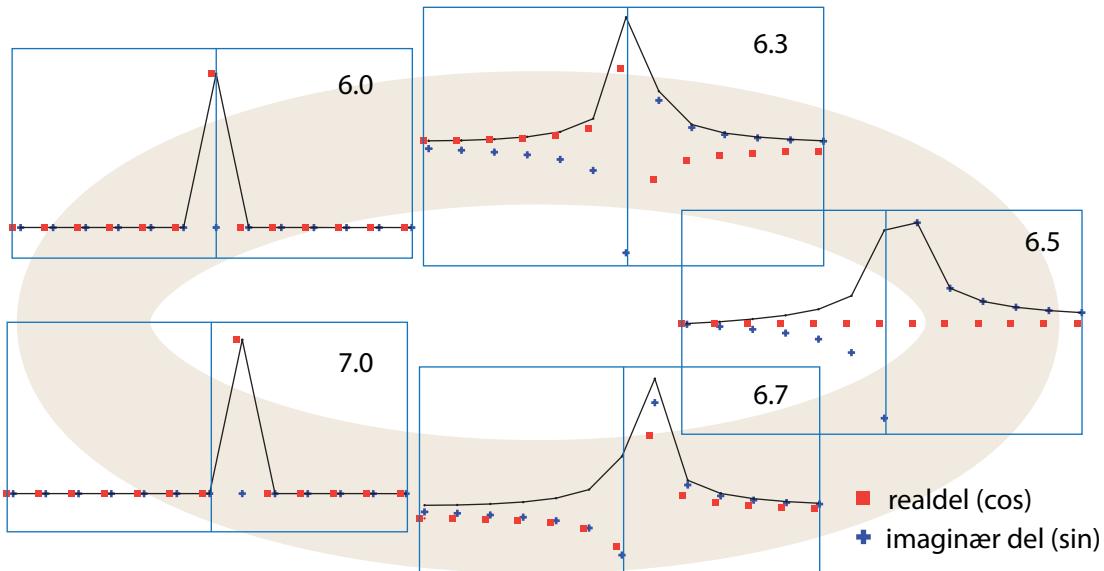
Figur 5.13 og 5.14 viser hvordan et fourierspekter ser ut dersom cosinus-signalet vi betrater ikke har nøyaktig et heltall perioder innenfor den totale samplingstiden. Til venstre i figur 5.13 er det eksakt 6 hele perioder innenfor de 512 samplingene som er foretatt. Til høyre er det 6.3 perioder innen det samme intervallet.

Figur 5.14 viser de første punktene i fourierspekteret. Røde firkanter svarer til cosinusleddene (reelle delen av fourierspekteret), mens de blå kryssene svarer til sinusleddene (imaginære delen av fourierspekteret). Den helstrukne sorte linjen svarer til amplituden bestemt fra ligning (5.23). Denne kontinuerlige kurven kan virke litt misvisende siden det ikke finnes noen verdier mellom de diskrete tallene.

Vi ser at fourierspekteret for et signal med seks hele periodetider innenfor samplings-perioden er som forventet. Den reelle delen av tall nummer 7 i rekken er det eneste tallet i frekvensspekteret forskjellig fra null. Dette er rimelig siden vi jo kan beskrive det samplede



Figur 5.13: Eksempel på signal med eksakt (6.0) og ikke eksakt (6.3) antall hele perioder innenfor den totale samplingstiden. Signalet er samplet i 512 tidspunkter.



Figur 5.14: De første punktene i Fourierspekteret av signaler som i figur 5.13. Amplituder for cosinusfunksjoner (reell del av fourierspekteret) er markert med røde firkanter, mens blå kryss gir amplituder for sinusbidragene (imaginær del). Se tekst for detaljer.

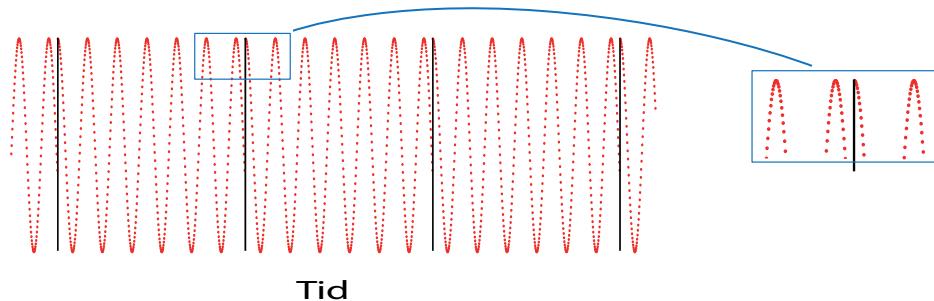
signalet ved hjelp av én av frekvensene $f_{k+1} = k/T$, nemlig den hvor $k = 6$. Vi trenger ikke noe mer!

Når signalet har 6.3 perioder innenfor den totale samplingstiden blir ting annerledes. Frekvensspekteret inneholder en rekke tall forskjellig fra null, men alle tallene ligger i nærheten av punkt nummer syv. Det er bidrag både fra den reelle og imaginære delen.

Frekvensspekteret alene kunne tolkes dithen at signalet, allerede da vi laget det, besto av flere bidrag som alle har nøyaktig et helt antall bølgelengder innenfor T . I virkeligheten vet vi at signalet vi startet ut med var helt rent med én eneste frekvens. Problemet er bare at samplingstiden som ble valgt svarer til 6.3 perioder innenfor den totale samplings-perioden (ikke et heltall perioder). Dette er en svakhet ved diskret fouriertransformasjon som vi må ha i bakhodet når vi vurderer et frekvensspektrum.

Pragmatisk sett er det ikke så rart at det blir slik. Det finnes ingen basisfunksjon i den diskrete transformasjonen som svarer eksakt til 6.3 perioder innenfor den totale samplingstiden. Men signalet SKAL kunne gjengis 100 prosent riktig innenfor det aktuelle tidsintervallet etter at vi foretar en invers fouriertransformasjon av frekvensspekteret. Da må vi mikse nærliggende frekvenser og ha korrekte innbyrdes faser for å kunne matche det opprinnelige signalet.

Dette kan vi kanskje forstå enda bedre dersom vi går tilbake til forutsetningene for



Figur 5.15: Signalet vi i virkeligheten analyserer når vi bruker fourierrekke-analyse på signalet til høyre i figur 5.13. Signalet er periodisk med periode lik avstanden mellom de sorte strekene, og strekker seg fra minus til plus uendelig.

fourierrekker, nemlig at signalet skal være periodisk med periode T . Det betyr at signalet vi faktisk analyserer når vi bruker en fourier-rekke-analyse av signalet med 6.3 perioder innenfor tiden vi sampler, ser ut som vist i figur 5.15. Da innser vi at et slikt signal kan ikke beskrives ved en enkel sinus! Og det er grunnen til at fourieranalysen inneholder flere komponenter!

Figur 5.14 viser også frekvensspekteret for signaler med 6.5, 6.7 og 7.0 perioder innenfor den totale samplingstiden. Hensikten er å vise at frekvensspekteret endres på en ikke-triviell måte når signalfrekvensen endres.

[♣] ⇒ Et talleksempl kan kanskje være av interesse i denne sammenheng:

Anta at vi digitaliserer (sampler) et tidsavhengig signal ved å sample 2^m ganger hvor m er et heltall, nærmere bestemt $N = 512 = 2^9$ punkter. Anta at vi har en samplingsfrekvens f_s på 1.0 kHz. I så fall vil tidsstrenge vi sampler være $T = N/f_s = 512$ ms lang og frekvensoppløsningen i en fouriertransformasjon blir $\Delta f = 1.9531\dots$ Hz.

Oppløsningen i frekvensspekteret er med andre ord bestemt av total samplingstid, mens den maksimale frevensen i frekvensdomenet er bestemt av den opprinnelige samplingsfrekvensen.

Anta at selve signalet vi samplet har frekvensen 50 Hz, slik at periodetiden T er 20 ms. Det samplede signalet inneholder da $512/20 = 25.6$ perioder. I fourierspekteret vil punkt nr 26 og punkt nr 27 dominere og ha nesten like stor absoluttverdi (antar at indeksene starter ved 1). Også noen nærliggende punkter i frekvensspekteret vil ha verdier forskjellig fra null.

Dersom vi hadde samplet signalet i $2^{14} = 16384$ punkter, ville total samplingstid vært 16.384 sekunder. Frekvensoppløsningen i frekvensbildet ville da blitt 0.061035... Hz, og 50 Hz signalet ville få en klart markant topp i punkt nr 820, men litt også i punkt 821 og i enda mindre grad punktene på begge sider av disse. Totalt sett ville da toppen i frekvensspekteret bli langt mer markant og "skarp" enn da vi bare samplet signalet i 0.512 s.

Vi konstaterer imidlertid i begge tilfeller at selv om signalet hadde en svært ren 50 Hz tidsvariasjon, og at vi samplet med 1000 Hz, som jo svarer nøyaktig til 20 samlinger per periode, er frekvensspekteret likevel ikke skarpt etter en FFT-transformasjon.

⇐ ♣]

5.7.1 Fouriertransformasjon av mer kompliserte signaler

Fouriertransformasjon brukes mye til såkalt frekvensanalyse hvor vi bestemmer hvilke frekvenskomponenter som finnes i et signal. Frekvensspekteret er nyttig fordi det ofte gir et "fingeravtrykk" av de fysiske prosessene som ligger bak signalet vi betrakter.

Eksempelvis viser figur 5.16 et frekvensspekter fra et lydsignal fra en tverrfløyte. I figuren er relative amplituder i frekvensspekteret angitt, se ligning (5.23). Vi mister da faseinformasjonen, men "styrken" på ulike frekvenskomponenter kommer godt fram.

Spekteret består i hovedsak av en rekke topper med litt ulik høyde. Toppenes plassering har en viss lovmessighet. Det finnes en frekvens f_0 (burde kanskje heller vært kaldt f_1), den såkalte *grunntonen*, slik at de resterende linjene i en gruppe av linjer tilnærmet har

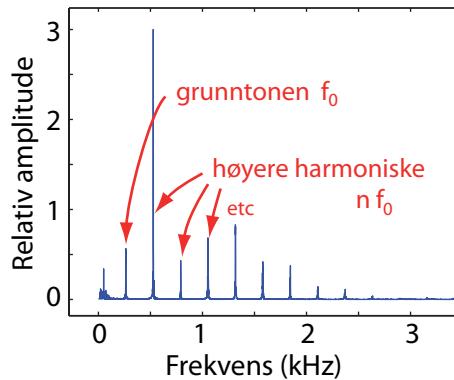
frekvensene $k f_0$, der k er et heltall. Vi sier at frekvensene $k f_0$ for $k > 1$ er *harmoniske av grunntonens*, og vi omtaler dem som “*overtoner*”.

I figur 5.16 ser vi også en linje ved en frekvens nær null. Det er en topp ved 50 Hz, som er nettfrekvensen på lysnettet. Dette signalet har på et eller annet vis sneket seg inn sammen med lyden fra fløyta, kanskje ved at elektronikken har plukket opp elektriske eller magnetiske felter et sted i signalgangen.

Frekvensspekteret viser at når vi spiller fløyte, vil lufta ikke bare svinge ved én bestemt frekvens, men ved flere frekvenser samtidig. Ulike instrumenter kan karakteriseres ved frekvensspekteret av lyden de genererer. Noen instrumenter gir lite overtoner/harmoniske, mens andre (f.eks. obo) gir mange!

Frekvensspekteret kan benyttes som utgangspunkt også ved syntese av lyd. Siden vi kjenner intensitetsfordelingen i frekvensspekteret til fløyta, kan vi i prinsippet starte med denne fordelingen og foreta en invers fouriertransformasjon for å generere lyd som høres ut omtrent som en fløyte.

Det må likevel bemerkes at vårt inntrykk av lyd ikke bare bestemmes av frekvensspekteret for et vedvarende lydsignal, men også av hvordan lyden starter og dør ut. I den sammenhengen er fouriertransformasjon til lite hjelp.



Figur 5.16: Et frekvensspekter av lyd fra en tverrfløyte. Amplitudeverdier (intensiteter) er vist.

La oss nevne noen få andre eksempler på bruk av fouriertransformasjon:

- Dersom vi spiller i et band og bruker ulike forsterkere, ønsker vi at signalet som kommer f.eks. fra en mikrofon skal bli forsterket uten å bli forvrengt. En type forvrengning viser seg ved at det dukker opp høyere harmoniske av de opprinnelige frekvensene. Ved å sende et rent sinussignal inn på forsterkeren, og digitalisere lyden etter at signalet har gått gjennom forsterkeren, kan vi ved hjelp av fouriertransformasjon undersøke om det er kommet noe ekstra signal til som ikke burde vært der.
- Ved å analysere lysintensiteten fra enkelte stjerner er det oppdaget periodiske variasjoner. Ved en fouriertransformasjonanalyse kan vi bestemme periodetiden og vi kan få et mål for intensitetsfluktasjonen som kan arkiveres for å se på langtidsendringer på sikt.
- Solflekkaktiviteten er selvfølgelig også en tidsavhengig størrelse som lar seg analysere ved hjelp av fouriertransformasjon.
- Fouriertransformasjon kommer også inn for å finne hvor høye frekvenser et system må kunne håndtere for at vi skal kunne sende/motta digitale pulser (firkantpulser) uten at pulsene skal deformeres så mye at det går ut over lesbarheten.

- Fouriertransformasjon kan også brukes om romlige forhold i stedet for tidsavhengige forhold. I optikken brukes nå ofte uttrykket fourieroptikk.
- En fjærpendel som beskrevet i kapittel 1 svinger harmonisk såfremt Hookes lov gjelder og friksjonen er neglisjerbar. Ved fouriertransformasjon av posisjon versus tid, kan vi oppdage når bevegelsen ikke lenger er harmonisk, og vi kan utnytte denne muligheten i ulike typer studier.

En mengde detaljer knyttet til fouriertransformasjon finner du på f.eks. Wikipedia.

5.8 Tidsbegrenset signal

Fra ligning (5.20) går det fram at en fouriertransformasjon egentlig er en sum (integrasjon) av produktet mellom signalet som skal transformeres og en ren sinus eller cosinus. Figur 5.17 viser med røde kurver to ulike signaler som er tidsbegrensede. Det er rett og slett vår tidsbegrensede kraft fra kapittel 2 som er vist.

Matematisk er kraften gitt som:

$$F(t) = F_0 \cos(\omega(t - t_0)) e^{-(t-t_0)/\sigma^2}$$

hvor σ angir varigheten på kraften (fra amplituden har hatt sin maksimale verdi til amplituden har sunket til $1/e$ av max). ω er vinkelfrekvensen til den underliggende cosinusfunksjonen, og t_0 er tiden der kraften har maksimal amplitude (toppen av kraftpulsen forekommer ved tiden t_0).

I a og b er kraften kortvarig (σ liten), mens i c og d varer kraften lengre tid (σ fem ganger så stor som i a og b).

I a og c har vi i tillegg til kraftpulsen i rødt, tegnet inn cosinussignalet med frekvens eksakt lik $\omega/2\pi$ med blå tynnere strek. I b og d har cosinussignalet 10 % høyere frekvens.

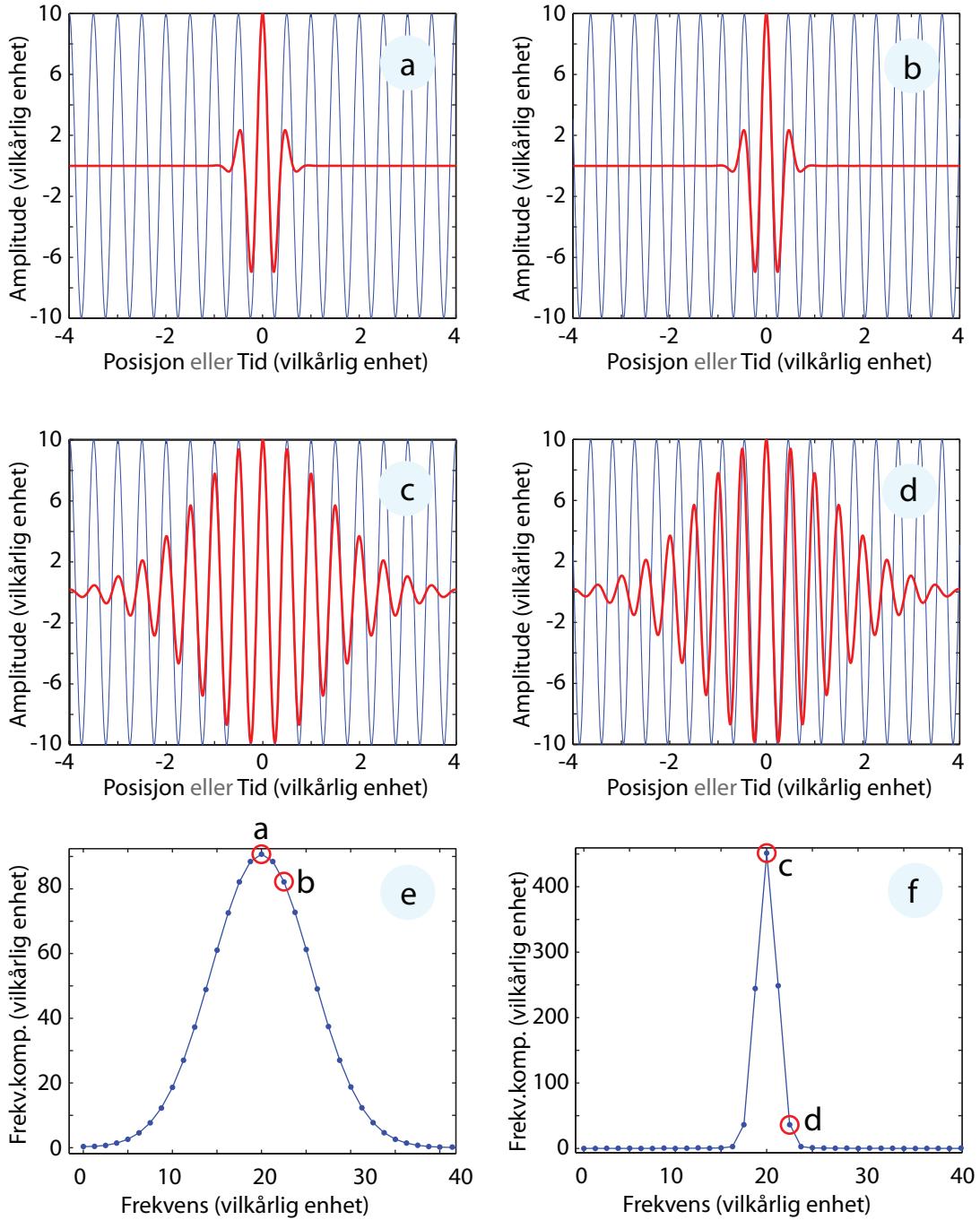
Vi ser at integralet (summen) av produktet mellom den røde og blå kurven i a og i b vil være omrent det samme. Derimot ser vi at det tilsvarende integralet for d må være betydelig mindre enn integralet for c siden kraft-signalet og cosinussignalet kommer i motfase litt vekk fra sentrum av pulsen i d.

Dersom vi foretar en fouriertransformasjon av selve den røde kurven i a (den kortvarige kraften), og tar absoluttverdien av fourierkoeffisientene, blir resultatet som i e. Fouriertransformasjonen av kraften i c (kraften som varer litt lengre), er vist i figurens nedre høyre hjørne f. Vi ser at fouriertransformasjonen fanger opp de prediksjonene vi kunne gjøre ut fra figurene a til d.

Merk at det kortvarige signalet ga et bredt frekvensspekter, mens signalet med flere perioder i den underliggende cosinus-funksjonen, ga et smalere frekvensspekter. Dette er på ny en manifestasjon av prinsippet vi har vært borti tidligere, og som har klare likheter med Heisenbergs uskarphetsrelasjon.

[♣ ⇒ I kapittel 2 sammenlignet vi frekvensresponsen til et svingende system med frekvensspekteret til kraften. Poenget var at dersom en påtrykt kraft virker lenge nok, vil systemets frekvensrespons være bestemt av Q-verdien til systemet. Derimot, dersom kraftpulsen varer svært kort tid, ville frekvensresponsen til systemet flytte mer ut. I ekstreme tilfeller vil bredden på frekvensresponsen til systemet bli identisk med bredden til frekvensspekteret av kraften alene. I figur 5.17 er det kun frekvensspekteret til kraften vi betrakter!]

⇐ ♣



Figur 5.17: Fouriertransformasjon av et cosinussignal som er konvolutert med en gaussisk funksjon. Bare en liten del av det totale frekvensspekteret er vist. Se teksten for detaljer.

5.9 Til ettertanke

Resultater slik vi ser i figur 5.17 kan lett føre til alvorlige feiltolkninger. Ia ser vi at kraften varer kun en meget kort tid (få periodetider). Resten av tiden er kraften rett og slett null (eller vi kunne satt den til eksakt lik null med ingen nevneverdig forskjell i frekvensspekteret).

Hva viser fouriertransformasjonen? Fra delfigur e kan vi se at det er om lag 30 frekvenskomponenter som er klart forskjellig fra null. Det betyr at vi må ha i størrelsesorden 30 ulike sinus- og cosinus-funksjoner *som varer ved HELE tiden* for å beskrive det opprinnelige signalet (jamfør ligning (5.21)).

Noen trekker da den slutning at *egentlig* så er kraften ikke null der den ser ut for å være null, men rett og slett summen av om lag 30 ulike sinus- og 30 ulike cosinusfunksjoner overalt i hele tiden. Dette er vrøvl!

Det er korrekt at vi kan beskrive kraften i delfigur a ved hjelp av alle disse sinus- og cosinusfunksjonene, men det er en ren matematisk greie og har lite med fysikk å gjøre. Ikke for det, det er en del fysikk og fysiske realiteter som samvarierer med bredden på frekvensspekteret. Men det finnes andre metoder å få fram dette poenget på uten at det impliserer at det er noe fysisk til stede den tiden kraften faktisk er lik null. Vi skal i et senere kapittel ta for oss såkalt wavelet-transformasjon, og da vil dette komme bedre fram.

[♣ ⇒ I mitt eget forskningsfelt, kvanteoptikk, ser vi hvor uheldig denne type kortslutning er. Noen sier at vi må “bruke mange ulike fotoner” for å skape en lyspuls, og at hvert foton må ha energien $E = hf$ hvor h er Plancks konstant og f frekvensen. Da tillegges det en fysisk virkelighet til hver enkelt fourierkoeffisient, mens det burde vært mer fokus på hva som er fysikk og hva som er matematikk. ← ♣]

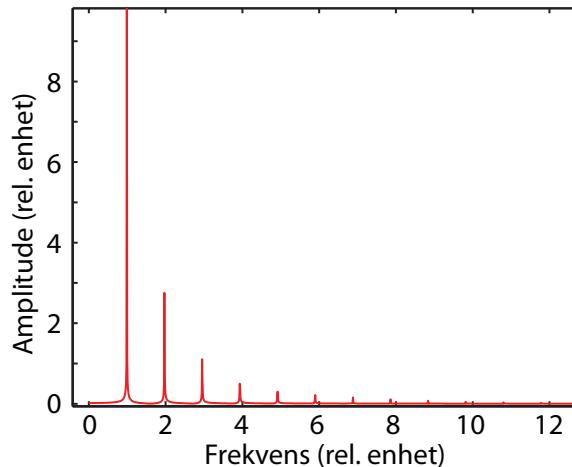
Et vesentlig poeng her er at all tidsinformasjon om et signal forsvinner så snart vi tar absoluttverdien av fourierkoeffisientene. Så lenge vi beholder komplekse fourierkoeffisienter, er tidsinformasjonen intakt, men den er ofte svært godt gjemt. Tidsinformasjonen er nemlig spredt ut over hele fourierspekteret. Det er bare en full tilbaketransformasjon fra frekvensbildet til tidsbildet som gir oss tidsinformasjonen tilbake. Fouriertransformasjon har derfor begrenset verdi for signaler som er null i enkelte tidsperioder eller som på annet vis helt endrer karakter i løpet av samplingstiden.

Også i en annen sammenheng er kan en fourieranalyse føre til uheldige konklusjoner. Figur 5.18 viser en fouriertransformasjon for en periodisk bevegelse. På sett og vis ligner denne figuren på figur 5.16 som viste frekvensspekteret av lyd fra en tverrfløyte. I den anledning omtalte vi de høyere harmoniske ved å si at “når vi spiller fløyte, vil lufta ikke bare svinge ved én bestemt frekvens, men ved flere frekvenser samtidig”. En slik formulerig er vanlig, men ikke uproblematisk.

Sier vi at “flere frekvenser finnes samtidig” i bevegelsen som ligger bak fourierspekteret i figur 5.18, passer utsagnet dårlig i forhold til fysikken som ligger bak! Figuren ble nemlig laget slik: Vi beregnet først en planets bane rundt Sola. Banen ble beskrevet ved et sett koordinater som funksjon av tid $\{x_i(t), y_i(t)\}$. Figur 5.18 er da rett og slett den fouriertransformerte av $\{x_i(t)\}$ for en tid som er mange runder rundt sola for den aktuelle planeten.

Grunnen til at vi får en rekke “harmoniske” i dette tilfellet, er at fouriertransformasjon er basert på harmoniske funksjoner, som svarer til sirkelbevegelser. Planetbevegelsen er imidlertid ikke sirkelformet, men elliptisk! Fouriertransformasjonen er rett og slett ikke spesielt egnet for analyser av elliptiske fenomener.

Onde tunger sier at dersom vi hadde hatt datamaskiner på Keplers tid og fouriertransformasjonen var tilgjengelig, ville vi fortsatt operert med middelalderens *episykler* den dag



Figur 5.18: Fouriertransformasjon av en periodisk bevegelse. Se teksten for forklaring.

i dag. For fourieranalysen vår i figur 5.18 viser nettopp at vi kan erstatte ellipsen med en rekke sirkelbevegelser med passe amplitude (og fase). De fleste vil imidlertid være enige i at det er bedre å bruke en beskrivelse av planetbevegelse basert på ellipser og ikke sirkler. Jeg skulle ønske at vi var like åpne for å droppe matematisk formalisme basert på fourieranalyse også i enkelte andre sammenhenger.

Fourieranalyse kan gjennomføres for praktisk talt alle fysiske tidsvariable systemer, siden settet med sinus- og cosinusfunksjoner som inngår i analysen danner et fullstendig sett med funksjoner. Pass på at du *ikke* av dette trekker slutningen at “når noe er mulig, så er det også gunstig”. I kapitlet om wavelet-transformasjon vil vi komme tilbake til denne problemstillingen, siden vi i waveletanalyse kan velge et helt andre basissett av funksjoner enn sinuser og cosinuser.

Kort oppsummert kan vi si:

Fouriertransformasjon er et meget godt hjelpemiddel, men har mer eller mindre samme basis som Middelalderens episykel-beskrivelse av planetbevegelser. Det er fullt mulig å beskrive planetbevegelser ved hjelp av episykler, men det er lite fruktbart å bruke en slik analyse. På tilsvarende vis er det en rekke fysiske fenomener som i dag beskrives ved formalisme basert på fourieranalyse, der denne formalismen egentlig er lite egnet. Den kan føre til fysiske bilder som villeder mer enn de er til hjelp for oss. Et eksempel finnes i kvanteoptikk.

5.10 Fouriertransformasjon, eksempel på et dataprogram

```
% Enkelt eksempelprogram for å vise hvordan fouriertransformasjon
% kan gjennomføres i praksis i Matlab. Eksemplet er en modifikasjon
% av et eksempelprogram på hjelpestidene i Matlab.

Fs = 1000; % Samplingsfrekvens
delta_t = 1/Fs; % Tid mellom hver sampling
N = 1024; % Antall samplinger
t = (0:N-1)*delta_t; % Tidsvektor

% Lager her et kunstig signal som en sum av et 50 Hz sinussignal
% og en 120 Hz cosinus pluss legger til et random signal:
x = 0.7*sin(2*pi*50*t) + cos(2*pi*120*t);
x = x + 1.2*randn(size(t));

plot(Fs*t,x) % Plotting av signalet i tidsbilet
title('Opprinnelig signal')
xlabel('tid (millisekunder)')

X = fft(x,N); % Fouriertransformasjon

frekv = (Fs/2)*linspace(0,1,N/2); % Frekvensvektor (for plot)

% Plotter bare lengden på frekvenskomponentene i frekvensspekteret.
% Velger å bare ta med frekvenser opp til halve samplingsfrekvensen.
figure; % Hindrer overskriving av forrige figur
plot(frekv,2*abs(X(1:N/2))) % Plotter halvparten av fourierspekteret
title('Absolutt-verdier av frekvensspekteret')
xlabel('Frekvens (Hz)')
ylabel('|X(frekv)|')
```

5.11 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du vite at:

- En integrerbar tidsavhengig kontinuerlig funksjon kan transformeres ved kontinuerlig fouriertransformasjon til et “frekvensbilde”, som siden kan entydig transformeres med en invers fouriertransformasjon tilbake til utgangspunktet.
- En diskret funksjon kan transformeres ved en diskret fouriertransformasjon til et “frekvensbilde”, som siden kan entydig transformeres med en diskret invers fouriertransformasjon tilbake til utgangspunktet.
- Frekvensbildet i en diskret fouriertransformasjon består av komplekse tall hvor realdelen representerer cosinus-bidrag ved de ulike frekvensene, mens imaginærdelen representerer sinus-bidragene. “Pythagoras” gir amplituden på signalet ved den aktuelle frekvensen. Arkustangens av forholdet mellom imaginær og realdel angir fasen til den aktuelle frekvenskomponenten (relativt til en $\cos(\omega t + \phi)$ -beskrivelse).
- For et reelt signal, er siste halvpart av fourierkoeffisientene kompleks konjugerte av den første halvparten, og “speiling” forekommer. Vi tar derfor oftest bare vare på den første halvparten av frekvensspekteret.
- Dersom et signal “samples” med en samplingsfrekvens f_s , vil vi bare kunne behandle signaler med frekvenser under halve samplingsfrekvensen på en entydig måte.
- For å unngå problemer med “folding”, må derfor et lavpassfilter benyttes for å fjerne signalkomponenter som kan ha høyere frekvens enn halve samplingsfrekvensen. Ved numeriske beregninger må vi selv passe på at “samplingsfrekvensen” er høy nok for signalet vi behandler.
- Fouriertransformasjon er et ypperlig hjelpemiddel ved studier av stasjonære tidsvariable fenomener i store deler av fysikken. Eksempelvis er fouriertransformasjon i utstrakt bruk ved analyse og syntese av lyd.
- Fouriertransformasjon egner seg (nesten) bare for å analysere signaler som har samme eller mindre samme karakter i hele samplingstiden. For transiente signaler som endrer karakter kraftig i løpet av samplingstiden, kan et fourierspekter være mer villedende enn til hjelp.
- Fouriertransformasjonen er mulig å gjennomføre (nesten) alltid, men egner seg dårlig for en del tidsforløp.

5.12 Oppgaver

Forståelses- / diskusjonsspørsmål

- Hvordan kan vi ved å ta utgangspunkt i et frekvensspekter lage en syntetisk lyd?
Vil en slik lyd lyde som et ordentlig instrument?
- For CD lyd er samplingsfrekvensen 44.1 kHz. Ved lydinnspilling må vi ha et lavpass-filter mellom mikrofonforsterker og samplingskretsene som fjerner alle frekvenser over ca 22 kHz. Hva ville kunne skje med lyden ved avspilling dersom vi ikke tok denne regelen høytidelig?
- Etter en fouriertransformasjon (FFT) bruker vi ofte bare å plotte en del av alle dataene som produseres. Nevn eksempler på hva som kan påvirke vårt valg.
- Dersom du fourieranalyserer lyd fra en CD-innspilling av et instrument og finner at grunntonen har frekvensen 440 Hz. Hvor finner du da den foldete frekvensen?
- Finn ut hvor mange samplinger det er hver periode i hver linje i figur 5.3 (angi dette som et desimaltall). Har tallene du kommer fram til relasjon til Nyquist-Shannons samplingsteorem? Påpek hvorfor vi får "speiling" omkring halve samplingsfrekvensen.

Regneoppgaver

- Vis at første punkt i en digital fouriertransform av et signal er lik halvparten av gjennomsnittsverdien til signalet vi startet ut med.
- Noen hevder at månefasene virker inn på alt fra vær til humøret til oss mennesker. Sjekk om du kan finne holdepunkter for at temperaturen (maksimums- og/eller minimumstemperaturen hvert døgn) varierer litt med månefasene (i tillegg til all annen variasjon).

Data kan du hente ned fra api.met.no for det stedet du er interessert i og for det tidsrommet du er interessert i. Alternativt kan du bruke en allerede nedhentet og litt forenklet fil *oslotemp.txt* på kurswebsidene. Filen gir temperaturen på Blindern i tiden 1. januar 2003 til og med 31. desember 2012. Fjerde kollonne i filen gir minimumstemperaturene, mens femte kollonne gir max-verdiene.

Forklar nøyne hvordan du kan trekke en sluttning om hvorvidt månefasen virker inn på temperaturen eller ikke.

Nedenfor er noen linjer Matlab-kode som viser hvordan data kan leses inn fra vår fil til et Matlabprogram (filen har fem kolonner):

```
filnavn = 'tempBlindern10aar.txt';
fileID = fopen(filnavn, 'r');
A = fscanf(fileID, '%d %d %f %f',[5,inf]);
minT = A(4,:);
maxT = A(5,:);
plot(minT,'-r');
hold on;
plot(maxT,'-b');
```

- Hent opp solflekk-data fra web og lag en figur lignende vår figur 5.1. Vær spesielt oppmerksom på å få korrekte verdier langs aksene i fourierspekteret. Er det samsvar mellom høydene i toppene i tidsbildet og amplitudene i frekvensspekteret? Nedenfor er noen linjer Matlab-kode som viser hvordan data kan leses inn i et Matlabprogram (to kolonner):

```

filnavn='soldata.txt';
fileID = fopen(filnavn, 'r');
A = fscanf(fileID,'%f %f', [2,inf]);
plot(A(1,:),A(2,:),'-b');

```

9. Anta at du skal lage et frekvensspekter liknende det i figur 5.16 for en lydsnutt du henter fra en CD. Samplingsfrekvensen er 44.1 kHz. Du lagrer $2^{14} = 16384$ datapunkter og foretar en “fast fourier transform” og ender opp med 16384 nye datapunkter som representerer frekvensspekteret. Hvordan skal du i programmet ditt gjøre om fra punkt-nummer til frekvens langs x-aksen når frekvensspekteret skal plottes?
10. Hva blir oppløsningen langs x-aksen i plottet i forgående oppgave? Sagt med andre ord: Hvor stor endring i frekvens får vi ved å gå fra ett punkt i frekvensspekteret til det neste? Ville oppløsningen vært den samme selv om vi bare brukte 1024 punkter som utgangspunkt for fouriertransformasjonen?
11. Skriv et fouriertransformasjons-program i Python eller Matlab (eller hvilket som helst programmeringsspråk) og sjekk at et harmonisk signal med eksakt 13 perioder innenfor 512 punkter gir et frekvensspekter som du forventer. La gjerne signalet være en rent sinussignal eller et kombinert sinus- og cosinussignal.
12. Modifiser programmet såvidt slik at signalet nå får 13.2 perioder innenfor de 512 punktene. Hvordan ser frekvensspekteret ut nå? Beskriv så godt du kan!
13. Modifiser programmet slik at du får 16 hele perioder med *FIRKANTsignal* innenfor 1024 punkter. Hvordan ser frekvensspekteret ut nå? Finn på internett et uttrykk for hvordan amplituden for ulike frekvenskomponenter skal være for et firkantsignal, og verifiser at du får omtrent det samme fra dine numeriske beregninger.
14. Modifiser programmet slik at du får 16 hele *sagtenner* (trekantsignal) innenfor de 1024 punktene. Beskriv også dette frekvensspekteret!
15. I et eksempel i kapittel 3 beregnet vi vinkelutslaget til en fysisk pendel ved store utslag. Gjennomfør disse beregningene for 3-4 ulike vinkelutslag og foreta en fourieranalyse av bevegelsen i hvert tilfelle. Kommenter resultatene.
16. Lag et dataprogram hvor du genererer et signal med lengde 1024 punkter. Anta at samplingsfrekvensen er 1 kHz, og beregn det samplede signalet for signalfrekvensen 200 Hz. Fouriertransformer signalet, og kontrollér at frekvensspekteret er slik du forventer. Gjenta det samme f.eks. for følgende signalfrekvenser: 400 Hz, 750 Hz, 1020 Hz, 1400 Hz, 1800 Hz og endelig 2100 Hz. Forsøk ut fra resultatene å finne ut hvordan speiling/folding arter seg når signalfrekvensen blir til dels betydelig høyere enn halve samplingsfrekvensen.
17. AM-radio (AM: Amplitude-Modulert). Beregn hvordan signalet som sendes fra en AM-sender ser ut, og finn frekvensspekteret av signalet. Det er enklest å gjøre dette for et radiosignal på langbølge-båndet (153 - 279 kHz). La bærebølgen ha frekvensen $f_b = 200$ kHz, og velg at talesignalet er en enkel sinus med frekvens (etter tur) $f_t = 440$ Hz og 4400 Hz. Signalet bør samples med en samplingsfrekvens $f_s = 3.2$ MHz og det kan være passe å bruke $N = 2^{16} = 65536$ punkter. AM-signalet er gitt ved:

$$f(t) = (1 + A \sin(2\pi f_s t)) \cdot \sin(2\pi f_b t)$$

hvor A angir normert amplitude på lydsignalet (for den aller sterkeste lyden som kan sendes uten forvrengning er $A = 1.0$. Bruk en litt mindre verdi, men test gjerne ut hvordan signalet påvirkes av A .)

Plott AM-signalet både i tidsdomenet og frekvensdomenet. Velg ut passe utsnitt i forhold til det fulle datasettet for å få fram det du ønsker å vise. Husk å sette på korrekte tidsangivelser langs x-aksen i tidsdomenet og korrekte frekvensangivelser langs x-aksen i frekvensdomenet.

Hver radiostasjon på mellombølge og langbølge får bare strekke seg ut over et frekvensbånd på 9 kHz totalt. Hvilke følger har dette for kvaliteten på lyden som overføres?

18. FM-radio (FM: Frekvens-Modulert). Beregn hvordan signalet som sendes fra en FM-sender ser ut, og finn frekvensspekteret av signalet. Bruk samme parametre som i forrige oppgave (selv om det ikke i praksis brukes FM på langbølge). FM-signalet kan gis på følgende måte:

```
f(t) = sin(fase(t)); % Prinsipielt, implementeres litt annerledes i Matlab
```

hvor fasen integreres opp i en løkke på følgende måte:

```
fase(1) = 0.0;
for i=1:(N-1)
    fase(i+1)=fase(i) + omega_b*delta_t*(1.0 + A*sin(omega_t*t(i)));
end;
```

hvor "omega_b" og "omega_t" er vinkelfrekvensen for bærebølgen og talesignalet henholdsvis. Tidsstrenge "t(i)" antas å være beregnet på forhånd (avstand mellom punktene er "delta_t", som bestemmes av samplingsfrekvensen).

A er igjen en normert amplitude for lydsignalet hvor også en såkalt modulasjonsgrad inngår. Du kan velge etter tur f.eks. $A = 0.2$ og 0.7 og se hvordan dette påvirker både tidsbildet og frekvensbildet.

Plott FM-signalet både i tidsdomenet og frekvensdomenet etter samme retningslinjer som i forrige oppgave. (Hint: Det kan være enklest å plotte tilfellet hvor talefrekvensen er 4400 Hz og at $A = 0.7$.)

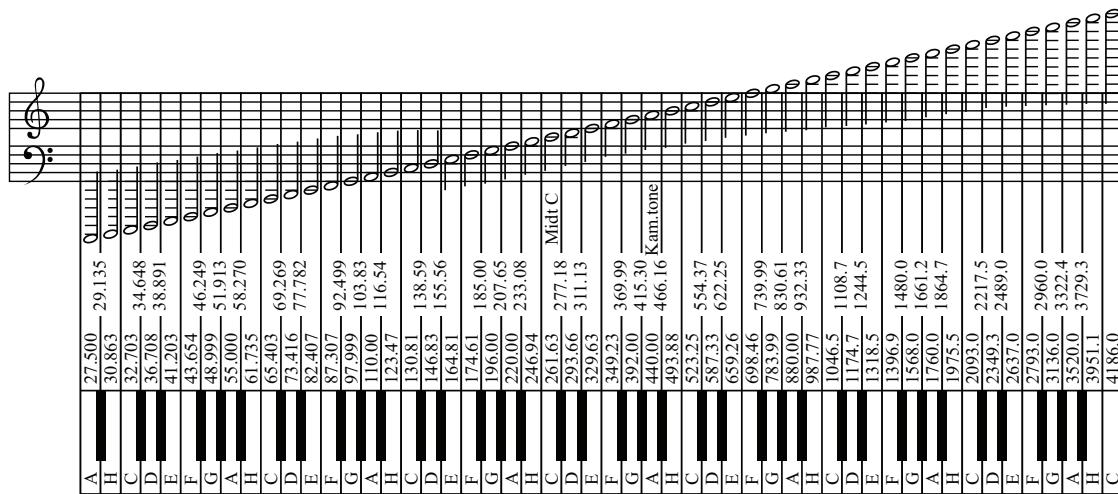
Er det noen klare forskjeller i hvordan frekvensbildet fremstår for FM-signaler sammenlignet med AM-signaler?

19. Bruk invers fouriertransformasjon for å generere en enkel sinus, og spill av lyden på datamaskinen. Helt konkret anbefales følgende: Bruk CD-samplingsfrekvensen $F_s = 44100$ Hz og $2^{16} = 65536$ punkter. Verdiene av signalet f bør ligge strengt innenfor intervallet $[-1,+1]$. Bruk den innebygde funksjonen *wavplay(f, Fs)* i Windows eller *sound(f, Fs)* i Linux. Forsøk å lage lyd med frekvensene 100 Hz, 440 Hz, 1000 Hz og 3000 Hz. Du kan gjerne lage et signal som består av flere samtidige sinuser også? Husk å skalere totalsignalet før vi bruker wavplay eller sound.
20. Les en lydfilen "transient.wav" og foreta fouriertransformasjon for å få fram frekvensspekteret. Lydfilen er tilgjengelig fra kursets websider, samplingsfrekvensen er $F_s = 44100$ Hz. Bruk gjerne 2^{17} punkter i analysen. En aktuell Matlabfunksjon er

```
s = 'lydfil1.wav';
[f,Fs,type] = wavread(s, [nstart nslutt]);
g = f(:,1); % Henter ut ett monosignal fra stereosignalet f
```

Dersom du lytter til lyden og dernest betrakter frekvensbildet, håper jeg at du reflekterer over det du har gjort. Fourieranalyse blir iblant misbrukt. Hva er problemet med analysen som er foretatt for det aktuelle lydsignalet?

21. Foreta frekvensanalyse for lyd fra en tuba og fra en piccoloflöyte (lydfiler tilgjengelig fra kursets websider). Bruk 2^{16} punkter i analysen (se forøvrig forrige oppgave). Bestem tonens plassering i en temperert skala ved å bruke figur 5.19. Frekvensspekteret viser forøvrig overtoner slik det er beskrevet i dette kapitlet (vi kommer tilbake til dette i senere kapitler).



Figur 5.19: *Toneskalaen for en temperert skala slik vi finner den på et piano. Frekvenser for tonene er gitt.*

22. “Åpen oppgave” (Det vil si at svært få føringer og tips er gitt): Fouriertransformasjon kan brukes i digital filtrering. Forklar prinsippet og hvordan dette kan gjennomføres i praksis. Lag et lite program som foretar selvvalgt digital filtrering av en virkelig lydfil, hvor det er mulig å lytte til lyden både før etter filtrering. (Vær litt omhyggelig i beskrivelsen av detaljer i det du gjør!)

Kapittel 6

Bølger



Det finnes en mengde ulike former for bølger, og de er til dels svært forskjellige. Likevel har de noe til felles.

I dette kapitlet vil vi først se hvordan en bølge vanligvis blir beskrevet matematisk. Det er til dels store likheter med formalismen vi brukte da vi beskrev svingninger. Likevel er det vesentlige forskjeller, ikke minst fordi "randbetingelser" spiller en sentral rolle for bølger.

I siste del av kapitlet starter vi ut med Newtons annen lov og utleder bølgeligningen for bølgebevegelse langs en streng. På tilsvarende måte utledes bølgeligningen for lydbølger.

Det er nesten en form for magi at en høyttaler kan produsere lokale variasjoner i lydtrykket, og at disse små variasjonene kan forplante seg kilometervis uten at et eneste luftmolekyl flytter seg mer enn noen mikrometre på grunn av lydbølgen. Vi vil i dette og det neste kapitlet forsøke å trenge gjennom mystikken og få en bedre forståelse av hva som driver bølgen framover.

Og skulle det være noen som ikke har sett en høyttaler i funksjon på nært hold, kan vi fortelle at høyttalermembranen (innenfor de bølgete områdene på fotografiet) skyves fram og tilbake og dyster på luftmolekyler når den genererer lyd. Svingninger i membranen gir lydbølger!

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 07022013.

6.1 Innledning

Alle har sett bølgeringer som brer seg ut over en vannoverflate (se figur 6.1). Vi er så vant til dette at vi knapt registererer det.

Men har du egentlig forstått det magiske med bølger? For hvordan kan det ha seg at bølgen vandrer bortover vannoverflaten uten at det er noe materie som forflytter seg med bølgehastigheten? Kaster vi en ball fra punkt A til B, forflytter ballen seg romlig med hele sin masse fra A til B. Når en bølge flytter seg fra A til B, er det imidlertid ingen tilsvarende masse som har forflyttet seg fra A til B. Hva i all verden er det da som får bølgen til å forplante seg bortover?



Figur 6.1: *Bølger som ringer på vann.*

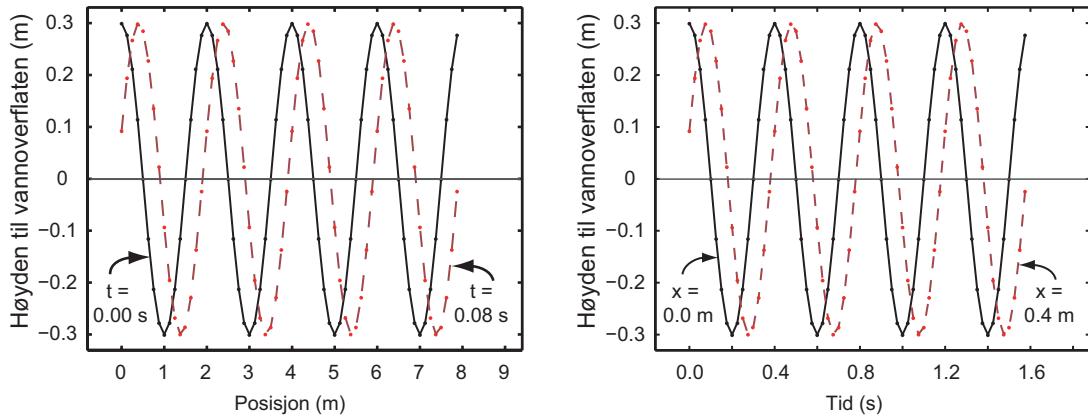
Bølger får vi når en svingning ett sted i rommet på et eller annet vis påvirker naboområdet slik at også det begynner å svinge, som i sin tur igjen fører til at nok et naboområde begynner å svinge osv. Når vi beskriver dette samspillet og fokuserer på forklaringen av fysikken som ligger bak bølgebevegelsen, arbeider vi med dynamikken for systemet. Vi starter likevel på samme måte som i kapittel 1 med “kinematikken”, det vil si den matematiske beskrivelsen.

En bølge kan vi anskueliggjøre på tre måter:

- Vi kan ta et øyeblikksbilde (“blitzbilde”) av hvordan bølgen ser ut ved ett valgt tidspunkt i ulike deler av rommet (som funksjon av posisjon).
- Vi kan registrere utslaget som funksjon av tid på *ett* sted i rommet idet bølgen passerer dette stedet, og plotte resultatet.
- Vi kan bruke en “film” (animasjon) som viser hvordan bølgen brer seg i rommet etter som tiden går

Figur 6.2 viser eksempler på de to første anskuelsesformene. Tenk deg at du står på en brygge og ser på dogne bølger som ruller foran deg. Du kan ta et bilde av bølgene, og får noe som svarer til venstre del av figur 6.2. Tar du et bilde litt senere, har bølgen flyttet litt på seg i mellomtiden (som antydet i figuren).

Tenk deg at det står en vertikal påle i vannet. Vannoverflaten vipper da opp og ned langs stolpen, og du kan registrere høyden som funksjon av tid. Det svarer til den høyre



Figur 6.2: En bølge kan angis som funksjon av posisjon ved en bestemt tid, eller som funksjon av tid for en bestemt posisjon. Se teksten for detaljer.

del av figur 6.2. Dersom det er to påler som står et lite stykke fra hverandre, vil ikke vannoverflaten være på topp samtidig på begge pålene, generelt sett.

For en *harmonisk* bølge (form som en sinus- eller cosinus-funksjon) vil de to førsste anskuelsesformene begge se ut som harmonisk svingning, den første som harmonisk svingning som funksjon av posisjon, den andre som harmonisk svingning som funksjon av tid. Vi vet fra tidligere at en harmonisk svingning er en løsning av en annenordens differensielligning. Dersom vi betrakter hvordan bølgen ser ut som funksjon av posisjon (ved ett tidspunkt), må utslaget f være en løsning av differensielligningen:

$$\frac{d^2 f}{dx^2} = -C_x f$$

Dersom vi betrakter bølgen som funksjon av tid ettersom den passerer ett sted i rommet, må utslaget være en løsning av differensielligningen:

$$\frac{d^2 f}{dt^2} = -C_t f$$

I disse ligningene indikerer x posisjon og t tid, og C_x og C_t er positive reelle konstanter som er ulike i de to tilfellene. Utslaget f derimot er angitt med akkurat samme symbol i begge ligningene siden det er det samme utslaget vi betrakter i begge anskuelsesformene. Utslaget kan f.eks. være lufttrykk ved lydbølger, eller elektrisk feltstyrke ved elektromagnetiske bølger eller meter for overflatebølger på havet. Da innser vi at utslaget er det samme uansett om vi betrakter bølgen som funksjon av posisjon i rommet eller som funksjon av tiden. Vi kan da kombinere de to ligningene og får:

$$\frac{d^2 f(x, t)}{dt^2} = \frac{C_t}{C_x} \frac{d^2 f(x, t)}{dx^2}$$

I dette uttrykket har vi også angitt at utslaget både avhenger av rom og tid. Og når en funksjon avhenger av flere uavhengige parametere samtidig, bruker vi *partiell derivasjon* og skriver:

$$\frac{\partial^2 f(x, t)}{\partial t^2} = \frac{C_t}{C_x} \frac{\partial^2 f(x, t)}{\partial x^2} \quad (6.1)$$

Foretar vi en omdøping: $C_t/C_x \rightarrow v^2$, får ligningen formen:

$$\frac{\partial^2 f(x, t)}{\partial t^2} = v^2 \frac{\partial^2 f(x, t)}{\partial x^2} \quad (6.2)$$

Denne ligningen kalles *bølgeligningen*.

Siden C -ene var positive, reelle konstanter, kan også v være positiv og reell.

♣ ⇒ Kommentar: Vi tar et kort sidesprang for å friske opp hva vi mener med partiell derivasjon.

Anta at vi har en funksjon $h = h(kx - \omega t)$ og at vi skal finne den partielt deriverte av denne funksjonen mhp x . Vi definerer en ny variabel $u = kx - \omega t$ og bruker kjerneregelen og finner:

$$\frac{\partial h}{\partial x} = \frac{dh(u)}{du} \cdot \frac{\partial u}{\partial x}$$

Det er først i det siste leddet vi for ordentlig får fram hva partiell derivasjon innebærer. Vi har:

$$\frac{\partial u}{\partial x} = \frac{\partial(kx - \omega t)}{\partial x}$$

Både x og t er variable, men når vi skal beregne den partielt deriverte mhp x , skal vi anse t som en konstant! Følgelig får vi:

$$\frac{\partial(kx - \omega t)}{\partial x} = k$$

På liknende måte kan vi gå fram for å finne partiell derivert for t . Da anses variablene x som konstant.

Partiell deriverte representerer derfor den deriverte av funksjonen under forutsetning at alle variable holdes konstant, bortsett fra den ene som vi skal beregne den partielt deriverte med hensyn på. ← ♣]

Ligningen vi kom fram til i ligning (6.1) er en annenordens partiell differensielligning, og den kalles for “bølgeligningen”. Denne ligningen vil vi stifte bekjentskap med ganske mange ganger i boka, så det kan være nyttig å forsøke å skjonne den ordentlig så raskt som mulig.

Da vi drøftet svingninger i kapittel 1, så vi at dersom vi kjenner startposisjonen og startfarten f.eks. til en pendel, kan vi beregne entydig hvordan svingningen blir i all fremtid (så sant selve differensielligningen for bevegelsen er kjent).

For bølger er det totalt annerledes. Selv om vi har nøyaktig samme bølgeligning, og har samme initialbetingelser, så er det uendelig mange ulike løsninger. Grunnen er at bølgen brer seg i rommet, og formen på rommet vil påvirke bølgen selv om den grunnleggende differensielligningen er den samme. Det er lett å forstå dersom vi tenker på dønninger på havet som kommer mot land. Bølgen lokalt vil variere kolossalt alt etter hvordan kysten ser ut lokalt med steiner, nes og viker. Løsning av bølgeligningen krever derfor at vi kjenner både initialbetingelser og *randbetingelser*. Og siden det finnes uendelig mange randbetingelser vi kan tenke oss, vil det også finnes uendelig mange løsninger. Men når vi først har gitt både initialbetingelser og fullstendig sett med randbetingelser, finnes det bare én løsning.

Siden det er så utrolig stor variasjonsmulighet for bølger, må vi ofte ty til forenklede løsninger for å i det minste få fram noen typiske trekk. Noen slike løsninger er faktisk en brukbar tilnærming til virkelige bølger i spesielle tilfeller. Den mest vanlige forenklede løsningen kalles for *plan bølge* og vi skal se litt nærmere på den nå.

6.2 Plan bølge

En plan bølge er karakterisert ved at utslaget er identisk i et helt plan normalt på retningen i rommet hvor bølgen brer seg. Dersom bølgen overalt i det tredimensjonale rommet brer seg i en retning parallelt med x-aksen, vil en plan bølge svare til at utslaget i bølgen ved en vilkårlig valgt tid, er identisk overalt i et uendelig plan vinkelrett på x-aksen.

For en plan lydbølge som beveger seg i x-retning vil dette i praksis si at for et hvilket som helst tidspunkt er det slik at det lokale lufttrykket har et maksimum overalt langs et plan vinkelrett på x-aksen. Vi kaller et slikt plan for en “bølgefront”. For plane bølger er bølgefrontene plane.

En plan harmonisk (monokromatisk) bølge kan f.eks. beskrives matematisk slik:

$$f(x, t) = A \cos(kx - \omega t) \quad (6.3)$$

I denne sammenheng kalles k for *bølgetallet* og ω for vinkelfrekvensen. Holder vi tiden konstant, f.eks. ved $t = 0$, og starter i $x = 0$, forflytter vi oss en bølgelengde når $kx = 2\pi$. Bølgelengden λ er derfor nettopp lik denne x -verdien, altså:

$$\lambda = \frac{2\pi}{k}$$

På tilsvarende måte kan vi holde posisjonen konstant, f.eks. ved å sette $x = 0$, og starte ved $t = 0$. Da ser vi at dersom vi skal endre tidsfunksjonen med en periode, må tiden øke inntil $\omega t = 2\pi$. Tidsforskjellen kaller vi *periodetiden* T og får:

$$T = \frac{2\pi}{\omega}$$

Det kan legges til at ordet “bølgetall” kommer av at k angir antall bølgelengder innenfor måleenheten vi bruker (“hvor mange bølgetopper det er i en meter”), men multiplisert med 2π .

Vi kan også anvende en liknende tankemåte for vinkelfrekvensen. I så fall kan vi si at vinkelfrekvensen er å betrakte som “(tids)periodetallet” som mäter hvor mange periodetider vi har innenfor den måleenheten vi bruker for tid (“hvor mange perioder vi har i svingningen i løpet av ett sekund”), men multiplisert med 2π .

Måleenhet for bølgetallet er inverse meter, dvs m^{-1} . Enhet for vinkelfrekvens er egentlig inverse sekund, dvs s^{-1} , men for å redusere faren for forveksling med frekvens, bruker vi ofte å angi vinkelfrekvenser i *radianer per sekund*.

6.2.1 Bølgens hastighet

La oss finne ut hvor fort bølgen vandrer i x-retningen. Tenk deg at du følger en topp som f.eks. svarer til at argumentet i cosinusfunksjonen er 6π . I så fall vil

$$kx - \omega t = 6\pi$$

$$x = \frac{\omega}{k}t + \frac{6\pi}{k}$$

Vi deriverer uttrykket for posisjonen mhp tiden for å se hvor raskt dette punktet forflytter seg, og får

$$\frac{dx}{dt} \equiv v = \frac{\omega}{k}$$

Hastigheten bølgen går med er altså lik forholdstallet mellom vinkelfrekvens og bølgelengde. Vi kan gjøre dette om litt ved å innføre bølgelengde og periodetid i stedet, og får:

$$v = \frac{2\pi/T}{2\pi/\lambda} = \frac{\lambda}{T}$$

Men vi vet at frekvensen er gitt som inversverdien til periodetiden, dvs $\nu = 1/T$. Setter vi inn dette, får vi en velkjent relasjon:

$$v = \lambda\nu \quad (6.4)$$

Hastigheten for en bølge som kan beskrives på den enkle formen gitt i ligning (6.3) er altså bølgelengden multiplisert med frekvensen.

6.2.2 Løsning av bølgeligningen?

Foreløpig har vi bare *påstått* at ligning (6.3) tilfredsstiller bølgeligningen. Vi vil nå sjekke dette, og får ved å dobbeltderivere ligning (6.3):

$$\frac{\partial^2 f(x, t)}{\partial t^2} = -\omega^2 f(x, t)$$

og

$$\frac{\partial^2 f(x, t)}{\partial x^2} = -k^2 f(x, t)$$

Vi ser da at:

$$\frac{\partial^2 f(x, t)}{\partial t^2} = \frac{\omega^2}{k^2} \frac{\partial^2 f(x, t)}{\partial x^2}$$

eller:

$$\frac{\partial^2 f(x, t)}{\partial t^2} = v^2 \frac{\partial^2 f(x, t)}{\partial x^2} \quad (6.5)$$

Vi ser altså at den plane bølgen gitt i ligning (6.3) tilfredsstiller bølgeligningen, men hva med initialbetingelser og grensebetingelser? Vel, her er det mer problematisk. Dersom en plan bølge skal kunne danne seg og holde seg slik, må vi initiere en bølge som faktisk har uendelig utstrekning og samme amplitude og startvariasjon i tid i hele dette uendelige planet. Det må heller ikke være noen grensebetingelser som påvirker bølgen i noe punkt. Dersom alle disse kravene var oppfylt, ville den plane bølgen forbli plan videre, men vi innser at dette er fysisk urealiserbart.

Dersom vi derimot starter med å betrakte en bølge mange, mange bølgelengder unna det stedet den ble generert, f.eks. lys fra Sola når lyset når Jorda, vil den såkalte "bølgefronten" være temmelig plan så lenge vi bare betrakter lyset over f.eks. en tenkt 1 x 1 m stor flate på tvers av lysretningen. Dersom vi da følger lyset noen få meter videre, vil bølgen oppføre seg omtrent som en plan bølge i dette begrensede volumet. Men dersom reflektert lys når inn i dette volumet, har vi ingen plan bølge lenger!

Plane bølger er derfor bare en idealisering som vi aldri kan oppnå i praksis. Planbølgebeskrivelsen kan likevel gi en relativt god beskrivelse over et begrenset volum når vi er langt unna ting og tang som kan påvirke bølgen på et eller annet vis.

Med "langt unna" menes at avstanden er stor relativt til bølgelengden, fra kilden til bølgene og til randbetingelser som forstyrrer bølgen.

6.2.3 Hvilken vei?

Vi fant ovenfor at en plan bølge beskrevet med ligningen:

$$f(x, t) = A \cos(kx - \omega t)$$

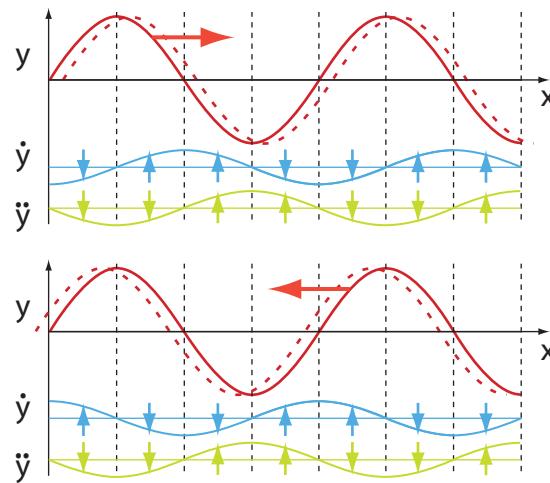
hadde en hastighet $v = +\omega/k$. Det vil si at bølgjen forplanter seg i positiv x-retning etter som tiden går. Det kan vi med litt øving lese direkte ut av argumentet til cosinusfunksjonen: Dersom vi skal holde oss på samme sted i en bølge (f.eks. en topp), må argumentet forbli uforandret etter som tiden går. Og øker tiden t , kan vi bare oppnå at argumentet beholder samme verdi dersom vi kompenserer med også å la x-verdien øke. Med andre ord, bølgens topp forflytter seg mot høyere x-verdier når tiden øker.

Bruker vi tilsvarende argumentasjon, kan vi lett vise at en bølge beskrevet ved:

$$f(x, t) = A \cos(kx + \omega t)$$

forplanter seg mot lavere x-verdier når tiden øker. Bildelig, for de av oss som er vant til at x-aksen øker mot høyre, kan vi si at bølgjer beskrevet på den første av disse måtene (med minus) beveger seg til høyre og bølgjer beskrevet på den andre måten (med pluss) beveger seg mot venstre.

Merk at hastigheten til bølgjen ikke beskriver hastighet på samme måte som hastigheten til en ball under kast. Hastigheten til ballen er definert som den tidsderiverte av posisjonen til ballen, altså til et fysisk legeme. For bølgjen er hastigheten definert som en mer *abstrakt størrelse*, nemlig f.eks. den tidsderiverte til posisjonen i rommet hvor bølgjen har sin maksimale verdi. For en lydbølge i luft, er hastigheten til bølgjen lik hastigheten f.eks. til et et sted i rommet hvor det lokale lufttrykket har et maksimum, det vil si hastigheten til en bølgefront. Vi kommer tilbake til mer kompliserte forhold siden.



Figur 6.3: Øyeblikksbilde av "utslag" (y i rødt), den tidsderiverte til utslaget (\dot{y} i blått), og den dobbeltderiverte til utslaget (\ddot{y} i grønt) i ulike posisjoner langs en bølge. Bølgene som sådan beveger seg til høyre (øverst) og til venstre (nederst). Stiplet posisjonskurve viser hvor bølgjen er en kort tid etter det stedet bølgjen er nå (heltrukket).

Figur 6.3 viser et øyeblikksbilde av "utslag", den tidsderiverte til utslaget, og den dobbeltderiverte til utslaget i alle posisjoner langs en bølge. Bølgene som sådan går mot høyre eller mot venstre slik pilene viser (øverst). Legg merke til at for en bølge som går

mot høyre vil den tidsderiverte av utslaget ligge en kvart periodetid “foran” “utslaget”, og den annenderiverte mph tid ytterligere en kvart periodetid “foran” den enkle tidsderiverte. For en bølge som går mot venstre gjelder egentlig akkurat det samme, men “foran” bølgen må nå forstås ut fra at bølgen beveger seg mot venstre.

[♣ ⇒ La oss forsøke å konkretisere disse betraktningene, men velger en bølge på en streng (f.eks. den vi får like etter vi har svinget den ene enden av en lang horisontal streng opp og ned noen ganger). “Utslaget” er i dette tilfellet meget konkret, fordi det jo rett og slett angir posisjonen til strengen på det stedet utslaget vurderes. Den tidsderiverte av utslaget vil da si vertikalhastigheten til det punktet langs strengen vi betrakter, og den dobbelt tidsderiverte for dette punktet vil da være den vertikale akselerasjonen til dette punktet. Selve bølgen beveger seg i horisontal retning.]

Når vi betrakter figur 6.3 for dette tilfellet, er det relativt enkelt å se og forstå hvilken retning den vertikale *hastigheten* til et punkt langs strengen har for ulike utslag langs strengen. Det er vanskeligere å gjennomskue at den vertikale akselerasjonen går som den gjør. Husk at akselerasjon er proporsjonal med *endring av hastighet* i en kort tid. Når en bit av strengen (et punkt på strengen) beveger seg innover mot likevektspunktet, er den akselerert innover mot likevektspunktet. Farten øker i denne perioden. Men når en bit av strengen har passert likevektspunktet og er på vei utover mot maksimalt utslag, bremses farten opp. Kreftene som virker, virker i retning innover mot likevektspunktet igjen. Forsøker du å tenke langs disse banene, er det ikke så vanskelig å forstå at fortegnet på akselerasjonen (for enkeltpunkter) er som den er.

Legg merke til at for et vilkårlig punkt langs strengen, er fortegnet for akselerasjonen til enhver tid motsatt fortegnet for posisjonen relativt til likevektspunktet. Vi håper du innser at dette er akkurat som det skulle være (ut fra hva vi lærte i kapittel 1). (Hint: Hvilkens bevegelse har egentlig en liten bit av en streng når en harmonisk bølge passerer langs strengen?)

Det bør legges til at svingninger på en streng i vår sammenheng kan betraktes som en plan bølge fordi det er et én-dimensjonalt system slik at “et plan vinkelrett på bølgens retning” i en viss forstand ikke finnes, men dette blir kanskje litt for abstrakt for noen (!?). ← ♣]

6.2.4 Andre bølgeformer

Hittil har vi betraktet harmoniske bølger, dvs bølger med sinusform. Kan også bølger med en annen form tilfredsstille bølgeligningen?

La oss forsøke en bølge beskrevet ved:

$$g(x, t) = G(kx - \omega t)$$

der G kan ha en hvilken som helst form (men G må være en deriverbar funksjon). Vi innfører en ny variabel $u = kx - \omega t$, partiell deriverer, bruker kjerneregelen og får for venstre siden av ligning (6.5):

$$\begin{aligned} \frac{\partial^2 g(x, t)}{\partial t^2} &= \frac{d^2 G(x, t)}{du^2} \left(\frac{\partial u}{\partial t} \right)^2 \\ &= \omega^2 \frac{d^2 G(x, t)}{du^2} \end{aligned}$$

For høyresiden får vi på liknende måte:

$$\frac{\partial^2 g(x, t)}{\partial x^2} = k^2 \frac{d^2 G(x, t)}{du^2}$$

Vi ser da at $g(x, t)$ faktisk tilfredsstiller bølgeligningen, forutsatt at k og ω er virkelige konstanter.

Det vil si at enhver bølge som kan beskrives ved en deriverbar funksjon og ett eneste argument ($kx - \omega t$), hvor k og ω er konstanter, er løsning av bølgeligningen.

6.2.5 Sum av bølger

Hva så dersom vi har en sum av to ulike funksjoner, der den ene har en litt annen kombinasjon av k og ω enn den andre. Sumfunksjonen er da gitt ved:

$$\begin{aligned} g(x, t) &= G_1(k_1 x - \omega_1 t) + G_2(k_2 x - \omega_2 t) \\ &= G_1(u_1) + G_2(u_2) \end{aligned}$$

Partiell derivering mhp tid gir:

$$\frac{\partial^2 g}{\partial t^2} = \omega_1^2 \frac{d^2 G_1(u_1)}{du_1^2} + \omega_2^2 \frac{d^2 G_2(u_2)}{du_2^2}$$

Og partiell derivasjon mhp posisjon gir:

$$\frac{\partial^2 g}{\partial x^2} = k_1^2 \frac{d^2 G_1(u_1)}{du_1^2} + k_2^2 \frac{d^2 G_2(u_2)}{du_2^2}$$

Skal denne sumfunksjonen passe inn i bølgeligningen

$$\frac{\partial^2 g}{\partial t^2} = v^2 \frac{\partial^2 g}{\partial x^2}$$

må vi kreve at derivering av tid skal være lik v^2 multiplisert med den deriverte mhp posisjon. Vi antar at dette kan tilfredsstilles, og får da ved innsetting og ordning av leddene:

$$\begin{aligned} (\omega_1^2 - v^2 k_1^2) \frac{d^2 G_1(u_1)}{du_1^2} \\ = -(\omega_2^2 - v^2 k_2^2) \frac{d^2 G_2(u_2)}{du_2^2} \end{aligned}$$

Siden G_1 og G_2 kan velges fritt, kan ikke denne ligningen tilfredsstilles generelt med mindre

$$(\omega_1^2 - v^2 k_1^2) = (\omega_2^2 - v^2 k_2^2) = 0$$

hvilket vil si at

$$v = \frac{\omega_1}{k_1} = \frac{\omega_2}{k_2}$$

Dette vil si at begge de to delbølgene må gå med nøyaktig samme fart!

Vi har da vist at *summen av to (eller flere) bølger som går med samme fart vil tilfredsstille bølgeligningen såfremt hver av delbølgene gjør det.*

Vi har også vist at *dersom en bølge består av flere komponenter som går med ulik fart, vil vi ikke kunne beskrive hvordan sumbølgen utvikler seg med bare en bølgeligning.*

6.2.6 Bølge beskrevet på kompleks form

Vi kan anvende kompleks beskrivelse for bølger på samme måte som vi gjorde det for svingninger.

En plan harmonisk bølge i x-retning kan på kompleks form skrives:

$$f(x, t) = Ae^{i(kx - \omega t + \phi)} \quad (6.6)$$

Tilsvarende kan vi beskrive en plan harmonisk bølge som beveger seg i en vilkårlig retning $\mathbf{k}/|\mathbf{k}|$ hvor \mathbf{k} er en såkalt bølgevektor, på følgende måte:

$$f(\mathbf{r}, t) = Ae^{i(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi)} \quad (6.7)$$

Siden f normalt skal være reell, må vi enten ta realverdien av uttrykkene ovenfor, eller sikre oss på annet vis. En elegant og vanlig måte å unngå dette problemet på, er å legge til den kompleks konjugerte ("c.c.") av uttrykket og dividere med 2:

$$f(\mathbf{r}, t) = \frac{1}{2}Ae^{i(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi)} + \text{c.c.} \quad (6.8)$$

Denne skrivemåten kan brukes både for reell og kompleks A .

6.3 Transversell og longitudinal

Det er flere ulike typer bølger. En klassifisering er basert på hvilken retning "utslaget" har i forhold til retningen bølgen brer seg. Men siden "utslaget" kan være nærmest hva som helst, og ikke nødvendigvis noe som forflytter seg i rommet, er det ofte misvisende å basere seg på en slik betraktnign. Det er sikkere å gjøre en klassifisering ut fra symmetriegenskaper til bølgen, noe vi forsøker å gjøre i det følgende.

For lydbølger er "utslaget" en trykkforandring. For lydbølger i luft, er det en trykkforandring i luft. Tilsvarende for lyd i andre materialer. Trykkforandringene lokalt oppstår ved at luftmolekyler beveger seg i samme retning (eller motsatt retning) som bølgen brer seg.

Det er lokal rotasjonssymmetri for lufttrykket rundt retningen som angir bølgens bevegelsesretning. Med det menes at det ikke er fysisk mulig lokalt å peke ut en spesiell retning på tvers av lydens bevegelsesretning der fysiske egenskaper knyttet til lydbølgen er forskjellig fra en annen retning på tvers av lydens bevegelsesretning. En slik bølge kalles *longitudinal* (langs-retningen).

Det er imidlertid ikke slik at luftmolekyler flytter seg fra f.eks. en høyttaler til øret mitt når jeg lytter til musikk. I tillegg til Brownske bevegelser svinger hvert enkelt luftmolekyl (statistisk sett) fram og tilbake i forhold til et likevektpunkt. Amplituden i denne svingningen kan gjerne være mindre enn en millimeter (for lyd i metaller en enda mindre lengde).

Hvordan kan en bølge bre seg fra en høyttaler til mitt øre når luftmolekylene underveis rører så lite på seg?

Grunnen er at luftmolekylene ett sted i rommet har en bevegelse som er *tidsforskjøvet* i forhold til luftmolekylene i et nærliggende område. Det er denne tidsforskyvingen (faseforskyvningen) som fører til at vi får en vandrende bølge. Vi skal siden leke oss litt med dette numerisk for å se hvor mange artige bølger vi kan oppnå selv om vi *starter* med en og samme form hver gang, men endrer på den relative bevegelsen i ulike deler av bølgen.

Transversale bølger er den andre hovedtypen bølger. Det mest kjente eksemplet er elektromagnetiske bølger. Da fysikerne på begynnelsen av 1800-tallet innså at lys måtte beskrives ved bølger (og ikke som partikler slik Newton hadde fått fysikere til å tro i over hundre år), hadde de problemer med å forklare polarisasjon. Grunnen er at de gikk ut fra

at lysbølgene var longitudinale slik de mente alle bølger var. Først da Fresnell med flere foreslo at lysbølgene var transversale, ble polarisasjon forstått.

En transversal bølge har et “utslag” vinkelrett på bølgens utbredelsesretning (transversal: “på tvers”). Med det mener vi at *den fysiske parameteren vi kaller “utslaget” ikke har lokal rotasjonssymmetri om aksen som angir bølgens bevegelsesretning*.

For elektromagnetiske bølger er elektrisk og magnetisk felt “utslaget”. Elektrisk og magnetisk felt er vektorer, og har en retning i rommet. At en elektromagnetisk bølge er transversal betyr da at elektrisk og magnetisk felt er rettet i en retning vinkelrett på utbredelsesretningen til bølgen. Da blir rotasjonssymmetrien automatisk brutt.

Merk at det er ingen forflytning av noe som helst materielt på tvers av en elektromagnetisk bølge! Mange forestiller seg at det er “noe” som forflytter seg på tvers av en elektromagnetisk bølge, omtrent som vannspeilet i en overflatebølge på vann. Det er feil. Tegner vi inn elektrisk felt som vektorpiler i punkter langs en linje langs utbredelsesretningen, vil riktig nok pilene skyte ut og trekke seg tilbake. Men vektorpilene er hjelpebidriller for vår tanke og har ingen eksistens i seg selv. Vi vil forsøke å rense opp i denne og flere andre meget vanlige misforståelser når vi omtaler elektromagnetiske bølger i et senere kapittel.

Noen bølger sier vi er en mellomting mellom longitudinelle og transverselle. Overflatebølger på vann er et eksempel. Her flytter vannmolekyler seg både fram og tilbake i bølgens utbredelsesretning og i en retning vinkelrett på.

6.4 Utledning av bølgeligningen

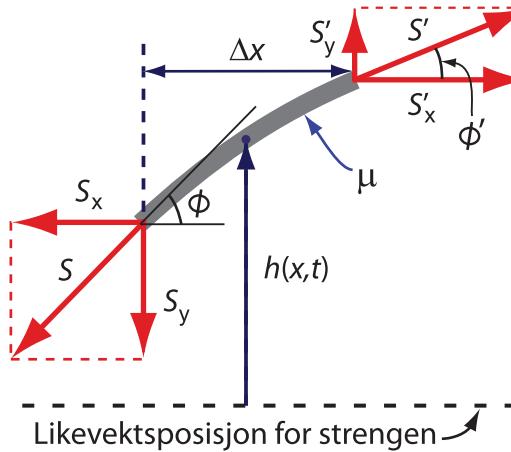
Vi har tidligere gitt et matematisk uttrykk for en bølge og (ved en kvasi baklengs argumentasjon) kommet fram til en differensialligning med bølger som løsninger. Nå skal vi starte med et fysisk system og utlede bølgeligningen derfra. Vi skal gjøre dette for svingninger på en streng og for lydbølger i luft/væske. Det er betydelig vanskeligere å gjøre en utledning for overflatebølger i vann, så på det området vil vi nøy oss med å gi mer omtrentlige løsninger uten utledning. Siden skal vi også utlede bølgeligningen for en elektromagnetisk bølge. Overflatebølger på vann og elektromagnetiske bølger vil vi først ta opp i senere kapitler.

6.4.1 Bølger på en streng

Utgangspunktet er en bølge langs en streng. Vi betrakter en liten del av strengen, nærmere bestemt en bit som er liten i forhold til den effektive bølgelengden. Figur 6.4 viser biten sammen med krefter som virker på den. Bølgen antas å forplante seg i horisontal retning (x -retning), og likevektstillingen til strengen når det ikke er noen bølger på den, er også horisontal. Bølgen antas å være rent transversal slik at utslaget utelukkende er i vertikal retning i figuren (y -retning). Det må bemerkes at utslaget i vertikal retning er *svært* lite i forhold til strengebitens lengde. Vi overdriver y -retningen i figuren for å få litt visuell hjelp når viktige relasjoner skal angis.

Det antas at strengens stivhet er så liten (i forhold til utslaget) at kreftene S og S' som virker i hver ende av strengbiten er *tangentielt* rettet langs strengen. Massesenteret til biten av strengen vil stadig vekk endre posisjon $h(x, t)$ i forhold til en midlere posisjon (likevektsposisjonen til strengen når det ikke er noen bølge der). Bevegelsen til strengbiten må kunne beskrives ved hjelp av Newtons annen lov.

Newton sann lov dekomponeres i horisontal og vertikal retning, og vi tar horisontalen først. Siden strengen antas å ha en ren transversell bevegelse, forskjer ikke strengbitens



Figur 6.4: Krefter som virker på en liten bit av en streng ved transversell bevegelse. Se teksten for detaljer.

massesenter seg (nevneverdig) i x -retning. Følgelig må summen av krefter i horisontal retning være lik null, med andre ord:

$$S_x = S \cos \phi = S' \cos \phi' = S'_x$$

Dette oppnås automatisk (til annen orden i ϕ) dersom $S = S'$, siden ϕ er en meget liten vinkel. (Husk at ifølge Taylorutvikling er $\cos \phi \approx 1 - \phi^2 + \dots$)

Newton sann lov anvendt i y -retning gir:

$$\sum F_y = ma_y \quad (6.9)$$

Strengen har en masse per lengde lik μ , og bitens lengde er Δx . Massen av denne biten er derfor $m = \mu \Delta x$.

$h(x, t)$ angir posisjonen til midtpunktet av strengbiten relativt til likevektsposisjonen når det ikke er noe bølge langs strengen. Anvendes også resultatet $S \approx S'$, følger det fra ligning (6.9):

$$S \sin \phi' - S \sin \phi = \mu \Delta x \left(\frac{\partial^2 h}{\partial t^2} \right)_{midtpunkt} \quad (6.10)$$

Indeksen på siste parantes indikerer at den dobbelt deriverte til massesenteret beregnes midt i intervallet Δx , dvs midt på strengbiten.

Siden vinklene ϕ og ϕ' er svært små, gir vanlig Taylorutvikling at:

$$\sin \phi \approx \phi \approx \tan \phi$$

og liknende for ϕ' . Sinus kan derfor erstattes med tangens i uttrykket ovenfor. Men tangens angir stigningstallet, som også kan skrives som $\partial h / \partial x$. Vi skal ha stigningstallet både i starten og slutten av strengbiten, og får:

$$\sin \phi' - \sin \phi \approx \left(\frac{\partial h}{\partial x} \right)_{(x+\Delta x)} - \left(\frac{\partial h}{\partial x} \right)_x$$

Dette kan omformes videre til:

$$\frac{\left(\frac{\partial h}{\partial x} \right)_{(x+\Delta x)} - \left(\frac{\partial h}{\partial x} \right)_x}{\Delta x} \Delta x \approx \left(\frac{\partial^2 h}{\partial x^2} \right)_{midtpunkt} \Delta x$$

Vær sikker på at du gjenkjenner den annen deriverte i uttrykket ovenfor!

Settes dette uttrykket inn i ligning (6.10), følger:

$$S \left(\frac{\partial^2 h}{\partial x^2} \right)_{midtpunkt} \Delta x \approx \mu \Delta x \left(\frac{\partial^2 h}{\partial t^2} \right)_{midtpunkt}$$

Begge derivasjonene refererer seg til samme punkt (midtpunktet), slik at denne indeksen kan droppes. Videre kan Δx forkortes bort. Med enkel manipulering av uttrykket følger da:

$$\frac{\partial^2 h}{\partial t^2} \approx \frac{S}{\mu} \frac{\partial^2 h}{\partial x^2}$$

Vi er nå så freidige at vi erstatter "tilnærmet lik" med "lik" og får:

$$\frac{\partial^2 h}{\partial t^2} = \frac{S}{\mu} \cdot \frac{\partial^2 h}{\partial x^2} \quad (6.11)$$

Vi har da vist at den transversale bevegelsen til en streng kan beskrives ved hjelp av bølgeligningen. Bølgen vil bevege seg med en hastighet:

$$v = \sqrt{\frac{S}{\mu}}$$

En løsning av denne svingeligningen kan da f.eks. være:

$$h(x, t) = A \cos(kx - \omega t + \phi)$$

hvor A er amplituden, k er bølgetallet, ω er vinkelfrekvensen og ϕ en vilkårlig fasevinkel. I første omgang kan alle disse fire størrelsene velges fritt, bortsett fra at k og ω må tilfredsstille relasjonen:

$$v = \sqrt{\frac{S}{\mu}} = \frac{\omega}{k}$$

Sagt på en annen måte: Det er tre frihetsgrader i bølgebevegelsen, og det er kanskje mest vanlig å angi disse som amplitude, frekvens og fase (fase angir i praksis valg av nullpunkt for tid). Det er initialbetingelsene som bestemmer disse, skjønt randbetingelsene spiller en enorm rolle, noe som fører til at løsningen i praksis kan bli stående bølger selv om initialbetingelsene alene tilsier noe helt annet.

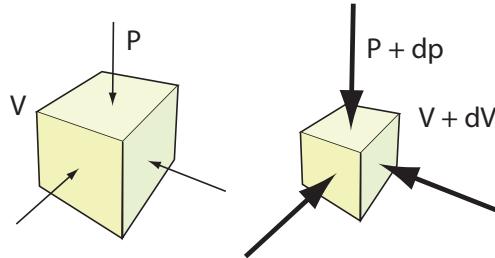
Før vi forlater bølgeligningen som beskriver bevegelsen til en streng, kan det være nyttig å minne om utgangspunktet for vår utledning:

- Newtons annen lov gjelder.
- Bølgen er rent transversell.
- Kraften som virker i hver ende av en bit av strengen er tangentielt rettet (dvs en temmelig ren geometriantakelse).
- Vinkelen mellom tangenten til strengen og likevektslinjen er meget liten i ethvert punkt langs strengen.

Ut fra disse enkle antakelsene følger det et elegant samspill mellom krefter, posisjon og tid som er ansvarlig for at bølgen beveger seg bortetter strengen. Du anbefales å tenke litt gjennom hvordan dette samspillet faktisk er. Hva er det som faktisk driver bølgen videre? Hva får utslaget til å øke, og hva får det til å avta? Det er ikke bare Mona Lisa som gjemmer på noe spennende!

6.4.2 Bølger i luft/væske

Utledning av bølgeligningen for bevegelse i luft/væske er mer komplisert enn den forgående. En grunn til dette er at vi nå opererer med et medium som fyller tre dimensjoner. For å gjøre utledningen overkommelig, begrenser vi oss til en plan, longitudinal bølge, som i effekt gjør at posisjonsendringer osv kan beskrives fullstendig selv med bare én romlig dimensjon (pluss tid).



Figur 6.5: Et volumelement med gass kan trykkes litt sammen dersom ytre trykk øker. Dersom dp er positiv, vil dV være negativ.

I vår sammenheng er den viktigste egenskapen til luft og væske at de er forholdsvis *kompressible*, det vil si, det går an å trykke sammen en viss mengde gass eller væske til et mindre volum enn den hadde opprinnelig. Luft kan trykkes sammen relativt lettere enn væske (og væske relativt lettere enn faste stoffer^{a)}). Figur 6.5 illustrerer nomenklaturen som brukes i utledningen nedenfor.

^{a)}Dette var grunnen til at vi ikke diskuterte kompressibilitet i den vibrerende strengen i avsnitt 4.4.1

Anta at en avgrenset mengde gass/væske med volum V ekspenderer eller trykkes sammen til et nytt volum $V + dV$. dV kan være både positiv og negativ, men tallverdien er liten sammenlignet med V . Den trykkforandringen som forårsaker volumendringen er dp . Er trykket opprinnelig P , blir den nå $P + dp$. Igjen antas det at dp kan være positiv og negativ, men tallverdien er alltid liten relativ til P .

Trykk måles i pascal (forkortes Pa) der

$$1 \text{ pascal} = 1 \text{ Pa} = 1 \text{ N/m}^2$$

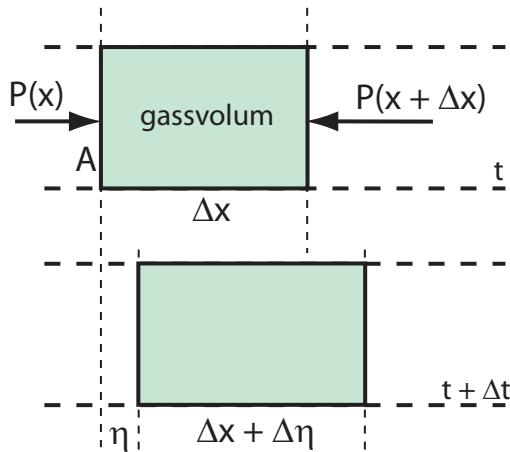
Hvor lett vi kan trykke sammen en gass eller væske beskrives av en materialkonstant kalt *kompressibilitetsmodulen*. Denne betegnes med K på norsk (B på engelsk, for “bulk compressibility module”). Modulen er definert slik:

$$K = -\frac{dp}{dV/V} \quad (6.12)$$

Kompressibilitetsmodulen er altså forholdet mellom en endring i trykk og den relative volumendringen denne medfører. Stor K svarer til at det må en stor trykkendring til for å få en gitt relativ volumendring. Med andre ord, stor K betyr at mediet er vanskelig å presse sammen.

La oss betrakte et volum gass eller væske der det forekommer trykkvariasjoner i én dimensjon, i retning x . Med det menes det at for et vilkårlig valgt plan vinkelrett på x -aksen og til enhver valgt tid, skal trykket overalt i planet være identisk.

For en slik modell, skjer all forflytning av molekyler parallelt med x -aksen når lufta presses sammen eller utvider seg. Det vil si at et gassvolum vil kunne endre lengde i x -retning, men aldri i y eller z -retning, som illustrert i figur 6.6. Effektiv forflytning av luftmolekyler angis med η , og η vil endre seg i x -retning (noe som henger sammen med



Figur 6.6: Ved en longitudinell bevegelse av et gass- eller væskevolum vil trykk og volum endre seg, men bare i én romlig dimensjon (her i x retning).

volum og trykkforandringer i x -retning).

Trykket P setter opp en kraft F som virker på tverrsnittet A det valgte volumelementet har i planet vinkelrett på x -aksen. Kraften er normalt ikke motsatt lik på de to sidene av volumelementet, noe som er indikert i figuren ved å bruke betegnelsene $P(x)$ og $P(x + \Delta x)$. Det virker derfor normalt en *netto* kraft på volumelementet, følgelig får denne gassen en akselerasjon gitt fra Newtons annen lov. Massen til volumelementet er lik massetettheten ρ multiplisert med volumet. Dersom positiv x -akse også defineres som positiv akseretning for F og akselerasjonen a , følger:

$$\begin{aligned} \Sigma F &= ma \\ PA - (P + \frac{\partial P}{\partial x} \Delta x)A &= (\rho \Delta x A) \frac{\partial^2 \eta}{\partial t^2} \end{aligned}$$

hvor vi har brukt den generelle relasjonen (Taylorutvikling):

$$P(x + \Delta x) = P(x) + \frac{\partial P}{\partial x} \Delta x$$

Følgelig:

$$\frac{\partial P}{\partial x} = -\rho \frac{\partial^2 \eta}{\partial t^2} \quad (6.13)$$

[♣ ⇒ En bemerkning: Her er det egentlig gjort en tilnærming allerede, idet vi opererer med en konstant massetethet ρ . Dersom vi hadde bakt inn at også massetetheten endres når trykk og volum endres, ville et annenordens ledd kommet i tillegg til de som inngår i ligningen nå. Dette ville under vanlige forhold gitt et lite korreksjonsledd som ikke betyr mye for bevegelsen til gassen, men for store trykkendringer osv ville ledet fått større betydning. ← ♣]

For å komme videre må vi å finne en kobling mellom trykkforandringer i P og effektiv utslag (posisjonsendringer) η til gassmolekylene. Denne sammenhengen kan hentes fra definisjonen av kompressibilitetsmodulen gitt i ligning (6.12). Multipliseres hele uttrykket med nevneren, følger:

$$dp = -K \frac{dV}{V}$$

Settes denne relasjonen inn i venstre side av i ligning (6.13), følger:

$$\frac{\partial P}{\partial x} = \frac{\partial(P_0 + dp)}{\partial x} = -K \frac{\partial \frac{dV}{V}}{\partial x} \quad (6.14)$$

I denne mellomregningen er det valgt å skrive trykket som en *konstant* gjennomsnittsverdi P_0 pluss en tids- og posisjonsvariabel endring i trykk dp som er lite sammenlignet med gjennomsnittsnivået.

Den relative volumendringen kan i sin tur relateres til utslaget til gassmolekylene η , siden (referer til figur 6.6):

$$\Delta\eta = \frac{\partial\eta}{\partial x} \cdot \Delta x$$

Følgelig:

$$\frac{dV}{V} = \frac{A \frac{\partial\eta}{\partial x} \Delta x}{A \Delta x} = \frac{\partial\eta}{\partial x}$$

Herav følger fra ligning (6.14):

$$\frac{\partial P}{\partial x} = -K \frac{\partial}{\partial x} \left(\frac{\partial\eta}{\partial x} \right) = -K \frac{\partial^2\eta}{\partial x^2}$$

Setter dette inn i ligning (6.13), følger:

$$-K \frac{\partial^2\eta}{\partial x^2} = -\rho \frac{\partial^2\eta}{\partial t^2}$$

eller

$$\frac{\partial^2\eta}{\partial t^2} = \frac{K}{\rho} \frac{\partial^2\eta}{\partial x^2} \quad (6.15)$$

Vi har altså kommet fram til bølgeligningen også for dette systemet. Forskyves luftmolekyler på en systematisk måte, som indikert i utledningen, vil utslaget til forflytningen av luftmolekyler bre seg som en bølge. Hastigheten til bølgen er gitt ved:

$$v = \sqrt{\frac{K}{\rho}} \quad (6.16)$$

Lyd hastigheten øker med andre ord dersom gassen/væsken er vanskelig å trykke sammen, men avtar med massetetheten til materien lyden brer seg gjennom.

[♣ ⇒ Kommentar: Det er interessant å merke seg at lydhastigheten i luft er lavere enn middelfarten til luftmolekylene mellom kollisjonene dem imellom i den Brownske bevegelsen. For nitrogen ved romtemperatur og en atmosfærisk trykk, er toppunktet i sannsynlighetsfordelingen for molekylene fart om lag 450 m/s. Interesserte kan lese mer om dette under "Maxwell-Boltzmann distribution" f.eks. på Wikipedia.
⇐ ♣]

6.4.3 Konkrete eksempler

Beregningen vi foretok for å komme fram til bølgeligningen for bevegelser i luft og væsker er ganske grov. Vi startet ut med Newtons annen lov, anvendte lovmessigheten som ligger i definisjonen av kompressibilitetsmodulen, pluss noen andre mindre betydningsfulle detaljer, og kom fram til bølgeligningen. Kan en så enkel beskrivelse gi et brukbart estimat av lydhastigheten?

La oss forsøke å beregne lydhastigheten i vann. Kompressibilitetsmodulen for vann (ved omtrent atmosfæretrykk) er gitt ved $K = 2.0 \cdot 10^9$ Pa. Tettheten til vann er $\rho \approx 1.0 \cdot 10^3$ kg/m³. Settes disse verdiene inn i uttrykket for lydhastigheten i ligning (6.16), er resultatet:

$$v_{vann} \approx 1.43 \cdot 10^3 \text{ m/s}$$

Tabellverdi for lydhastighet i vann er 1402 m/s ved 0 °C, og 1482 m/s ved 20 °C. Med andre ord er overensstemmelsen faktisk god!

La oss så forsøke å beregne lydhastigheten i luft. Da oppstår et problem ved at kompressibilitetsmodulen vanligvis ikke gis som en generell tabellverdi, siden verdien avhenger av hvilket trykk vi betrakter. Vi starter i stedet med gassloven:

$$PV^\gamma = \text{konstant}$$

hvor $\gamma = C_p/C_v$ der C_p er spesifikk varme ved konstant trykk, og C_v er spesifikk varme ved konstant volum. Det forutsettes at de endringene som skjer i volum og trykk foregår slik at vi ikke tilfører energi til gassen (adiabatiske forhold). For lyd med normal lydintensitet, er dette kravet rimelig godt tilfredsstilt, men ikke for svært kraftig lyd.

Foretas en generell derivering av gassloven, følger:

$$dP V^\gamma + P d(V^\gamma) = 0$$

$$V^\gamma dP + \gamma V^{\gamma-1} dVP = 0$$

Kombineres dette med ligning (6.12), følger:

$$K = -\frac{dP}{\frac{dV}{V}} = \gamma P$$

Varmekapasitetene for luft hentes fra tabeller, hvilket gir:

$$\gamma = \frac{C_p}{C_v} = 1.402$$

En atmosfærer trykk er 101325 Pa, følgelig får vi et mål for kompressibilitetsmodulen for luft under en atmosfærer trykk (og adiabatiske forhold):

$$K = 1.402 \cdot 101325 \text{ Pa}$$

Fra tabeller kan massetetheten for luft ved en atmosfærer trykk og ca 20 °C hentes ut ($\rho = 1.293 \text{ kg/m}^3$). Da er det endelig mulig å beregne lydhastigheten i luft:

$$v_{\text{lyd i luft}} = 331 \text{ m/s}$$

Tabellverdi er 344 m/s.

De tallene som er brukt refererer ikke alle til 20 °C og en atmosfærer trykk. Det er derfor ikke så rart at beregningene ikke gir fullt klaff. Likevel er den beregnede verdien "bare" om lag fire prosent for lav. Det indikerer at våre beregninger og formelen vi kom fram til for lydhastighet i gasser/væsker, er rimelig god!

[♠ ⇒ Kommentar: Også for metaller er det i tabeller oppgitt kompressibilitetsmodul, og beregner vi lydhastigheten i metaller ved å bruke samme formel som for gasser og væsker, får vi verdier som er i nærheten av den korrekte, men med langt større avvik enn for luft og vann. Ekspempelvis beregner vi lydhastigheten i stål til å være 4510 m/s, mens den i virkeligheten er om lag 5941 m/s. Tilsvarende for aluminium gir beregningen 5260 m/s, mens virkelig verdi er 6420 m/s. Vi ser altså som nevnt at det er større sprik her mellom beregnet og virkelig lydhastighet.]

Forøvrig bør vi merke oss at i metaller kan lyden gjerne forplante seg som en transversal bølge i stedet for eller i tillegg til en longitudinal, spesielt når formen til metallstykket er spesiell. Lydhastigheten til en transversal bølge i et metall avhenger av stivheten til metallet, med den følge at transversale bølger ofte har lavere bølgehastighet enn for longitudinale bølger. Slår vi på en metallstav, får vi ofte både transversale og longitudinale bølger samtidig, og de longitudinale har ofte en høyere frekvens enn de transversale (etter at stående bølger har dannet seg). ← ♠]

6.4.4 Trykkbølger

I utledningen ovenfor så vi at effektiv bevegelse til gass- eller væskemolekyler kan følge en bølgeligning. Det er interessant å se hvor stor forskyvning molekylene har når en bølge passerer, men vanligvis er det mer interessant å beskrive bølgen i form av *trykkforandringer*. Overgangen kan gjennomføres som følger.

Ovenfor ble det angitt at molekylene effektivt forflytter seg en avstand η longitudinalt i bølgens retning. Forflytningen tilfredsstiller bølgeligningen (6.15):

$$\frac{\partial^2 \eta}{\partial t^2} = \frac{K}{\rho} \frac{\partial^2 \eta}{\partial x^2}$$

En løsning kan da f.eks. være den enkle bølgen:

$$\eta(x, t) = \eta_0 \cos(kx - \omega t) \quad (6.17)$$

hvor bølgetall k og vinkelfrekvens ω må tilfredsstille:

$$\frac{\omega}{k} = \sqrt{\frac{K}{\rho}}$$

hvor størrelsene er som definert over.

Ved overgang til trykkbølger, brukes på ny definisjonen av kompressibilitetsmodulen (ligning (6.12)), som gir:

$$dp = -K \frac{dV}{V}$$

Ut fra figur 6.6 følger da:

$$dp = -K \frac{((\Delta x + \frac{\partial \eta}{\partial x} \Delta x) - \Delta x)A}{\Delta x A}$$

$$dp = -K \frac{\partial \eta}{\partial x} \quad (6.18)$$

Størrelsen dp er endring i trykk *i forhold til gjennomsnittsverdien*, og er en funksjon av posisjon og tid. Størrelsen dp kan gjerne omdøpes slik:

$$dp \equiv p(x, t) \quad (6.19)$$

Ved å kombinere ligningene (6.17), (6.18) og (6.19), følger da:

$$p(x, t) = -K \frac{\partial \eta}{\partial x} = -K \eta_0 (-\sin(kx - \omega t)) \cdot k$$

Og endelig:

$$p(x, t) = kK \eta_0 \sin(kx - \omega t) \quad (6.20)$$

Resultatet viser at såfremt forskyvningen av molekylene beskriver en bølgebevegelse, vil også trykkvariasjonen gjøre det samme. Det er en faseforskjell mellom disse bølgene, men mer viktig er sammenhengen mellom amplitudene. Dersom amplituden for forskyvning til molekylene er η_0 , er amplituden for trykkbølgen $kK \eta_0$, altså bølgetallet multiplisert med kompressibilitetsmodulen multiplisert med forskyvningsamplituden.

Denne relasjonen kan brukes for å bestemme gjennomsnittlig forskyvning til molekylene i enkelte sammenhenger, noe vi kommer tilbake til i oppgaver i et senere kapitel.

6.5 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Skrive opp en standard bølgeligning (for en plan bølge).
- Gjøre rede for amplitude, bølgetall, bølgelengde, periodetid, frekvens, fase, bølgens hastighet og formelen $f\lambda = v$.
- Gi et matematisk uttrykk for en harmonisk plan bølge såvel som en vilkårlig formet plan bølge, som beveger seg i en angitt retning. For en harmonisk plan bølge bør vi også kunne gi en matematisk beskrivelse basert på Eulers formel.
- Gjøre rede for hvordan en bølge kan anskueliggjøres enten som funksjon av tid eller som funksjon av posisjon.
- Gjøre rede for forskjellen mellom longitudinal og transversal bølge, og gi minst ett eksempel på hver.
- Utlede bølgeligningen for en transversal svingning på en streng.
- Kjenne hovedtrekkene i utledningen av bølgeligningen for en trykkbølge gjennom f.eks. luft (lydbølge).
- Kunne beregne omtrentlig lydhastigheten i vann ved å bruke mekaniske egenskaper til vann.

6.6 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Sett opp ett eksempel på svingeligningen og ett eksempel på bølgeligningen. Hvilke typer opplysninger må vi kjenne til for å finne en konkret løsning av hver av disse to typene differensielligninger?
2. Er hastigheten til bølger beskrevet som i ligning (6.6) avhengig av amplituden? Som vanlig: Begrunn svaret.
3. Ved tordenvær ser vi oftest lynet før vi hører tordenen. Forklar dette. Det finnes en enkel regel for å anslå hvor langt unna lynet befant seg. Hva er regelen, og hvor kommer tallene fra?
4. Anta at en lang snor henger fra et høyt tak og nesten ned til gulvet. Anta at snora gis en transversal bølgebevegelse i nedre ende, og at bølgen så brer seg mot taket. Vil bølgehastigheten være konstant på vei opp mot taket? Som vanlig: Begrunn svaret.
5. Dersom du strekker en gummistrikk og klimprer på den, hører du en slags tone med en eller annen tonehøyde. Anta at du strekker enda litt mer, og klimprer på ny. (Prøv selv!) Hvordan er tonehøyden nå i forhold til den forrige? Forklar resultatet! (Hint: Lengden på en vibrerende streng er lik halve bølgelengden til grunntonene.)
6. Da vi omtalte lydbølger sa vi (med en modifiserende kommentar) at hvert enkelt luftmolekyl svinger fram og tilbake i forhold til et likevekspunkt. Dette er på en måte totalt feil, men likevel har bildet en viss berettigelse. Forklar.

7. Forskjellen på en longitudinal og en transversal bølge er i kapitlet knyttet opp mot symmetri. Hvordan?
8. Til slutt i underkapittel 4.4.1 ble det gitt en oversikt over de essensielle antakelsene som ble gjort i utledningen av bølgeligningen for bevegelser langs en streng. Forsök å sette opp en tilsvarende liste for utledningen av bølgeligningen i luft/vann.
9. Siste 3/4 av utledningen i delkapittel 4.4.2 behøver man ikke kunne gjennomføre til eksamen i vårt kurs. Hva tror du hensikten var da vi likevel tok med denne utledningen her i boka? Hva bestemmer faktisk lydhastigheten i luft/vann?
10. Drøft lydbølger med hensyn på energi.

Regneoppgaver

11. Sjekk om funksjonen $y(x, t) = A \sin k(x + vt)$ tilfredsstiller bølgeligningen.
12. Hva karakteriserer en plan bølge? Nevn to eksempler på bølger som ikke er plane, og gi et eksempel på en (tilnærmet) plan bølge.
13. Sett opp et matematisk uttrykk for en plan bølge som beveger seg i negativ z-retning.
14. Er dette en plan bølge: $S = A \sin(\mathbf{k} \cdot \mathbf{r} - \omega t)$? Her er \mathbf{k} bølgevektor og \mathbf{r} er en vilkårlig valgt posisjonsvektor, ω vinkelfrekvens og t tid. A er en reell skalar. Begrunn svaret.
15. Forklar med egne ord hvordan vi kan se av de matematiske uttrykkene at en bølge $A \cos(kx - \omega t)$ beveger seg mot høyere x-verdier etter som tiden går, mens bølgen $B \cos(kx + \omega t)$ beveger seg motsatt vei.
16. En stående bølge kan beskrives ved $g(x, t) = A \sin(kx) \sin(\omega t)$. Vis ved direkte innsetting at en stående bølge også er en løsning av bølgeligningen for $v = \omega/k$. (Vi kommer tilbake til stående bølger i neste kapittel.)
17. Hvor lang er bølgelengden til lydbølger ved 100 Hz i luft og vann? Hva er de tilsvarende bølgelengdene for lyd med frekvens 10 kHz?
18. Når vi tar ultralydbilder av fostre, hjerte osv, er bildekvaliteten avhengig av at bølgelengden ikke er mer enn ca 1 mm. Lydbølger i vann/vev har en hastighet på om lag 1500 m/s. Hvilken frekvens må ultralyden ha? Er ordet "ultralyd" en ok betegnelse?
19. Hvor lang er bølgelengden for FM kringkasting ved 88.7 MHz? Og hvilken bølgelengde har mobiltelefonen din dersom den opererer på 900 eller 1800 MHz?
20. Et ungt menneske-øre kan høre frekvenser i området 20 Hz til 20 kHz. Hvor stor er bølgelengden i luft for disse yttergrensene? (Lydhastigheten i luft er om lag 340 m/s.)
21. En 2 m lang streng av metall har masse $3 \cdot 10^{-3}$ kg. Strengen holdes horisontalt med den ene enden festet. Den andre enden glir over en glatt, rund kant, og er festet til et lodd med vekt 3 kg.
 - a) Beregn hastigheten på en transversal bølge langs strengen.
 - b) Endrer bølgehastigheten seg dersom vi endrer lengden på den horisontale delen av strengen (det vil si med hvor mye av de 2 m som befinner seg mellom det fastspente punktet og den runde kanten)?

- c) Hvor lang måtte den horisontale delen av strengen være for at strengen skulle kunne svinge med en frekvens på 280 Hz dersom du klimpret på den? (Hint: Anta at strengen da er en halv bølgelengde lang.)
- d) Hvor tung måtte loddet være for at frekvensen i forrige punkt ble dobbelt så høy som i stad (anta uforandret lengde)?
22. Lag et program i Matlab eller Python som sampler lydlydsignalet som kommer inn på mikrofoninngangen på en PC når en mikrofon er koblet til, og plot signalet med riktig tidsangivelse langs x-aksen. I matlab kan følgende funksjon brukes:

```
s = waverecord(N,fs,'int16');
```

hvor N er antall punkter som skal samples, fs er samplingsfrekvensen i Hz, og 'int16' lar du stå som den står. De samplede signalet blir liggende i arrayen s. Bruk gjerne fs = 44100 dersom det finnes lydsignaler med meget høy frekvens (opp mot 20 kHz). For mer lavfrekvente signal kan du gjerne bruke fs = 11025.

Tilsvarende funksjon i Python får du finne ved å søke på internett dersom det er aktuelt.

Sample lyden når du synger en dyp “aaaaaa”. Er lydbølgen harmonisk? Hvilken frekvens har den?

Du kan dersom du ønsker det foreta en fourieranalyse av signalet. I så fall er det viktig at du sørger for å få korrekt frekvensangivelse langs frekvensaksen.

23. Lag en animering av bølgen $A \sin(kx - \omega t)$ i Matlab eller Python. Velg selv verdier for A , k , ω og variasjonsområde for x og t . Når du har fått denne animasjonen til å gå, kan du forsøke å animere bølgen $A \sin(kx - \omega t) + A \sin(kx + \omega t)$. Beskriv hva du ser!

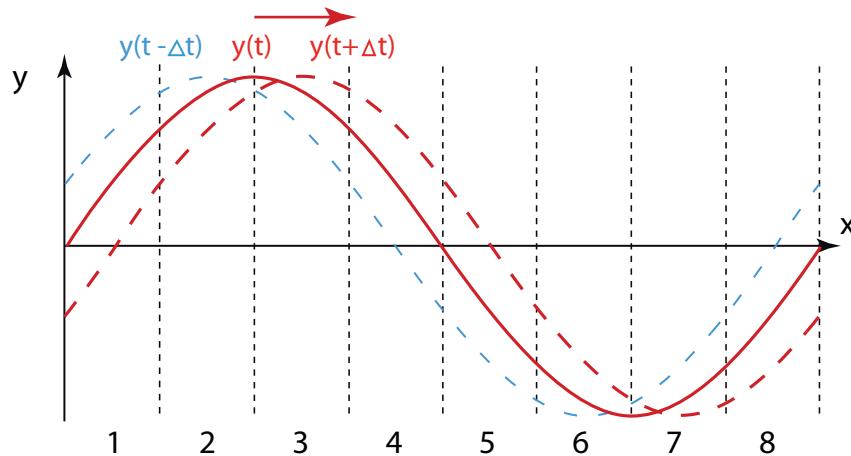
Her er en kodesnutt som kanskje kan være til hjelp:

```
function bolgeanimering1
% Versjon 6.feb.2013
clear all;
k = 3;
omega = 8;
N = 1000;
x = linspace(0,20,N);
y = linspace(0,20,N);
p = plot(x,y,'-','EraseMode','xor');
axis([0 20 -2.5 2.5])
for i=1:200
    t = i*0.01;
    y = 2.0*sin(k*x-omega*t);
    set(p,'XData',x,'YData',y)
    drawnow
    pause(0.02); % For å forsinke fremvisningen
end
```

24. I figur 6.7 er det vist en bølge langs en streng (et lite utsnitt) ved tre nærliggende tidspunkter. Ta utgangspunkt i figuren og forklar:

- Hvilken retning peker nettokraften for hvert av de åtte segmentene av strengen ved tiden t (ser bort fra gravitasjon).
- Forklar i detalj hvordan du resonerte for å finne kraften, spesielt for segment 2, 4, 5, 6 og 7.

- Hvilken retning har hastigheten for disse segmentene ved tiden t ?
- Det kan i første omgang virke som om det er en konflikt mellom krefter og hastighet. Forklar den tilsvarende konflikten.
- Det siste poenget har sammenheng med forskjellen mellom Aristoteles fysikk og Newtons fysikk. Kjenner du til forskjellen?
- Hvordan går det med energien til et element langs strengen når bølgen vandrer forbi?
- Hva ligger i uttrykket av typen "bølgen bringer med seg energi"?



Figur 6.7: En bølge langs en bit av en streng ved tre nærliggende tidspunkt.

25. Les kommentar-artikkelen "What is a wave?" av John A. Scales og Roel Snieder i Nature vol. 401, 21. oktober 1999 side 739-740. Hvordan definerer disse forfatterne en bølge?

Kapittel 7

Lyd



Fire typer bølger preger vår hverdag: Bølger på en streng, lydbølger, elektromagnetiske bølger og bølger på vann. Det er store forskjeller på disse bølgetypene, men likevel har de selvfølgelig også mye til felles.

I dette kapitlet vies lydbølger størst oppmerksomhet. Blant annet gjennomgås noen karakteristiske trekk for musikkinstrumenter og toneskalaen. Aller først diskuteres imidlertid generell refleksjon av bølger som fører til stående bølger, og sist i kapitlet beskrives dobblerskift og sjokkbølger.

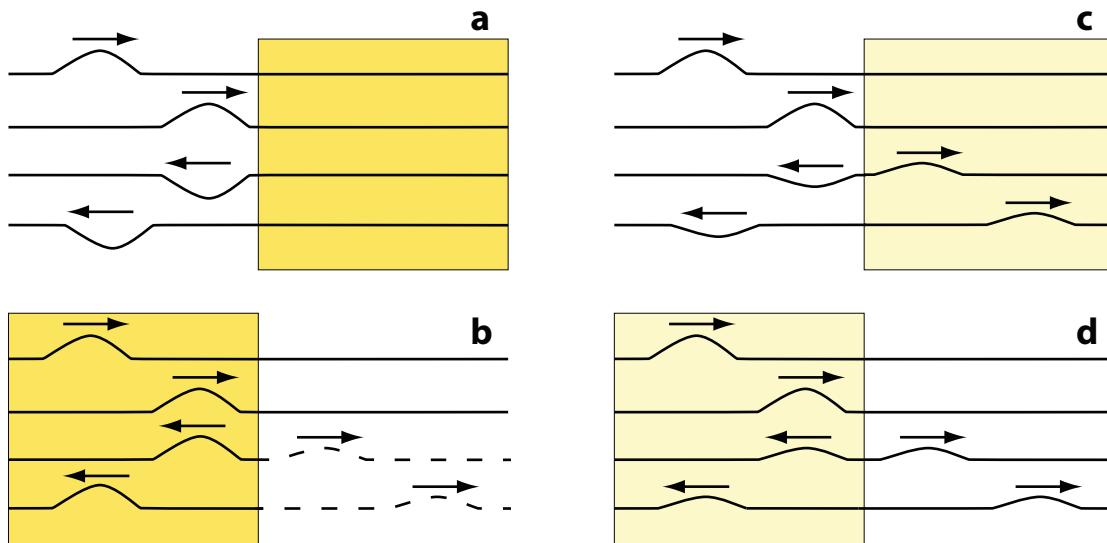
Lydintensiteter og decibelskalaen er også sentralt stoff i dette kapitlet. Her møter vi for første gang forskjellen mellom intensitet angitt i et rent fysiske mål (W/m^2 og dB(SPL)) og intensitet slik vi mennesker opplever den (dB(A) m.fl.).

Vi møter helt tilsvarende todeling når vi i et senere kapittel skal vurdere lysintensiteter.

Mye morsom fysikk ligger bak en god gitar! Båndene på halsen viser f.eks. at toneskalaen er basert på logaritmiske forholdstall

7.1 Refleksjon av bølger

Lydbølger reflekteres av en betongvegg, lysbølger reflekteres av et speil, bølgene langs en gitarstrenge reflekteres der strengen er festet i endene. Refleksjon av bølger er et tema vi møter gang på gang i boka i mange ulike situasjoner. Matematisk sett oppstår refleksjoner på grunn av såkalte "randbetingelsene". Som nevnt tidligere, kan en differensialligning for en bølgebevegelse se identisk ut i mange ulike sammenhenger, men likevel skiller løsningene seg drastisk fra hverandre fordi randbetingelsene er forskjellige. Det første, og antakelig det enkleste eksemplet på dette, er en bølgebevegelse langs en streng med endelig lengde, der fysiske forhold ved "randen" (enden på strengen) spiller en avgjørende rolle for bølgebevegelsen.



Figur 7.1: En transversal bølge når et grensesjikt mellom to medier. I a går bølgen fra et område med liten impedans (hvitt) til mye større impedans (gult). Bølgen er tegnet inn for fire etterfølgende tidspunkt. Bølgen blir fullstendig reflektert og utslaget får motsatt fortegn av den innkommende. I b reduseres impedansen i grensesjiktet (bølgen kommer fra et høyimpedansområde og møter et område med mye lav impedans). Refleksjon av energi er her bortimot fullstendig, men betrakter vi bare amplituder, er effekten mer beskjeden (indikert ved stiplet linje). I c og d har vi tegnet inn et tilfelle hvor amplituden i den reflekterte og transmitterte bølgen er like store.

Anta at vi har en streng festet til en massiv struktur i den ene enden. Vi gir en liten transversal "puls" i motsatt ende av strengen (se figur 7.1). Pulsen vil bevege seg langs strengen med hastigheten $\sqrt{\frac{S}{\mu}}$ hvor S er strammingen og μ er masse per lengde. Pulsns form opprettholdes.

Når pulsen kommer fram til den fastspente enden av snora, må nødvendigvis utslaget i denne enden alltid være lik null. Det betyr at pulsen tett opp mot endepunktet vil bli sammenpresset og kraften på tvers av strengen vil øke betydelig. Når endepunktet ikke vil flytte på seg, virker kraften tilbake på strengen og skaper en misbalanse mellom utslag og "tvershastigheter" som gjør at bølgen rett og slett snur og vender bakover igjen langs strengen. Bølgen som går bakover vil imidlertid ha utslag til motsatt side enn den opprinnelig (innkommende) pulsen. Det går ikke noe energi tapt (i første tilnærming) siden tap i form av friksjon krever at friksjonskraften har virket over en viss vei, mens vi har antatt at endepunktet ligger helt fast.

En annen ytterlighet er at strengen ender fritt. Det kan f.eks. oppnås ved å holde i en

steng i den ene enden og la den andre henge fritt nedover og ende i løse luften (ser bort fra luftmotstand). Dette er imidlertid ingen god modell siden strekket i strengen da ikke er definert. Det er bedre å ha en steng med stor masse per lengde som knyttes sammen med en steng med svært mye mindre masse per lengde, og utsetter hele strukturen for en temmelig veldefinert strekkraft.

Sendes nå en puls innover langs den strengebiten som har mest masse per lengde, vil pulsen bevege seg normalt helt til den når grensen mellom de to typer strenger. Kraften som virker fra den tykke til den tynne strengbiten, vil gi den tynne biten et betydelig større utslag enn om tykkelsen på strengen var den samme overalt. Det blir igjen et mismatch mellom utslag og hastighet, og resultatet i dette tilfellet er at vi får reflektert pulsen, men nå med utslag til samme side som den opprinnelige pulsen. I dette tilfellet vil imidlertid også en del av bølgen (og energien) gå videre langs den tynne strengen. Dersom masse per lengde er svært liten i den tynne delen sammenlignet med den tykke, vil nesten all energi reflekteres (tilfelle b i figur 7.1).

Betegnelsene “en massiv struktur”, og “en tynnere eller tykkere snor” (i betydning masse per lengde) er ikke presis ordbruk. I blant er det lettere å bruke ordet “impedans” i slike sammenhenger. Regelen er da:

Når en bølge treffer et grensesjikt til et nytt medium hvor bølgehastigheten (i betydning fasehastigheten) reduseres, sies det nye mediet å ha en høyere impedans enn det forgående.

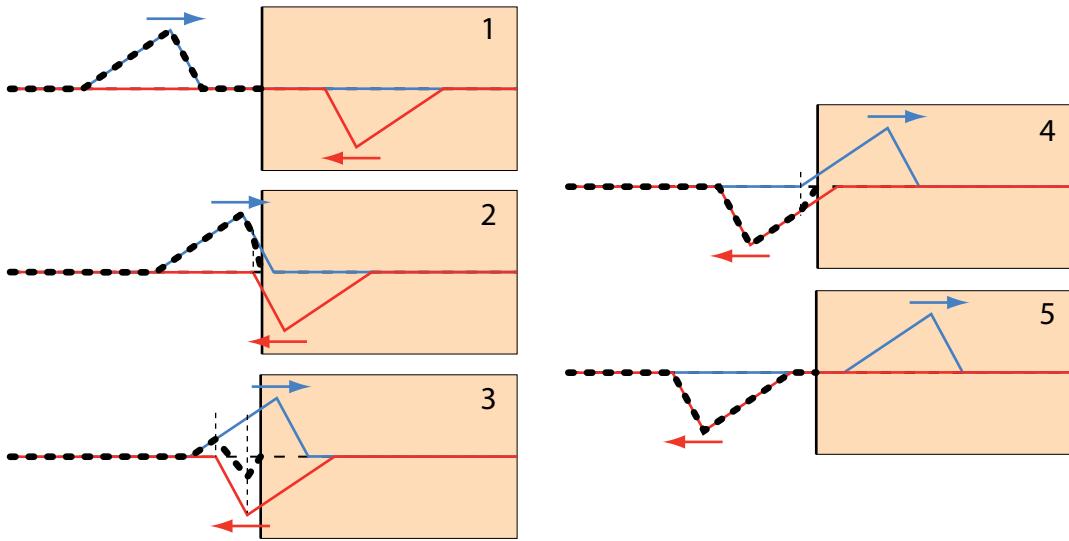
Med utgangspunkt i denne ordbruken, kan reglene for refleksjon og transmisjon av bølger ved et grensesjikt angis slik:

Det kan vises både eksperimentelt og teoretisk at:

- Bølger som treffer et grensesjikt der impedansen til mediet øker, deler seg slik at den delen av bølgen som blir reflektert får utslag med motsatt fortegn av den innkommende. Den transmitterte bølgen får utslag med samme fortegn som den innkommende på hver sin side av grensesjiktet.
- For bølger som treffer et grensesjikt der impedansen til mediet minker, vil den reflekterte bølgen ha utslag med samme fortegn som den innkommende. Den transmitterte bølgen har også utslag med samme fortegn som den innkommende.
- Hvor mye som reflekteres og transmitteres avhenger av relativ impedansendring i forhold til impedansen i mediet bølgen opprinnelig kommer fra. Er impedansendringen null, blir ingenting reflektert. Dersom forholdet mellom den største og minste impedansen er uendelig stor, vil all energi bli reflektert.

I figur 7.1 er bølgeformen med vilje ikke tegnet inn mens bølgen treffer grensesjiktet mellom de to mediene. Detaljert bølgeform kan finnes ved å bruke en fremgangsmåte som er skissert i figur 7.2. I figuren er det totalrefleksjon som illustreres. Korrekt bølgeform før grensesjiktet tegnes inn, og vi lar bølgen nærme seg grensesjiktet. En virtuell bølge identisk med den reelle, men speilvendt til denne om grensesjiktet, tegnes inn. Fortegnet på den virtuelle bølgen endres dersom totalrefleksjonen skjer mot et medium med svært stor impedans (f.eks. ved at en steng spennes fast til en massiv struktur). Formen på den virkelige bølgen under og etter refleksjon finnes ved å addere opprinnelig bølgeform med den virtuelle bølgen (summen er markert med tykk, stiplet linje i figuren). Etter hvert er det bare den virtuelle bølgen som befinner seg til venstre for grensesjiktet, og da følger den videre bølgeutviklingen bevegelsen til den virtuelle bølgen alene.

Denne modellen kan lett modifiseres for det tilfellet at bølgen når inn mot et medium med mye lavere impedans slik at det blir tilnærmet total refleksjon uten bytte av for-



Figur 7.2: En metode for å konstruere tidsutviklingen til en bølgeform for en transversal bølge som reflekteres ved et grensesjikt mellom to medier. Se teksten for detaljer.

tegn. Modellen kan også modifiseres for å kunne håndtere tilfeller der noe av bølgen blir reflektert og noe transmittert.

Vi kommer siden i boka tilbake til en langt mer detaljert beskrivelse av refleksjon og transmisjon for elektromagnetiske bølger som treffer et grensesjikt mellom to medier.

7.2 Akustisk impedans

Vi brukte ovenfor ganske vage formuleringer så som større eller mindre impedans da vi diskuterte refleksjon av bølger ved overgangen fra et medium til et annet. Når det gjelder lyd kan vi innføre en størrelse kalt ”akustisk impedans” for å bli mer presise. Det er flere varianter av akustisk impedans.

“Karakteristisk akustisk impedans” Z_0 er definert som:

$$Z_0 = \rho c \quad (7.1)$$

hvor ρ er massetettheten til mediet (kg/m^3), og c er lydhastigheten (m/s) i dette mediet.

Z_0 er materialavhengig og måles i Ns/m^3 eller $\text{Pa s}/\text{m}$.

Den karakteristiske impedansen for luft ved romtemperatur er om lag $413 \text{ Pa s}/\text{m}$. For vann er den om lag $1.45 \cdot 10^6 \text{ Pa s}/\text{m}$, dvs. ca 3500 ganger større enn den karakteristiske impedansen til luft.

Forskjeller i karakteristisk akustisk impedans bestemmer hvor mye av en bølge som blir transmittert og reflektert når en ”plan bølge” når et plant grensesjikt mellom to medier.

Den store forskjellen i karakteristisk akustisk impedans mellom luft og vann betyr at lyd i luft bare i liten grad vil transmitteres inn i vannet, og lyd i vann bare i liten grad vil trenge ut i luften. Det meste av lyden vil reflekteres ved overflaten mellom luft og vann.

I forrige kapittel fant vi at lydhastigheten i luft eller vann var gitt med:

$$v = \sqrt{K/\rho}$$

hvor K er kompressibilitetsmodulen og ρ er massetettheten. Kombinerer vi dette uttrykket med definisjonen av karakteristisk impedans i ligning (9.34), får vi:

$$Z_0 = K/c \quad (7.2)$$

Dersom vi husker definisjonen på kompressibilitetsmodulen fra forrige kapittel, får vi da:

$$Z_0 = \frac{-dp}{c \cdot dV/V}$$

Vi ser da (dersom vi legger godviljen til) at Z_0 blir stor dersom det må stort trykk til for å få en volumendring, som igjen er avhengig av hvor mye molekylene faktisk forflytter seg.

Dette har en viss analogi til impedans i elektromagnetismen. Der er impedansen Z til en krets/komponent er gitt ved

$$Z = V/I$$

hvor V er spenningen over kretsen og I er strømmen som spenningen fører til.

Det en analogi mellom lydtrykk og spenning, og det er en analogi mellom volumendring (som skyldes netto forflytning av molekyler) og strøm.

For et spesifikt system, f.eks. et musikkinstrument, kan vi definere “**akustisk impedans**” slik:

$$Z = \frac{p}{vS} \quad (7.3)$$

hvor p er lydtrykket, v er partikkel-hastigheten (den som ligger på toppen av Brownske bevegelser), og S er tverrsnittet hvor lydtrykk og partikkelhastigheten gjelder (f.eks. ved munnstykket til en trompet).

Ved en slik formulering er det enda enklere å se analogien mellom vår akustiske impedans og impedans i elektromagnetismen. Det er derfor ikke så rart at definisjonen av akustisk impedans sammenlignes med Ohms lov, og at Z iblant kalles “lydmotstanden” eller “lydimpedansen”.

Akustisk impedans er ikke bare avhengig av materialet, men også av fysisk utforming. Størrelsen forteller hvor stort lydtrykk som genereres som følge av vibrasjon av molekyler i et akustisk medium ved en gitt frekvens og for en bestemt gjenstand eller geometri. Den akustiske impedansen Z er oftest betydelig mer frekvensavhengig enn den karakteristiske akustiske impedansen. Akustisk impedans er en svært nyttig størrelse innen akustikk, f.eks. når det skal regnes på hvor mye effekt som må til for å få et visst lydnivå f.eks. i en konsertsal.

Vil du lære mer om akustisk impedans, kan følgende artikkel kanskje være av interesse: “What is acoustic impedance and why is it important?” på <http://www.phys.unsw.edu.au/jw/z.html> (per. 13. feb. 2013).

7.2.1 Ultralydbilder

Fra ligningene (9.34) og (7.2) ser vi at karakteristisk akustisk impedans vil endre seg med massettethet og kompressibilitetsmodulen. Nøyaktig sammenheng er ikke så lett å få med seg ut fra disse ligningene siden lydhastigheten også avhenger av de samme størrelsene.

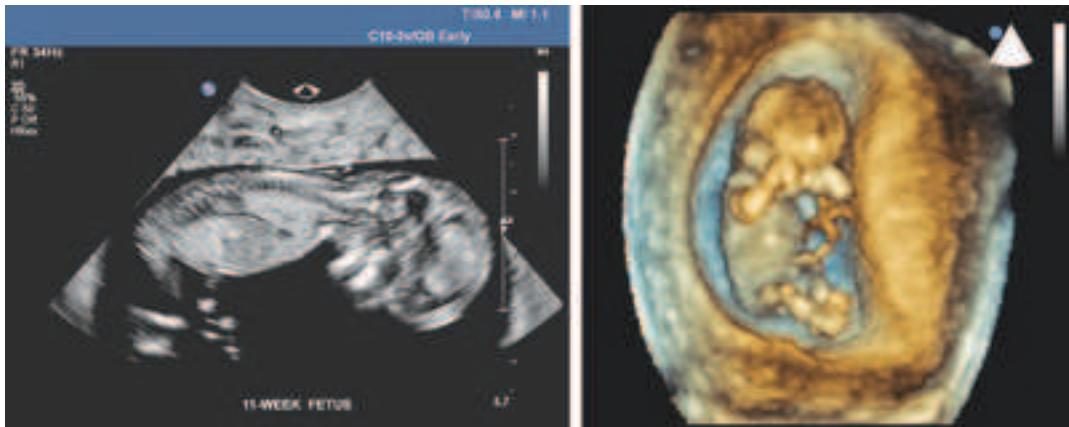
Uansett, det er forskjeller i karakteristisk akustisk impedans f.eks. mellom blod og hjertemuskel. Det er forskjell i karakteristisk akustisk impedans mellom et foster og fostervæsken. Dersom vi sender lyd inn mot kroppen, vil derfor litt av lyden bli reflektert fra grenseflatene mellom blod og hjertemuskel, og mellom fostervæsken og fosteret. Det er imidlertid aller størst forskjell mellom karakteristisk akustisk impedans for luft og kropp. Skal vi derfor få lyd effektivt inn og effektivt ut av kroppen, må vi smøre på et geleaktig materiale mellom lydproben og kroppen ved ultralydundersøkelser. Dette materialet bør ha omrent samme karakteristiske akustiske impedans som vevet lyden skal gå inn i.

Lyden vil etter refleksjon i grenseflater mellom ulike impedanser kunne oppfanges som et ekko, forutsatt at lydpulsen vi startet ut med allerede er avsluttet før ekkoet kommer tilbake. Ved å analysere ekkoet som funksjon av tid, vil vi kunne bestemme avstander.

Og dersom vi kan sende lyd i vel definerte retninger, vil vi også kunne danne bilder av det som er inne i kroppen.

Det er mye nydelig fysikk i utformingen av lydproben ved ultralydundersøkelser. Vi kan styre strålen i to retninger ved å bruke interferens fra mange uavhengige lydgivere på overflaten til lydproben på en aktiv måte. Styring av lydstrålen foretas ved å endre fasen til lyden på en systematisk måte for hver enkelt liten transducer på ultralydproben. Fokusering for å redusere diffraksjon kan også gjøres ved liknende triks. Vi kommer tilbake til dette i senere kapitler.

Figur 7.3 viser et par ultralydbilder av et foster.



Figur 7.3: To ulike ultralydbilder av fostre. Til venstre er det et snittbilde (2D) av et 11 ukers gammelt foster. Til høyre er det et 3D bilde av et foster i første trimester. Bildene er gjengitt med velvillig tillatelse fra Vingmed.

Det bør legges til at det er store likheter mellom ultralydundersøkelser f.eks. av fostre, og kartlegging av havbunnen ved oljeleting. I det siste tilfellet brukes gjerne en rekke lydgivere (og mikrofoner) langs en lang kabel som slepes langs bunnen. Ekko fra ulike lag med ulik akustisk impedans er utgangspunkt for å finne ut hvor man kan forvente olje og hvor det ikke er olje og hvor dypt oljen ligger.

Mange fysikere her i landet, utdannet ved NTNU, UiO eller andre steder, har vært med på å utvikle utstyr for ultralydundersøkelser og seismiske undersøkelser. Firmaet Vingmed har vært verdensledende på utvikling av utstyr for ultralyddiagnostikk. På liknende måte har vi også vært helt i teten også ved seismiske undersøkelser. Det er mye morsom fysikk bak disse metodene, og det kommer nok til å dukke opp også andre anvendelser av disse prinsippene i årene som kommer. Kanskje du er en av dem som finner nye anvendelsesområder?

7.3 Stående bølger

Når en bølge vandrer langs en streng som er fast knyttet til en massiv gjenstand i en ende, vil bølgen bli reflektert fra endepunktet og vandre tilbake langs strengen i motsatt retning og med motsatt amplitud som den innkommende bølgen. Dersom bølgene varer ved, vil innkommende og reflektert bølge addere seg til hverandre (superposisjonsprinsippet). La den innkommende bølgen være en harmonisk bølge beskrevet på følgende form:

$$y(x, t) = A \cos(\omega t + kx)$$

for $x \geq 0$. Det vil si at bølgen kommer inn "fra høyre" (store x) og vandler mot origo. I origo er punktet der strengen knyttes til en massiv gjenstand, som fører til en reflektert

bølge gitt ved:

$$y_r(x, t) = -A \cos(\omega t - kx)$$

Vi har valgt å beskrive bølgene på den litt uvanlige måten for å sikre at amplituden i origo er nøyaktig motsatt av hverandre, slik at de to bidragene bestandig slokker hverandre ut i origo.

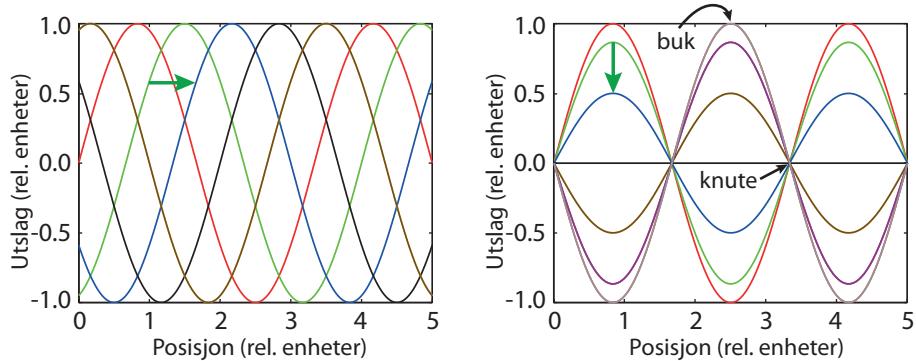
Totalt utslag er ifølge superposisjonsprinsippet summen av innkommende og reflektert bølge:

$$y_{sum} = A \cos(\omega t + kx) + (-A \cos(\omega t - kx))$$

for $x \geq 0$.

Fra ren matematikk vet vi at

$$\cos a - \cos b = -2 \sin\left(\frac{a+b}{2}\right) \sin\left(\frac{a-b}{2}\right)$$



Figur 7.4: En vandrende (venstre) og en stående bølge (høyre) som funksjon av posisjon ved en rekke ulike tidspunkt. Grønn pil viser hvordan bølgen endrer seg fra ett tidspunkt (grønn kurve) til et etterfølgende tidspunkt (blå kurve).

Anvender vi denne relasjonen for vår sum av en innkommende og en totalreflektert bølge på en streng, følger:

$$y_{sum} = -2A \sin(kx) \sin(\omega t) \quad (7.4)$$

I dette uttrykket har vi tatt hensyn til relativ fase, slik at utslaget er konstant lik null i endepunktet der refleksjonen foregår (origo i vår beskrivelse).

Det viktige med ligning (7.4) er at koblingen mellom posisjon og tid er brutt. Maksimalt utslag i en viss posisjon oppnås for de *tidspunktene* der $\sin(\omega t) = \pm 1$, og tidspunktene har ikke noe med posisjon å gjøre. Tilsvarende er *posisjonene* der maksimalt utslag forekommer ene og alene bestemt av ledet $\sin(kx)$, som altså ikke endrer seg med tiden. Disse karakteristiske trekkene er vist i figur 7.4.

MERK: I beskrivelsen har vi *ikke* lagt noe krav til de tre hovedparametrerne som må til for å beskrive en bølge: Amplitude, fase og frekvens. Uansett hvilke verdier vi velger for alle disse tre parametrerne, får vi stående bølger etter en total refleksjon som beskrevet ovenfor.

Derimot, dersom en streng er fastspent i *begge ender*, vil begge ender bli en knute (ikke noe utslag). I så fall må kx være lik f.eks. 0 i ene enden (slik som i vår beskrivelse) og $n\pi$ i andre enden (hvor n er et heltall).

En streng som er fastspent i *begge* ender, får med andre ord stående bølger som er “kvantisert” mhp frekvens (og bølgelengde). Kvantiseringen er et resultat av ren geometri i randbetingelsene. Dette er et generelt resultat som gjelder løsninger av bølgeligningen generelt. “Kvantisering” i kvantefysikken har ofte et lignende opphav matematisk sett. Som vi ser er det også på dette området analogi mellom klassisk bølgelære og kvantefysikk.

7.4 Musikkinstrumenter og frekvensspekter

Noen musikkinstrumenter, så som en tromme, gir transiente lyder, mens andre instrumenter gir mer eller mindre vedvarende “toner”. En tone kan karakteriseres som dyp/mørk eller høy/lys. Tonehøyden avhenger av frekvensen til grunntonen. Frekvensen kan bestemmes eksperimentelt f.eks. ved hjelp av fouriertransformasjon av et signal i tidsbildet.

Både strenge- og blåseinstrumenter er basert på at strengen eller “luftstrengen” har en bestemt lengde, og at vi får stående bølger hvor bølgelengden har et bestemt forhold til strengens eller luftstrenge lengde.

En gitarstreng vil danne stående bølger hvor strengens lengde er lik et helt antall halve bølgelengder.

I forrige kapittel ble det vist at bølgelengden λ multiplisert med frekvensen f er lik bølgehastigheten v , med andre ord:

$$v = \lambda f$$

Når så strengens lengde L er relatert til bølgelengden λ slik:

$$L = n \frac{\lambda}{2}$$

hvor n er et naturlig tall, følger det at strengen kan svinge med frekvensene

$$f = \frac{v}{2L} n$$

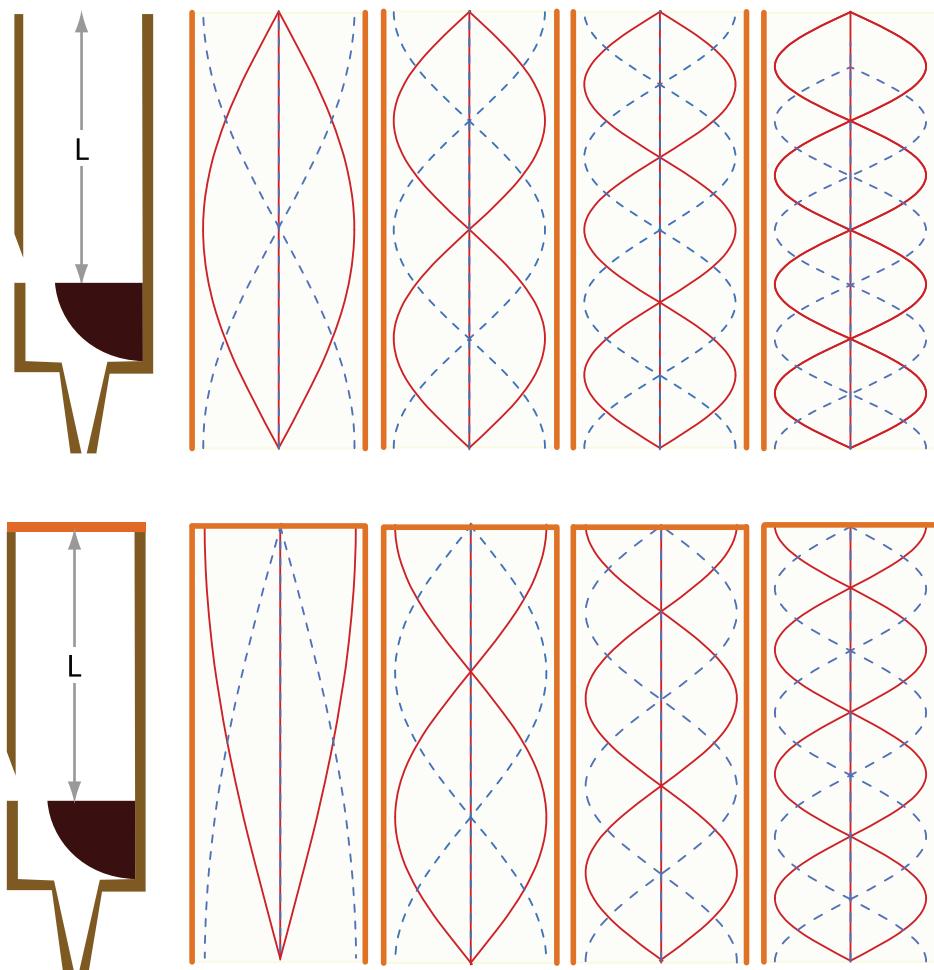
I enden av strengen har de stående bølgene en knute (null utslag), men strengen vil ha en eller flere buker alt etter hvilken “harmoniske” strengen svinger på.

Ved å klimpre på strengen på ulike steder langs strengen, kan vi framelske enkelte harmoniske lettere enn andre. For eksempel vil vi ikke framelske svingninger av strengen med to buker og en knute midt på dersom vi klimprer midt på strengen.

For blåseinstrumenter av typen blokkfløyte, tverrfløyte og åpne orgelpiper er det litt annerledes. Luften har maksimalt forflytnings-amplitude i bølgen nærmest munstykket (der lufta strømmer inn i selve luftstrenget i instrumentet). Luftstrenge ender åpent ut mot fri luft i motsatt ende i disse instrumentene, og endepunktet svarer omtrent til et nytt punkt med maksimum forflytnings-amplitude i de stående bølgene som dannes (se øvre del av figur 7.5).

Trykkbølgen inne i luftstrenget viser et forløp faseforskjøvet (i posisjon) en kvart bølgelengde sammenlignet med forflytningsbølgen. Trykkbølgen er null på stedet luftstrømmen går inn i luftsøylen og stedet luftsøylen går ut i åpent rom. I figuren er det vist fire eksempler på stående bølger som tilfredsstiller disse kravene. Det er henholdsvis 1, 2, 3 og 4 halve bølgelengder innenfor lengden av den vibrerende luftsøylen.

I en lukket orgelpipe derimot, er det en knute (null forflytnings-amplitude) for svingninger av luftmolekylene i den lukkede enden av luftstrenget og maksimal forflytnings-



Figur 7.5: En åpen (øverst) og en lukket (nederst) orgelpipe. I en åpen orgelpipe vil forflytnings-amplituden til luftmolekylene ha en buk (maks utslag) både der luften kommer inn og i den åpne enden av pipen (stiplet blå kurve). Trykkbølgen har imidlertid en knute ved start og slutt (heltrukket rød kurve). I en lukket orgelpipe vil forflytnings-amplituden til luftmolekylene ha en buk (størst utslag) der luften kommer inn og en knute (minst utslag) i den lukkede enden av pipen (stiplet blå kurve). Trykkbølgen har imidlertid minst amplitude ved start og maksimal amplitude i den lukkede enden (heltrukket rød kurve). Grensebetin-gelsene kan tilfredsstilles ved flere ulike bølgelengder. På figuren er svingningene vist for de fire lengste bølgelengdene (laveste frekvensene).

amplitude i motsatt ende (stiplet blå kurve i nedre del av figur 7.5). Trykksvingningene følger motsatt forløp, dvs maksimal trykkamplitude i den lukkede enden og null trykkamplitude i den åpne enden (rød kurve). I blåseinstrumenter så som en klarinett, obo, trompet, trekkbasun osv. er det ingen åpen spalt like ved der lufta blåses inn. For slike instrumenter ville vi kanskje teoretisk sett forventet maksimal trykkamplitude i den enden lufta kommer inn, og null trykkamplitude der lufta går ut av instrumentet. I praksis er det likevel ofte slik at forflytningsamplituden er størst ved munnstykket og trykkamplituden nær null, det vil si temmelig likt som for en åpen orgelpipe.

Ut fra figur 7.5, hvor grensebetingelsene er som beskrevet foran, er regelen for et instrument med åpen luftstreng i begge ender:

$$L = \frac{\lambda}{2}n$$

hvor n fortsatt er et positivt heltall.

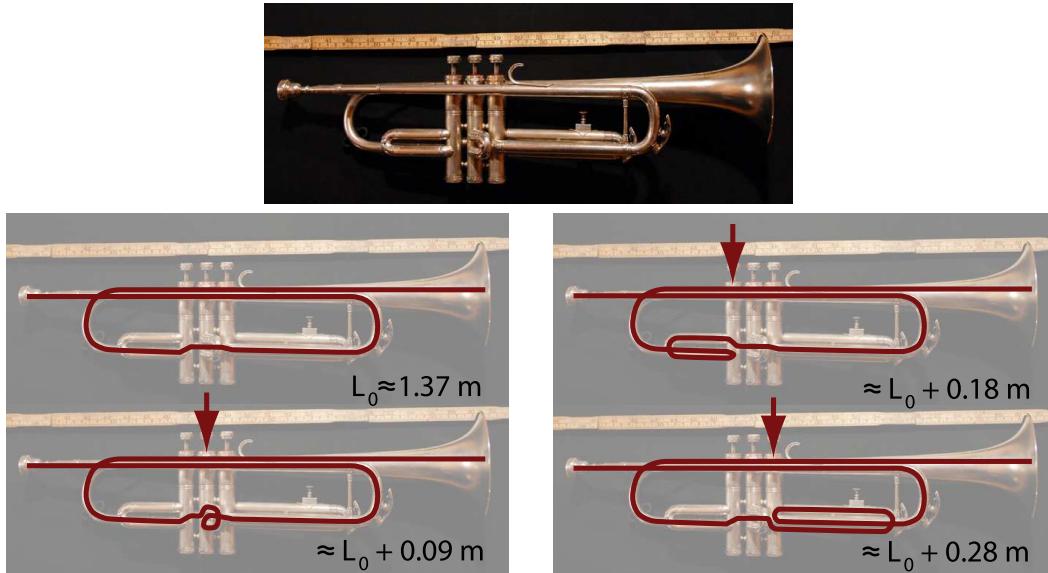
For et instrument med åpen luftstreng i den ene enden, men stengt i den andre, vil relasjonen i prinsippet bli:

$$L = \frac{\lambda}{2}n - \frac{\lambda}{4}$$

Bølgelengden for grunntonens (dvs $n = 1$) for en lukket orgelpipe er altså fire ganger lengen på pipen, men bare dobbelt så lang som pipen når den er åpen. Lydbølgen for grunntonens har altså en dobbelt så lang bølgelengde for en lukket orgelpipe som en tilsvarende lang åpen orgelpipe.

Rekken av frekvenser som kommer ut av et instrument med en lukket og en åpen ende vil ifølge relasjonene ovenfor være 1, 3, 5, 7 ganger grunnfrekvensen. For andre instrumenter vil frekvensene være 1, 2, 3, 4 ganger grunnfrekvensen.

Det er forresten interessant å se hvoran vi endrer tonehøyde (frekvens) på ulike instrumenter. For en gitar er det åpenbart at vi endrer lengden på den svingende delen av strengen. Siden strammingen stort sett ikke endres når vi klemmer en streng inn mot et bånd i halsen på gitaren, er hastigheten på bølgene uforandret. Dersom vi da reduserer lengden på strengen, vil bølgelengden også endres tilsvarende, og frekvensen gå opp ifølge sammenhengen $v = f\lambda$.



Figur 7.6: Luftsøylen i en trompet er svakt traktformet fra munnstykket til ytre åpning. Med ventiler kan lengden av luftsøylen endres. For en B-trompet (gruntonen er en B når ingen ventiler er trykket inn) er lengden på luftsøylen omtrent så lang som angitt.

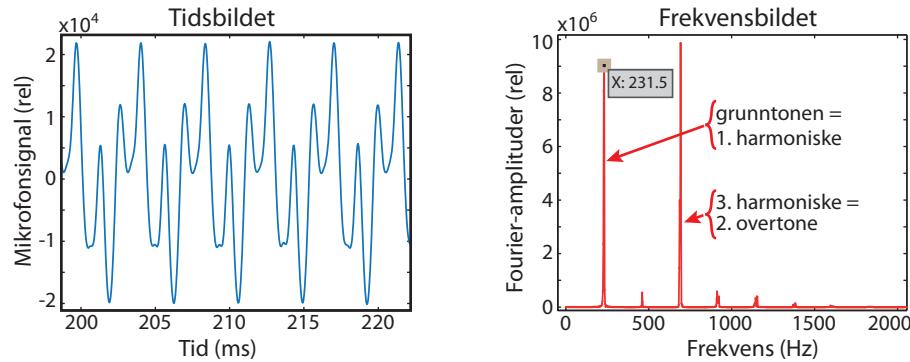
I messingblåseinstrumenter, så som en trompet, endres lengden på luftsøylen i instrumentet når ventilene trykkes inn. For en trompet går lufta en liten ekstra sløyfe når midterste ventil er trykket ned. Luften går om lag en dobbelt så lang ekstra sløyfe dersom bare første ventil trykkes ned og om lag tre ganger den ekstra sløyfen dersom bare den tredje ventilen trykkes ned. I figur 15.13 er målene for effektiv luftsøylepunktellengde gitt for ulike

enkeltventiler neddtrykket. Flere ventiler kan trykkes ned samtidig, og da blir den totale luftsøyleforlengelsen lik summen av alle de ekstra sløyfene som blir innkoblet.

[♣ ⇒ Det er mange andre detaljer i utforming av musikkinstrumenter. For eksempel vil intensiteten for de overharmoniske komponentene avhenge av hvorvidt lufrøret er smalt eller vidt, og hvorvidt lufrøret har samme diameter hele veien eller om lufrøret er traktformet. Moderne instrumenter sørger også for at lyden kommer ut i omgivelsene på en mer effektiv måte enn instrumenter laget for flere hundre år siden. Joe Wolfe ved The University of New South Wales i Sydney har meget informative websider som de interesserte kan studere videre (webadressen var i februar 2013 www.phys.unsw.edu.au/jw/basics.html).

Senere i boka brukes wavelet-transformasjon for å analysere lyd. Det vil da gå frem at fouriertransformasjon ofte gir et alt for “dødt” og lite nyansert bilde. I virkeligheten er det ikke slik at de harmoniske eksisterer med samme intensitet hele tiden. Intensitetsfordelingen for de harmoniske varierer, og det er litt av grunnen til at lyd fra virkelige musikkinstrumenter ofte har mer liv over seg enn syntetisk produsert lyd. ⇐ ♣]

I figur 7.7 er det vist et eksempel på en vedvarende lyd fra en trompet, både betraktet i tidsbildet og frekvensbildet. I dette tilfellet er grunntonen og høyere harmoniske til stede samtidig, og størrelsesforholdene mellom dem kommer fram i fourierspekteret (frekvensbildet).



Figur 7.7: Eksempel på tidsbilde og frekvensbilde av lyden fra en B-trompet hvor det spilles ”C” (som i virkeligheten er en B, se neste underkapittel). Det er opplagt at tidssignalet ikke er en ren sinus, men en blanding av flere. Frekvensspekteret viser nettopp dette. Merk at grunnfrekvensen er en del av harmoniske rekken, mens grunnfrekvensen ikke regnes med i nummereringen av såkalte ”overharmoniske”.

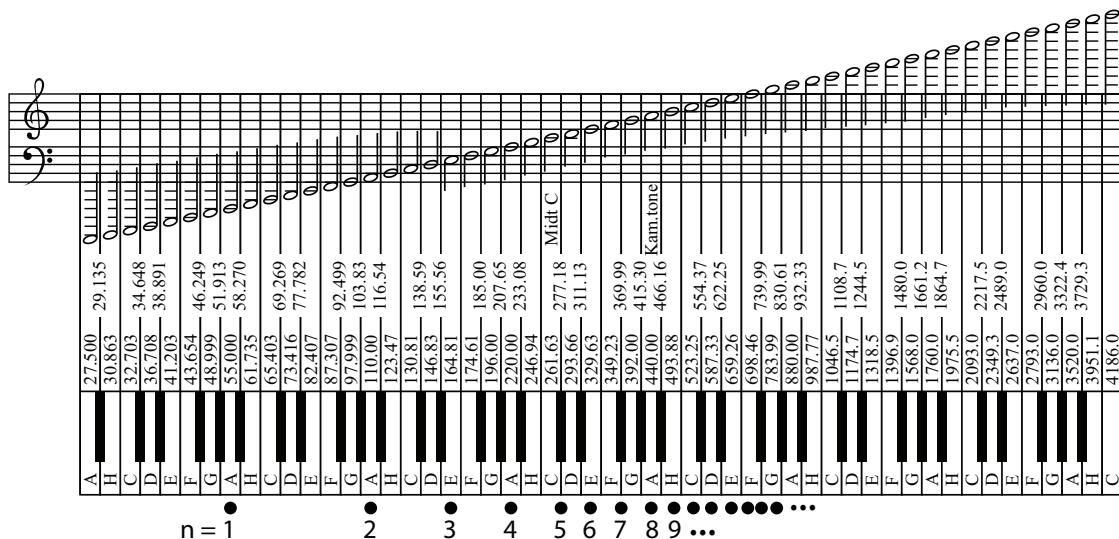
7.4.1 Toneintervaller

I vår kultur bruker vi toner i en skala bestående av 12 halvtoner som til sammen utspenner et frekvensområde hvor frekvensen øker med en faktor 2.0. Med det menes at for en tone C finnes det en ny tone C en oktav høyere, som har dobbelt så høy frekvens som den opprinnelige.

Halvtонene mellom er valgt slik at det er et konstant frekvensforhold mellom en tone og halvtonen under. Siden det skal være 12 slike trinn for å oppnå en oktav, følger det at forholdet mellom frekvensen til en tone og halvtonen under, må være

$$2^{1/12} \approx 1.0595$$

forutsatt at alle trinnene er like store. En skala definert på denne måten kalles *temperert*. Figur 7.8 viser frekvensene i en temperert skala dersom vi tar utgangspunkt i at enstrøken A skal ha frekvensen 440.00 Hz.



Figur 7.8: Tonene på et piano sammen med beregnet frekvens i en temperert skala. Figuren er tegnet på ny med grunnlag i en illustrasjon på: <http://amath.colorado.edu/outreach/demos/music/MathMusicSlides.pdf> nedlastet den 18. feb. 2012.

To toner fra f.eks. en violin kan lyde spesielt godt sammen dersom frekvensforholdet mellom dem er lik en heltallsbrøk. Forholdet mellom frekvensen til en E relativt til C-en under i en temperert skala er om lag 1.260. Dette er nært 5:4. Tilsvarende er frekvensen til en F relativt til C-en under lik 1.335, som er nært 4:3. Endelig kan vi nevne at forholdet mellom en G og C-en under er 1.4987 som er svært nært 3:2.

Det går an å lage en skala hvor tonetrinnene er nøyaktig lik heltallsbrøkene nevnt ovenfor. En slik skala kalles “renstemt”. Visse kombinasjoner av toner lyder da vakkre enn i en temperert skala, men ulempen er at vi ikke kan transponere en melodi (forskyve alle tonene med et gitt antall halvtonger) og beholde samme vakre klangen.

[♣ ⇒ I figur 7.8 er det laget noen interessante markeringer nederst. Dersom vi starter med en lav A med frekvens 55 Hz (n=1), vil første harmoniske (n=2) ha dobbelt så høy frekvens (110 Hz). Forskjellen mellom grunntonen og første harmoniske er da en hel oktav.

Andre harmoniske (n=3) vil ha frekvensen $3 * 55 \text{ Hz} = 165 \text{ Hz}$ som nesten svarer til en E, og tredje harmoniske (n=4) vil ha frekvensen $4 * 55 \text{ Hz} = 220 \text{ Hz}$, som er neste A. Det ble altså to harmoniske innenfor en og samme oktav.

Fortsettes det på samme måte, følger det at det er fire harmoniske innenfor neste oktav og åtte innenfor den etterfølgende oktaven. Med andre ord vil de høyere harmoniske etter hvert bli liggende tettere enn halvtongene ligger. Det er grunnen til at vi nesten kan spille en hel skala uten bruk av ventiler, ved å presse instrumentet til å gi lyd først og fremst ved de høyere harmoniske.

På en trompet oppnås grunntonen (som svarer til n=1) dersom leppene bare presses moderat hardt sammen. Frekvensen til grunntonen til lyden kan økes i ssprang (n øker) ved å stramme/presse leppene mer og mer. Luften som slipper gjennom leppene vil da komme i tettere småstøt enn om leppene er mer avslappet. ⇐ ♣]

I figur 7.7 så vi at frekvensen på grunntonen var ca. 231.5 Hz. Dette skulle være en B, og for de som kjenner toneskalaen vet vi at en B er halvtonen som ligger mellom A og H. Fra figur 7.8 ser vi at dette er som det skal være. Ved å variere litt på leppestramming kan tonen fra trompeten varieres en del (selv kan jeg variere frekvensen mellom ca 225 og 237 Hz for den aktuelle B-en). Gode musikere utnytter denne fintuningingen av tonehøyden når de spiller.

7.5 Svevelyd

Når vi lytter til to samtidige lyder med omtrent samme frekvens, kan det iblant høres ut som om *styrken* på lyden varierer på en regelbundet måte. Et slikt fenomen kalles “sveving” eller “svevelyd”. På engelsk kalles fenomenet “beat” fordi lyden liksom slår mot en i en fast takт.

Matematisk kan dette vises på omtrent samme måte som ved utledningen av uttrykket for en stående bølge. For vårt nye fenomen er det imidlertid ikke interessant å følge bølgens utbredelse i rommet. Det interessante er å vurdere hvordan lyden høres ut på ett sted i rommet.

Utgangspunktet er to svingninger som funksjon av tid og summen av disse:

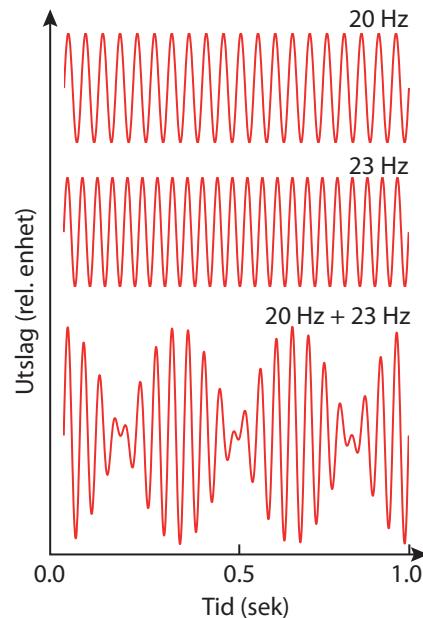
$$y_{sum} = A \cos(\omega_1 t) + A \cos(\omega_2 t)$$

En lignende sumformel som forrige gang gir:

$$y_{sum} = 2A \cos\left(\frac{\omega_1 + \omega_2}{2}\right) t \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t$$

Dersom de to (vinkel)frekvensene er omtrent like store, kan det innføres en middelverdi og differansverdi angitt som $\bar{\omega}$ og $\Delta\omega$ i formelen. Da følger:

$$y_{sum} = 2A \cos(\bar{\omega}t) \cos\left(\frac{\Delta\omega}{2}\right) t \quad (7.5)$$



Figur 7.9: Når to lydsignaler med nær samme frekvens adderes, vil intensiteten på lydsignalene variere i tid på en karakteristisk måte.

Dette er igjen et produkt av to ledd som er uavhengige av hverandre. For så små frekvensforskjeller at øret ikke klarer å skille dem i tonehøyde, vil det første leddet i ligning (7.5) svare til omtrent samme lydopplevelse som om hver enkelt av de to lydene var alene. Det siste leddet gir imidlertid en svingning med langt lavere frekvens enn de opprinnelige. Lytter vi for eksempel til to samtidige, omtrent like sterke lyder med frekvens

400 og 401 Hz, vil det siste leddet være et ledd av typen $\cos(\pi t)$. En gang per sekund vil dette leddet være lik null. Når det skjer, vil den totale lyden forsvinne. Lytter-opplevelsen er da en lyd omtrent som enkeltlydene når de høres hver for seg, men styrken på lyden vil sveve med en frekvens på 1 Hz. Det er denne pulsasjonen i lydstyrken som kalles sveving.

I figur 7.9 er det vist et eksempel på sveving. Der er det to signaler med henholdsvis 20 og 23 Hz som blandes, og vi følger hver av signaletene og summen over en tid på ett sekund. Vi ser at i sumsigalet er det tre “perioder” med kraftig og svak lyd innenfor det sekundet vi betrakter. Merk at i ligning (7.5) inngår *halve* differansen av frevensene som adderes. I vårt tilfelle svarer dette til 1.5 Hz. Hvorfor blir det likevel tre “perioder” i intensiteten i figur 7.9? Dette er en detalj du bør merke deg og forstå, for den sniker seg inn i flere ulike sammenhenger. (Hint: Hvor mange ganger er sinus lik null i løpet av én periode?)

I neste kapittel behandles en lignende summasjon av bølger. Summen blir da en “bølgegruppe”, et nyttig begrep når “fasehastighet” og “gruppehastighet” skal drøftes.

♣ ⇒ Det er flere pussigheter knyttet til svevelydfenomenet. Fourieranalyserer vi et signal beskrevet med ligning (7.5), får vi to topptopp i fourierspekteret som svarer til ω_1 og ω_2 , og KUN dette. Vi ser *ikke* noe signal svarende til differansfrekvensen. Hvorfor hører vi svevelyden da og ikke to separate toner?

Dersom de to frevensene er lengre fra hverandre, vil vi nettopp høre to separate toner, og ingen svevelyd. Hvor langt frevensene må ligge fra hverandre for at svevelyden skal forsvinne, er beskrevet i boka til akustikeren Tor Halmrast: “Klangen. Kompendium i lydlære 1 + 2”. Institutt for musikkvitenskap, Universitetet i Oslo, 2013. Kompendiet/boka kan du be om ved å sende en mail til torhalm@online.no. Boka til Tor Halmrast gir et vell av interessante opplysninger om lyd, musikk og akustikk! Mange av de temaene vi tar opp i dette kapitlet er beskrevet langt mer utførlig i Tor Halmrast’s bok.

Forøvrig anbefaler vi at du gjør følgende: Lag et signal som består av to sinuser med samme amplitude og frevensene hhv 100 og 110 Hz. Beregn signalet over minst hundre perioder for 100-Hz-signalet. Foreta en fouriertransformasjon og se på resultatet. Beregn så *kvadratet av signalet* ($(\sin(\omega_1 t) + \sin(\omega_2 t))^2$).*($\sin(\omega_1 t) + \sin(\omega_2 t)$) og foreta en fouriertransformasjon av *dette* signalet. Studer nøyde de linjene som da framkommer, og forsøk å finne et system i galskapen.

Dette lille numeriske eksperimentet er interessant fordi mange fysiske detektorer for bølgefenomener, egentlig ikke responderer på det momentane utslaget i bølgen, men på kvadratet av utslaget. Kan frekvensspekteret av kvadratet av signalet gi oss en indikasjon på hvorfor vi hører en svevelyd? ← ♣]

7.6 Lydintensitet

Lyd kan være så svak at vi ikke hører den, eller så kraftig at det gir smerteformennelse. Forskjellen ligger i lydens *intensitet*, og lydintensitet er definert som:

Lydintensiteten er tidsmidlet energi per flate og tid som transporteres i lydens retning.
Alternativt kan lydintensiteten defineres som den tidsmidlete energi per flate og tid som passerer en flate normalt på bølgens utbredelsesretning.
Måleenheten er watt per kvadratmeter: W/m².

Det er også mulig å operere med en “momentan” intensitet (i motsetning til tidsmidlet), men denne vil avhenge av både posisjon og tid. For lydbølger vil den lokale, momentane intensiteten være gitt ved:

$$I_m(\vec{r}, t) = \vec{p}(\vec{r}, t) \cdot \vec{v}(\vec{r}, t)$$

hvor \vec{p} er det lokale trykket (egentlig trykkforskjellen relativt til middelverdien), og \vec{v} er her *den lokale hastigheten til luftmolekylene* ved samme sted og tid (altså *ikke* lydhastigheten!).

♣ ⇒ Kommentar: En huskeregel kan være grei å ha: I stedet for å omtale mengde energi, kan vi fortelle hvilket arbeid lydbølgen har mulighet til å gjøre. Arbeid er kraft ganger vei, og kraften som kan virke på et tverrsnitt med areal A er det lokale trykket i lydbølgen multiplisert med arealet (egentlig trykkforskjellen som finnes i lydbølgen multiplisert med arealet).

Arbeid er “kraft ganger vei”, og dersom bølgen beveger seg en avstand Δx i en tid Δt , følger da:

$$\text{Momentan intensitet} = \frac{\text{Arbeid som kan gjøres}}{\text{Areal og Tid}}$$

$$I_m = \frac{p \cdot A \cdot \Delta x}{A \cdot \Delta t}$$

$$I_m = p \frac{\Delta x}{\Delta t} \approx pv$$

Så langt huskeregelen. $\Leftarrow \spadesuit$

I forrige kapittel ble en lydbølge blant annet beskrevet ut fra lokal forflytning η til molekylene, slik:

$$\eta(x, t) = \eta_0 \cos(kx - \omega t)$$

hvor η_0 er amplituden i luftmolekylene forflytning omkring et likevekspunkt (på toppen av Brownske bevegelser!). Bølgetallet k og vinkelfrekvensen ω må tilfredsstille:

$$v = \frac{\omega}{k} = \sqrt{\frac{K}{\rho}}$$

der v nå er lydhastigheten, K er kompressibilitetsmodulen og ρ er massetettheten.

Den samme bølgen kunne også beskrives som en trykkbølge slik:

$$p(x, t) = kK\eta_0 \sin(kx - \omega t)$$

Hastigheten til molekylene som deltar i bevegelsen er den tidsderiverte av forflytningsbølgen η :

$$\frac{\partial \eta}{\partial t} = \omega \eta_0 \sin(kx - \omega t)$$

Den momentane intensiteten er nå lokal hastighet til molekylene multiplisert med lokalt trykk. Bølgen antas å være longitudinal og den brer seg i x -retning, slik at to vektorer har samme retning. Følgelig:

$$I_m = p \frac{\partial \eta}{\partial t} = k\omega K \eta_0^2 \sin^2(kx - \omega t) \quad (7.6)$$

Den tidsmidlede intensiteten blir da:

$$I = \frac{1}{2} k\omega K \eta_0^2 = k\omega K \eta_{rms}^2 = 4\pi^2 \frac{K}{v} (f \eta_{rms})^2$$

siden tidsmidlet av \sin^2 er lik $1/2$. Her er η_{rms} lik root mean square utslaget for luftmolekylene, dvs $\eta_{rms} = \eta/\sqrt{2}$. [Minner igjen om at vi nå omtaler den kollektive forflytningen av molekylene på toppen av de mer “individuelle” Brownske bevegelsene.]

Det kan være nyttig å eliminere kompressibilitetsmodulen K og heller bruke forflytningsamplitude og trykkamplitude, sammen med massetetthet, lydhastighet, bølgelengde og frekvens. Med litt triviell manipulering av uttrykkene ovenfor, kan vi vise at:

$$I = \frac{(p_{rms})^2}{\rho v} \quad (7.7)$$

hvor p_{rms} er root mean square verdien av trykkvariasjonen, ρ er massetettheten i luft, og v er lydhastigheten i luft.

Videre kan det vises at:

$$I = 4\pi^2 \rho v (f \eta_{rms})^2 \quad (7.8)$$

hvor λ er bølgelengden til lyden i luft, dvs $\lambda = v/f$ der f er frekvensen til lyden.

Ligning (7.7) viser det interessante at lyd med ulike frekvenser vil ha samme intensitet dersom trykkamplituden er den samme.

Ligning (7.8) viser at lyd med samme intensitet, men ulik frekvens, har en forflytningsamplitude η_{rms} som er proporsjonal med bølgelengden.

Det er langt enklere å måle trykkamplituder enn forskyvning av molekyler. Derfor er ligning (7.7) den versjonen som kanskje brukes mest når lydintensiteter skal måles og angis.

[♣ ⇒ Før det gis eksempler på intensitetsverdier, returnerer vi til ligning (7.6) en kort stund. Ligningen viser momentanverdien for energitransport som funksjon av posisjon og tid. Uttrykket er bestandig positivt (siden $\sin^2 > 0$). Det er et viktig særtrekk for bølger! Molekylene som bringer bølgen framover svinger fram og tilbake, men middelposisjonen ligger fast og flytter seg ikke med bølgen. [Når vi ser bort fra Brownske bevegelser.] Likevel transporterer det energi fra bølgens kilde og utover, energi som normalt aldri kommer tilbake til kilden.

Det kan derfor være interessant å integrere opp all energi per tid som sendes ut fra kilden til bølgen. Det kan vi f.eks. gjøre ved å se på total energi per tid som går gjennom et kuleskall rundt bølgekilden. Enhet for en slik oppintegret intensitet er watt.

En menneskestemme yter ved normal samtale en total effekt på om lag 10^{-5} W. Hyler man, kan effekten komme opp i om lag $3 \cdot 10^{-2}$ W. Med andre ord er det ikke rare effekten som skal til for å få en brukbar lydbølge.

Tallene for menneskestemmen kan virke underlig når vi vet at et stereoanlegg gjerne kan gi effekter på 6 - 100 W. Nå er det riktignok slik at et stereoanlegg som anvendes ved 100 W normalt gir langt kraftigere lyd enn en menneskestemme kan yte. Likevel er forskjellen påfallende.

Årsaken til den store forskjellen kommer av at bare en liten del av effekten som tilføres høyttalerne omsettes til lydenergi: Bare noen få prosent for vanlige høyttalere. For spesielle horn-høyttalere kan effektiviteten komme opp i om lag 25 %. Resten av energien går til varme.

⇐ ♣]

7.6.1 Lydintensitet vs avstand og tid

Når lyd forplanter seg i luft, er det lite energi som blir borte underveis. Det betyr at omrent samme mengde energi som passerer et kuleskall med radius r_1 også vil passere et kuleskall lengre ute med radius r_2 . Den lokale lydintensiteten er mengde energi per flate og tid. Siden et kuleskall har en flate lik $4\pi r^2$, betyr det at intensiteten vil avta som $1/r^2$ hvor r er avstanden til kilden.

Nå er det sjeldent slik at lyden brer seg likt over et helt kuleskall. Avstanden til bakken er vanligvis betydelig kortere enn lydutbredelsen i horisontalplanet. Relasjonen

$$\frac{I(r_2)}{I(r_1)} = \left(\frac{r_1}{r_2}\right)^2$$

gjelder imidlertid rimelig bra også for begrensede romvinkler^a.

^aSå lenge interferensfenomener ikke spiller en vesentlig rolle

Det betyr at dersom vi på en konsert står 10 meter fra høyttalerne, vil intensiteten være 400 ganger større der enn for tilhørere 200 meter unna.

En lydpuls vil imidlertid dempes med tiden. Trykkbølgene fører til svingninger i gjenstander, og mange gjenstander har en innebygget friksjon der lydenergien blir omgjort til varme. Ulike materialer demper lyd mer eller mindre effektivt. En glatt betongmur settes ikke mye i svingninger av lydbølger, og lyden blir reflektert uten stort tap av energi. Vegger dekket med mineralull eller andre materialer som lettere settes i svingninger av lydbølger, kan dempe lyden mye mer effektivt.

Vegger og interiør i et rom kan føre til store forskjeller i demping. Derved påvirkes den såkalte "etterklangstiden". I Trefoldighetskirkken i Oslo, med mursteinvegger og lite

tekstiler, er etterklangstiden så lang at musikk med raske passasjer blir grøtete å lytte til, spesielt når det er få tilhørere. I et rom med mye tekstiler og møbler og mennesker i forhold til rommets volum, vil lyden dø ut betydelig raskere. I et ekkofritt rom er gulv, veggger og tak dekket av dempende materialer, og etterklangstiden er ekstremt kort. For konsertlokaler og teaterlokaler betyr det mye for den totale lydopplevelsen at etterklangstiden er tilpasset til de lydbildene som forekommer. Bygningsakustikk er en egen del av fysikken, der gode fagfolk er vanskelig å finne og derfor ettertraktet. Mange interessante detaljer finnes i boka til akustikeren Tor Halmrast: "Klangen. Kompendium i lydlære 1 + 2" nevnt til slutt i delkapitlet om svevelyd.

7.7 Desibel-skalaen

Lydintensitet kan angis i watt per kvadratmeter, slik som beskrevet ovenfor. Det er imidlertid en nokså uegnet målestokk. En grunn til dette er at menneskelig hørsel har en mer logaritmisk enn lineær respons. Med det menes at øret oppfatter endringer i lydstyrken ut fra prosentvis endring i forhold til nivået lyden allerede ligger på. Øker lydintensiteten fra 10^{-5} til 10^{-4} W/m^2 , oppfattes endringen omtrent like stor som om lydintensiteten økte fra 10^{-3} til 10^{-2} W/m^2 .

Det er derfor innført en logaritmisk skala for lydintensitet, den såkalte Desibel-skalaen. Lydintensiteten I relativt til en referanseintensitet I_0 er gitt i antall desibel på følgende måte:

$$\beta = L_I = (10 \text{ dB}) \log \frac{I}{I_0} \quad (7.9)$$

Enheten "bel" er oppkalt etter Alexander Graham Bell, som oppfant telefonen. "Desi" kommer fra 10-er faktoren som er lagt inn for å få enkle verdier å jobbe med. Desibel-skalaen brukes i mange deler av fysikken, ikke bare når vi beskriver lydstyrke.

I prinsippet kan vi velge hvilken som helst referanseverdi, og kan da f.eks. si at lydintensiteten 10 m unna høyttalerne i eksemplet ovenfor, er 26 dB høyere enn lydintensiteten 200 meter unna (sjekk at du forstår hvordan 26-tallet fremkommer).

I en del sammenhenger er det behov for å angi lydintensitet i en absolutt skala. Det kan oppnås ved å bruke en veldefinert referanseverdi angitt i et absolutt mål. For lyd brukes ofte følgende absolutte skala:

$$L_{Iabs} = (10 \text{ dB(SPL)}) \log \frac{I}{I_{abs.ref}} = (10 \text{ dB(SPL)}) \log \frac{p^2}{p_{abs.ref}^2} \quad (7.10)$$

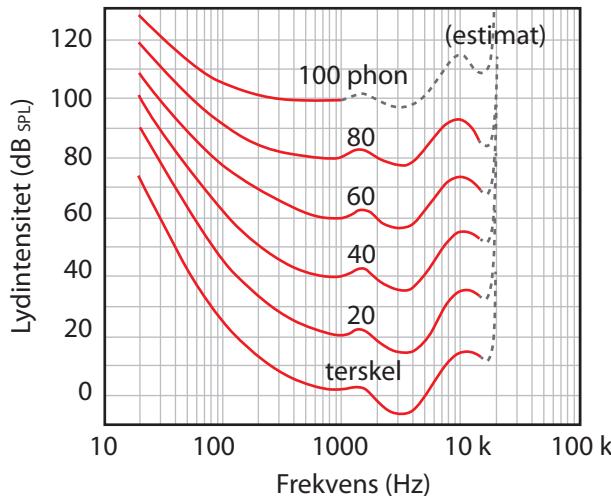
SPL står for sound pressure level, og referanseverdien er 1000 Hz lyd med lydtrykk $p_{rms} = 20 \mu\text{Pa}$ (rms). Dette lydtrykket svarer omtrent til en intensitet på 10^{-12} W/m^2 , og representerer omtrent den laveste intensiteten en 1000 Hz lyd kan ha for at et menneske skal oppfatte den. Dette tilsvarer omtrent lydintensiteten tre meter unna en flyvende mygg.

Overgangen fra intensiteter til kvadrater av lydtrykk følger av ligning (7.7).

I praksis sløyfes ofte betegnelsen SPL når lydstyrke angis. Dette er uheldig, for når vi f.eks. sier at lydintensiteten er 55 dB, er utsagnet i prinsippet ufullstendig, fordi referansen ikke er spesifisert. Hadde vi i stedet sagt at lydintensiteten er 55 dB(SPL), ville det gå fram at referansenivået er som angitt ovenfor, og at lydnivået da er spesifisert i en absolutt skala.

Det er imidlertid enda flere forhold vi må ta hensyn til når lydintensiteter skal angis.

Definisjonen i ligning (7.10) er først og fremst meningsfull for lyd med frekvensen 1000 Hz. Øret oppfatter ikke lyd med ulike frekvenser som like intense, selv om antall watt per kvadratmeter er uforandret. Vi har vanskeligere for å høre lave og høye frekvenser enn midlere frekvenser. Figur 7.10 viser lik-opplevd-lydstyrke konturer for ulike frekvenser, det vil si fysisk intensitet som må til for å gi samme opplevde intensitet når frekvensen varierer. Flere kurver er innegnet, for den relative endringen med frekvensen varierer noe med hvor kraftig lyden i utgangspunktet er.



Figur 7.10: *Lydnivåer ved ulike frekvenser som gir omtrent samme opplevd lydintensitet (se teksten). Figuren er en litt omarbeidet versjon av en som fantes på http://en.wikipedia.org/wiki/Equal-loudness_contour den 18. feb. 2012.*

Enheten phon angir lydstyrker for rene toner. 1 phon svarer til 1 dB(SPL) ved frekvensen 1000 Hz. Lydintensiteten som svarer til et gitt antall phon varierer svært mye med frekvensen til de rene tonene. Eksempelvis ser vi av figur 7.10 at en ren 20 Hz lyd med lydstyrke 100 dB(SPL) oppleves som like intens som en ren 1000 Hz lyd med 40 dB(SPL). Vi ser videre at lydintensiteten ved 100 Hz må ligge på om lag 25 dB(SPL) for at den skal være hørbar. Videre vil en lydintensitet på 40 dB(SPL) ved 1000 Hz svare til intensiteten 55 dB(SPL) for lyd med frekvensen 10000 Hz.

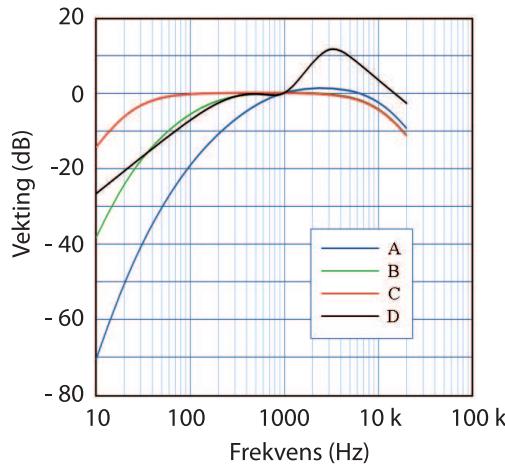
Kurvene er angitt av International Organization for Standardization (ISO), og er en oppdatert kurve fra 2003. Årstallet tyder på at det ikke er enkelt å fastlegge slike kurver så lenge det er betydelige individuelle variasjoner. Folk med tydelige hørselskader er nok ikke brukt ved fastsetting av kurver som dette!

Det sier seg selv at desibelskalaen slik den ble presentert i ligning (7.10) ikke kan brukes for å angi opplevd lydintensitet når lyden er sammensatt av flere frekvenser. Av den grunn er det innført enda flere intensitetsmål, men nå utformet slik at ulike frekvenser er vektet etter hvordan øret oppfatter lyd. Det finnes ulike vektfunksjoner, noe som gir opphav til dB(A)-skala, dB(B)-skala m.m. Figur 7.11 viser eksempler på de vanligste vektkurvene.

Kurvene viser at lave frekvenser ikke teller like mye som midlere frekvenser når dB(A)-mål skal fastsettes, sammenlignet med en ren dB-skala som definert i ligning (7.9).

♣ ⇒ Disse kurvene er slik å forstå: Anta at en lyd består av et rent 100 Hz signal og et rent 1000 Hz signal. Anta at begge bestanddelene hver for seg er like sterke i dB(SPL)-skalaen, f.eks. 80 dB(SPL) hver. Lydintensiteten for det sammensatte signalet ville da i en dB(SPL)-skala bli:

$$\begin{aligned} L &= (10 \text{ dB(SPL)}) \log \frac{p_{tot}^2}{p_{abs.ref}^2} = (10 \text{ dB(SPL)}) \log \frac{p_{100Hz}^2 + p_{1000Hz}^2}{p_{abs.ref}^2} \\ &= (10 \text{ dB(SPL)}) \log 2 \frac{p_{1000Hz}^2}{p_{abs.ref}^2} = 3 + 80 \text{ dB(SPL)} = 83 \text{ dB(SPL)} \end{aligned}$$



Figur 7.11: Vektkurver som brukes når vi skal angi opplevd lydstyrke i et signal som har mange ulike samtidige frekvenser. Kurvene gir opphav til dB(A)-mål, dB(B)-mål osv. Figuren er en noe omarbeidet versjon av en som fantes på <http://en.wikipedia.org/wiki/A-weighting> den 18. feb. 2012.

I en dB(A)-skala ville imidlertid beregningen se slik ut: Bidraget fra 1000 Hz signalet skal vektes med en vektfaktor 1.0, det vil si effektivt som 80 dB(SPL). Bidraget fra 100 Hz signalet skal imidlertid vektes med en faktor -20 dB, det vil si at vi må trekke 20 dB vekk fra de 80 dB lyden ville hatt i en dB(SPL)-skala, når den trekkes inn i en dB(A)-skala. 80 dB(SPL) svarer til at

$$\frac{p^2}{p_{abs.ref}^2} = 10^8$$

og 60 dB(vektet) svarer til at

$$\frac{p^2}{p_{abs.ref}^2} = 10^6$$

Totalt får vi da:

$$\begin{aligned} L &= (10 \text{ dB(A)}) \log \frac{p_{tot,vektet}^2}{p_{abs.ref}^2} = (10 \text{ dB(A)}) \log \left(\frac{p_{100Hz,vektet}^2}{p_{abs.ref}^2} + \frac{p_{1000Hz,vektet}^2}{p_{abs.ref}^2} \right) \\ &= (10 \text{ dB(A)}) \log(10^6 + 10^8) = 80.04 \text{ dB(A)} \end{aligned}$$

Lyden ved 100 Hz har med andre ord langt mindre å si for opplevd lydintensitet sammenlignet med lyd ved 1000 Hz.

$\Leftarrow \spadesuit$

Ofte ser vi tabeller med lydstyrker i ulike omgivelser, og et eksempel kan være:

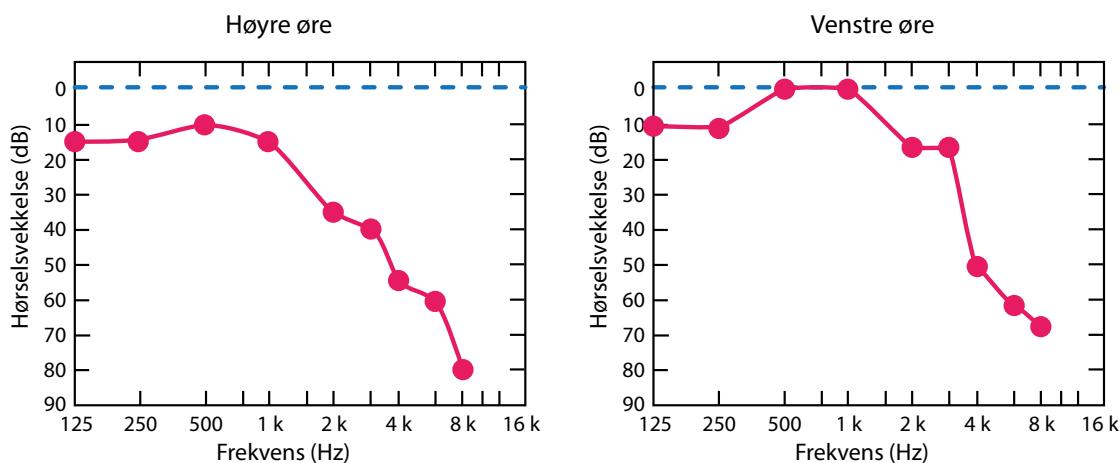
Hørselsgrensen ved 1000 Hz . . .	0 dB(A)
Hvisking	20 dB(A)
Svak radio hjemme	40 dB(A)
Samtale	60 dB(A)
Bytrafikk generelt	70 dB(A)
Kraftig musikk	100 dB(A)

Oftest angis det ikke hvilken dB-skala det virkelig er snakk om, men den mest vanlige i slike sammenhenger er dB(A)-skalaen. Egentlig er denne ikke den beste når lydstyrken er stor, så når lyd fra jetmotorer måles, brukes gjerne dB(D) i stedet. I prinsippet bør vi angi lydstyrker i dB(A), dB(B) etc i stedet for bare dB, først og fremst for å poengtere at tallene inngår i en absolutt skala, og at det er foretatt en vektning av bidrag fra ulike frekvenser slik at vi avspeiler *opplevd lydintensitet* og ikke bare et rent fysisk effektmål.

For store lydintensiteter vet vi at:

- 85 dB(A) gir hørselskader ved langvarig eksponering
- 120 dB(A) gir hørselskader ved akutt eksponering
- 130 dB(A) gir smerter ("Smertegrensen")
- 185 dB(A) gir vevskader.

Tall som dette varierer fra kilde til kilde, og må tas med en viss klype salt. Det er imidlertid helt klart at kraftig lyd kan ødelegge hårene i kontakt med basillarmembranen i det indre øret. Alt for mange angrer på at de har latt seg friste til å lytte til så kraftig musikk at hørselskaden ble permanent. Merk forøvrig at ved svært kraftig lyd, ristes rett og slett vanlig vev i kroppen i stykker slik at kroppen som sådan degenererer helt. Kraftig lyd er ikke å spøke med!



Figur 7.12: Eksempel på audiogram tatt opp hos en audiolog. Kurven viser aldersbetinget nedsatt hørsel hos en 60 år gammel mann. Se tekst for forklaring.

Vi kan få testet vår hørsel hos en audiolog, eller ved å bruke tilgjengelige dataprogrammer og datamaskinenes lydkort (men presisjonen er da ofte så som så). Ja, det finnes til og med programmer for iPhone for denne type test. Resultatet fra en hørseltest angis ofte som et såkalt audiogram, og et eksempel er gitt i figur 7.12. Et audiogram er laget slik at dersom vi har normal hørsel, skal audiogrammet være en horisontal, rett kurve på 0 dB-nivå (som den blå stippled linjen i figuren). Har personen nedsatt hørsel for enkelte frekvenser, vil kurven ligge under 0-nivået. Avstanden til nulllinjen angir hvor stor forskjell det er i følsomhet hos testpersonen ved den aktuelle frekvensen sammenlignet med normalen.

Figur 7.12 viser at personen som er testet har normal hørsel for 500 og 1000 Hz på venstre øret, men har hørseltap på alle andre frekvenser. Hørseltapet er hele 80 dB på høyre øret ved 8 kHz. Det betyr at personen praktisk talt er døv ved høye frekvenser. Dette er et eksempel på såkalt aldersbetinget nedsatt hørsel. Det er ikke rare at eldre har problemer med å forstå samtaler mellom folk, for det viktigste frekvensområdet i denne sammenheng er mellom 500 og 4000 Hz.

[♣ ⇒ Kommentar: Dere har tidligere jobbet med fouriertransformasjon av lyd. Dersom fourierspekteret med egnet kalibrering gir et mål for lydintensiteten ved ulike frekvenser, burde du ved hjelp av kurvene i figur 7.11 kunne regne deg fram til dB(A)-verdier, dB(B)-verdier m.m. Som du skjønner kan du lage ditt eget lydmålingsinstrument! (Men kalibrering må til!) ← ♣]

Til slutt defineres enda en dB-skala som er mye brukt i fysikk, nemlig dBm-skalaen. Dette er en absolutt skala hvor I_0 er valgt lik 1 mW. dBm-skalaen brukes i mange deler av fysikken, ofte knyttet til elektronikk, men sjeldent ved angivelse av lydnivå. Målet brukes gjerne for å angi f.eks. utstrålt effekt eksempelvis fra en antenn. Dersom en kilde gir fra seg 6 dBm, betyr det at utstrålt effekt er

$$10^{6/10} \text{ mW} = 4 \text{ mW}$$

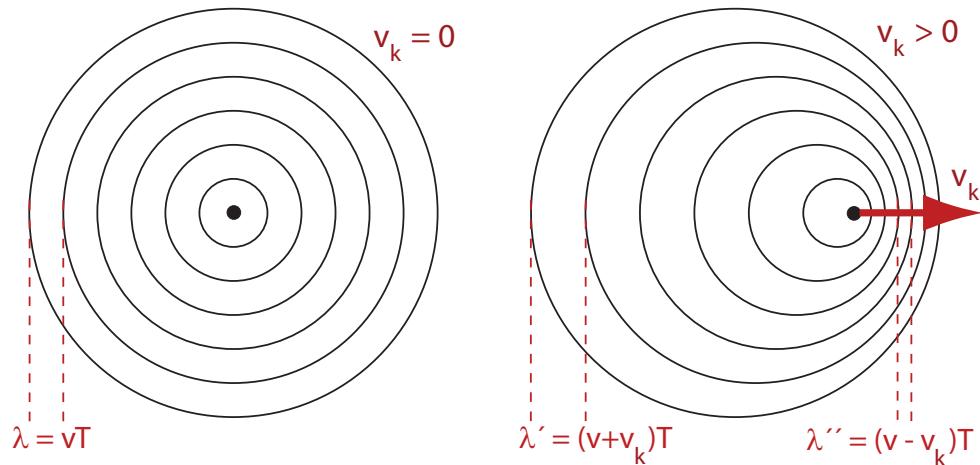
7.8 Doppler-effekt

De fleste av oss kjenner til at lyden fra en sykebil under utrykning endrer tonehøyde når bilen passerer oss. Fenomenet kalles *doppler-effekt*. Vi skal nå utlede et matematiske uttrykk for frekvensendringen vi observerer.

Lydbølger går med en viss hastighet *i forhold til mediet bølgen beveger seg i*. Uansett hvilken hastighet kilden har, og uansett hvilken hastighet en observatør har, går lydbølgene ufortrødent gjennom f.eks. luft med hastigheten $v = \sqrt{K/\rho}$ (størrelser definert tidligere).

Til venstre i figur 7.13 er det vist bølgefrontene karakterisert ved maksimum i lufttrykkbølgene fra en kilde som er i ro. Lyden brer seg jevnt ut i alle retninger, og så lenge kilden ikke flytter på seg, vil alle bølgefrontene ha samme sentrum. Til høyre i samme figur er det vist bølgefrontene når kilden til lyden har flyttet seg mellom hver gang en trykkbølge startet ut. Deretter går hver av trykkbølgene ufortrødent videre med lydhastigheten (f.eks. i luft).

Det betyr at en observatør som er plassert slik at lydkilden nærmer seg, vil oppleve at bølgetoppene kommer oftere (flere bølgetopper per sekund) enn om kilden var i ro. For en observatør som er plassert slik at lydkilden fjerner seg, blir det motsatt. Det betyr at frekvensen en observatør opplever, vil være forskjellig i de to situasjonene.



Figur 7.13: Lydbølger brer seg ut med samme hastighet i alle retninger i mediet lydbølgene går gjennom. Bølgetoppene ligger like langt fra hverandre dersom kilden er i ro i forhold til luften. Beveger kilden seg i forhold til luften med hastigheten v_k , ligger bølgetoppene tettere på den ene siden enn på den andre. Lydhastigheten er angitt som v .

Når observatøren står i ro i forhold til luften, vil lydbølgene strømme mot ham med lydhastigheten v . Når den effektive bølgelengden er som vist i høyre del av figuren, følger

det at frekvensen som observatøren hører, f_o er:

$$f_o = \frac{v}{\lambda_{eff}}$$

Når lydkilden med periodetid T og frekvens $f_k = 1/T$ nærmer seg observatøren med hastigheten v_k , følger:

$$\begin{aligned} f_o &= \frac{v}{(v - v_k)T} \\ f_o &= \frac{1}{1 - v_k/v} f_k \end{aligned} \quad (7.11)$$

hvor v er lydhastigheten i luft. For en observatør hvor kilden fjerner seg, blir minustegnet erstattet med pluss.

Denne versjonen av Doppler-effekt kan beskrives ved at bølgehastigheten i forhold til observatøren er lik lydhastigheten i luft, mens effektiv bølgelengde er forskjellig fra en situasjon hvor både kilde og observatør er i ro.

En annen vri av Doppler-effekten får vi når kilden står i ro, men observatøren beveger seg. Da er hastigheten til bølgetoppene relativt til observatøren forskjellig fra lydhastigheten i luft generelt. Bølgelengden er imidlertid uendret.

Frekvensen som observatøren da opplever, vil være proporsjonal med effektiv hastighet til bølgetoppene i forhold til observatøren, sammenlignet med hastigheten bølgene hadde nådd observatøren med dersom han og kilden stod i ro. For en stillestående kilde, og observatør i bevegelse med hastigeten v_o i retning mot kilden, følger da:

$$f_o = (1 + v_o/v) f_k \quad (7.12)$$

hvor f_k igjen er frekvensen til kilden.

Det er fullt mulig å kombinere de to variantene av Doppler-effekt vi har behandlet ovenfor, slik at vi får et mer generelt uttrykk som gjelder for situasjoner hvor både observatør og kilde er i bevegelse i forhold til luften der lyden brer seg utover.

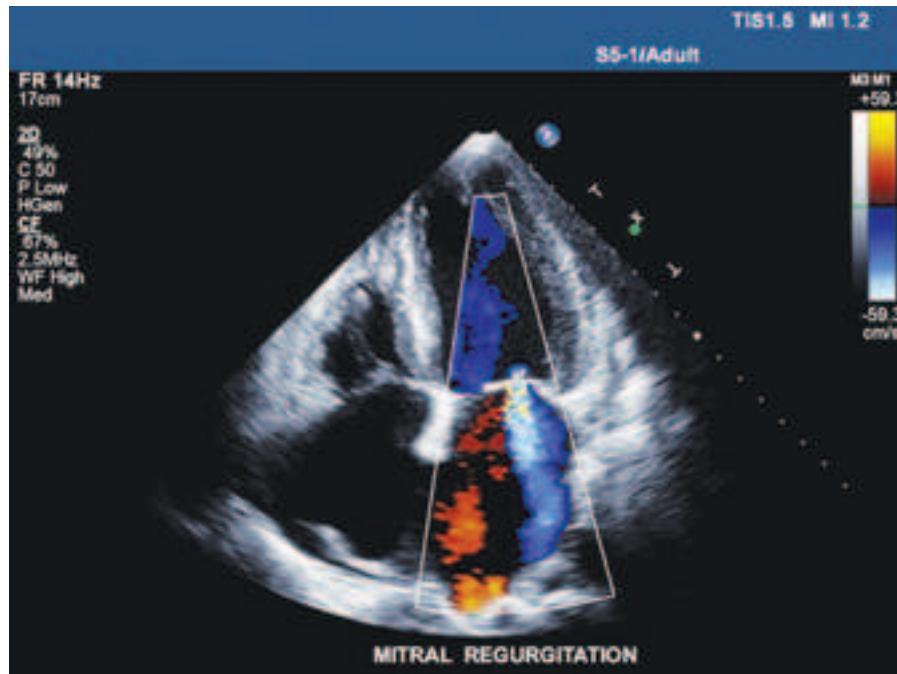
I ligning (7.12) kan frekvensen f_k erstattes med den frekvensen en observatør (indeks o) hadde opplevd dersom kilden (indeks k) var i bevegelse, altså med f_o fra ligning (7.11). Resultatet blir:

$$f_o = \frac{v + v_o}{v - v_k} f_k \quad (7.13)$$

Her er v lydhastigheten i luft (f.eks. 344 m/s). v_k og v_o er hhv kildens og observatørens hastighet relativt til luften lyden går gjennom. I likningen ligger det følgende fortegnskonvensjon: Dersom kilden beveger seg mot observatøren med en hastighet v_k relativt til luft, er v_k positiv. Dersom observatøren beveger seg i retning mot kilden med hastigheten v_o i forhold til lufta, er v_o positiv.

Merk at fortegnet er gitt ut fra relativ bevegelse mellom kilde og observatør slik som angitt ovenfor, mens selve storrelsen på hastighetene er angitt i forhold til luft (eller det mediet som lydbølgen brer seg gjennom).

Dopplerskift utnyttes i dag i ultralyddiagnostikk. I figur 7.14 er det vist et ultralydbilde av et hjerte, og i en viss sektor av bildet er det lagt til en markering i farger som viser om blodcellene beveger seg mot oss eller fra oss. Bildet viser en pasient som har en hjerteklaff som ikke lukker ordentlig når hjertekammeret trekker seg sammen.



Figur 7.14: Ultralydbilde av et hjerte, sammen med ultralyd-doppler markering av blodstrømmen i en viss fase av hjerterytmen. Bildet viser at hjerteklaffen ikke lukker seg ordentlig under hjertekammerkontraksjonen. Bildet er gjengitt med tillatelse fra Vingmed. Det er vanskelig å forstå et enkeltbilde som dette. Video av slike undersøkelser er ofte letttere å skjønne. En video som anbefales finner du på <http://library.thinkquest.org/05aug/01883/ultrasound3.htm> (per. 17. feb. 2013).

7.8.1 Doppler for elektromagnetiske bølger

Doppler-effekt for lydbølger preges av en konstant lydhastighet i forhold til mediet lyden går gjennom. For elektromagnetiske bølger er situasjonen en helt annen. Lyshastigheten er på en litt uforståelig måte knyttet opp til hele vårt rom/tidsbegrep, og lyshastigheten i vakum er den samme uansett hvilken hastighet kilden beveger seg med og uansett hvordan en observatør beveger seg. Når bølgelengder måles, observeres lengdekontraksjoner på grunn av relativistiske effekter, og tidsdilatasjon/kontraksjon som følge av relativistiske effekter. Utledningen av Doppler-effekt for elektromagnetiske bølger blir derfor noe mer komplisert enn for lyd, og vi nøyer oss med å bare gjengi det aktuelle uttrykket.

Dopplerskift for elektromagnetiske bølger i vakum er gitt ved:

$$f_o = \sqrt{\frac{c+v}{c-v}} f_k \quad (7.14)$$

Her er c lyshastigheten, og v hastigheten til *kilden relativt til observatør*, $v > 0$ dersom de to nærmer seg hverandre. f_k er som før frekvensen til bølgene som kilden sender ut.

Denne relasjonen viser at lys fra fjerne galakser vil observeres med en lavere frekvens dersom galaksene fjerner seg fra oss. Effekten er velkjent og går under navnet "rødforskyvning" i de observerte spektrene.

Rødforskyvningen er sterkt ved observasjon av lys fra fjerne galakser, siden disse (i tråd med Big Bang modellen for universet) beveger seg fra oss med stor hastighet. Effekten er så sterkt at deler av det synlige spekteret er forskjøvet inn i det infrarøde området.

Dette er én grunn til at romteleskopet James Webb har detektorer i det infrarøde området.

7.9 Sjokkbølger

Fra høyre del av figur 7.13 går det fram at trykkbølgene ligger tettere framfor en lydkilde som beveger seg relativt til luft enn om kilden hadde stått i ro. Det lå imidlertid implisitt en antakelse i figuren, nemlig at lydkilden aldri tar igjen lydbølgene den genererer. Med andre ord, lydkilden beveger seg med en hastighet mindre enn lydhastigheten i luft (eller mediet vi betrakter).

Hva skjer dersom lydkilden beveger seg *raskere* enn lydhastigheten? Forholdene blir da som vist i figur 7.15. For å komme fra tilfellet nederst i figur 7.13 til 7.15, må vi imidlertid gjennom en situasjon der kilden beveger seg akkurat like fort som lyden. I denne situasjonen vil trykkbølgene på forsiden av kilden bli liggende oppå hverandre, og vi kan få enorme trykkvariasjoner innenfor relativt korte avstander. Denne situasjonen kalles “lydmuren”.

Det skal betydelig energi til for å trenge gjennom trykkbølgjen vi kaller lydmuren. Intensiteten i sjokkfronten kan komme opp i 160-170 MW/m². Og kanskje enda viktigere: Gjenstanden som skal “gå gjennom lydmuren” må være kraftig bygget for å tåle påkjenningene når trykkvariasjonene blir meget store over objektet. Lydintensiteten i sjokkbølgjen er om lag 200 dB, slik at personer om bord i et fly som går gjennom lydmuren må skjermes betydelig for å ikke få varige skader.

Merk: Det er ikke lyden av motoren på flyet som gir opphav til sjokkbølgjen. Det er rett og slett trykkbølgjen som skyldes at flykroppen trenger seg fram gjennom lufta. Motorlyden kommer som et tillegg til denne hovedbestandelen til trykkbølgjen.

Lydhastigheten i luft angis gjerne som 340 eller 344 m/s. Omgjort til kilometer per time får vi ca 1230 km/t. Jagerfly kan fly raskere enn dette, og bryter da lydmuren på vei mot de høyeste hastighetene.

Hastigheten til supersoniske fly angis i antall Mach, hvor:

$$v \text{ målt i Mach} = \frac{v_{fly}}{v_{lyd}}$$

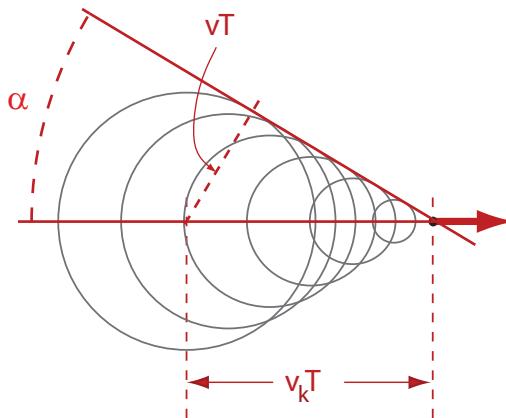
Concorde-flyet hadde en normal marsjhastighet over Atlanteren på ca 1.75 Mach, men hadde en toppfart på ca 2.02 Mach. Romferjen hadde en hastighet på opp mot 27 Mach. Husk forresten i denne sammenheng at lydhastigheten i tynn luft langt oppe i atmosfæren er forskjellig fra lydhastigheten ved bakken.

Fra figur 7.15 går det fram at sjokkbølgjen danner overflaten til en kjegle etter flyet som lager bølgen. Åpningsvinkelen i kjegleflaten er gitt ved:

$$\sin \alpha = \frac{v_{lyd} \cdot t}{v_{fly} \cdot t} = \frac{v_{lyd}}{v_{fly}}$$

Når et supersonisk fly flyr høyt oppe i lufta, vil flyet passere en observatør på bakken flere sekunder før observatøren hører lyden fra flyet. Først når sjokkbølgjen når bakken der observatøren er, vil hun/han høre flyet, og da gjerne som et lite smell idet trykkbølgjen på kjegleoverflaten passerer. Smellet som høres svarer derfor *ikke* til tidspunktet da flyet går gjennom lydmuren, men bare til tidspunktet sjokkbølgjen som danner seg når observatøren.

For Concorde-flyet hadde sjokkbølgjen et trykk på ca 50 Pa ved bakken når flyet fløy i en høyde på 12000 meter. Det var da lett å høre smellet fra sjokkbølgjen en kort tid etter at flyet hadde passert. På samme måte kunne vi i Los Angeles distriktsdistriktet høre et smell når romferjen kom inn for landing på landingsstripen i ørkenen litt nordøst for byen.



Figur 7.15: Bak et supersonisk fly danner det seg en sjokkbølge som har form som en kjegleflate med flyet i spissen. Vinkelen på kjegleflaten avhenger av hvor mye raskere flyet beveger seg enn lydhastigheten.

Historisk sett var det det amerikanske Bell X-1 rakettmotodrevne flyet som første gang brøt lydmuren. Det skjedde 14. oktober 1947. Flyet oppnådde da en fart på 1.06 Mach.

7.9.1 Eksempel: Helikoptere

Det er få som tenker på helikoptere når vi snakker om overlydhastighet, men det må vi faktisk gjøre. Et Black Hawk helikopter har rotorer som roterer om lag 258 ganger per minutt. Det tilsvarer ca 4.3 rotasjoner per sekund.

Rotorbladene har en lengde på 27 fot, som svarer til om lag 9 meter.

Hastigheten ytterst på rotoren for et stillestående helikopter (med rotoren i gang) er da:

$$\frac{2\pi r}{1/4.3} \text{ m/s} = 243 \text{ m/s}$$

Dersom helikopteret kjører med en hastighet på 100 km/t i forhold til lufta, vil hastigheten ytterst på rotoren i forhold til lufta bli 360 m/s på en side av helikopteret. Dette er omtrent lik lydhastigheten!

Konstruktører av helikoptere må faktisk balansere rotorhastighet og rotasjonshastighet og fart på en slik måte at vi unngår problemer med lydmuren. Det at hastigheten til ytterkanten av rotoren ikke holder samme hastighet i forhold til lufta gjennom en hel omdreining, gjør at vi har litt å gå på i forhold til et supersonisk fly.

Det kan forresten være artig å regne ut radiell akselerasjonen for et punkt ytterst på en helikopterrotor. Med utgangspunkt i tallene ovenfor, følger:

$$a_r = \frac{v^2}{r} = \frac{243^2}{9} \text{ m/s}^2$$

$$a_r = 6561 \text{ m/s}^2 \approx 670 \text{ g}$$

Det er med andre ord enorme krefter som virker på rotoren, og materialet må være feilfritt for å unngå ulykker. Det er ikke uvanlig at et rotorblad koster over 1 million kroner pr. stk.

7.10 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for generelle trekk ved refleksjon og transmisjon av bølger ved en grenseflate mellom to medier med ulik impedans.
- Gjøre rede for betingelser for å få stående bølger, og hva som karakteriserer slike bølger.
- Gjøre rede for hva som bestemmer tonehøyden for noen ulike musikkinstrumenter og hvordan vi kan oppnå ulike tonehøyder med ett og samme instrument.
- Kunne beregne frekvensen (omtrentlig) for en svingende streng og for et blåseinstrument.
- Gjøre rede for hva vi mener med et frekvensspekter, grunnfrekvens og harmoniske, når lyd analyseres ved hjelp av f.eks. fouriertransformasjon.
- Gjøre rede for en temperert skala og kunne beregne frekvensen til en hvilken som helst tone på et piano.
- Forklare hva som menes med en svevelyd, og kunne utlede et matematisk uttrykk som viser at sveving har noe med lydintensiteten å gjøre.
- Kunne beregne (når formler er oppgitt) amplitude for bevegelse av luftmolekyler og amplitude for trykkbølgen ved harmonisk lydbølge med angitt dB-verdi.
- Gjøre rede for dB, dB(SPL), dB(A) og dBm-skalaen.
- Gjøre rede for årsaken til Dopplerskift i ulike sammenhenger, kunne utlede formler som gjelder for Dopplerskift i luft, og kunne gjennomføre beregninger basert på disse formlene.
- Gjøre rede for sjokkbølger, spesielt "lydmuren" ved supersoniske fly o.l.

7.11 Oppgaver

Forståelses- / diskusjonsspørsmål

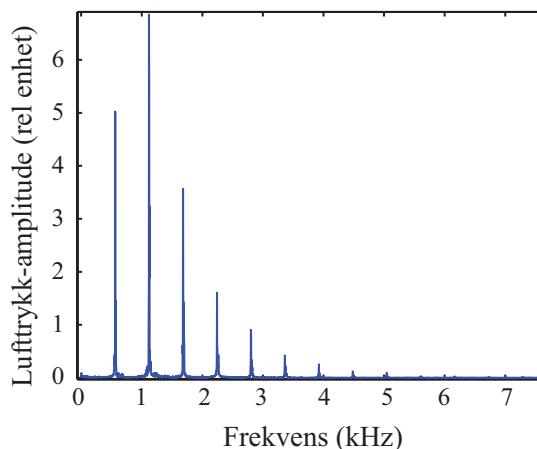
1. Ved ultralydundersøkelser av f.eks. et foster, må det være minst like mye reflektert lyd fra grenseflaten mellom livmorvegg og fostervannet som mellom fostervannet og fosteret. Hvorfor vil ikke reflektert lyd fra den første grenseflaten ødelegge bildet av fosteret?
2. Noen pianostemmere baserer stemmingen på en medbrakt frekvensteller alene. Mange mener at det ikke gir noe god stemming. Kan du gi en rimelig forklaring på en slik skepsis?
3. Forsøk å gi en verbal beskrivelse av hva som skjer fysisk sett idet vi *begynner* å blåse luft inn i en orgelpipe og helt til lyden blir stabil.
4. Vi kan lage en tone ved å blåse luft gjennom et rett rør. Ved å endre strammingen av leppene, kan vi få til ulike tonehøyder. Hvordan henger det sammen? Hvordan

er bølgemønsteret inne i røret ved noen av lydene som kan frembringes? Hvordan vil du tro at frekvensspekteret ser ut?

5. Kan vi få en stående bølge ved å addere to bølger som beveger seg i motsatt retning av hverandre, der den ene har større amplitud enn den andre, men samme frekvens? Kan vi få en stående bølge dersom vi adderer to bølger som beveger seg i motsatt retning av hverandre, der den ene har større frekvens enn den andre, men samme amplitud?
6. Er stående bølger alltid kvantisert? Forklar.
7. I musikken er en oktav karakterisert ved at frekvensen f.eks. til en høy C er dobbelt så stor som frekvensen til en C en oktav lavere. Dersom vi har en korrekt stemt gitar, og vi vil leke oss med å stramme en streng slik at den kommer en oktav høyere enn den normalt skal være. Hvor mye mer stramming må da til? [Er dette en selskapslek som kan anbefales?]
8. En fiolinist berører en streng så vidt på midten mens hun stryker buen over strengen. Hva oppnår hun med dette trikset?
9. Når lyd går fra luft til vann, hvilken av følgende størrelser holder seg konstant: Bølgelengde, bølgehastighet, frekvens, utslag (i posisjon) for molekylene som bringer lyden videre?
10. På en trompet kan vi spille forskjellige toner ved å trykke på ventiler som fører til at lufta går gjennom bøyler (av ulik lengde) som forlenger effektiv lengde på luftstrenget innenfor instrumentet. Hvordan kan vi spille forskjellige toner på et "posthorn" eller lignende instrumenter der vi ikke kan forlenge effektiv lengde? Kan vi spille samme type melodier på et slikt instrument som f.eks. på en trumpet?
11. Dersom vi inhalerer helium og prater, får vi en "Donald Duck stemme" som er lys og skrikende. Hva er grunnen til det? [Husk at å inhalere for mye helium kan gi helsekader og død, så vær forsiktig dersom du forsøker å teste ut dette i praksis!]
12. Når vi spiller på en akustisk gitar, blir klangen forskjellig alt etter om vi klimper på strenene helt ned mot tverrbåndet hvor strengen slutter sammenlignet med dersom vi klimper på strengen nær hullet i gitaren (eller enda nærmere midt på strengen). Hva er grunnen til forskjellen i tonekvalitet? Og hvordan vil du karakterisere forskjellen?
13. Går det an å si som så: Å legge til X dB i lyden svarer til å multiplisere intensiteten til den opprinnelige lydbølgen med et bestemt faktor?
14. Fortell kort hva som er forskjellen mellom dB, dB(SPL), dB(A) og dBm.
15. Ved en orgelkonsert merket en lytter seg at etter at organisten hadde avsluttet spillingen, tok det likevel noen få sekunder før lyden forsvant helt. Hva er grunnen til at lyden gradvis går mot null? Og hvor blir det av den energien som var i den opprinnelige lyden?
16. Anta at vi står stille et sted og hører en fabrikkpipe varsle at arbeidsdagen er slutt. Det blåser nokså friskt fra fabrikkpipa mot oss. Vi vil merke et Dopplerskift i lyden?
17. Innbyggere i Los Angeles kunne merke når romfergen var på vei mot landing i ørkenstrøkene litt nord-øst for byen. Forklar hvilket fenomen vi kunne basere oss på, og hvordan geometrien var når vi observerte dette fenomenet.

Regneoppgaver

18. En orgelpipe er 3.9 m lang. Orgelpipen er åpen i enden. Hvilken tone antar du at orgelpipen gir fra seg (sammenlign med figur 7.8).
19. Lengden på den fri delen av strengene på en gitar er 65 cm (dvs den delen som kan svinge). Klemmer vi ned G-strengen i femte båndet, får vi en C. Hvor må det femte båndet være plassert på gitarhalsen? G-en har en frekvens om lag 196.1 Hz og C-en om lag 261.7 Hz.
20. Bruk info og svar fra forrige oppgave. For hver halvtone vi går opp fra der vi er, må frekvensen øke med en faktor 1.0595. Beregn posisjonen til første båndet, og til sjette båndet. Er avstanden mellom båndene (målt i antall millimetre) identiske langs gitarhalsen? Vis at avstanden mellom båndene er gitt ved 0.0561 ganger lengden til strengen da den var klemt inn i forrige bånd.
21. Sjekk frekvensene som er angitt i figur 7.8. Dersom vi brukte lydanalyse vha. fouriertransformasjon for å bestemme frekvensen, hvor lang tid måtte vi da ha samlet lyden for å få en slik presisjon? Er dette en realistisk måte å bestemme frekvensen nøyaktig på? Er det mer realistisk å angi frekvensen med fem gjeldende siffer for de høyeste frekvensene?
22. Anta (foreløpig) at intensiteten til lyden som kommer fra et kor er proporsjonalt med antall sangere. Hvor mye kraftigere, angitt i en desibelskala, vil et kor på 100 korister lyde sammenlignet med et kor på fire personer (en kvartett)?
23. Figur 7.16 viser frekvensspekteret til en trompetlyd.
 - Angi frekvens og relativ trykk-amplitude for de fem første harmoniske.



Figur 7.16: Frekvensspekteret til en trompet.

- b) Hva er frekvensen til den femte overtonen?
- c) Anta at grunntonens lydintensitet lik 50 dB(SPL). Beregn lydintensiteten i dB(SPL) for hele trompetlyden (nok å ta med de første fire (eller fem) harmoniske).
- d) Beregn lydintensiteten i dB(A) for hele trompetlyden (nok å ta med fire harmoniske).
24. Anta at en person ligger på en strand og lytter til en CDspiller som står 1 meter fra hodet og at musikken har en intensitet på 90 dB. Hvor kraftig lyd vil en naboen på stranden som er fire meter unna høytaleren høre musikken? Dersom naboen klager på lydnivået, hva kan den første personen gjøre for å bedre på forholdet? Presenter gjerne en beregning som kan belegge ditt forslag.

25. To strenger på et instrument blir begge stemt til å svinge ved 440 Hz. Etter noen timer merker vi at de ikke lenger har samme frekvens, for vi hører en svevetone på 2 Hz når vi lar begge strengene svinge samtidig. Anta at en av strengene fortsatt svinger ved 440 Hz. Hvilken (hvilke) frekvens(er) kan den andre strengen ha? Hvor mye har strammingen endret seg på den strengen som har mistet sin stemming?
26. I denne oppgaven skal vi sammenholde lydintensiteter, forflytningsamplituder og trykkamplituder. Husk å kommentere resultatene du kommer fram til i hvert del-spøsmål!
- Hvor stor er forskyvningssamplituden for luftmolekylene når lydintensiteten er 0 dB(SPL) ved 1000 Hz? Gjenta samme beregning for lyd med intensitet 100 dB(SPL).
 - Hvor stor er lydtrykkamplituden (både i pascal og i atmosfæres trykk) når lydintensiteten er 0 dB(SPL) ved 1000 Hz? Gjenta beregningen for lyd med intensitet 100 dB(SPL).
 - Hvor stor er forskyvningssamplituden og trykkamplituden for lyd med frekvensen 100 Hz og intensiteten 100 dB(A) (!)?
 - Det er en øvre grense for hvor stort lydtrykkamplituden kan være dersom lydbølgen skal være tilnærmet harmonisk (sinusformet). Hvilken grense er dette? Hvor kraftig ville lyden være ved denne grensen (angitt i dB(SPL))?
27. Anta at du kjører bil i 60 km/t og hører at en politibil med sirener nærmer seg bakfra og kjører forbi. Du merker den vanlige endringen i lyd idet bilen passerer. Anta at politibilen kjører i 110 km/t og at den øvre frekvensen i sirenene har en frekvens på 600 Hz dersom vi hadde lyttet til sirenene i politibilen. Hvilke frekvenser opplever vi å høre før og etter at politibilen har kjørt forbi oss?
28. Anta at et jagerfly tar av fra Bodø flyplass og når 1.75 Mach allerede ved 950 m høyde. Hvilken vinkel har sjokkbølgen? Hvor lang tid tar det fra flyet passerer direkte over en person på bakken før personen merker sjokkbølgen? Se bort fra endringer i lydhastighet med høyden.
29. Ved en ultralydundersøkelse av et foster benyttes dopplereffekten for å måle hastigheten til hjertebevegelsen i fosteret. Lyden har en frekvens på 2.000000 MHz, men lyden tilbake har en frekvens på 2.000170 MHz. Hvor stor hastighet hadde den delen av fosterets hjerte hvor lyden ble reflektert fra, i den korte perioden der denne målingen ble foretatt. Lydhastigheten i fosteret er om lag 1500 m/s.
30. Krabbetåken er en gass-sky som kan observeres også med små teleskop. Den er restene etter en supernova som ble sett på jorda 4. juli 1054. Gass i de ytterste lagene av skyen har en rød farge som kommer av varm hydrogengass. På jorden har hydrogen-alfa-linjen $\text{H}\alpha$ en bølgelengde på 65562.82 Å. Når lyset fra Krabbetåken studeres, har $\text{H}\alpha$ -linjen en *bredde* på 56.942 Å.
- Beregn hvilken hastighet gassen i ytre del av Krabbetåken beveger seg med. [Anta at lyshastigheten er 3.0×10^8 m/s, og at relativistisk dopplerskift for elektromagnetiske bølger tilnærmet kan gis som $f_{\text{observ}} = (1 - v/c)f_{\text{kilde}}$ dersom kilden beveger seg med hastigheten v vekk fra observatøren.]
 - Anta at gassen i den ytre delen av skyen har beveget seg med samme hastighet helt siden supernovaen “gikk i lufta”. Estimer størrelsen av Krabbetåken slik den ser ut nå. Angi svaret både i meter og i lysår.
 - Vinkeldiametren til Krabbetåken når vi ser den fra Jorden er om lag 5 bueminutter. Et bueminutt er $1/60$ av en grad. Estimer avstanden (i lysår) til Krabbetåken.
 - Når fant egentlig eksplosjonen av stjernen sted (om lag).
 - I virkeligheten er ikke Krabbetåken sfærisk. Sett fra Jorda ser krabbetåken mer

elliptisk ut med største og minste vinkeldiameter på hhv 420 og 290 busekunder. Selv i dag kjenner vi ikke avstanden til Krabbetåken særlig nøyaktig. Kan du gi en god grunn til unøyaktigheten ut fra den beregningen du har foretatt?

31. En pianostemmer stemmer først alle tre C-strengene (som alle blir aktivisert av én tangent) slik at de får frekvensen 261.63 Hz. [Hun starter egentlig ut med en annen frekvens, men la oss ta dette utgangspunktet her.] Hun ønsker nå å stemme F-strengene ved å ta utgangspunkt i C og bruke "renstemming" der frekvensen til F er nøyaktig $\frac{4}{3}$ av frekvensen til C. Dette gjør hun for samtlige tre F-strenger som anslås når vi trykker på tangenten. Hun skjevstemmer så én av de tre F-strengene ved å lytte til svevelyd-frekvensen hun får når hun trykker på tangenten. Ved å stille inn svevelyd-frekvensen korrekt, oppnår hun at strengen får korrekt frekvens i en temperert skala (og kan justere frekvensen på de to andre F-strengene etter denne første). Hvilken svevelyd-frekvens må hun velge?
32. I figur 7.8 er det angitt at høy C har frekvens 523.25 Hz og høy F 698.46 Hz.
 - a) Hvor mange halvtonetrimm er det mellom disse tonene?
 - b) Er frekvensene i samsvar med hva vi vet om en temperert skala?
 - c) Hva ville frekvensen for F-en vært dersom det var en renstemt skala?
 - d) Bruk Matlab eller Python for å plotte tidsbildet av de to tonene (når vi ikke tar med overharmoniske), både for temperert og renstemt skala. Bruk gjerne samme ampliutde på de to signalene som adderes. Pass på at du tar med tilstrekkelig mange perioder til at forskjellen kommer godt fram.
 - e) Angi en "oppskrift" for hvordan en pianostemmer skal få korrekt temperert stemming av F-en dersom vi antar at C-en er korrekt allerede?
33. Bruk tallene for lengden av luftsøylen i en trompet gitt i figur 15.13 for å sjekke:
 - a) At grunntonen er om lag en B (angi ca frekvens).
 - b) Sjekk at den ekstra veilengden luftsøylen får når ventil 1 er trykket ned, svarer omtrent til en heltone sammenlignet med ingen ventiler trykket ned. Går frekvensen opp eller ned når vi trykker på en ventil?
34. Gjør en fouriertransformasjon frekvensanalyse av lyden fra to ulike musikkinstrumenter (ta opp lyd selv via mikrofon og lydkort på en PC, på en mobiltelefon, eller bruk wav-filer som gjøres tilgjengelig fra websidene til kurset vårt). Bestem frekvensen på lyden (grunntonen) og finn hvilken tone på skalaen den tilsvarer. Angi omtrent hvor mange harmoniske du finner.



Kapittel 8

Vannbølger og dispersjon



Illustrasjonen er et av Katsushika Hokusai's 36 berømte tegninger av bølger (med Mount Fuji i bakgrunnen).

Bølger på vann og hav har fascinert mennesker til alle tider. Det finnes et eventyrlig spekter av bølgeformer, og fysikken bak er så kompleks at det selv i dag er nesten umulig å gjøre beregninger på ville bølger som i denne illustrasjonen. De bølgene vi behandler i dette kapitlet er uhyre enkle til sammenligning. Likevel håper vi at selv de enkle beskrivelsene inneholder lovmessigheter du kan ha nytte av å kjenne til.

I dette kapitlet er det en del stoff som er "av orienterende art" som det er greit å kjenne til, men som vi behandler så overfladisk at vi ikke kan gjennomføre detaljerte beregninger på dem. Derimot presenterer vi lovmessigheter for gravitasjonsdrevne og overflatehinnedrevne overflatebølger på vann som vi kan bruke for å få fram artig fysikk. Blant annet kan dispersjon og forskjeller mellom fasehastighet og gruppehastighet belyses på en flott måte ved slike bølger. Vi avslutter med numeriske beregninger av bølgebevegelse. Bruker du litt tid på den beskrivelsen og gjennomfører noen numeriske beregninger, kan gevinsten bli en ganske dyp forståelse av det nydelige samspillet i tid og rom som ligger bak bølger.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 21022013.

8.1 Innledning

Før vi ser konkret på vannbølger, tar vi en kort rekapitulasjon om svingninger og bølger generelt. En fellesnevner for alle slike fenomen er at:

- Det finnes en likevekts tilstand for systemet når svingninger og bølger har dødd ut.
- Det finnes en “gjenopprettende kraft” (“restoring force” på engelsk) som forsøker å dra systemet tilbake mot likevekts tilstanden når det ikke er der.
- Det finnes en “treghetskraft” som gjør at systemet svinger forbi likevekts tilstanden selv om kraften her er lik null.

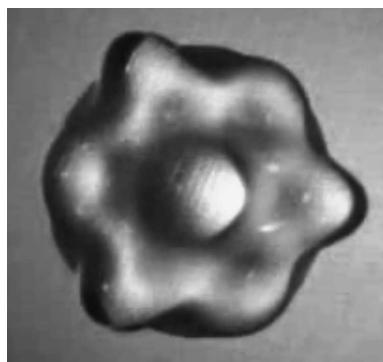
For en svingende pendel er det gravitasjon, og for bølger på en streng er det strammingen i strengen som er den gjenopprettende kraften. For lydbølger i luft eller væske er det trykkforskjeller som er gjennopprettende kraft pga sammentrykking av deler av volumet. “Treghetskraften” i alle disse eksemplene er den som kommer til uttrykk i Newtons første lov.

For overflatebølger på vann er det *to* gjenopprettende krefter, nemlig gravitasjon og overflatespenning.

Bølger på havet kjenner vi alle til. Mindre kjent er oscillasjoner i små vanndråper hvor overflatespenningen er den dominerende gjenopprettende kraften. Når en dråpe drypper fra en kran, vil den oscillere mens den faller. Eksempler på dette finner du i referanse 1 og 2 bakerst i kapitlet.

Når en vanndråpe ligger på en plate, og platen vibrerer, kan vi også få flotte oscillasjoner i vanndråpen. Eksempler på denne type oscillasjoner i en vanndråpe (som kan oppfattes som stående bølger i en vanndråpe) kan du finne i referanse 3 (videosnutt).

Samme fenomen kan vi observere når vi heller litt vann i gropen på en gammeldags elektrisk komfyr, forutsatt at platen er så varm at dråpen flyter på toppen av en luftpute (damp-pute) som danner seg. Vi kan få nydelige kvantiserte oscillasjoner der vi f.eks. har en trekantform som svinger fram og tilbake slik at totalbildet kan se ut som en sekskantet stjerne (se figur 8.1). Litt variasjon i varme eller størrelse på dråpen kan føre til at den nokså plutselig endrer svingemønster til å være en firkant der hjørnene skytes ut og trekkes tilbake på en slik måte at det hele minner om en åttekantet stjerne.



Figur 8.1: Et bilde av en oscillatorende vanndråpe. Tidsoppløsningen er så dårlig at vi på på samme bilde ser dråpen når den har trekantform med ene spissen mot venstre, og dråpen da den har trekantform med spissen mot høyre. Dråpen oscillatorer mellom disse formene. Bevegelsen kan anses som en stående bølge i dråpen. Bildet er hentet fra videoen i referanse 3.

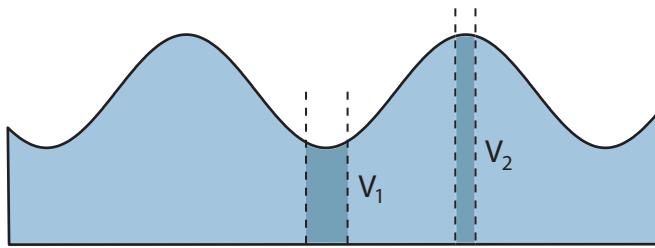
Hensikten med denne beskrivelsen er å minne om at klassisk fysikk er full av kvantiserte tilstander, på en analog måte til hva som finnes i atomær skala beskrevet i kvantefysikken.

Vi har allerede i tidligere kapitler sett andre eksempler på kvantisering i makroskopisk skala, så som svingninger på en streng, lydbølger i et musikkinstrument osv.

Grunnen til kvantisering ligger i randbetingelsene for systemet. For bølger på en gitarstreng kommer kvantiseringen av at amplituden i endepunktene må være lik null. Dette er helt analogt til kvantiseringer i kvantefysikken (f.eks. for en "partikkel i boks").

8.2 Bølgebeskrivelser

Vi skulle ideelt sett vist hvordan vi kommer fram til bølgeligningen også for overflatebølger i vann, men det er en såpass omfattende oppgave at den er utelatt i denne boka. Det henviser i stedet til bøker i hydrologi eller geofysikk. Vi vil likevel ta med noen detaljer. I figur 8.2 er det vist én mulig modell det er mulig å ta utgangspunkt i (modellen er utgangspunkt for utledningen i f.eks. boka til Persson, referanse 4 bak).



Figur 8.2: Én modell for bølger går ut på at vertikale volumsnitt langs bølgefronten beholder sitt volum uansett om vi er i en bølgedal eller bølgetopp. Overflaten til snittene vil endres, men det ser vi bort fra i første omgang.

Her tenker vi oss at et vertikalt volumelement parallellt med bølgefronten har samme volum uansett om det befinner seg i en bølgedal eller bølgetopp. I figuren vil det si at $V_1 = V_2$. Men siden trykket er lik lufttrykket over vannoverflaten (tilnærmet det samme over alle volumelement), og trykket øker med dybden inne i vannet, er trykket i en gitt høyde over bunnen høyere i volumelementet som svarer til bølgetoppen sammenlignet med det i bølgedalen. På denne måten kan vi anse bølgen som en longitudinal trykkbølge som beveger seg med bølgehastigheten.

I kapittel 4 ble lydbølger i luft og vann beskrevet som trykkbølger. Modellen i figur 8.2 ligner litt på den beskrivelsen, men er likevel ganske forskjellig!

For lydbølger anså vi gassen eller væsken som kompressible, det vil si at dersom vi øker trykket, vil volumet gå ned. Kompressibilitetsmodulen stod svært sentralt i utledningen. Gravitasjon var imidlertid ikke inni bildet overhodet.

Når overflatebølger på vann modelleres, ser vi helt bort fra kompressibiliteten. Vi antar tvert om at vannet er ikke-kompressibelt. Uansett trykkendringer beholder et volumelement samme volum.

I overflatebølger vil stort trykk svare til at volumelementet er trykt sammen på tvers av bølgefronten, det vil si i et volumelement under en bølgetopp.

Vi kan undre oss over hvorvidt det er rimelig å operere med helt forskjellige modeller for lydbølger og overflatebølger, og selvølgelig finnes det overgangssoner hvor disse beskrivelsene vil måtte gå over i hverandre. Imidlertid, det er fysisk sett gode grunner til å operere med ulike modeller.

Når det gjelder lydbølger er vi mest interessert i frekvenser i det hørbare området (og evt ultralyd). Det vil si fra ca 20 Hz og oppover. Periodetiden er 50 millisekunder eller mindre (til dels mye mindre). Dersom lyd skulle ført til overflatebølger slik vi skal beskrive i dette kapittelet, måtte vi forflytte betydelige vannmengder opp til flere meter i løpet av 25 millisekunder eller mindre! Det ville kreve enorme krefter (ifølge Newtons annen lov).

Derimot kan vi forflytte store vannmengder noen få mikrometer innen 25 millisekunder, og til og med enda kortere tider (høyere lydfrekvenser). De kreftene som skal til i denne sammenhengen er oppnåelige.

Overflatebølger på vann har en langt lavere frekvens (i alle fall for store bølgehøyder). Da får vi tid til å flytte store mengder vann fra en bølgebunn til en bølgetopp med de kreftene som er tilgjengelig.

Det er altså tidsskalaen og Newtons annen lov som medfører at vi opererer med helt forskjellige modeller for lydbølger i vann og gravitasjonsdrevne overflatebølger på vann.

En bedre modell

Modellen gitt i figur 8.2 gir ikke noe god beskrivelse av overflatebølger når alt kommer til alt. For bedre beskrivelse tar vi gjerne utgangspunkt i en av basisligningene for fluidmekanikk, nemlig Navier-Stokes ligning:

$$\rho \left(\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} \right) = -\nabla p + \nabla \cdot \mathcal{T} + \vec{\mathcal{B}}$$

hvor ρ er massetetthet, \vec{v} er strømingshastigheten, p hydrostatisk trykk, \mathcal{T} er en stressfaktor (kan for eksempel inkludere overflatespenning), og $\vec{\mathcal{B}}$ er volumkrefter ("body forces" på engelsk) som virker per enhetsvolum i væsken. ∇ er del-operatoren.

Det kan være nyttig å se nøyne på Navier-Stokes ligning og gjenkjenne at den i det store og hele er en videreutvikling av Newtons annen lov for et veske-kontinuum.

Navier-Stokes ligning er ikke-lineær, noe som betyr at løsninger av denne ligningen ikke nødvendigvis følger superposisjonsprinsippet. Dersom to funksjoner hver for seg er løsninger av ligningen, vil ikke nødvendigvis summen av disse funksjonene være løsning av ligningen. Et annet karakteristisk trekk ved ikke-lineære ligninger er at de kan ha kaotiske løsninger, det vil si løsninger hvor vi ikke kan forutsi hvordan løsningen vil utvikle seg i tid (på en rent deterministisk måte). Enhver liten endring i initialbetingelser eller randbetingelser, vil kunne føre til at løsningen etter en tid vil kunne ha svært forskjellige verdier. Dette kalles gjerne "sommerfugleffekten". En sommerfugls bevegelse kan føre til at værutviklingen etter en stund er helt annerledes enn om sommerfuglen ikke hadde fløyet som den gjorde.

[♣ ⇒ Det er en del morsomme rent matematiske utfordringer knyttet til Navier-Stokes ligning den dag i dag, men det skal vi ikke ta opp her.

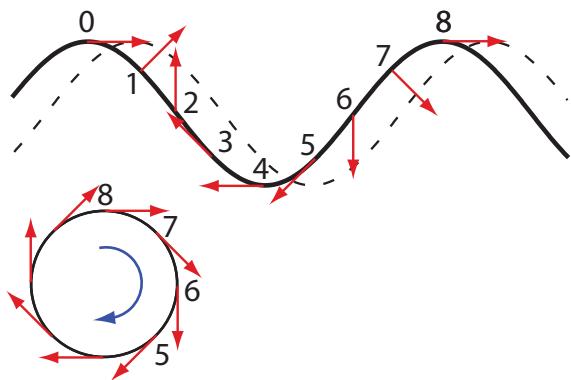
Hoverpoenget mitt er å påpeke at det er et vell av ulike fenomener knyttet til bevegelse i fluider, og utrolig mange fysikere og matematikere har vært interessert i vannbølger. Her kan nevnes Newton, Euler, Bernoulli, Laplace, Lagrange, de la Coudraye, Gerstner, Cauchy, Poisson, Fourier, Navier, Stokes, Airy, Russel, Boussinesq, Koertweg, de Vries, Zabusky, Kruskal, Beaufort, Benjamin, Feir og flere. Vi snakker om monsterbølger, tsunamier, solitære bølger osv. Her er det en lang tradisjon, også i norske forskningsmiljøer, og fortsatt masse å ta fatt i!

I vår tid har datamaskinene blitt så slagkraftige, og det er utviklet så mange numeriske metoder for bruk innen matematikk og fysikk, at vi nå kan ta tak i bølgebeskrivelser på en helt annen måte enn vi kunne for få decennier tilbake. Som et eksempel på den utviklingen som har foregått, kan det nevnes at professor Ron Fedkiw (født 1968) som arbeidet med Computer Sciences ved Standford University, fikk en Oscar-pris i 2008 for sitt arbeid med å animere realistiske vannbølger for bruk i filmindustrien (bl.a. i filmen "Poseidon"). For dere som er studenter i dag og som blir vant til å bruke numeriske metoder i løsning av matematiske og fysiske problemstillinger, er jo dette ekstra morsomt. Dere vil etter studiet ha ferdigheter som gjør at også dere med overkommelig innsats kan ta tak i realistiske animeringer av liknende omfang som Ron Fedkiw! (Interesserte kan lese mer om Ron Fedkiw på Wikipedia, hvor det også er lenker til Oscar-utdelingene og til Fedkiw's hjemmeside). ← ♣]

8.2.1 Enkel bølgebeskrivelse

La oss nå gi en tegner-og-forteller beskrivelse av selve bølgene. Figur 8.3 viser et vertikalt snitt på tvers av bølgefrontene. Den heltrukne bølgen viser bølgen i ett øyeblikk, den stiplede bølgen viser bølgen en kort tid etter. Bølgen beveger seg altså mot høyre.

I figuren er det tegnet inn piler som viser hvilken retning vannet må bevege seg for å komme fra bølgen slik den er nå, til slik den skal bli. Pilene i øvre halvdel er nokså naturlige å forstå, mens pilene i nedre halvdel kanskje er vanskeligere å henge på. Vi husker imidlertid at bølgen tross alt ikke medfører en netto transport av vann i bølgens retning, følgelig må vann som beveger seg forover i en del av bølgen, bevege seg bakover i en annen del av bølgen. Og vann som beveger seg oppover i en del av bølgen må bevege seg nedover i en annen del. Tar vi utgangspunkt i disse kjennsgjerningene, faller pilenes retning på plass rimelig greit.



Figur 8.3: Øvre del indikerer hvilken retning vannet i overflaten beveger seg når bølgen ruller mot høyre. I nedre del er det tegnet inn posisjon og hastighet til ett og samme volumelement idet en bølgetopp passerer. Det aktuelle bølgeelementet er det som er ved posisjon 8 i starten, men som i neste øyeblikk befinner seg i på den delen av bølgen som er inndikert med punkt 7 i øvre del. I neste øyeblikk har den en plassering i bølgebildet som svarer til punkt 6, osv. Resultatet er at volumelementet vi følger synes å bevege seg med i urretningen etter som tiden går.

Merk at vannet må bevege seg både langs bølgens utbredelsesretning og på tvers av denne. Det betyr at bølgen er en blanding av en longitudinal og en transversal bølge.

Tegner vi inn bevegelsesretning og relativ posisjon for ett og samme lille volumelement ved ulike tidspunkt mens en bølgetopp passerer, får vi et diagram som i nederste del av figuren. Det synes altså som om vannet i overflaten beveger seg langs en vertikal sirkel på tvers av bølgefronten.

Lenger ned i vannet vil sirkelbevegelsen gå over fra å være nær sirkelformet (som i overflaten) til å bli en mer og mer flatklemt ellipse, som vist i figur 8.4. Helt nede ved bunnen er bevegelsen nesten en ren horizontal bevegelse fram og tilbake. Det kan vi se når vi snorkler på bunnen av sjøen. Tare og andre vekster svinger litt dovent fram og tilbake etter som bølger passerer på overflaten.

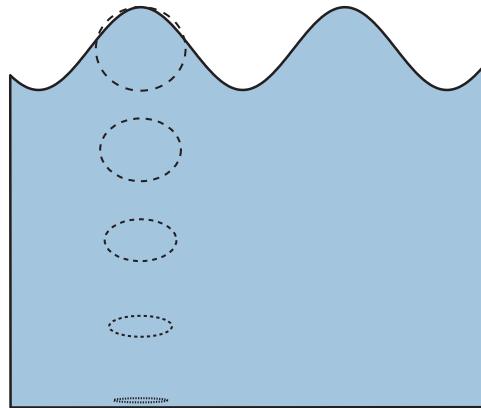
Denne beskrivelsen gjelder imidlertid bare for grunt vann, det vil si for vann som ikke er svært dypere enn bølgelengden mellom bølgetoppene.

For dypere vann vil bølgene på overflaten bare forplante seg et stykke nedover, men nær bunnen merkes ikke bølgene på overflaten overhodet.

Det er mulig å se den sirkulære bevegelsen ved å sprute små dråper med farget olje inn i vannet, forutsatt at massettettheten til disse dråpene er omtrent som for vann. Vi kan da i vannbølgetanker, som i kjelleren på Abels hus ved UiO og i Sintefs bølgetanker,

følge bevegelsen til dråpene. Jeg er imidlertid blitt fortalt at det er vanskeligere å vise disse sirkulære bevegelsene enn vi får inntrykk av gjennom lærebøker.

Når vi tegner inn sirkler og ellipser i ulike dyp, er det også en beskrivelse som lett kan misoppfattes. Hvordan skal vi se for oss sirklene og ellipsene for etterfølgende volumelementer i bølgeretningen? Her må det være en form for synkronisering som ikke kommer fram av figuren og som nødvendigvis må gi en mer utfyllende beskrivelse enn enkle skisser viser.



Figur 8.4: Når vi skal inndikere hvordan vannet beveger seg mellom overflaten og bunnen, brukes enkle skisser som denne. Imidlertid gir skisser som dette et vel enkelt bilde av hva som skjer.

Sinusformen er forresten ikke den beste modellen for overflatebølger på vann. Ofte er bølgetoppene spissere enn bunnene, slik som indikert i figur 8.5. Jo høyere amplitude, desto spissere blir bølgetoppen. Det er imidlertid en grense for denne utviklingen. Når bølgetoppen blir større enn om lag 1/7 av bølgelengden, blir bølgen gjerne ustabil og kan f.eks. gå over til en brytende bølge. Ved grensen er vinkelen mellom oppadgående og nedadgående del av bølgetoppen om lag 120 grader (en vinkel som selvfølgelig ikke gjelder helt inn på selve toppunktet).

8.2.2 Fasehastigheten til vannbølger

Selv om vi ikke har vist hvordan bølgeligningen faktisk vil se ut for overflatebølger, kan vi sette opp et tilnærmet uttrykk for én side av løsningene, nemlig fasehastigheten til vannbølger. Uttrykket er:

$$v_f^2(k) = \left[\frac{g}{k} + \frac{T k}{\rho} \right] \tanh(kh) \quad (8.1)$$

hvor k er bølgetallet, g tyngdens akselerasjon, T overflatespenningen, ρ massetettheten og h dybden på vannet. Formelen gjelder for en praktisk talt flat bunn (sammenlignet med bølgelengden).

Det første ledet inni parentesen angir gravitasjonens bidrag til den gjenopprettende kraften, mens siste ledet angir overflatespenningens bidrag. Første ledet svarer altså til såkalte "gravitasjonsdrevne bølger", mens siste ledet svarer til hva vi kaller "kapillærerbølger".

Siden bølgetallet k inngår i nevneren i det ene ledet og i telleren i det andre, betyr det at gravitasjonsleddet vil dominere for små bølgetall (dvs for lange bølgelengder), mens overflatehinne-leddet vil dominere for høye bølgetall (små bølgelengder). Det kan være

interessant å finne den bølgelengden der de to leddene er omtrent like store. Vi setter da:

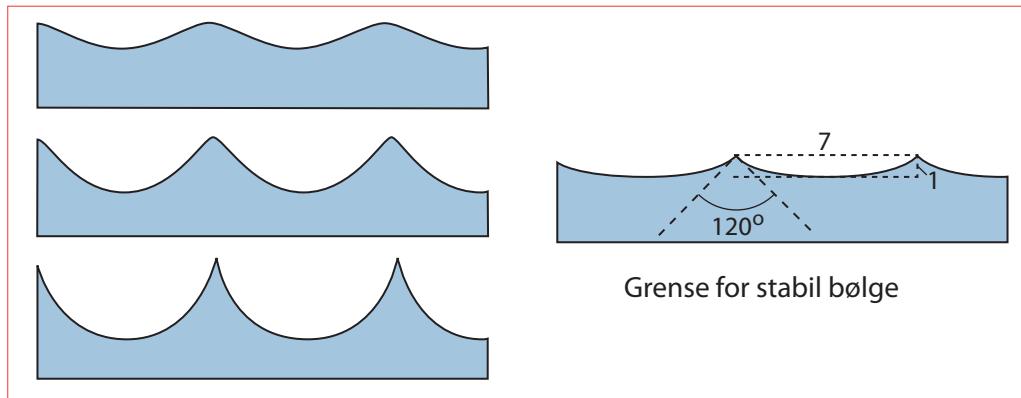
$$\frac{g}{k_c} = \frac{T k_c}{\rho}$$

Indeksen c indikerer et "kritisk" bølgetall hvor de to bidragene er like. Resultatet er:

$$\frac{1}{k_c^2} = \frac{T}{g\rho}$$

Siden $k = 2\pi/\lambda$ følger endelig:

$$\lambda_c = 2\pi \sqrt{\frac{T}{g\rho}}$$



Figur 8.5: *Bølgeformen er gjerne slik at toppen er spissere enn bunnen. Effekten blir tydeligere etter som amplituden øker. Når topp-til-topp amplituden er 1/7 av bølgelengden, får vi en grenseverdi hvor ytterligere økning av amplitudene ofte gir ustabil bølge.*

For vann ved omtrent 1 atmosfære, får vi verdien:

$$\lambda_c \approx 1.7 \text{ cm}$$

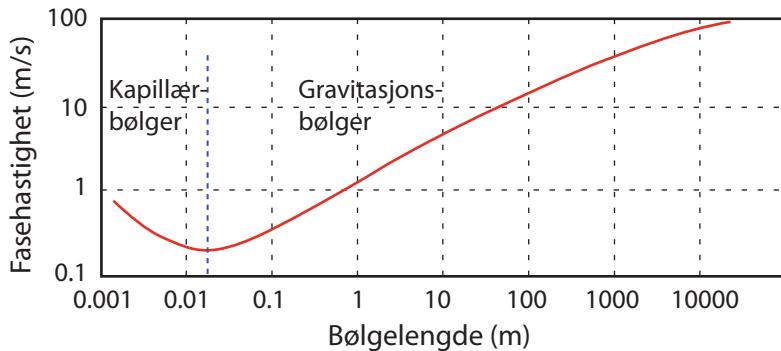
Med andre ord vil overflatehinnen dominere fasehastigheten for bølger med bølgelengde godt mindre enn 1.7 cm, mens gravitasjon vil dominere for bølgelengder godt større enn 1.7 cm.

Fasehastigheten er faktisk minst når bølgelengden er om lag 1.7 cm. Den er da bare 0.231 m/s. Både for mindre og større bølgelengder øker fasehastigheten, og ved svært lange bølgelengder kan fasehastigheten komme opp i over 100 m/s. Figur 8.6 viser beregnet fasehastighet for bølgelengder fra 1 mm til 10 km. Beregningene er stort sett basert på at vanndybden er stor i forhold til bølgelengden (noe som ikke lett lar seg gjøre i praksis her på Jorda for de lengste bølgelengdene!).

Vi skal straks se nærmere på uttrykket for fasehastighet, men vil først minne om noen særtrekk ved tangenshyperbolikus-funksjonen. Hele spekteret av hyperbole trigonometriske funksjoner kan defineres på en nokså analog måte som vanlig sinus, cosinus osv (som alle kan beskrives ved eksponensialfunksjoner med komplekse eksponenter). For hyperbolikusfunksjonene ser uttrykkene slik ut:

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2}$$



Figur 8.6: Fasehastigheten for overflatebølger på vann, forutsatt at dybden er betydelig større enn bølgelengden (ikke lett å realisere i praksis for øvre del av diagrammet). Figuren er omarbeidet etter en figur hentet fra <http://hyperphysics.phy-astr.gsu.edu/hbase/waves/watwav2.html> i februar 2010.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

I fortsettelsen konsentrerer vi oss om hvordan tangenshyperbolikus oppfører seg når argumentet er mye mindre eller mye større enn 1. Da gjelder:

$$\tanh(x) \approx x \text{ for } |x| < 1$$

$$\tanh(x) \approx 1 \text{ for } x > 1$$

I ligning (8.1) er argumentet for tanh lik hk . Argumentet kan også skrives

$$hk = \frac{2\pi h}{\lambda}$$

Det er da naturlig å skille mellom "grunt vann" karakterisert ved $h < \lambda/20$ og "dypt vann" karakterisert ved at $h > \lambda/2$. Disse grensene fører nemlig til at grunt vann betingelsen svarer til:

$$hk < \frac{2\pi\lambda}{20\lambda} = \frac{\pi}{10} < 1.0$$

og dypt vann betingelsen svarer til:

$$hk > \frac{2\pi\lambda}{2\lambda} = \pi > 1.0$$

Tiden er da inne for å drøfte noen hovedtrekk i ligning (8.1). For grunt vann først, og bølgelengder godt over 1.7 cm (slik at vi ser bort fra overflatehinne-leddet) følger:

$$v_f^2(k) = \frac{g}{k} \tanh(kh) \approx \frac{g}{k} kh = gh$$

$$v_f(k) = \sqrt{gh}$$

Vi ser at fasehastigheten er uavhengig av bølgelengden (bølgetallet). Videre merker vi oss at fasehastigheten avtar når dybden avtar.

Dette gir en artig effekt. Når bølger kommer fra storhavet inn mot en langgrunn strand, vil bølger som kommer på skrå innover bevege seg raskest i den delen der dybden er størst. Det vil si den delen av bølgen som er lengst ute vil gå raskere enn den delen av bølgen som

er lenger inne. Vanligvis fører dette til at bølgefrontene etter hvert blir temmelig parallelle med strandlinjen, uansett hvilken retning bølgene hadde før de nærmet seg stranda.

For en kyst med dypt vann helt inn til fjellklippene ned mot vannet, er det ikke noe tilsvarende effekt, og bølgene kan komme inn mot klippene i hvilken som helst retning.

For bølger på dypt vann er fasehastigheten (under forutsetning at overflatespenningen spiller en negilsjerbar rolle):

$$v_f^2(k) = \frac{g}{k} \tanh(kh) \approx \frac{g}{k} 1 = \frac{g\lambda}{2\pi}$$

$$v_f(k) = \sqrt{\frac{g}{2\pi}} \sqrt{\lambda} \approx 1.25 \sqrt{\{\lambda\}} \text{ m/s}$$

hvor $\{\lambda\}$ betyr måltallet for λ (uten benevning) målt i antall meter.

På dypt vann er altså fasehastigheten avhengig av bølgelengden (bølgelaget), noe vi kaller *dispersjon*. Øker bølgelengden med to dekader, vil fasehastigheten øke med en dekade. Dette gjenspeiles omtrentlig i figur 8.6.

Noe å tenke på

Det kan være artig å vite at havbølgen med størst bølgelengde her på Jorda har en bølgelengde på hele 20 000 kilometer. Den ruller og går hele tiden. Kan du gjette hvilken bølge det da er snakk om? Vil du karakterisere den som en overflatebølge som er gravitasjonsdrevet? I så fall, faller den inn under beskrivelsen vår ovenfor, og vil den ha en bølgehastighet som er gitt ut fra våre formler? Du kan tenke litt på det!

[♣ ⇒ Da vi behandlet ligning (8.1), sa vi at for bølgelengder godt over ca 1.7 cm, dominerte gravitasjon ved bølgebevegelsen. For kapillærbølger med bølgelengde klart mindre enn 1.7 cm, dominerte overflatehinnen. Disse tallene gjelder ved jordoverflatene.

En vanndråpe vil ha en form som bestemmes av både gravitasjon og overflatehinnen. Når gravitasjonen effektivt forsvinner, så som f.eks. i vektløs tilstand i Spacelab, er det mulig å lage vanndråper som er nærmest perfekt kuleformede, selv med en diameter opp mot 10 cm. Bølger på overflaten av slike vannkuler vil i vektløs tilstand være dominert av overflatespenningen selv ved bølgelengder større enn 1.7 cm. ← ♣]

8.3 Fase- og gruppehastighet

Vi så nettopp at fasehastigheten til gravitasjonsdrevne overflatebølger i vann er avhengig av bølgelengden når vanndybden er stor relativt til bølgelengden. Fenomenet kalles *dispersjon* og dukker opp i svært mange fysiske sammenhenger. Et kjent eksempel er fargespekteret vi får når vi sender hvitt lys gjennom et prisme. Dette skyldes at ulike bølgelengder har ulik fasehastighet gjennom glasset, med andre ord at glasset har dispersjonsegenskaper for lys.

Før vi drøfter dispersjon mer formelt, skal vi tilbake til den enkle matematiske beskrivelsen av bølger.

Vi bruker ofte ordet "bølge" temmelig ukritisk, og tenker ofte ikke over at en virkelig fysisk bølge *må* ha en begrenset utstrekning i tid og rom. Det betyr at når vi beskriver en bølge f.eks. med følgende uttrykk:

$$y(x, t) = A \cos(kx - \omega t)$$

er dette bare en *tilnærmet beskrivelse* av virkeligheten innenfor et begrenset x- og t-intervall. Dersom vi skal gi en mer realistisk matematisk beskrivelse av en bølge som har

begrenset utstrekning i tid og rom, må vi bruke et mer komplisert uttrykk. Dersom vi ønsker å bruke en fourieranalyse, kan bølgen da beskrives som en sum av mange bølger med ulik frekvens og/eller ulik bølgelengde.

8.3.1 Aller enkleste tilnærming

Den aller enkleste varianten av en sammensatt kurve vi kan tenke oss, er en bølge sammensatt av bare to frekvenskomponenter. Heller ikke denne enkle varianten er en god beskrivelse av en bølge som er begrenset i tid og rom, men den kan i det minste brukes for å få fram et nyttig begrep. Vi legger altså sammen to monokromatiske, plane bølger slik:

$$y(x, t) = A_1 \cos(k_1 x - \omega_1 t) + A_2 \cos(k_2 x - \omega_2 t)$$

Uttrykket her og den videre behandlingen av dette, minner mye om behandlingen av svevelyd i forrige kapittel. For svevelyden la vi bare sammen svingeledd, mens vi nå skal legge sammen bølger.

Fra matematikken (se f.eks. Rottmann) kjenner vi til at:

$$\cos a + \cos b = 2 \cos\left(\frac{a-b}{2}\right) \cos\left(\frac{a+b}{2}\right)$$

Anvendes denne relasjonen for vår bølge for spesialtilfellet at $A_1 = A_2 = A$, får vi:

$$y(x, t) = 2A \cos(\bar{k}x - \bar{\omega}t) \cos\left(\frac{\Delta k}{2}x - \frac{\Delta \omega}{2}t\right) \quad (8.2)$$

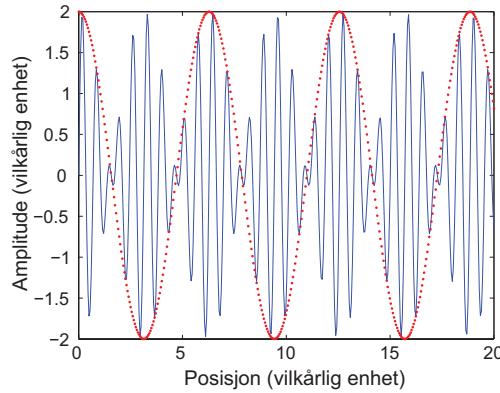
hvor $\bar{k} = (k_1 + k_2)/2$ og $\Delta k = k_1 - k_2$ og tilsvarende for ω . Det betyr at den første er en gjennomsnittsverdi av de to bølgetallene vi starter ut med, og den siste er differansen mellom dem.

Dersom de to bølgetallene er nokså like hverandre, vil totalamplituden få en “svevelyd”-karakter, som vist i figur 8.7:

Dette er fortsatt en bølge som har uendelig utstrekning i tid og rom, men den har likevel visse karakteristiske trekk som også gjelder for mer realistiske bølger. Det karakteristiske er at:

- Den sammensatte bølgen består av en “underliggende” bølge med “konstant” frekvens, men med en langsomt varierende amplitude.
- Amplitudevariasjonen i den underliggende bølgen definerer en tenkt “omhyllingskurve”, som også kan betraktes som en bølge.
- Den underliggende bølgen og omhyllingskurvebølgen kan bevege seg med ulik hastighet.

Det er vanskelig å forstå hvordan en bølge beveger seg ved å bare se på stillestående figurer, slik som figur 8.7. Det er atskillig enklere å forestille seg tidsutviklingen dersom vi



Figur 8.7: Et øyeblikksbilde av en bølge som består av to nærliggende frekvenskomponenter. Sveving-omhyllingskurven er også inntegnet (i rødt). Denne figuren kan synes å være nærmest identisk med den tilsvarende figuren i forrige kapittel da vi behandlet svevelyd. Forskjellen mellom disse fenomenene kommer først fram ved en animasjon der vi får med oss utviklingen både i tid og rom.

lager en animasjon. På neste side er det gitt et enkelt Matlab-program hvor vi kan følge hvordan bølgen romlig sett utvikler seg med tiden.

Programmet kan f.eks. kjøres med følgende parametre:

1. $k_2 = (5/4)k_1$ og $\omega_2 = (5/4)\omega_1$
2. $k_2 = (5/4)k_1$ og $\omega_2 = (4/4)\omega_1$
3. $k_2 = (5/4)k_1$ og $\omega_2 = (6/4)\omega_1$

Ligning (8.2) består av et produkt av to bølger.

Den først er den underliggende bølgen $\cos(\bar{k}x - \bar{\omega}t)$. Det er en bølge med frekvens $\bar{\omega}$ og hastighet $v_f = \bar{\omega}/\bar{k}$. Denne hastigheten kaller vi “fasehastigheten”.

Den andre bølgen er $\cos(\frac{\Delta k}{2}x - \frac{\Delta \omega}{2}t)$. Denne bølgen er definert av den såkalte “omhyllingskurven”, og har hastigheten $v_g = (\Delta\omega/2)/(\Delta k/2) = (\Delta\omega)/(\Delta k)$. Denne hastigheten kaller vi “gruppehastigheten”. I dette enkle tilfellet kan vi også angi en frekvens for omhyllingskurven: $\Delta\omega/2$.

Uttrykkene for hastighet har selvfølgelig basis i det vi vet fra før at en bølge av typen $f(kx - \omega t)$ der f er en vilkårlig funksjon, forplanter seg med en hastighet lik ω/k .

Bruker vi animasjonsprogrammet med parametervalg 1 ovenfor, vil fasehastigheten og gruppehastigheten være identisk, men for parametervalg 2 er gruppehastigheten større enn fasehastigheten og for valg 3 er gruppehastigheten mindre enn fasehastigheten. Vi anbefaler sterkt at du studerer disse animasjonene (eller tilsvarende på web) for å forstå forskjellen mellom gruppe- og fasehastighet. Du vil da observere følgende tilfeller:

I tilfelle 1 har en bølge med vinkelfrekvens ω_1 samme hastighet (fasehastighet) som en bølge med vinkelfrekvens ω_2 . Når en bølge forplanter seg gjennom et medium der bølgehastigheten for bølger er identisk for alle bølgelengder vi betrakter, sier vi at mediet er ikke-dispersivt. I slike medier er fasehastighet og gruppehastighet like store. I animasjonen av bølgen gitt i ligning (8.2), vil den kortbølgede strukturen i dette tilfellet bevege seg

akkurat like fort som omhyllingskurven. Det vil si at bildet vi ser i figur 8.7 vil forskyve seg mot høyre med tiden, uten noe som helst endring i form.

I tilfelle 2 ovenfor, hvor gruppehastigheten er større enn fasehastigheten, svarer dette til at hastigheten til en monokromatisk bølge avtar med frekvensen til bølgen. I tilfelle 3 ovenfor øker hastigheten for en bølge med frekvensen. I begge tilfeller gjelder det at hastigheten til bølger endrer seg med frekvensen i det aktuelle frekvensområdet vi betrakter. Medier som har en slik egenskap kaller vi *dispersive* medier.

Når gruppehastigheten er forskjellig fra fasehastigheten, vil omhyllingskurven vandre med en annen hastighet enn enkeltbølgeloppene. Det betyr at enkeltoppene vandler i forhold til omhyllingskurven. Det kommer fint fram dersom du kjører Matlab programmet i neste avsnitt for de tre variantene vi nevnte ovenfor.

8.3.2 Program-listing

Her følger et Matlab-program vi kan kjøre for å se hvordan de underliggende bølgene beveger seg i forhold til omhyllingskurven ved ingen dispersjon, normal dispersjon og anomal dispersjon.

```
function bolgeanimering3x

clear all;
% Velg a = b = 1.0 for ren monokromatisk bølge
% Velg a = 1.25 for å få to ulike frekvenser, og:
% 1) b = 1.25 ingen dispersjon
% 2) b = 1.50 anomal dispersjon
% 3) b = 1.15 normal dispersjon
a = 1.25;
b = 1.50;
k1 = 8;
k2 = 8*a;
w1 = 8;
w2 = 8*b;
N = 400;
x = linspace(0,20,N);
y = zeros(N,1);

% Plotter først sammensatt kurve sammen med omhyllingskurven
t = 0.0;
y = sin(k1.*x-w1*t) + sin(k2.*x-w2*t);
conv = 2*cos(((k2-k1)/2).*x - ((w2-w1)/2)*t);
plot(x,y,'-b');
hold on;
plot(x,conv,'.r');
xlabel('Posisjon (vilkårlig enhet)');
ylabel('Amplitude (vilkårlig enhet)');
figure;
```

```
% Her følger så animeringen av hvordan bølgen utvikler seg med tiden
p = plot(x,y,'-','EraseMode','xor');
axis([0 20 -2.5 2.5])
hold on
for i=1:200
    t = i*0.1;
    %y(j) = sin(k1.*x-w1*t);
    y = sin(k1.*x-w1*t) + sin(k2.*x-w2*t);
    set(p,'XData',x,'YData',y)
    drawnow
    wkmean = (w1+w2)/(k1+k2);
    wkdelta = (w2-w1)/(k2-k1);
    % Plotter rød stav for å vise fasehastigheten
    plot(t*wkmean,2.3,'.r')
    % Plotter grønn stav for gruppehastighet
    plot(t*wkdelta,-2.3,'.g');
    % Forsinker fremvisningen for at bølgen ikke skal gå for fort.
    pause(0.20);
end
```

8.3.3 Kommentarer til den enkle beskrivelsen

Ovenfor brukte vi en sum av to plane bølger med hver sine vel definerte frekvenser, som et pedagogisk hjelpemiddel i et forsøk på å få fram forskjellen mellom fase- og gruppehastighet. Resultantbølgen ble beskrevet i ligning (8.2), og det var denne beskrivelsen som lå til grunn for Matlab-animasjonen av hvordan bølgen beveger seg.

Dersom du imidlertid setter dette uttrykket inn i bølgeligningen:

$$\frac{\partial^2 u(x, t)}{\partial t^2} = v^2 \frac{\partial^2 u(x, t)}{\partial x^2} \quad (8.3)$$

vil du finne at y slett ikke er noen løsning av bølgeligningen! Hva betyr egentlig dette?

Ligning (8.2) beskriver en bølge som strekker seg uendelig langt i x-retning og for alle tider. Omhyllingskurven har samme form hele tiden.

Dette er normalt ikke tilfelle ved dispersjon.

I en beskrivelse som faktisk er en løsning av bølgeligningen (8.3) vil formen til omhyllingskurven endre seg med tiden. Selv om det også i det tilfellet ville være mulig å definere en "omtrentlig" gruppehastighet, er det generelt sett vanskeligere å gjøre det enn i vår forenklede beskrivelse. Det er da på tide å se på en mer generell beskrivelse av gruppehastighet.

8.3.4 Normal dispersjon og anomal dispersjon

I optikken vet vi at brytningsindeksen til glass varierer med bølgelengden for lys (se figur 8.8 for ulike typer glass). Brytningsindeksen øker når frekvensen øker (bølgelengden avtar).

Fasehastigheten til elektromagnetiske bølger (lys) i glass er gitt ved

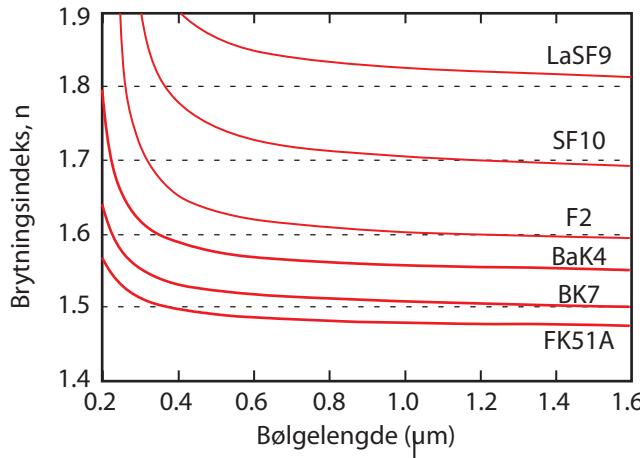
$$v_f = c_{glass} = c/n(\lambda)$$

hvor c er lyshastigheten i vakuum, $c_{glass} = v_f$ er lyshastigheten i glass, som også er fasehastigheten til lys i glass. $n(\lambda)$ er brytningsindeksen som altså er bølgelengdeavhengig.

For lys i glass i figur 8.8 betyr dette at fasehastigheten avtar når frekvensen øker. En slik sammenheng kaller vi *normal dispersjon*.

En litt annen grafisk fremstilling blir ofte brukt for å få fram lovemessigheter ved dispersjon. I et slikt diagram angis vinkelfrekvensen ω som funksjon av bølgetallet k . For en vanlig monokromatisk bølge $A \cos(kx - \omega t)$ er hastigheten (dvs fasehastigheten) gitt ved:

$$v_f = \frac{\omega}{k}$$



Figur 8.8: *Brytningsindeksen for lys i ulike former for glass. Figuren er omarbeidet noe fra en hentet fra web under oppslagsordet "Refractive Index" på Wikipedia (pr 20. feb. 2009).*

Dette gir:

$$\omega = v_f k$$

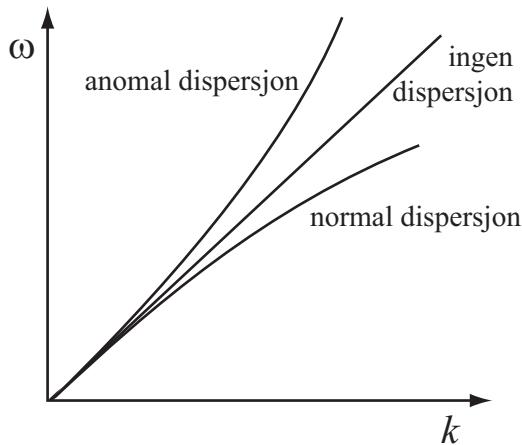
Dersom hastigheten v_f er uavhengig av bølgelengde (eller frekvens om vi ønsker å uttrykke det slik), vil en kurve som angir ω som funksjon av k være en rett linje. Vi kaller denne sammenhengen $\omega(k)$ for *dispersjonsrelasjonen* for det aktuelle mediet. For dispersive medier vil en kurve i et ω versus k diagram være en krum linje, slik som vist i figur 11.5.

Det kan være nyttig å anskueliggjøre hvordan fase- og gruppehastigheten kommer inn i forhold til disse kurvene. I figur 8.10 er dette illustrert for normal dispersjon.

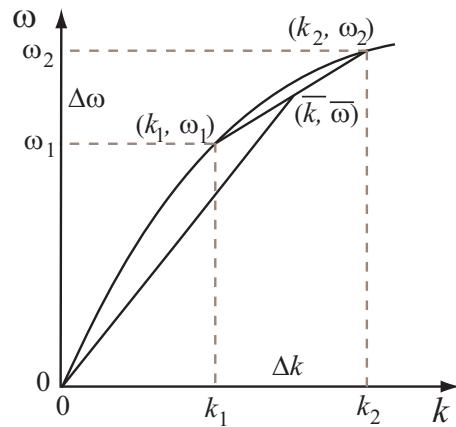
Da vi drøftet ligning (8.2) som gjalt for vår aller enkleste tilnærming til dispersjon, så vi at gruppehastigheten ble angitt som $v_g = (\Delta\omega)/(\Delta k)$. Det kan derfor være fristende å definere gruppehastighet v_g i det mer generelle tilfellet ved uttrykket:

$$v_g = \frac{\partial \omega}{\partial k} \quad (8.4)$$

Det kan vises matematisk at en slik definisjon faktisk svarer til hastigheten til "omhyllingskurven" til en sammensatt bølgepakke, i alle fall der omhyllingskurven har en gaussisk form. Det svarer til nettopp det vi forbinder med "gruppehastighet".



Figur 8.9: Sammenhengen mellom vinkelfrekvensen ω og bølgetallet k for et gitt medium, kalles dispersjonsrelasjonen for mediet. Vi skiller mellom tre ulike klasser av medier, slik som angitt i figuren.



Figur 8.10: En grafisk anskueliggjøring av hvordan fasehastighet og gruppehastighet er definert i forhold til en kurve som angir mediets dispersjonsrelasjon. Fasehastigheten er gitt ved stigningstallet til den rette linjen fra origo mot midten av korden mellom punktene som svarer til første og andre frekvens. Gruppehastigheten er gitt ved stigningstallet på korden. Jamfør med ligning (8.2). Dersom Δk går mot null, følger det at gruppehastigheten blir lik den deriverte til kurven.

Det at gruppehastigheten er den deriverte av dispersjonsrelasjonen $\omega(k)$ åpner opp for interessante sammenhenger. Det vil vi benytte oss av flere ganger i denne boka.

Aller først kan vi nå finne et uttrykk for hvordan gruppehastigheten varierer med brytningsindeksen. Utgangspunktet er da følgende to velkjente relasjoner:

$$v_f(\omega) = \frac{c_0}{n(\omega)} = \frac{\omega}{k}$$

Herav følger:

$$k = \frac{\omega}{v_f(\omega)} = \frac{\omega n(\omega)}{c_0}$$

Med litt “fysiker-matematikk” følger da:

$$\frac{1}{v_g} = \frac{dk}{d\omega} = \frac{n(\omega)}{c_0} + \frac{\omega}{c_0} \left(\frac{dn}{d\omega} \right)$$

$$\frac{1}{v_g} = \frac{n(\omega)}{c_0} \left(1 + \frac{\omega}{n} \left(\frac{dn}{d\omega} \right) \right)$$

$$v_g = v_f(\omega) \frac{1}{1 + \frac{\omega}{n} \left(\frac{dn}{d\omega} \right)} \quad (8.5)$$

Ved normal dispersjon er $\frac{dn}{d\omega} > 0$ hvilket innebærer at $v_g < v_f$, dvs at gruppehastigheten er mindre enn fasehastigheten.

Hvorfor er lyshastigheten i glass mindre enn i vakuum?

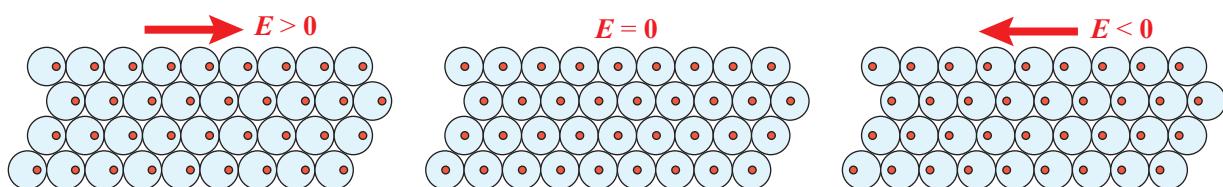
Det kan her være på sin plass med et lite sidesprang, siden det i praksis har vist seg at relativt få vet hvorfor lyshastigheten er mindre i glass enn i vakuum. En klar indikasjon får vi ved å granske uttrykket for lyshastigheten gjennom et medium. Dette uttrykket gis gjerne i bøker i generell elektromagnetisme, men blir også gjennomgått i detalj i neste kapittel i vår bok. Uttrykket er:

$$c = c_0/n = \frac{1}{\sqrt{\epsilon_0 \epsilon_r \mu_0 \mu_r}}$$

hvor ϵ_0 er permittiviteten i det tomme rom, ϵ_r er den relative permittiviteten, μ_0 er den magnetiske permeabiliteten til vakuum, og μ_r er den relative magnetiske permeabiliteten. I glass, som er diamagnetisk, er μ_r tilnærmet lik 1, og vi får:

$$c = \frac{1}{\sqrt{\epsilon_0 \epsilon_r \mu_0}} = c_0 \frac{1}{\sqrt{\epsilon_r}}$$

Når vi da husker at ϵ_r er et mål for hvor mye polarisering vi kan oppnå når vi putter et materiale inn i et elektrisk felt, skjønner vi at polarisering av glass er grunnen til at lys går saktere gjennom glass enn i vakuum.



Figur 8.11: Når en elektromagnetisk bølge passerer et stykke glass, vil det elektriskefeltet i bølgen få elektronene i elektronskyen rundt hver atomkjerne til å forskyve seg slik det forventes ut fra Coulomb-kraften.

Figur 8.11 indikerer hva som skjer når en elektromagnetisk bølge passerer gjennom glass. Det elektriskefeltet vil veksle på en harmonisk måte, og iblant ha en verdi i én retning (på tvers av lysets bevegelsesretning), iblant null, og iblant motsatt retning av den førstnevnte. Elektronene i glassatomene vil bli påvirket av det elektriskefeltet, og hele "elektronskyen" rundt hver atomkjerne vil forflytte seg litt i forhold til kjernen slik det er indikert i figuren. I virkeligheten er forskyvningen ytterst liten, siden det elektriskefeltet fra den elektromagnetiske bølgen normalt er lite i forhold til de elektriske kraftene

mellan kjerne og elektron.² Likevel skjer det selv ved svakt lys en kollektiv forflytning av elektronene i forhold til kjernene som faktisk har betydning.

Den kollektive forflytningen resulterer i at glasset nærmest kan ansees som en antenn med oscillerende strømmer. Denne oscillasjonen i ladninger fører til at det sendes ut elektromagnetiske bølger med samme frekvens som bølgen som startet det hele. Vi har imidlertid sett i kapittel 3 (tvungne svingninger) at det er en faseforskjell mellom bevegelsen og den påtrykte kraften når vi ikke er eksakt ved resonansfrekvensen. Det er *kombinasjonen av den opprinnelige bølgen og den faseforskjøvne reemitterte bølgen fra de oscillerende elektronskyene*, som sørger for at lyshastighetene i glass er mindre enn i vakuum.

Det sier seg selv da, at når den elektromagnetiske bølgen har passert glasset og kommer ut i luft (nær vakuum i vår sammenheng), vil det ikke være noe nevneverdig polarisering av mediet, og bølgen vil ikke bli forsinket av den reemitterte bølgen. Lyshastigheten går da selvfølgelig igjen tilbake til (nær) lyshastigheten i vakuum.

Dersom vi igjen tenker tilbake på kapittel 3 og tvungne svingninger, husker vi også resonansfenomenet. Ved visse frekvenser ble utslaget ekstra stort som følge av den påtrykte kraften. Dersom du gransker figur 8.8 kan du se klare antydninger til at noe spesielt skjer for bølgelengder litt under 200 nm ($0.2 \mu\text{m}$). Da er vi inne i UV-området, og vi nærmer oss resonans i elektronoscillasjonene rundt kjernene. Ved å tenke på resonanskurvets form i kapittel 3, kan du da forhåpentligvis også forestille deg hva som skjer dersom vi passerer resonans og kommer til enda kortere bølgelengder. Da vil kurver i et diagram likende figur 8.8 ha helling motsatt vei, og vi får såkalt anomal dispersjon. For enkelte materialer vil resonansfrekvensen ligge ved mye lengre bølgelengder, og da kan vi oppnå anomal dispersjon til og med for vanlig synlig lys.

Dette er ett av mange sider av fysikken der de enkle lovmessighetene i kapittel 2 og 3 dukker opp. Enkle prinsipper er ofte en del av forklaringen også ved mer kompliserte fenomener.

[♣ ⇒ En liten, morsom, historisk fortelling i denne sammenheng:

I Newtons partikkelfordeling for lys (“corpuscular model of light”) ble brytning av lys forklart ved at partiklene gikk *raskere* gjennom glass enn i luft, mens bølgebeskrivelsen gir motsatt prediksjon. Måling av lyshastigheten i glass ble derfor i en periode ansett som en viktig test på hvorvidt en bølgemodell eller partikkelfordeling var best med eksperimenter. Vi kan imidlertid ikke måle lyshastigheten i en sammenhengende, monokromatisk bølge. Vi må ha en “struktur” i bølgen som vi kan gjenkjenne for å kunne måle lyshastigheten. Dette tilsvarer at vi måler gruppehastigheten.

Det var imidlertid ingen som var i stand til å måle lyshastigheten til lys på denne måten på 1700-tallet og begynnelsen av 1800-tallet. Foucault var den første som gjennomførte eksperimentet. Det var i 1850, og resultatet viste at lyshastigheten var mindre i glass enn i luft, noe som støttet bølgemodellen for lys. På dette tidspunktet hadde imidlertid de fleste fysikere motvillig gått bort fra Newtons partikkelfordeling for lys. Eksperimenter av Thomas Young (dobeltpalteksperimentet i 1801) og et arbeid til Fresnel i 1818, som først ble imøtegått av Poisson, men siden styrket av et eksperiment gjennomført av Arago, hadde etter hvert overbevist fysikerne om at bølgemodellen for lys ga en bedre beskrivelse enn partikkelfordelingen. Du kan f.eks. lese om “Aragos flekk” (Arago spot på engelsk) på Wikipedia dersom du har lyst. ⇐ ♣]

²I et eksperiment med kraftig pulset laser i Tyskland i 2013, er likevel det elektriske feltet så kraftig at det rører mange elektronene helt løst fra kjernen. Da omdannes glasset fra å være en isolator til en god elektrisk leder i løpet av femtosekunder!

8.4 Bølger i vann

Vi har tidligere i ligning (8.1) gitt et uttrykk for fasehastigheten til overflatebølger i vann, men gjengir formelen på ny som en oppfriskning.

$$v_f^2(k) = \left[\frac{g}{k} + \frac{TK}{\rho} \right] \tanh(kh)$$

Her er k som vanlig bølgetallet, g tyngdens akselerasjon, T overflatespenningen, ρ massetetheten og h er dybden på vannet. Uttrykket kan utledes dersom vi starter med bare å ta hensyn til gravitasjon og overflatespenning, og vi ser bort fra viskositet, vind, og en liten, men endelig kompressibilitet til vannet.

Vi har tidligere funnet uttrykk for fasehastigheten for gravitasjonsdrevne bølger for grunt og dypt vann. Nå tar vi for oss også gruppehastighet og beskriver tre av de fire mulige enkle spesialtilfellene litt mer inngående:

1. **Tyngdedrevne bølger med liten dybde relativt til bølgelengden, dvs produktet $hk \ll 1$:**

Bølgelengde antas å være stor relativt til den kritiske (1.7 cm), og fra ligning (8.1) følger:

$$\begin{aligned} v_f^2(k) &\approx \frac{g}{k} hk \\ v_f &\approx \sqrt{gh} \end{aligned}$$

Dette har vi vist tidligere, men la oss også se på gruppehastigheten. Vi bruker da relasjonen $v_f = \omega/k$ og får:

$$\begin{aligned} \frac{\omega}{k} &= \sqrt{gh} \\ \omega &= \sqrt{gh} k \\ v_g &= \frac{d\omega}{dk} = \sqrt{gh} = v_f \end{aligned}$$

Altså:

$$v_g = v_f$$

Dette er enbetydende med at det ikke er noen dispersjon.

2. **Gravitasjonsdrevne bølger på dypt vann.**

I dette tilfellet fant vi:

$$v_f^2(k) \approx \frac{g}{k}$$

Setter på ny inn $v_f = \omega/k$, og får:

$$\frac{\omega^2}{k^2} = \frac{g}{k}$$

Dette gir følgende dispersjonsrelasjon:

$$\omega \approx \sqrt{gk}$$

Gruppehastigheten blir da:

$$v_g = \frac{d\omega}{dk} = \frac{1}{2} \sqrt{\frac{g}{k}}$$

$$v_g \approx \frac{1}{2} v_f \quad (8.6)$$

Vi ser altså at gruppehastigheten er omtrent lik halvparten av fasehastigheten.



Figur 8.12: Foto av en båt med bølger som danner en vifte bak seg. Se videre omtale i neste delkapittel. Fotografiet er hentet fra en masteroppgave til de Vries i 2007.

Skipsbølger faller ofte inn i denne kategorien. Enkeltbølgene synes å rulle forttere enn "plogen" eller "viften" som følger etter båten (se figur 8.12). Det fører til at enkeltbølgene ruller på en måte forbi "viften" og blir borte like etterpå. Vi skal se nærmere på dette om litt.

3. Korte rippler i dypt vann.

Her er bølgelengden på bølgene liten relativt til den kritiske bølgelengden på 1.7 cm. Samtidig er bølgelengden mye mindre enn dybden på vannet. Da får vi overflatespenning-drevne bølger og

$$v_f^2(k) \approx \frac{Tk}{\rho} \cdot 1 = \frac{\omega^2}{k^2}$$

Dispersjonsrelasjonen blir da:

$$\omega \approx \left(\sqrt{\frac{T}{\rho}} \right) k^{\frac{3}{2}}$$

Gruppehastigheten i dette tilfellet blir:

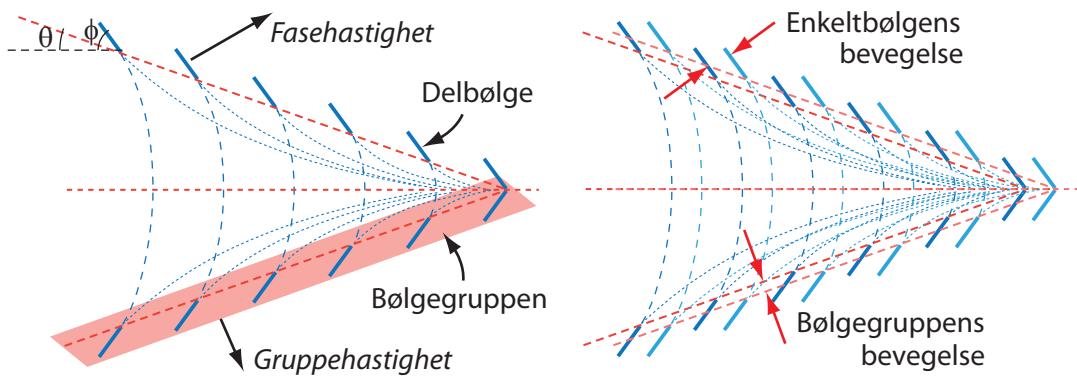
$$v_g = \frac{d\omega}{dk} = \left(\sqrt{\frac{T}{\rho}} \right) \frac{3}{2} k^{\frac{1}{2}} = \frac{3}{2} \sqrt{\frac{Tk}{\rho}}$$

$$v_g = \frac{3}{2} v_f$$

I dette tilfellet er faktisk gruppehastigheten større enn fasehastigheten (svarer til anomal dispersjon). I dette tilfellet vil liksom de enkelte bølgene synes å dukke opp fra intet på forsiden av gruppen, og så bevege seg ”bakover” gjennom gruppen. Relativt til vannet vil likevel også enkeltbølgene hele tiden forplante seg vekk fra kilden som skapte bølgene (så lenge vi ikke har refleksjon), men illusjonen om å vandre bakover kommer av at gruppehastigheten er enda større enn fasehastigheten.

8.4.1 Skipsbølger, et eksempel

Mange er ikke vant til å identifiesere hva som menes med en gruppe med bølger og hva som menes med enkeltbølger. Venstre del av figur 8.13 forsøker å vise dette. Figuren henviser til fotografiet i figur 8.12. Viften med mange enkeltbølger som går litt på tvers av ytterkanten av viften, danner gruppen av bølger. Denne viften brer seg utover med en hastighet som er gruppehastigheten. Hver enkeltbølge vil imidlertid vandre i en annen retning enn viften som sådan, og med en annen bølgehastighet som nå er fasehastigheten.



Figur 8.13: Til venstre: Identifisering av gruppen som beveger seg med gruppehastighet og enkeltbølger som beveger seg med fasehastighet i bølger fra en båt. Figuren er en video-reføring av en figur i masteroppgaven til de Vries. Til høyre: Detalj som viser hvor langt bølgegruppen og hvor langt en enkeltbølge har beveget seg på tvers av hver sine bølgefronter i løpet av en viss tid. Figuren viser klart at gruppehastigheten er lavere enn fasehastigheten for disse vannbølgene.

Vi har tidligere utledet at for bølger på dypt vann, er gruppehastigheten om lag halvparten av fasehastigheten (se ligning (8.6)). Det betyr at enkeltbølgene beveger seg raskere enn gruppen. Enkeltbølgene synes derfor bare å oppstå nærmest av ingenting på innsiden av gruppen, og rulle tvers gjennom gruppen, og forsvinner nesten på uforståelig vis når ytterkanten til gruppen er nådd.

Dersom du har padlet i kano og sett med litt nervøsitet hvor raskt enkelbølger nærmer seg kanoen etter at en båt har passert, har du kanskje undret deg over at de bølgene som så så skumle ut bare liksom ble borte av seg selv før de nådde fram til kanoen. Først mye senere enn vi først fikk inntrykk av, når bølgen kanoen. Bølgene når kanoen først når selve gruppen nå fram, og gruppen vandrer bare halvparten så fort som enkeltbølgene.

I høyre del av figur 8.13 er bølgene tegnet inn ved ett tidspunkt og en litt senere tid. Da kommer det klart fram at gruppen har vandret en mye kortere avstand enn enkelt-

bølgene i den perioden vi studerer. Det hører med til historien at bølgebildet bak båten er stasjonært i forhold til båten. Når båten har beveget seg 10 meter framover, har også hele bølgemønsteret etter båten beveget seg 10 meter framover.

[♣ ⇒ **Tilleggskommentar:**

Det er skjedd mye på forskningsfronten når det gjelder fase og gruppehastighet siden ca 1980. Mye av dette er knyttet til lys.

Vi har i lang tid visst at når lys går gjennom glass (eller en vanndråpe for den saks skyld), går lys med ulike bølgelengder med ulik hastighet. Bryningssindeksen er bølgelengdeavhengig, $n(\lambda)$, noe som igjen er uttrykk for dispersjon. Lys i glass viser normal dispersjon.

De siste få decennier er det imidlertid utviklet en rekke spesielle materialer, og noen av disse har svært varierende fasehastighet for lys med ulike bølgelengder. Vi kan derfor få svært varierende fase- og gruppehastigheter, og det er til og med materialer hvor fasehastigheten har en retning, og gruppehastigheten går i motsatt retning.

Det er også laget materialer og eksperimentelle forhold der vi kan bremse ned lys kollosalt, til og med “stoppe” det for kortere perioder, for så å starte det opp igjen (søk på Lene Hau ved Harvard University, så får du innblikk i et spennende forskningsfelt. Lene er dansk og er en yndling for dansk presse).

I enkelte materialer mener noen å vise at en lyspuls kan gå raskere gjennom materialet enn lys i vakum, og at vi i prinsippet truer Einsteins relativitetsteori i så måte. Når vi ser nærmere på hva som skjer, ser vi at påstanden om “raskere enn lyshastigheten i vakuum” kan diskuteres. Det kommer helt an på hvordan vi definerer ting og tang, men Einsteins relativitetsteori står ikke akkurat i fare ut fra disse eksperimentene, generelt sett. Hva fremtiden vil bringe, er det vanskeligere å ha noe formening om!

Dispersjon er også aktuell for materiebølger i kvantefysikken. Gruppehastighet er definert ut fra dispersjonsrelasjoner hvor $\omega(k)$ er beskrevet og vi anvender $v_g = d\omega/dk$. For materiebølger er bølgelengden gjennom de Broglie-relasjonen knyttet til bevegelsesmengden, og frekvensen til energi. For materiebølger har vi derfor dispersjon dersom energien ikke øker som forventet med bevegelsesmengden.

Dispersjon har nær sammenheng mellom andre grunnleggende deler av fysikken, blant annet gjennom den såkalte Kramers-Kronig relasjonen som viser at dispersjon er relatert til hvor stor absorpsjonen er for ulike bølgelengder i mediet. Til en viss grad er dette knyttet til tvungne svingninger og Q-verdier, som vi har omtalt tidligere, men vi får ikke tid til å gå mer i dybden på dette. ⇐ ♣]

8.5 Lyspuls gjennom et medium

Det er i dag en rivende utvikling innen f.eks. optikk / materialvitenskap / kvantefysikk. Vi er jo i en situasjon der internett får en stadig større innflytelse på samfunnet. Eksempelvis antar vi at salg av filmer på DVD og BlueRay vil avta betydelig innen få år, og at folk i stedet laster ned filmer fra nettet når de ønsker å se den, uten å lagre den etterpå. Det betyr at kapasiteten på internett må økes betraktelig, og det kan bare gjøres ved å videreutvikle all teknologi som er involvert.

Det er utviklet nye kilder for lys (bl.a. såkalte “quantum dots”), og vi forsøker å redusere dispersjon og tap i optiske fibre, lage nye lysledere (bl.a. såkalte “photoinc crystals”) som er materialvitenskap på sitt beste.

Vi bruker kortere og kortere pulser av lys når man koder signaler som sendes over systemet, men det er mange typer begrensinger som finnes. Vi skal ganske snart diskutere en begrensning som ligger i det såkalte “båndbreddeteometet”, men aller først skal vi se hva som skjer når en lyspuls sendes inn mot et dielektrikum, f.eks. glass.

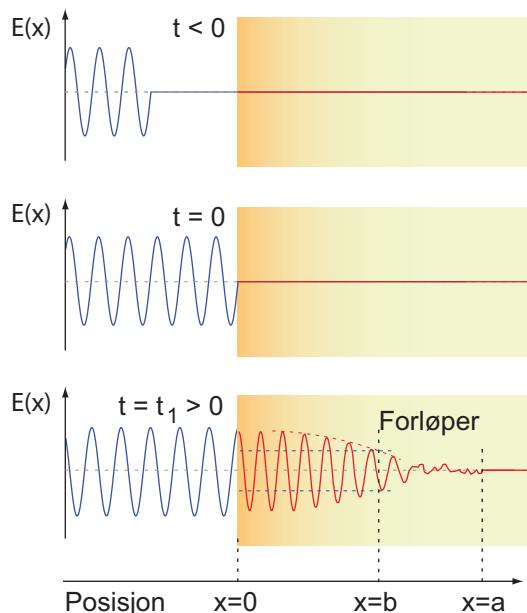
Lyshastigheten i glass er lik lyshastigheten i vakum dividert på brytningsindeksen:

$$c(n) = \frac{c_0}{n}$$

Det er *fasehastigheten* vi da snakker om. Men hva med gruppehastigheten? Ligning (8.5) viser at gruppehastigheten er lavere enn fasehastigheten (for vanlige glasstyper, ved normal dispersjon).

Det er imidlertid finere detaljer som kommer inn om vi går prosessen nærmere inn på klingen. Vi skal ikke gjøre beregninger i denne sammenheng, men nøye oss med å gi et kvalitativt bilde for å vise hva som foregår.

Figur 8.14 viser et tenkt forløp der en lyspuls kommer inn mot et stykke massivt glass. Vi ser stort sett bort fra refleksjon. I øvre del av figuren er lyspulsen på vei mot glassoverflaten, og i den midterste delen av figuren er lyspulsen akkurat kommet til grenseflaten. Tidspunktet der bølgen akkurat når fram til grenseflaten velges som nullpunkt for tiden.



Figur 8.14: En lyspuls kommer normalt inn mot en grenseflate mellom luft og glass. Vi følger hvordan det elektriske feltet ser ut i tre ulike tidspunkt, først litt før lyspulsen har nådd glasset, dernest idet lyspulsen når glasset, og endelig en såpass lang tid etter at pulsfronten har nådd glasset at bølgen like innenfor glassoverflaten har stabilisert seg på en "steady state" tilstand. Se teksten for øvrige detaljer.

Det antas at bølgen er tilnærmet plan hvor elektrisk felt i ethvert tidspunkt er identisk over et vilkårlig valgt plan parallelt med grenseflaten (og normalt på den horisontale retningen bølgen brer seg i). Vi velger å tegne inn øyeblikksbilder av det elektriske feltet i lyset i ulik avstand fra grenseflaten.

Det er ikke noe uventet som skjer så lenge bølgen er i luft (nesten samme forhold som i vakuum). Lufta er et tilnærmet ikke-dispersivt medium, og den skarpe bølgefrenten vil bevege seg uten å bli fordreid, helt inntil den når grenseflaten mot glass.

Litt etter at bølgen har nådd grenseflaten, kan vi f.eks. finne at øyeblikksbildet av elektrisk felt har en fordeling som vist i nederste del av figuren. Vi skiller mellom tre ulike soner inne i glasset (forutsatt at vi har ventet passe lengre):

1. Lengst vekk fra grenseflaten har vi et område hvor bølgen ennå ikke har nådd fram. Her er alle forhold identisk med det de var før bølgen nådde glasset.
2. I en midlere sone har pulsen nådd fram, men det elektriske feltet er ganske lite og nærmest kaotisk. Det er ikke noe regelmessig frekvens å finne. Vi sier at dette er forløperen til pulsen (“precursor” på engelsk).
3. Det elektriske feltet i området nærmest grenseflaten har en tydelig bølgekarakter. Bølgelengden er kortere enn i lufta utenfor, og amplituden noe redusert relativt til i luft (pga refleksjon). Amplituden er nokså konstant i området aller nærmest grenseflaten mot luft, men avtar når vi nærmer oss den midtre sonen der det ikke var noe veldefinert bølge (forløper-fasen).

Det interessante er at dersom t_1 er tiden mellom det øyeblikket lyspulsen nådde grenseflaten, til det øyeblikksbildet vi har i nedre del av figuren, så trenger forløper-sonen inn til en avstand $a = c_0 t_1$. Det betyr at bølgefronten går tvers gjennom glasset med samme hastighet som lyshastigheten i vakuum! Men det elektriske og magnetiske feltet i lyspulsen blir i starten delvis absorbert av elektronene / atomene i glasset. Elektronene i glasset hadde en viss bevegelse allerede før lyspulsen nådde fram, men det var ikke noe ordentlig samordning av bevegelsen.

Etter at det elektriske feltet i lyspulsen har virket en tid, vil elektronene i glasset gradvis få en mer samordnet bevegelse (svært liten bevegelse, men den betyr mye likevel fordi den er samordnet). Vi får da en situasjon som den vi illustrerte i figur 8.11. Fra da av vil vi få en økning i det elektriske feltet inne i glasset, fordi den samordnede elektronbevegelsen gir gjenutsending av elektromagnetisk felt (lys). Vi dørper avstanden fra grenseflaten til det stedet der amplituden til det elektriske feltet har nådd halvparten av hva det siden fikk av “steady state” verdi for b . Vi kan da definere enda en type hastighet, nemlig den såkalte “*signal-hastigheten*” ved:

$$v_s = \frac{b}{t_1}$$

For vanlige materialer (f.eks. glass, hvor frekvensen på lyset er langt fra absorpsjonsmaksim) er signal-hastigheten nokså nær gruppehastigheten. Vi ser av figuren at gruppehastigheten er mindre enn lyshastigheten i vakuum, men at det altså skjer noe i mediet allerede etter at pulsfronten har forplantet seg gjennom med en hastighet lik lyshastigheten i vakuum.

Merk at det fenomenet vi her betrakter har nær sammenheng med hva vi fant da vi studerte tvungne svingninger. Vi oppnår vanligvis ikke maksimalt utslag i tvungne svingninger ved å “dytte” til systemet bare én periode. Vi måtte la den ytre kraften virke over flere perioder for å få en steady state virkning. Slik er det også med elektronene i glasset. De vil ikke straks bevege seg særlig godt i takt med det ytre elektriske feltet som kommer fra lyspulsen. Men etter en viss karakteristisk tid (knyttet til Q-verdien for systemet), vil de nå sin steady-state verdi. Ulike stoffer vil respondere ulikt for ulike påtrykte frekvenser. I deler av fysikken snakker vi om en *relaksasjonstid* i slike sammenhenger.

Vi minner igjen om at responsen i et system med tvungne svingninger ofte ikke er helt i fase med den opprinnelige kraften. Slik er det også med elektronbevegelsen i glasset relativt til det elektriske feltet fra det opprinnelige elektriske feltet i lyset. Som allerede nevnt er det denne faseforskyvningen som til syvende og sist er årsak til at fasehastigheten til lyset blir lavere i et dielektrikum (f.eks. glass) enn den er i vakuum.

8.6 Numerisk beregning av tidsutvikling for en bølge.

Bølgeligningen er en partiell differensialligning, og vi skal nå se hvordan vi kan finne løsningen av en en-dimensjonal bølgeligning når initialbetingelser og randbetingelser er gitt. Vi har flere hensikter med en slik gjennomgang. Aller viktigst er å få fram den underliggende algoritmen fordi den kan gi en bedre forståelse for bølgebevegelse generelt.

Utgangspunktet er en generell en-dimensjonal bølgeligning:

$$\frac{\partial^2 u}{\partial t^2} = v^2 \frac{\partial^2 u}{\partial x^2}$$

I en numerisk løsning beskrives løsningen bare i diskrete posisjoner og tidspunkt:

$$u(x, t) \rightarrow u(x_i, t_j) \equiv u_{i,j}$$

hvor

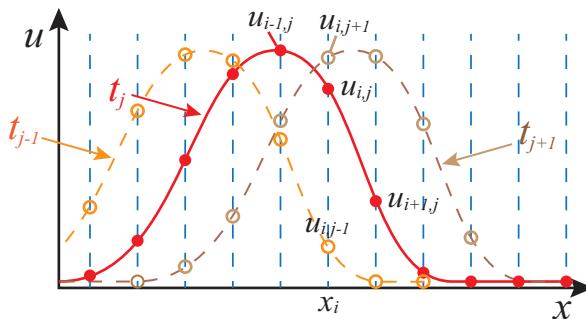
$$x_i = x_0 + i\Delta x$$

der $i = 0, 1, 2, \dots, N$, og

$$t_j = t_0 + j\Delta t$$

der $j = 0, 1, 2, \dots, M$.

Figur 8.15 illustrerer hvordan en bølge beskrives numerisk. For hvert tidspunkt beskriver en tallrekke utslaget i de valgte posisjonene vi betrakter. I figuren er deler av posisjonsdatapunktene vist for tre ulike tidspunkt.



Figur 8.15: Ved bruk av numeriske metoder beskrives en bølge bare i diskrete posisjoner i rommet og i diskrete tidspunkt. Her er en og samme bølge angitt i tre ulike tidspunkt. Første indeks angir posisjonsnummer, og andre indeks angir tidspunkt-nummer i beskrivelsen.

I kapittel 3 ble det vist at den andre deriverte kan uttrykkes på differens-form som følger:

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &\equiv u_{xx}(x_i, t_j) = \\ &\frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{\Delta x^2} \end{aligned}$$

På kortform kan dette skrives slik:

$$u_{xx,i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} \quad (8.7)$$

På tilsvarende måte kan vi angi den dobbeltderiverte med hensyn på tid:

$$u_{tt,i,j} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta t^2} \quad (8.8)$$

Hele bølgeligningen får da formen:

$$u_{tt,i,j} = v^2 u_{xx,i,j} \quad (8.9)$$

Ligning (8.8) settes inn i ligning (8.9), og leddene ordnes. Resultatet er:

$$u_{i,j+1} = u_{i,j} + (u_{i,j} - u_{i,j-1}) + (\Delta t v)^2 u_{xx,i,j}$$

Uttrykket viser at dersom vi kjenner bølgen i ett tidspunkt og i det forrige, kan vi beregne utslaget til bølgen ved neste tidspunkt i vår beskrivelse. Dette er en viktig formel som vi skal dvele en del ved:

Algoritmen for å beregne hvordan en bølge utvikler seg i tid og rom, er gitt ved ligningen:

$$u_{i,j+1} = u_{i,j} + (u_{i,j} - u_{i,j-1}) + (\Delta t v)^2 u_{xx,i,j} \quad (8.10)$$

Disse leddene er faktisk ganske enkle å forstå:

- Første ledd til høyre for likhetstegnet sier at vi må ta utgangspunkt i nåværende utslag til et punkt i bølgen når vi skal beregne utslaget i neste tidspunkt.
- Andre ledd svarer til at vi antar at den tidsderiverte til utslaget i vårt gitte punkt på bølgen vil være omtrent den samme i neste tidssteg som den var i det forrige. Dette er “treghetsleddet” svarende til Newtons første lov.
- Tredje ledd sier at dersom bølgen i vårt gitte punkt bujer (ofte: Buler vekk fra likevektstilstanden), finnes det en “gjenopprettende kraft” som forsøker å dra systemet tilbake mot likevektstilstanden. Se figur 8.15. Denne gjenopprettende kraften har en nær sammenheng med fasenhastigheten til bølgen. I uttrykket inngår fasenhastigheten i annen potens. Fasenhastigheten er derfor bestemt av hvor kraftig *naboområdet* påvirker bevegelsene til et vilkårlig valgt punkt i bølgen. Algoritmen kan anskuelig gjøres slik det er gjort i figur 8.16

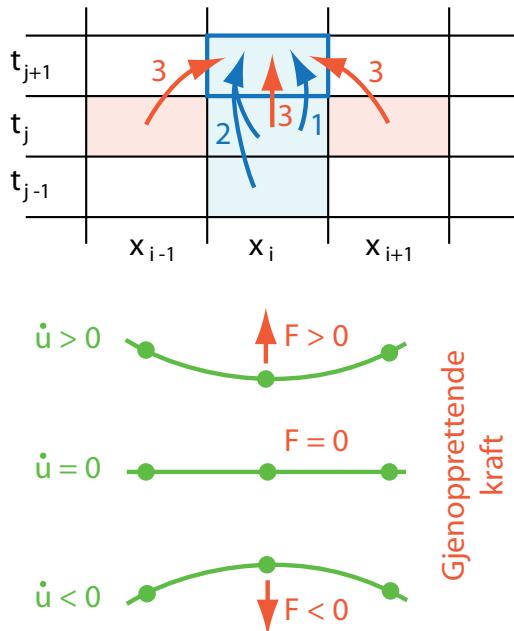
Algoritmen i ligning (8.10) viser at dersom vi kjenner til bølgen i ett tidspunkt t_j og bølgen slik den var et lite tidssteg før dette tidspunktet, t_{j-1} , så kan vi beregne bølgen slik den vil være i det kommende tidspunktet t_{j+1} . Det er utfordringer som må løse for hvordan vi bruker initialbetingelser og grensebetingelser, men det kommer vi tilbake til straks.

Ligning (8.10) er vel det enkleste uttrykket å ta utgangspunkt i når vi skal *forstå* grunnlaget for en beregningsalgoritme. Uttrykket egner seg ikke så godt for utformingen av selve programkoden. Da er det en fordel å sette ligning (8.7) inn i ligning (8.10), og resultatet blir med litt ordning av leddene:

$$\begin{aligned} u_{i,j+1} = & 2 \left(1 - \left(\frac{\Delta t v}{\Delta x} \right)^2 \right) u_{i,j} - u_{i,j-1} + \\ & \left(\frac{\Delta t v}{\Delta x} \right)^2 (u_{i+1,j} + u_{i-1,j}) \end{aligned} \quad (8.11)$$

Problem ved kantene av beregningsområdet

Ligning (8.11) er det sentrale uttrykket vi bruker for å beregne hvordan en bølge utvikler



Figur 8.16: Den sentrale algoritmen som kan brukes for å beregne tidsutviklingen til en en-dimensjonal bølge når vi kjenner initialbetingelser og randbetingelser. Nytt utslag i et punkt bestemmes av utslaget nå i samme punkt (1), hastigheten utslaget endret seg i like før nå (2) og hvilken vei den gjenopprettende kraften fra naboene til punktet virker.

seg i tid, men uttrykket inneholder noen viktige detaljer vi må se nærmere på. Når vi skal starte beregningene, har vi antatt at vi kjenner initialbetingelsene langs den delen av bølgen vi beskriver ved start av beregningene. For eksempel er utslaget ved tidspunktet $j = 0$ gitt ved $\{u_{i,0}\}$ for $i = 0, 1, 2, \dots, N$. Men i ligning (8.11) inngår også $x_{i+1,0}$ og $x_{i-1,0}$. Punktene $x_{-1,0}$ og $x_{N+1,0}$ eksisterer ikke, så algoritmen vår må gjøre noen kunstgrep for å behandle disse leddene. Vi må med andre ord angi såkalte “randbetingelser” for beregningsområdet vårt (engelsk: “boundary conditions”). Dette problemet gjelder ethvert tidspunkt i beregningene ut over initialbetingelsene..

I praksis kan det være nærliggende å finne randbetingelser som er perfekte for beregningene vi ønsker å gjøre. Mest vanlige randbetingelser er “åpen rand” og “fullstendig reflekterende rand”. I det første tilfellet setter vi $x_{-1,j} = x_{N+1,j} = 0$, i siste tilfellet setter vi f.eks. $x_{-1,j} = x_{0,j}$ og $x_{N+1,j} = x_{N,j}$. For en konkret beregning må vi selv velge hvordan vi skal angi randbetingelsene, og svaret avhenger i mange tilfeller sterkt av det fysiske systemet vi forsøker å beskrive.

For en bølge med null utslag ut mot randen, kan vi uten feil betrakte tidsutviklingen til bølgen helt til bølgen har brent seg til kanten av beregningsområdet. Ved å gjøre beregningsområdet stort nok, og begrense oss i hvor lenge vi betrakter bølgeutviklingen, kan beregninger av lokaliserte bølger bli bra selv uten å bekymre seg for randeffekter.

Problem med start-tidspunkt

Et annet ledd i ligning (8.11) som skaper problemer, er ledet $u_{i,j-1}$. Dersom vi starter beregningene ved tiden $t = 0$, eksisterer det ikke noe $u_{i,-1}$. Vi får derfor problemer med å starte selve beregningene.

På den annen side må vi ved alle differensialligninger ta utgangspunkt i initialbetingelsene (eller tilsvarende) for å komme fram til den spesielle løsningen vi søker. For en bølge betyr det at initialbetingelsene f.eks. kan være angitt som utslag i alle posisjoner ved

$t = 0$, sammen med den tidsderiverte av utslaget i alle posisjoner ved samme tidspunkt. Ut fra disse opplysningene kan vi regne oss fram til posisjoner ved starttidspunktet og tilnærmete posisjoner ett tidssteg før starttidspunktet.

Det er også andre måter å angi initialbetingelser på, og prosedyrer som kan følges for å utnytte initialbetingelsene. Vi holder oss bare til utslag og den tidsderiverte av utslaget, begge som funksjon av posisjon.

Den tidsderiverte av utslaget i et punkt i kan angis på følgende måte:

$$\dot{u}_{i,j} \equiv \left(\frac{\partial u}{\partial t} \right)_{i,j} \approx \frac{u_{i,j} - u_{i,j-1}}{\Delta t}$$

Følgelig:

$$u_{i,j-1} = u_{i,j} - \Delta t \dot{u}_{i,j} \quad (8.12)$$

For $j = 0$ får vi:

$$u_{i,-1} = u_{i,0} - \Delta t \dot{u}_{i,0} \quad (8.13)$$

Samlet

Anta at initialbetingelsene er gitt ved utslag $\{u_{i,0}\}$ i alle posisjoner langs bølgen og den tidsderiverte av utslaget $\{\dot{u}_{i,0}\}$ i alle posisjoner langs bølgen ved starttidspunktet. Da kan ligning (8.11) i kombinasjon med ligning (8.13) brukes for første tidssteg i beregningene. Dernest kan ligning (8.11) brukes for de resterende tidsstegene så mange tidssteg vi måtte ønske. Underveis må det tas hensyn til randbetingelsene.

Vi har ikke her laget såkalte dimensjonsløse uttrykk slik vi ofte finner i numeriske beregninger. Grunnen er som nevnt i kapittel 2, at for fenomenene som tas opp i denne boka er de aktuelle måleenhetene såpass nær opp til SI-enheter at faren for tap av numerisk presisjon er mindre enn f.eks. i kjernefysikk/elementærparikkelfysikk på den ene siden og kosmologi på den andre.

Det er også en annen bevisst grunn i vår sammenheng å beholde ligningene slik de er, uten å bruke dimensjonsløse uttrykk. Det er enklere å gjennomskue hva de ulike leddene i ligningene (8.10) og (8.11) representerer når vi unngår å bruke dimensjonsløse uttrykk.

8.6.1 Et bølge-eksempel

La oss som et eksempel beregne hvordan en gaussisk klokkeformet bølge beveger seg på en streng. Initialbetingelsene er et øyeblikksbilde av bølgen slik den er i ett tidspunkt (både mhp posisjon og hastighet!), og vi skal så følge den videre utviklingen i tid.

Utslaget som funksjon av posisjon langs strengen er gitt analytisk ved:

$$u(x, t) = A e^{-\left(\frac{x-vt}{2\sigma}\right)^2} \quad (8.14)$$

Den tidsderiverte av utslaget er da:

$$\frac{\partial u}{\partial t} \equiv \dot{u} = A e^{f(x,t)} \frac{df}{dt} = A e^{-\left(\frac{x-vt}{2\sigma}\right)^2} (-2) \left(\frac{x-vt}{2\sigma} \right) \left(-\frac{v}{2\sigma} \right) = \frac{(x-vt)v}{2\sigma^2} A e^{-\left(\frac{x-vt}{2\sigma}\right)^2}$$

$$\frac{\partial u}{\partial t} \equiv \dot{u} = \frac{v}{2\sigma^2} (x-vt) u \quad (8.15)$$

Vi velger å beskrive bølgen på en streng som er lang i forhold til bredden på klokkefunksjonen, og vi velger å følge bølgen bare så lenge at vi ikke støter nevneverdig mot endene av beregningsområdet vårt. Vi bruker i programmet en fullstendig fastlåsing av endepunktene underveis i beregningene.

Vi velger følgende parametere $A = 1.0$, $\sigma = 2.0$, $v = 0.3$, og lar x dekke området fra og med -20 til og med +20 i 400 ekvidistante steg. Vi forsøker oss med en Δt lik 0.1 og følger bevegelsen i 300 tidssteg. Ingen enheter er gitt, men vi antar at alle enheter er SI-enheter.

Et dataprogram skrevet i Matlab er gitt nedenfor. Programmet foretar beregningene ut fra de uttrykkene som er gitt ovenfor.

```
function bolgeanimasjonX

% Genererer posisjonsarray
delta_x = 0.1;
x = -20:delta_x:20;
n = length(x);

% Genererer posisjoner ved t=0
sigma = 2.0;
u = exp(-(x/(2*sigma)).*(x/(2*sigma))); % Gaussisk form
plot(x,u,'r');
figure;

% Genererer div parametre og tidsderivert av utslag ved t=0
v = 0.5;delta_t = 0.1;
faktor = (delta_t*v/delta_x)^2;
dudt = (v/(2*sigma*sigma))*x.*u;

% Angir effektive initialbetingelser:
u_jminus1 = u - delta_t*dudt;
u_j = u;

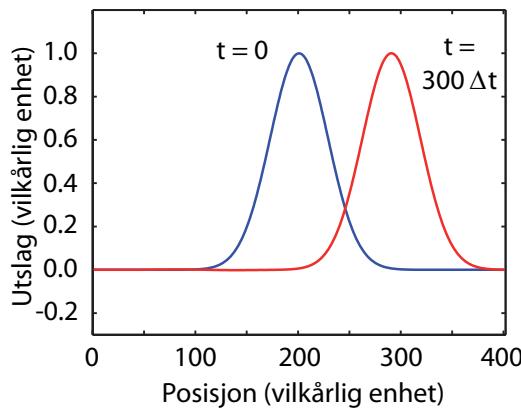
for t = 1:1000
    u_jplus1(2:n-1) = (2*(1-faktor))*u_j(2:n-1) - ...
        u_jminus1(2:n-1) + faktor.*((u_j(3:n)+u_j(1:n-2));
    % Håndtering av randproblem
    u_jplus1(1) = (2*(1-faktor)).*u_j(1) - u_jminus1(1) + faktor.*u_j(2);
    u_jplus1(n) = (2*(1-faktor)).*u_j(n) - u_jminus1(n) + faktor.*u_j(n-1);

    plot(u_j);
    axis([0 n+1 -0.3 1.2])
    drawnow;

    u_jminus1 = u_j;
    u_j = u_jplus1;
end;
```

Figur 8.17 viser bølgen ved startpunktet for tid og 300 tidssteg senere. Vi ser at bølgen beveger seg mot høyre (positiv v) og at bølgens form beholdes uforandret.

I en oppgave sist i kapittelet blir du bedt om å undersøke hvordan bølgen utvikler seg dersom vi bruker en \dot{u} som enten er for liten eller for stor i forhold til hva den burde vært ifølge ligning (8.15). Vi ber også i enda en oppgave om at du modifiserer koden slik at du kan håndtere et tilfelle der bølgen treffer på et grensesjikt mellom to medier med ulik impedans (ulik fasehastighet). Det anbefales sterkt at du gjennomfører disse oppgavene, siden det kan gi en betydelig bedre forståelse av bølger.



Figur 8.17: Eksempel på bølgen ved start av beregningene og 300 tidssteg senere.

8.7 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for forskjeller i overflatedrevne bølger på vann og lydbølger gjennom vann.
- Gjøre rede for de to forskjellige “gjenopprettende krefter” ved overflatebølger på vann.
- Angi et omtrentlig kriterium for hvorvidt det er overflatespenningen eller gravitasjonen som dominerer i et gitt tilfelle.
- Gi eksempler på overflatespenningsdrevne bølger og gravitasjonsdrevne bølger.
- Gjøre rede for en modell hvor vi forklarer/beskriver bølger ved at vannmolekyler følger en sirkulær bevegelse.
- Finne tilnærmede uttrykk for fasenhastigheten og gruppehastighet til bølger både på grunt og dypt vann ve å ta utgangspunkt i formelen

$$v_f^2(k) = \left[\frac{g}{k} + \frac{Tk}{\rho} \right] \tanh(kh)$$

- Gjengi hovedtrekkene i figur 8.6.
- Gjøre rede for forskjellen på fase- og gruppehastighet generelt, og sette opp / utlede et matematisk uttrykk for å demonstrere forskjellen (f.eks. slik det er gjort i ligning (8.2)).
- Gi eksempler på dispersive fysiske systemer, både system med normal dispersjon og anomal dispersjon.
- Gjennomføre numeriske beregninger av tidsforløpet for en en-dimensjonal bølge.
- Gjøre rede for algorimens innhold ved slike beregninger.

8.8 Referanser

1. R.E.Apfel, Y.Tian et al. : Free Oscillations and Surfactant Studies of Superdeformed Drops in Microgravity. Phys. Rev. Lett. 78 (1997) 1912-1915. (Stor vanndråpe analysert i romferjen Columbia.)
2. H.Azuma og S. Yoshihara: Three-dimentional large-amplitude drop oscillations: Experiments and theoretical analysis. J.Fluid Mech. 393 (1999) 309-332.
3. Se f.eks. <http://www.youtube.com/watch?v=YcF009w4HEE> (Het Leidenfrost-effect: de dansende drappel (versie 2)) eller siste halvpart av videoen på <http://www.youtube.com/watch?v=b7KpHGgfHkc> (JuliusGyula_HotPot 1.3). Begge var tilgjengelig på web 21. februar 2013.
4. Jonas Persson: Vågrörelselära, akustik och optik. Studentlitteratur 2007.

8.9 Oppgaver

Forståelses- / diskusjonsspørsmål

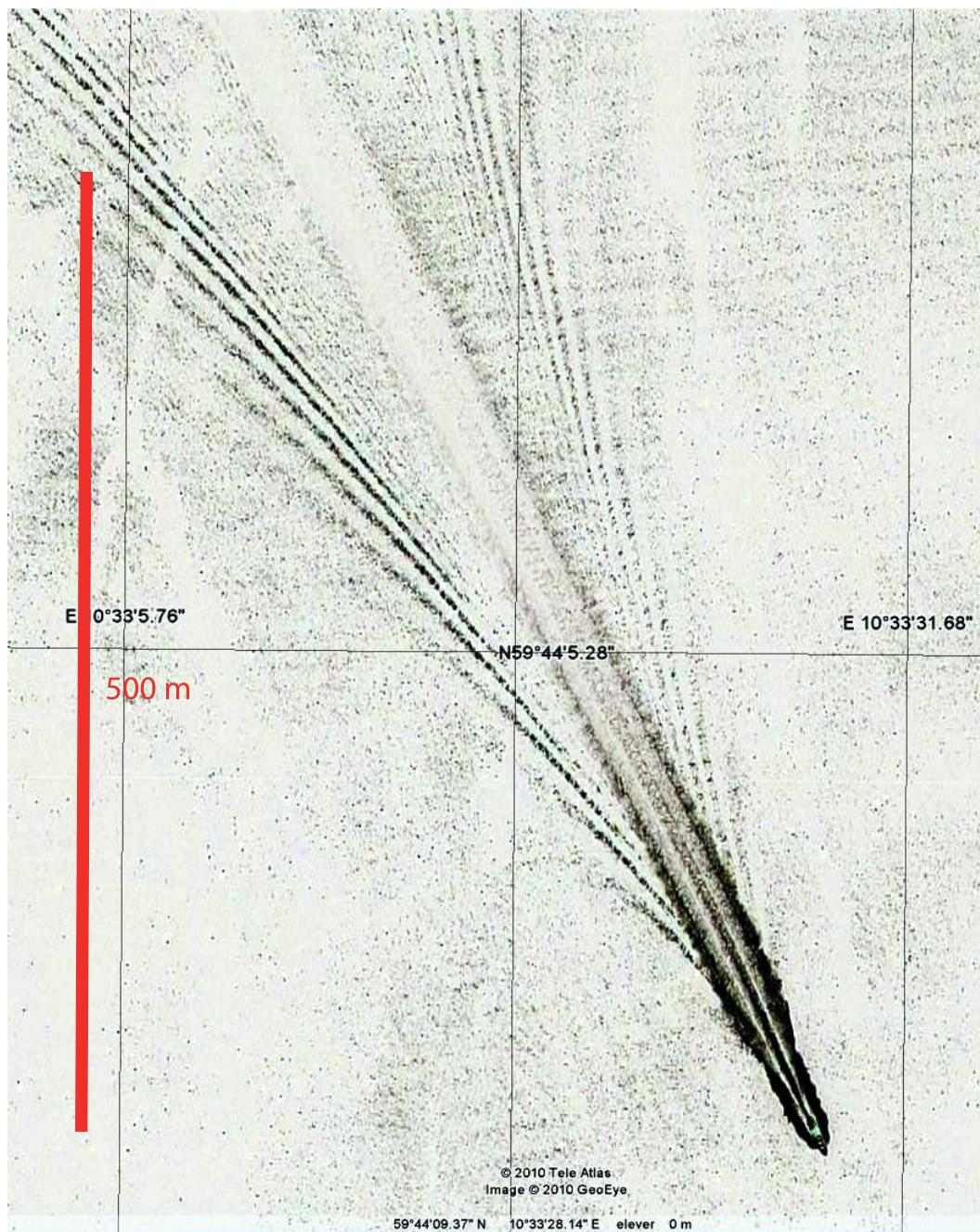
1. Er overflatebølger på vann transversale eller longitudinale bølger? Forklar.
2. Forsök å begrunne hvorfor vi ikke merker noe effekt av overflatebølger på vann i en dybde som er stor i forhold til bølgelengden.
3. Forklar hvorfor bølger ruller inn med bølgetoppene parallelt med vannkanten på en langgrunn strand.
4. Hva kjennetegner dispersjon? Hva er en dispersjonsrelasjon? Er det dispersjon som ligger bak fenomenet at bølger ofte kommer inn tilnærmet parallelt med en sandstrand?
5. Hva mener vi med et dispersivt medium? Hvordan vil dispersjon påvirke bølgebevegelsen til a) en harmonisk bølge, og b) en ikke-harmonisk bølge?
6. I denne boka har vi hittil addert to sinusuttrykk (eller cosinusuttrykk) i tre ulike sammenhenger: Stående bølger, svevelyd-frekvens og for å illustrere dispersjon. Gjør rede for forskjellene i fysiske forhold som gjør at disse matematiske beskrivelsene kommer ulikt ut i de tre tilfellene.
7. Hva er forskjellen mellom normal og anomal dispersjon?

Regneoppgaver

8. Sjekk ved egen utregning at bølgelengden er om lag 1.7 cm når overflatebølger på vann er like mye styrt av overflatespenning som av gravitasjon. Overflatespenning for rent vann ved 25 °C er $7.197 \cdot 10^{-2}$ N/m.
9. Bestem fasehastigheten for overflatebølger på "dypt" vann ved en bølgelengde på 1.7 cm. (Tips: Bruk info fra forrige oppgave.)
10. I figur 8.18 (siste side i dette kapittelet) er det vist et utsnitt fra et flyfoto tatt utenfor Fagerstrand i Oslofjorden (tatt fra Google Earth). Det er satt inn en stav som angir målestokken. Bildet viser bølger fra en båt. Beregn hastigheten båten kjørte med under antakelsen at det er dypt vann. Angi hastigheten i m/s, km/t og i knop. (1 knop = 1 nautisk mil per time, 1 nautisk mil = 1852 m = 1 bueminutt i nord-sør retning langs jordoverflatene.)
11. Sett opp et matematisk uttrykk (basert på bølgetall og vinkelfrekvens) for en plan, monokromatisk harmoniske bølge. Angi fasehastighet og gruppehastighet i den grad de er definert.
12. Sett opp et matematisk uttrykk for en stående bølge (som en sum av to plane harmoniske bølger). Angi fasehastighet og gruppehastighet i den grad de er definert.
13. To bølger summeres, nærmere bestemt bølgene $y_1 = A \cos(6x - 12t)$ og $y_2 = A \cos(8x - 14t)$ der det antas at x er måletallet for posisjon, og t er måletallet for tiden i sekunder.
 - Finn både fasehastigheten og gruppehastigheten for den kombinerte bølgen.
 - Er mediet bølgene går i dispersivt? I så fall, er det normal eller anomal dispersjon?
 - Bestem avstanden mellom to etterfølgende toppunkter i omhyllingskurven ved ett tidspunkt.
 - Bestem hvor lang tid det tar fra at vi har et toppunkt på omhyllingskurven til neste gang vi har et toppunkt på samme sted.
 - Hvor mange bølgelengder i den virkelige bølgen er det mellom to topper i omhyllingskurven dersom bølgen betraktes ved ett tidspunkt? Sammenlign dette med antall virkelige bølgelengder som passerer en observatør fra han merker at omhyllingskurven har et maksimum til neste maksimum.
14. Lag ditt eget program for å beregne numerisk løsninger av bølgeligningen. Ta gjerne utgangspunkt i programmet gjengitt under punkt 8.6.1 ovenfor. Test at en bølge beskrevet ved ligningene (8.14) og (8.15) kommer ut som vist i figur 8.17. Gjør så følgende endringer:
 - Endre den tidsderiverte av utslaget i startøyeblikket til den negative av hva det skulle ha vært. Gjennomfør beregningene og beskriv hva du observerer.
 - Reduser den tidsderiverte av utslaget i startøyeblikket til det halve i forhold til hva det skulle ha vært. Gjennomfør beregningene og beskriv hva du observerer.
 - Bruk i stedet den dobbelte tidsderiverte av utslaget i stedet for den korrekte i startøyeblikket. Gjennomfør beregningene og se hva du observerer denne gang. Vær nøy med å peke på både amplituder og faser.
 - Hvordan vil du lage initialbetingelsene for å simulere stående bølger? [Du kan gjerne teste dette ut, men du MÅ ikke gjøre det.]
 - Hvilken slutning kan du trekke av alle beregningene i denne oppgaven? Ved en pendelbevegelse kan vi velge posisjon og fart helt uavhengig av hverandre, og får alltid en svingbevegelse som er lett å forstå. Er det likadan for bølger?

15. Modifisér programmet du brukte i forrige oppgave slik at det kan behandle det tilfellet at en plan en-dimensjonal bølge langs en streng møter et materiale med en annen fasehastighet. Bølgen skal kunne fortsette inn i det nye materialet og evt også bli reflektert i punktet der strengen endrer egenskap (kan svare til at strengen endrer masse per lengde). Forsök både med en 30 % økning i fasehastighet og en 30 % reduksjon i fasehastighet. Beskriv resultatene og kommenter om resultatene stemmer overens med det som er beskrevet i kapittel 5 eller ikke.
16. Lag noen enkle skisser som viser hvordan du FØR du gjør beregningene (eller hører om resultatet fra medstudenter) forestiller deg at en gitarstreng svinger. Lag *dernest* et dataprogram som beregner bevegelsen til en gitarstreng minst et par svingeperioder etter at strengen ved hjelp av et plekter eller en fingernegl er dradd ut til siden i ett punkt i en avstand ca 1/3 av strenglengden fra en ende, og sluppet derfra (etter å ha vært i ro). Ta gjerne utgangspunkt i programmet gjengitt under punkt 8.6.1 ovenfor. Beskriv bevegelsen.

[Sjekk gjerne ETTER du har gjort beregningene, om det er samsvar mellom dine beregninger og YouTube-filmer av en gitarstreng tatt opp med meget hurtig videokamera. Best overensstemmelse med beregningene får du ved å se en video på YouTube "Slow motion: rubber string pulled and released" av Pavel Radzivilovsky. En stikk har ingen stivhet, og vi har ikke med noe ledd i våre beregninger som svarer til stivhet. Derfor blir resultatet vårt svært nært det som kan observeres som en stående bølge på en strikk.]



Figur 8.18: Bilde av bølgemønsteret bak en båt i Oslofjorden utenfor Fagerstrand. (Bildet er i negativt format for at mønsteret skal komme letttere fram.)

Kapittel 9

Maxwells ligninger og elektromagnetiske bølger



James Clerk Maxwell

James Clerk Maxwell is one of the most influential scientists of all time. Albert Einstein acknowledged that the origins of the special theory of relativity lay in Clerk Maxwell's theories, saying "The work of James Clerk Maxwell changed the world forever".

Clerk Maxwell's research into electromagnetic radiation led to the development of television, mobile phones, radio and infra-red telescopes. The largest astronomical telescope in the world, at Mauna Kea Observatory in Hawaii, is named in his honour.

Photo: James Clerk Maxwell as a young man, courtesy of The Master and Fellows of Trinity College, Cambridge.

www.bbc.co.uk/history/people/james_clerk_maxwell

Jeg må innrømme at jeg får en form for "andaktsfølelse" når jeg jobber med Maxwells ligninger. Det er for meg en edelstein i fysikkens utvikling.

Bragden til Maxwell regnes som en av de aller største i fysikkens historie. Men jeg synes også det er interessant å huske på at Maxwell bygget på en rekke arbeider fra fysikere og matematikere før ham. Selv synes jeg f.eks. at historien om Faraday også er svært fascinerende.

Fysikkens utvikling er avhengig av et elegant samspill mellom eksperimenter, matematikk/modellering og evnen til å oppdage sammenhenger på tvers av ulike fenomener. Alt dette mestret Maxwell.

La oss dukke inn i Maxwells verden og forsøke å forestille oss hans begeistring da han så at elektromagnetismen faktisk kunne føre til bølger!

Bildemontasjen er fra BBCs websider (adresse gitt nederst), lastet ned 3.3.2013.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes.

Versjon 03032013.

9.1 Innledning

Av alle bølgefenes som betyr noe for oss mennesker, er lydbølger og elektromagnetiske bølger i en særstilling. Teknologisk sett er det elektromagnetiske bølger som rangerer aller høyest.

Vi kommer i mange av de resterende kapitlene i denne boka til å møte elektromagnetiske bølger i litt ulike sammenehenger. Det er derfor naturlig at vi går litt i dybden i beskrivelsen av stoffet, for det er slett ikke slik at all elektromagnetisme kan reduseres til elektromagnetiske bølger. Det betyr at vi må være nøyne når vi behandler dette stoffet for å ikke trekke feilslutninger. Presisjonsnivået må være høyt!

Dette kapitlet er det mest matematiske av alle kapitlene i boka. Vi starter med Maxwells ligninger på integralform og viser hvordan de kan omformes til differensialform. Dernest vises det at Maxwells ligninger under visse forutsetninger kan lede til en enkel bølgeligning. Elektromagnetiske bølger er transverselle, det medfører at kompleksiteten er noe større enn for longitudinale lydbølger.

Kapitlet forutsetter at leseren tidligere har vært gjennom et kurs i elektromagnetisme, og kjenner til matematiske begreper så som linjeintegral og flateintegral. Det er en stor fordel å også kjenne Stokes teorem og divergensteoremet og vektorfelt-matematikken knyttet til divergens, gradient og rotasjon, og leseren bør helst kjenne til forskjellen mellom skalarfelt og vektorfelt før hun/han gir seg i kast med kapitlet..

Som nevnt er kapitlet sterkt preget av matematikk. Vi har likevel forsøkt å peke på fysikken bak matematikken, og vi anbefaler at du bruker mye tid også på den delen. Det er en utfordring å gripe lovmessigheten i Maxwells ligninger fullt ut!

Det er erfaringmessig mange misforståelser knyttet til elektromagnetisme. En vanlig misforståelse, utrolig nok, er at en elektromagnetisk bølge er et elektron som svinger opp og ned på tvers av den retningen bølgen brer seg. Andre misforståelser er vanskeligere å rydde bort. For eksempel er det mange som tror at løsningen av bølgeligningen er “plane bølger”, og at Poynting vektor alltid forteller om energitransporten i bølgen. Vi bruker en del tid på å drøfte slike misforståelser og håper at også dette stoffet kan være nyttig for noen.

Bak i kapitlet er det tatt med en huskeliste for de matematiske operasjonene som brukes såvel som en huskeliste angående hvordan elektrisk og magnetisk felt og fluksstetheter forholder seg til hverandre. Det kan være nyttig lesning og oppfriskning av tidligere kunnskap.

La oss da sette i gang med Maxwells fabelaktige systematisering (og utvidning) av alle kjente elektriske og magnetiske lover som fantes i 1864!

9.2 Maxwells ligninger på integralform

Fire ligninger forbinder elektriske og magnetiske felt:

1. Gauss lov for elektrisk felt:

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q_{innenfor}}{\epsilon_r \epsilon_0} \quad (9.1)$$

2. Gauss lov for magnetfelt:

$$\oint \vec{B} \cdot d\vec{A} = 0 \quad (9.2)$$

3. Faraday-Henrys lov:

$$\oint \vec{E} \cdot d\vec{l} = - \left(\frac{d\Phi_B}{dt} \right)_{innenfor} \quad (9.3)$$

4. Ampère-Maxwells lov:

$$\oint \vec{B} \cdot d\vec{l} = \mu_r \mu_0 \left(i_f + \epsilon_r \epsilon_0 \left(\frac{d\Phi_E}{dt} \right)_{innenfor} \right) \quad (9.4)$$

Vi regner med at du kjenner disse lovene fra før og går derfor ikke i stor detalj om hvordan de skal oppfattes eller hva symbolene betyr. I de to første ligningene integreres fluksen ut av en lukket flate og sammenholdes med kilden i volumet innenfor (elektrisk monopol, dvs. ladning, og magnetisk monopol, som ikke finnes). Vektoren $d\vec{A}$ er positiv dersom den vender ut av volumet den lukkede flaten avgrenser.

I de to siste ligningene beregnes linjeintegralet for elektrisk eller magnetisk felt langs en linje som omspenner en åpen flate. Linjeintegralet sammenholdes med fluks av magnetisk fluksstetthet eller elektrisk fluksstetthet samt fluks av elektriske strømmer av frie ladninger gjennom den åpne flaten. Fortegnene er da bestemt ut fra høyrehåndsregelen (når fire fingre på høyre hånd peker i den retningen vi integrerer langs randen, peker tommelen i den retningen som svarer til positiv fluks).

Alle disse detaljene regnes som kjent.

Symmetrien kommer best fram dersom siste ligning skrives på følgende form:

$$\oint \vec{H} \cdot d\vec{l} = \left(i_f + \left(\frac{d\Phi_D}{dt} \right)_{innenfor} \right) \quad (9.5)$$

Her er det brukt følgende relasjon mellom magnetisk feltstyrke H og magnetisk fluksstetthet B :

$$\vec{H} = \vec{B}/(\mu_r \mu_0)$$

hvor μ_0 er (magnetisk) permeabilitet i vakuum og μ_r er relativ permeabilitet.

Det er også brukt følgende relasjon mellom elektrisk feltstyrke E og elektrisk fluksstetthet D (også kalt "forskyvningsvektor"):

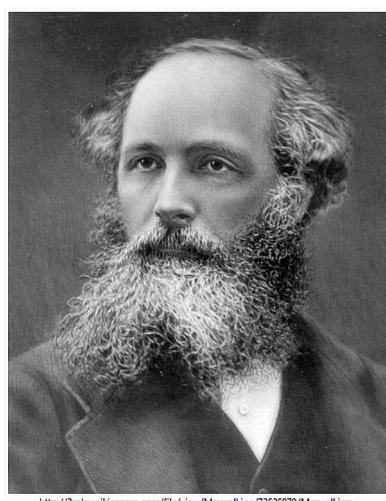
$$\vec{E} = \vec{D}/(\epsilon_r \epsilon_0)$$

hvor ϵ_0 er (elektrisk) permittivitet i vakuum og ϵ_r er relativ permittivitet.

Venstresiden av ligningene (10.1) og (9.5) er da linjeintegraler av feltstyrker (\vec{E} og \vec{H}), mens høyresiden er den tidsderiverte av fluksen gjennom den avgrensede flaten, pluss elektrisk strøm-fluks av frie ladninger. Fluksen er fluksstettheter (\vec{B} og \vec{D}) integrert over flaten.

Innholdet i Maxwells ligninger kan gis med ord omrent som så:

- Det er to kilder til elektrisk felt. Den ene kilden skyldes elektriske ladninger (kan betraktes som monopoler). Elektrisk felt fra ladninger er radielt rettet bort fra eller direkte mot ladningen, alt etter ladningens fortegn. (Dette er innholdet i Gauss lov for elektrisk felt.)
- Den andre kilden for elektrisk felt er et tidsvarierende magnetfelt. Elektrisk felt som oppstår på denne måten har en rotasjon, det vil si at feltlinjene har en tendens til å danne sirkler på tvers av retningen der magnetfeltet endres i tid. Hvorvidt det blir sirkler eller en annen orientering i rommet avhenger av grensebetingelser. (Dette er innholdet i Faradays lov.)
- Det er to bidrag til magnetfelt også, men det finnes ikke magnetiske monopoler. Magnetfelt vil derfor aldri strømme ut i radiell retning fra et kildepunkt på tilsvarende måte som elektriske feltlinjer nær en elektrisk ladning. (Dette er innholdet i Gauss lov for magnetfelt.)
- Derimot kan magnetfelt oppstå omrent som for elektrisk felt ved at et elektrisk felt varierer i tid. En alternativ måte å lage et magnetfelt på, er å ha fri ladninger i bevegelse som danner en netto elektrisk strøm. Begge disse kildene til magnetfelt gir felt som har en tendens til å danne lukkede kurver på tvers av retningen tidsvariasjonen i elektrisk felt eller netto elektrisk strøm er rettet. Hvilken form disse lukkede kurvene får i praksis er derimot helt avhengig av randbetingelsene. (Dette er innholdet i Ampère-Maxwells lov.)



<http://2mka.wikispaces.com/file/view/Maxwell.jpg/73625079/Maxwell.jpg>

Figur 9.1: James Clerk Maxwell. (Opphav til bildet gitt med liten skrift.)

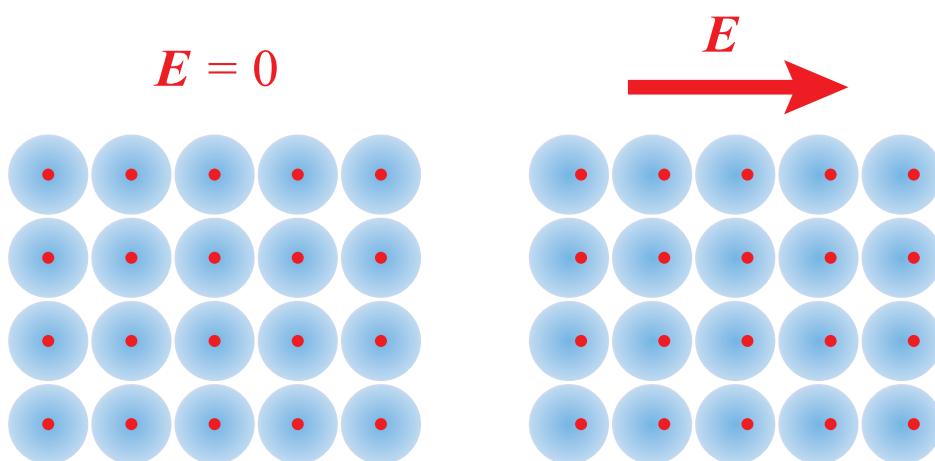
Det var fysikeren og matematikeren James Clerk Maxwell (1831 - 1879, figur 9.1) som mestret å samle all kunnskap om elektriske og magnetiske lover i én helhetlig formalisme. Hans publikasjon “A Dynamical Theory of Electromagnetic Field” ble publisert i 1865 og regnes som en like stor bragd som Newtons lover og Einsteins relativitetsteori(er). (Den

originale 54 siders lange artikkelen kunne lastes ned fra referanse 1 på:
http://en.wikipedia.org/wiki/A_dynamical_theory_of_the_electromagnetic_field
 den 26. februar 2013.)

Det kan kanskje være på plass med en liten repetisjon av noen detaljer her. Vi vil senere i kurset se at magnetisk permeabilitet og i særdeleshet elektrisk permittivitet spiller en vesentlig rolle for elektromagnetiske bølger. Verdiene i vakuum μ_0 og ϵ_0 er temmelig uinteressante. De har først og fremst sammenheng med hvordan vi har valgt enheter for elektrisk og magnetisk felt.

De relative verdiene er derimot av langt mer interesse. Den relative (magnetiske) permeabiliteten har sammenheng med hvor mye magnetfelt vi genererer i et materiale når det utsettes for et ytre magnetfelt. I et diamagnetisk materiale vil det genereres et litte lite magnetfelt i materialet, og feltet er rettet motsatt av det ytre magnetfeltet. I et paramagnetisk materiale genereres det også et litte lite magnetfelt i materialet, men nå i samme retning som det ytre feltet. Det genererte magnetfeltet i selve materialet er bare i størrelsesordenen 10^{-5} ganger det ytre magnetfeltet i begge tilfellene. I et ferromagnetisk materiale genereres et betydelig magnetfelt inne i materialet som følge av det ytre magnetfeltet, og i samme retning som dette. Det er mange detaljer knyttet til disse prosessene, og vi går ikke inn på disse her.

Siden de fleste stoffene vi kommer i kontakt med i dette kurset enten er diamagnetiske eller paramagnetiske, kan vi stort sett sette den relative permeabiliteten lik 1.0 og se bort fra magnetfeltets vekselvirkning med materialer i prosessene vi kommer til å diskutere.



Figur 9.2: I et materiale, f.eks. glass, vil et ytre elektrisk felt lett kunne gi en polarisering av ladningsfordelingen i hvert enkelt atom i materialet. Denne polariseringen fører til et elektrisk felt inne i materialet som er rettet motsatt vei av det ytre elektriske feltet.

For det elektriske feltet er det annerledes. Den relative (elektriske) permittiviteten sier oss noe om hvor stort elektrisk felt som oppstår inne i et materiale når det utsettes for et ytre elektrisk felt. I figur 9.2 er det gitt en skjematiske fremstilling av hva som skjer. Et ytre elektrisk felt vil føre til at elektronskyen omkring en atomkjerne forskyves litt. Men siden atomene er små og mange og ladningen på elektronene betydelig, kan det genererte elektriske feltet inne i materialet lett komme opp i samme størrelsesorden som det ytre elektriske feltet (f.eks. halvparten så stort).

Merk at det ikke er snakk om transport av fri ladninger! Det er bare snakk om en lokal polarisering av ladningsfordelingen av hvert enkelt atom, men dette gir tross alt en polarisering av hele materialet. Merk forøvrig at vi her snakker om "polarisering" i en bestemt betydning. Vi kommer snart til å snakke om polarisering i en helt annen sammenheng, så her kreves det at man er oppmerksom for å ikke blande sammen ulike begreper med samme navn!

9.3 Over til differensialform

Vi skal nå vise hvordan vi kan gå fra Maxwells ligninger på integralform til differensialform. Integralformen kan anvendes på makroskopiske geometrier, for eksempel for å finne magnetfelt fem meter vekk fra en rett leder der det går en netto elektrisk strøm. Differensialformen gjelder for et lite område i rommet. Hvor "lite" dette er, kan diskuteres. Maxwells ligninger ble utviklet før vi hadde noe god kunnskap om atomer og stoffers oppbygging på det mikroskopiske plan. Maxwells ligninger på differensialform anvendes ofte i praksis på *en midlere lengdeskala som er liten i forhold til den makroskopiske verden og likevel stor i forhold til atomære størrelser.*

Ved overgangen fra integral- til differensialform anvendes to matematiske relasjoner som gjelder for vilkårlige vektorfelt \vec{G} generelt:

Stokes' teorem (mer korrekt Kelvin-Stokes' teorem, siden teoremet først ble kjent gjennom et brev fra Lord Kelvin. George Stokes (1819-1903) var en britisk matematiker/-fysiker. Lord Kelvin (1824-1907), som egentlig het William Thomson, var en matematisk fysiker omtrent samtidig med Stokes.)

Stokes teorem:

$$\oint \vec{G} \cdot d\vec{l} = \int_A (\nabla \times \vec{G}) \cdot d\vec{A} \quad (9.6)$$

Teoremet gir en sammenheng mellom et linjeintegral av et vektorfelt og fluksen av rotasjonen til vektorfeltet gjennom flaten som linjen avgrenser.

Den andre relasjonen vi benytter er *Divergensteoremet* (som ble oppdaget av Lagrange og gjenoppdaget av flere andre siden. Joseph Louis Lagrange (1736-1813) var en italiensk/fransk matematiker og astronom.):

Divergensteoremet:

$$\int \nabla \cdot \vec{G} dv = \oint_A \vec{G} \cdot d\vec{A} \quad (9.7)$$

Divergensteoremet gir sammenhengen mellom divergens til et vektorfelt i et volum og fluksen av vektorfeltet gjennom flaten som avgrenser volumet.

Gauss lov for elektrisk felt:

Vi starter med Gauss lov for elektrisk felt.

$$\epsilon_r \epsilon_0 \oint \vec{E} \cdot d\vec{A} = Q_{innenfor}$$

Bruker vi divergensteoremet, følger:

$$\oint \epsilon_r \epsilon_0 \vec{E} \cdot d\vec{A} = \int \nabla \cdot (\epsilon_r \epsilon_0 \vec{E}) dv = Q_{innenfor}$$

Vi velger nå et så lite volum at $\nabla \cdot (\epsilon_r \epsilon_0 \vec{E})$ er tilnærmet konstant over hele volumet. Denne konstanten kan i så fall settes utenfor integraltegnet, og integralet over volumet gir rett og slett det lille volumet Δv vi betrakter. Følgelig:

$$\int \nabla \cdot (\epsilon_r \epsilon_0 \vec{E}) dv \approx (\nabla \cdot \vec{D}) \Delta v = Q_{innenfor}$$

$$\nabla \cdot \vec{D} = \frac{Q_{innenfor}}{\Delta v} = \rho$$

hvor ρ er ladningstettheten lokalt. Følgelig har vi kommet fram til Gauss lov for elektrisk felt på differensialform:

$$\nabla \cdot \vec{D} = \rho \quad (9.8)$$

Gauss lov for magnetfelt:

Samme fremgangsmåte leder oss til Gauss lov for magnetfelt på differensiell form:

$$\nabla \cdot \vec{B} = 0 \quad (9.9)$$

Faraday-Henrys lov:

Vi vil nå omforme Faradays lov. Utgangspunktet er altså:

$$\oint \vec{E} \cdot d\vec{l} = - \left(\frac{d\Phi_B}{dt} \right)_{innenfor}$$

Benytter vi Stokes' teorem, får vi:

$$\oint \vec{E} \cdot d\vec{l} = \int_A (\nabla \times \vec{E}) \cdot d\vec{A} = - \left(\frac{d\Phi_B}{dt} \right)_{innenfor}$$

Magnetfeltfluksen gjennom flaten kan også skrives på denne måten:

$$\Phi_B = \int_A \vec{B} \cdot d\vec{A}$$

Følgelig:

$$\begin{aligned} \int_A (\nabla \times \vec{E}) \cdot d\vec{A} &= - \frac{d}{dt} \int_A \vec{B} \cdot d\vec{A} \\ &= - \int_A \frac{\partial \vec{B}}{\partial t} \cdot d\vec{A} \end{aligned}$$

Her har vi antatt at flaten dA ikke endrer seg med tiden. Vi har i tillegg skiftet fra vanlig derivert til partiell derivert siden magnetisk fluksstetthet \vec{B} avhenger både av tid og romlige forhold, men vi antar at de romlige forholdene i seg selv ikke endres i tid. For små nok areal A vil igjen hovedleddet i integranden kunne settes utenfor integrasjonen, og vi ender opp med:

$$\nabla \times \vec{E} = - \frac{\partial \vec{B}}{\partial t} \quad (9.10)$$

Dette er Faradays lov på differensiell form.

Ampère Maxwells lov:

Samme fremgangsmåte kan benyttes for å vise den siste av Maxwells ligninger på differensiell form, nemlig Ampère-Maxwells lov. Resultatet er:

$$\nabla \times \vec{H} = \vec{j}_f + \frac{\partial \vec{D}}{\partial t} \quad (9.11)$$

hvor \vec{j}_f er elektrisk strømtetthet av fri ladninger.

Samlet:

La oss til slutt sette opp alle Maxwells ligninger på differensiell form samlet:

$$\nabla \cdot \vec{D} = \rho \quad (9.12)$$

$$\nabla \cdot \vec{B} = 0 \quad (9.13)$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (9.14)$$

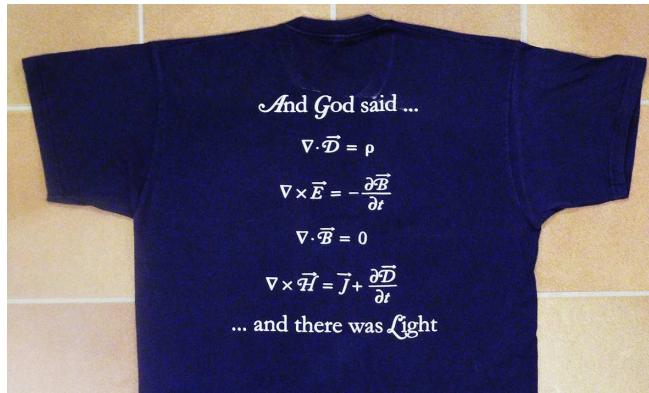
$$\nabla \times \vec{H} = \vec{j}_f + \frac{\partial \vec{D}}{\partial t} \quad (9.15)$$

Maxwells ligninger sammen med Lorentzkraften

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B})$$

danner det fullstendige grunnlaget for klassisk elektrodynamisk teori.

Einstein hadde bilder av Newton, Maxwell og Faraday på kontoret sitt, noe som indikerer hvor viktig han syntes deres arbeider var. Det er derfor ikke så rart at Fysikkforeningen ved UiO har valgt Maxwells ligninger på deres T-skjorter (se bilde 9.3) som et symbol på et høydepunkt i fysikk, et høydepunkt både mhp. hvor slagkraftig ligningene er og et høydepunkt i matematisk eleganse! (Det skal dog nevnes at den matematiske elegansen visstnok ikke var like blankpolert på Maxwells tid som den er i dag.)



Figur 9.3: Maxwells ligninger på T-skjorte.

9.4 Utledning av bølgeligningen

Bølgeligningen kan utledes fra Maxwells ligninger ved først og fremst å benytte de to siste ligningene sammen med en generell relasjon som gjelder for ethvert vilkårlig vektorfelt \vec{G} (jamfør med f.eks. Karl Rottmann: Matematisk formelsamling):

$$\nabla \times (\nabla \times \vec{G}) = -\nabla^2 \vec{G} + \nabla(\nabla \cdot \vec{G}) \quad (9.16)$$

Med ord sier relasjonen at “rotasjonen til rotasjonen til et vektorfelt er lik minus Laplaceoperatoren anvendt på vektorfeltet pluss gradienten til divergensen av vektorfeltet” (pusteøvelsen slutt).

Vi anvender denne relasjonen på elektrisk felt, og får:

$$\nabla \times (\nabla \times \vec{E}) = -\nabla^2 \vec{E} + \nabla(\nabla \cdot \vec{E})$$

Vi gjenkjenner rotasjon til elektrisk felt i uttrykket på venstre side av likhetstegnet. Denne erstattes ved Faradays lov. Samtidig bytter vi høyre og venstre side og bytter fortegn og får:

$$\nabla^2 \vec{E} - \nabla(\nabla \cdot \vec{E}) = -\nabla \times \left(-\frac{\partial \vec{B}}{\partial t} \right) \quad (9.17)$$

I høyresiden bytter vi rekkefølgen av derivering med hensyn på tid og derivering med hensyn på posisjon, og får:

$$= \frac{\partial}{\partial t} (\nabla \times \vec{B})$$

Vi anvender dernest Ampère-Maxwells lov og får:

$$= \frac{\partial}{\partial t} \left(\mu_r \mu_0 \left(\frac{\partial \vec{D}}{\partial t} + \vec{j}_f \right) \right) \quad (9.18)$$

hvor det også er benyttet at:

$$\vec{B} = \mu_r \mu_0 \vec{H}$$

For venstresiden av ligning (9.17) benyttes Gauss lov for elektrisk felt for å erstatte divergensen av elektrisk felt i det andre leddet på venstresiden med ladningstetthet ρ dividert med total permittivitet.

$$\nabla^2 \vec{E} - \frac{\nabla \rho}{\epsilon_r \epsilon_0} \quad (9.19)$$

Ved å sette høyre side (9.19) lik venstre side (9.18), samt flytte noen ledd over på motsatt side av likhetstegnet, ender vi opp med:

$$\nabla^2 \vec{E} - \epsilon_r \epsilon_0 \mu_r \mu_0 \frac{\partial^2 \vec{E}}{\partial t^2} = \frac{\nabla \rho}{\epsilon_r \epsilon_0} + \mu_r \mu_0 \frac{\partial \vec{j}_f}{\partial t} \quad (9.20)$$

Dette er en ikke-homogen bølgeligning for elektrisk felt. Kildeleddene er på høyre side av likhetstegnet.

I områder hvor gradienten til ladningstettheten ρ er lik null (altså ingen endring i elektrisk ladningstetthet), samtidig som det ikke er noen tidsvariasjon i elektrisk strømtetthet \vec{j}_f av frie ladninger, reduseres den inhomogene ligningen til en enkel bølgeligning:

$$\nabla^2 \vec{E} - \epsilon_r \epsilon_0 \mu_r \mu_0 \frac{\partial^2 \vec{E}}{\partial t^2} = 0$$

eller i en mer vanlig form:

$$\frac{\partial^2 \vec{E}}{\partial t^2} = \frac{1}{\epsilon_r \epsilon_0 \mu_r \mu_0} \nabla^2 \vec{E} \quad (9.21)$$

Vel, for å være ærlig så er ikke dette en helt vanlig bølgeligning slik vi har sett det tidligere siden vi har Laplaceoperatoren anvendt på elektrisk felt på høyresiden. Med visse forenklinger får vi imidlertid den for oss vanlige bølgeligningen:

$$\frac{\partial^2 \vec{E}}{\partial t^2} = c^2 \frac{\partial^2 \vec{E}}{\partial z^2} \quad (9.22)$$

hvor

$$c = \frac{1}{\sqrt{\epsilon_r \epsilon_0 \mu_r \mu_0}} \quad (9.23)$$

er bølgehastigheten (fasehastigheten) for den elektromagnetiske bølgen. Det er ingen dispersjon i vakuum, men i et dielektrisk materiale kan dispersjon forekomme dersom ϵ_r (og/eller μ_r) er bølgelengdeavhengig.

Det kan bemerkes at for lys gjennom glass, opererer vi med en brytningsindeks n hvor $c = c_0/n$, altså at lyshastigheten i glasset er lik lyshastigheten i vakuum dividert på brytningsindeksen. Glass er diamagnetisk og $\mu_r \approx 1.0$. Da ser vi av uttrykkene ovenfor at

$$n \approx \sqrt{\epsilon_r}$$

hvor den relative permittiviteten også kalles dielektrisitetskonstanten.

Vi skal se nærmere på enkelte detaljer i neste underkapittel, men la oss først undersøke hvordan bølgeligningen ser ut for magnetfelt. Vi starter også da ut med ligning (9.16), men anvender den på magnetisk fluksstetthet \vec{B} . Første mellomresultat blir da:

$$-\nabla^2 \vec{B} + \nabla(\nabla \cdot \vec{B}) = \nabla \times (\nabla \times \vec{B})$$

Vi bruker så Ampère-Maxwells lov for å erstatte rotasjonen til \vec{B} med den tidsderiverte av elektrisk fluksstetthet \vec{D} pluss strømtetthet av fri ladninger. Som ved utledningen for elektrisk felt bytter vi så rekkefølge av en tidsderivasjon og en romlig derivasjon, og får et ledd hvor rotasjonen til \vec{E} inngår. Vi anvender så Faradays lov, og setter også inn at divergensen til \vec{B} er lik null (Gauss lov for magnetfelt) for endelig å ende opp med følgende differensialligning for \vec{B} :

$$\nabla^2 \vec{B} - \epsilon_r \epsilon_0 \mu_r \mu_0 \frac{\partial^2 \vec{B}}{\partial t^2} = -\mu_r \mu_0 \nabla \times \vec{j}_f \quad (9.24)$$

Vi ser at magnetisk fluksstetthet også tilfredsstiller en ikke homogen bølgeligning, og kildeleddet her er rotasjonen til strømtetthet av fri ladninger. I områder av rommet hvor det ikke finnes noe kildeledd, får vi en homogen bølgeligning som under visse forenklinger kan skrives:

$$\frac{\partial^2 \vec{B}}{\partial t^2} = c^2 \frac{\partial^2 \vec{B}}{\partial z^2} \quad (9.25)$$

hvor bølgens hastighet c er nøyaktig den samme som gitt i ligning (16.2) som gjaldt for bølger av elektrisk felt. Det anbefales at du gjennomfører overgangen mellom ligning (9.24) og ligning (9.25) på egen hånd slik at du ser hvilke forenklinger det er snakk om!

9.5 Én løsning av bølgeligningen

Ligningene (16.1) og (9.25) viser at vi har fått en bølgeligning for \vec{E} og en helt tilsvarende for \vec{B} . Vi kunne ledes til å tro at vi kunne få løsninger av typen:

$$E = E_0 \cos(kz - \omega t)$$

$$B = B_0 \cos(kz - \omega t)$$

Så enkelt er det likevel ikke, for vi har med vektorielle størrelser å gjøre. Samspillet mellom \vec{E} og \vec{B} følger ikke av bølgeligningene alene. Vi må gå tilbake til Maxwells ligninger direkte for å finne den sammenhengen. Et resultat av dette er at det elektriske feltet ikke kan peke i samme retning som bølgen beveger seg. Dersom bølgen beveger seg i z -retning, slik vi har indikert hittil, vil \vec{E} ikke kunne ha noen komponent i z -retning!

La oss forsøke oss med følgende løsning av bølgeligningen for \vec{E} :

$$\vec{E} = E_0 \cos(kz - \omega t) \vec{i} \quad (9.26)$$

med andre ord at elektrisk feltstyrke er rettet i $\pm x$ -retning.

Dette er en *plan* bølge fordi elektrisk felt i et vilkårlig tidspunkt vil være identisk over et helt uendelig stort plan karakterisert med en gitt z -verdi (altså i et plan vinkelrett på z -aksen).

Vi vil så bestemme magnetfeltet som svarer til et slikt valg av elektrisk felt. Én mulighet er å bruke en forenklet Ampère-Maxwells lov:

$$\nabla \times \vec{B} = \mu\epsilon \frac{\partial \vec{E}}{\partial t}$$

hvor vi har antatt at kildeleddene i ligningene (9.20) og (9.24) er lik null. Den totale permeabiliteten er gitt som $\mu = \mu_r \mu_0$ og tilsvarende for permittiviteten ϵ . Ampère-Maxwells lov er likevel ikke det gunstigste valget i denne sammenheng, siden rotasjonen til den ukjente \vec{B} inneholder seks ledd vi etter tur må si noe om.

Det er enklere å starte med Faradays lov:

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

Setter vi inn prøveløsningen \vec{E} , ligning (9.26), får vi (på determinant form):

$$\begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ E_x & 0 & 0 \end{vmatrix} = -\frac{\partial \vec{B}}{\partial t}$$

$$\frac{\partial E_x}{\partial z} \vec{j} - \frac{\partial E_x}{\partial y} \vec{k} = -\frac{\partial \vec{B}}{\partial t}$$

Det andre leddet er lik null siden E ikke avhenger av y . Den partiell deriverte av E beregnes, og vi får:

$$\frac{\partial \vec{B}}{\partial t} = k \cdot E_0 \sin(kz - \omega t) \vec{j}$$

Integrasjon gir:

$$\vec{B} = \frac{k}{\omega} E_0 \cos(kz - \omega t) \vec{j} + \vec{B}_s$$

Integrasjonskonstanten \vec{B}_s er et statisk magnetfelt. Det kan eksistere ved siden av et tidsvariabelt felt, men vi er mest interesserte i bølgdedelen og setter det statiske feltet lik null. Videre vet vi at hastigheten til bølgen er gitt ved:

$$c = \frac{\omega}{k}$$

Dette gir oss et endelig uttrykk for magnetfeltbølgen som svarer til den valgte elektriske feltbølgen gitt i ligning (9.26):

$$\vec{B} = B_0 \cos(kz - \omega t) \vec{j} \quad (9.27)$$

hvor

$$E_0 = cB_0 \quad (9.28)$$

Vi har da vist at prøveløsningen gitt i ligning (9.26) faktisk er en mulig løsning av bølgeligningen som også lar seg kombinere med en magnetfeltbølge gitt i ligning (9.27). Disse to ligningene er simultane løsninger av Maxwells ligninger, og kan oppsummeres slik:

En plan elektromagnetisk bølge i et område langt fra kilden og langt fra områder med frie ladninger og områder hvor ladningstettheter, permittiviteter og permeabiliteter endrer seg, kan ha formen:

$$\vec{E} = E_0 \cos(kz - \omega t) \vec{i}$$

$$\vec{B} = B_0 \cos(kz - \omega t) \vec{j} \quad (9.29)$$

hvor

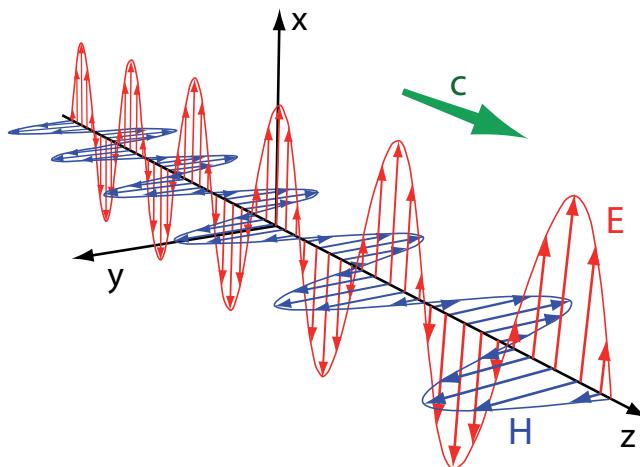
$$E_0 = cB_0$$

Figur 9.4 viser et øyeblikksbilde av en elektromagnetisk bølge med egenskaper som gitt i ligningene (9.29). En slik statisk figur gir ikke et godt bilde av bølgen. Det kan derfor være lurt å betrakte en animasjon for å få en forståelse av tidsutviklingen. Det er flere animasjoner av en enkel elektromagnetisk bølge på weben. Se f.eks. <http://www.phy.ntnu.edu.tw/ntnujava/index.php?topic=35> (tilgjengelig 26. februar 2013).

Bølgen vi har beskrevet er plan fordi elektrisk felt ved et gitt øyeblink er identisk overalt i et uendelig plan normalt på bølgeretningen z . En annen måte å si dette på er at "bølgefronten" er plan. En bølgefront kan karakteriseres som en flate i rommet hvor bølgen har identisk fase (dvs. argumentet til sinus- eller cosinusfunksjonen er identisk i et gitt øyeblink).

Mange forveksler "plan" med det faktum at elektrisk felt peker i samme retning hele tiden ($\pm x$ -retning i vårt tilfelle). Med andre ord kan vi forledes til å tro at elektrisk felt ligger i et plan. Det er imidlertid ikke tilfelle. Figurer som 9.4 gir bare feltverdier langs z -aksen og ikke i noen andre punkter! Velger vi et punkt vekk fra z -aksen vil elektrisk felt fortsatt være rettet i x -retning, men denne vektoren vil da *ikke* ligge i samme planet som vektoren som går gjennom z -aksen. Vi kommer tilbake til dette poenget om litt.

Det at elektrisk felt overalt er rettet i $\pm x$ -retning er likevel et karakteristisk trekk ved den løsningen vi har kommet fram til. Vi sier at bølgen er *lineært polarisert* i x -retning. Vi



Figur 9.4: Et øyeblikksbilde av den enkleste formen for elektromagnetisk bølge, næmlig en plan bølge. En slik bølge kan oppnås langt fra kildene til bølgen og langt fra materialer som kan perturbere bølgen. Figurer av denne typen gir erfaringmessig en rekke misoppfatninger. Disse blir diskutert i siste del av dette kapitlet.

kommer tilbake til polarisasjon i et senere kapittel, men nevner allerede her at en annen løsning av Maxwells ligninger er en såkalt sirkulært polarisert bølge. For en slik løsning vil de elektrisk feltvektorene i et øyeblikksbilde tilsvarende figur 9.4 se ut som trinnene i en vindeltrapp, og selve pilspissene vil danne en "skrulinje" med akse i z-aksen. Også magnetfeltet vil danne en skrulinje. Også i dette tilfellet vil elektrisk felt og magnetfelt stå vinkelrett på hverandre og vinkelrett på den retningen bølgen beveger seg.

Du kan finne en fin animasjon av elektromagnetiske bølger med ulike polariseringer (kombinert med virkningen av et lineært polarisasjonsfilter) på samme webside som nevnt ovenfor (<http://www.phy.ntnu.edu.tw/ntnujava/index.php?topic=35>). Du må dreie på figuren for å få fram hvordan bølgen er orientert i rommet.

Forøvrig kommer vi senere i kapitlet tilbake til en viktig drøfting av gyldighetsområdet til de enkle elektromagnetiske bølgene vi hittil har beskrevet.

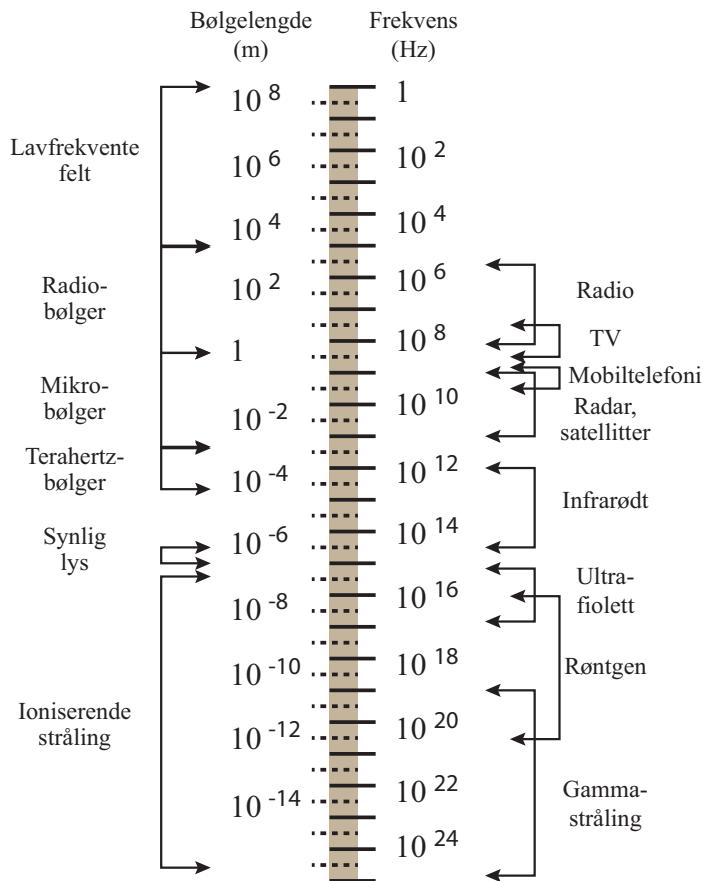
Det var den tyske fysikeren Heinrich Hertz (1857-1894) som først demonstrerte hvordan vi kunne sende og motta elektromagnetiske bøller. Det skjedde i 1887 da Hertz var 30 år gammel.

9.6 Det elektromagnetiske spekteret

I utledning av bølgeligningen for elektromagnetiske bøller, hadde vi (i første omgang) ingen begrensinger i hvilke frekvenser og bølgelengder vi opererte med. I prinsippet kunne mer eller mindre "alle" frekvenser komme på tale med de tilsvarende bølgelengdene.

Det viser seg også i praksis at vi kan generere elektromagnetiske bøller ved et vidt spekter av frekvenser (og bølgelengder). Figur 9.5 viser en omtrentlig oversikt over hvilke frekvensområder / bølgelengdeområder vi opererer i, hva vi kaller bølgene ved ulike frekvenser, og hva slike bøller brukes til. Vi sier at figurer som 9.5 presenterer "det elektromagnetiske spekteret".

Figurer av denne type må tas med en stor klype salt. Mange tror at det er snakk



Figur 9.5: Elektromagnetiske bølger kan eksistere i et imponerende variasjonsområde av frekvenser (og tilsvarende bølgelengder). Oversikter som dette kan imidlertid gi inntrykk av en større grad av likhet mellom ulike fenomener enn det er i praksis. Vi kommer tilbake til dette blant annet når vi omtaler forskjellen på nærfelt og fjernfelt senere i kapitlet.

om pene, pyntelige plane bølger ved hver av de angitte frekvensene, men det er det ikke. Bølgenes utbredelse i tid og rom, energitransport (eller mangel på sådan) og flere andre faktorer varierer mye fra en frekvens til en annen. Vi kommer tilbake til dette litt senere i dette kapitlet.

9.7 Energitransport

Da vi diskuterte lyd så vi at en lydbølge kunne frakte energi bort fra kilden, selv om molekylene som bidro bare svingte i størrelsesorden en mikrometer fram og tilbake omkring samme punkt (når vi ser bort fra Brownsk/diffusjons-bevegelsen til molekylene).

På en lignende måte kan en elektromagnetisk bølge frakte med seg energi, noe vi alle kjenner til når vi slikker sol på påskefjellet eller på en badestrand om sommeren.

Et elektrisk felt har en energitetthet gitt ved:

$$u_E(z, t) = \frac{1}{2} E(z, t) D(z, t)$$

På samme måte er energitettheten til et magnetfelt gitt ved:

$$u_H(z, t) = \frac{1}{2} H(z, t) B(z, t)$$

Når en plan elektromagnetisk bølge (slik vi har beskrevet den foran) passerer oss, vil den momentane energitettheten bli:

$$\begin{aligned} u_{tot}(z, t) &= \frac{1}{2}E(z, t)D(z, t) + \frac{1}{2}H(z, t)B(z, t) \\ &= \frac{1}{2}E_0 \cos(\omega t) \cdot \epsilon E_0 \cos(\omega t) + \frac{1}{2}B_0 \cos(\omega t) \cdot \frac{B_0}{\mu} \cos(\omega t) \end{aligned}$$

Vi har her droppet å skrive ut innholdet inni parantesen for cosinusfunksjonen (for å spare plass).

Men vi vet at $E_0 = cB_0$. Dessuten ønsker vi å se på *tidsmidlet* energitetthet, og vi vet at middelverdien av $\cos^2(\omega t)$ er lik en halv. Følgelig finner vi for tidsmidlet energitetthet:

$$\bar{u}_{tot} = \frac{1}{4}\epsilon E_0^2 + \frac{1}{4\mu} \left(\frac{E_0}{c} \right)^2$$

Energitetthet er energi per volum. Hvor mye energi vil da passere en flate A vinkelrett på bølgens bevegelsesretning i løpet av en tid Δt ? En slik størrelse definerer vi som bølgens (tidsmidlete) intensitet:

$$I = \text{intensitet} = \frac{\text{Energi passert}}{\text{Areal Tid}} = u_{tot} \cdot c$$

Uttrykket har bare relevans når vi betrakter en lang tid i forhold til den tiden en bølgelengde trenger for å passere flaten vår. Innsatt for energitettheten vi fant i stad, får vi:

$$I = \frac{1}{4} \left(c\epsilon E_0^2 + c \frac{1}{c^2 \mu} E_0^2 \right)$$

Men vi vet at

$$c = \frac{1}{\sqrt{\epsilon \mu}}$$

Følgelig blir

$$\frac{1}{c^2 \mu} = \epsilon$$

og vi ser at energibidraget fra det elektriske feltet er nøyaktig like stort som energibidraget fra magnetfeltet!

Følgelig er intensiteten i en elektromagnetisk bølge gitt ved:

$$I = \frac{1}{2}c\epsilon E_0^2 = \frac{1}{2}cE_0 D_0 \quad (9.30)$$

Ved å benytte oss av det kjente forholdstallet mellom elektrisk og magnetisk felt, kan resultatet også skrives slik:

$$I = \frac{1}{2}c \frac{1}{\mu} B_0^2 = \frac{1}{2}cH_0 B_0 \quad (9.31)$$

Dersom vi velger å angi størrelsen på elektrisk felt og magnetfelt ved å bruke effektivverdier i stedet for amplitudeverdier, kan ligningene (9.30) og (9.31) skrives på formen:

$$I = c\epsilon E_{eff}^2 = cE_{eff} D_{eff} \quad (9.32)$$

og

$$I = \frac{c}{\mu} B_{eff}^2 = cH_{eff} B_{eff} \quad (9.33)$$

Disse to uttrykkene kan være nyttige når vi skal finne elektrisk feltstyrke eller magnetisk fluksstetthet (eller magnetisk feltstyrke) for en gitt intensitet.

En liten digresjon: Betegnelsen "effektivverdi" har rot i vekselstrøm i en ledning. Vi kan da angi amplitudeverdi på en harmonisk variasjon i strøm og spenning, men vi kan også angi tilsvarende verdi av likestrøm og likespenning som gir samme overførte effekt. Det er disse likestrøm/spenningsverdiene som kalles effektivverdier. I vårt tilfelle med elektromagnetiske bølger er det egentlig kunstig å trekke inn likestrømmer og denslags, men likevel anvender vi effektivverdier på tilsvarende måte som for vekselstrømmer og vekselspenninger i en ledning.

Vi kan utlede også et annet uttrykk som forbinder elektrisk og magnetisk felt for en elektromagnetisk bølge i fjernfeltet. Tar vi utgangspunkt i ligningene (9.32) og (9.33), og benytter relasjonen $B = \mu H$, får vi:

$$c\epsilon E_{eff}^2 = \frac{c}{\mu} B_{eff}^2 = c\mu H_{eff}^2$$

Herav får vi:

$$\frac{E_{eff}}{H_{eff}} = \sqrt{\mu/\epsilon}$$

I vakuum får vi da:

$$\frac{E_{eff}}{H_{eff}} = \sqrt{\mu_0/\epsilon_0} \equiv Z_0 = 376.7\Omega \quad (9.34)$$

hvor Z_0 kalles (den iboende, indre) impedansen til det tomme rom.

Uttrykkene har et større gyldighetsområde enn det som ligger bak utledningen vår. Vi må imidlertid være varsom med å bruke uttrykkene for elektromagnetiske bølger i områder nær kilder og nær materialer som kan forstyrre bølgene. Vi omtaler såkalte nærfelt og fjernfelt litt senere i dette kapitlet.

9.7.1 Poynting vektor

Det er en mer elegant måte å angi energifluksstetthet på (svarende til intensitet) enn uttrykkene vi ga i forrige avsnitt. Det elegante er at plane elektromagnetiske bølger er transversale slik at elektrisk og magnetisk vektor er rettet vinkelrett på hverandre og vinkelrett på bølgens bevegelsesretning.

Vi så at dersom elektrisk felt var rettet i x -retning og magnetfelt i y -retning, beveget bølgen seg i z -retning. Vi vet at for kryssproduktet gjelder $\vec{i} \times \vec{j} = \vec{k}$, slik at vi muligens kan utnytte denne relasjonen på en smart måte.

Vi forsøker å beregne:

$$\begin{aligned} \vec{E} \times \vec{B} &= E_0 \cos(\phi) \vec{i} \times \frac{E_0}{c} \cos(\theta) \vec{j} \\ &= \frac{cE_0^2}{c^2} \cos^2(\theta) \vec{k} \\ &= \mu(c\epsilon E_0^2) \cos^2(\theta) \vec{k} \end{aligned}$$

Tidsmidlet er:

$$\overline{\vec{E} \times \vec{B}} = \mu \left(\frac{1}{2} c \epsilon E_0^2 \right) \vec{k} = \mu I \vec{k}$$

Når vi så vet at $B = \mu H$, følger det:

$$\vec{I} = \overline{\vec{E} \times \vec{H}} \quad (9.35)$$

Her har vi innført en intensitetsvektor som peker i samme retning som energistrømmen. Vi opererer også med en momentan energitetthetstrøm (en momentan intensitet), og kaller denne Poynting vektor. Denne betegnes gjerne med symbol S eller P . Vi velger første variant og skriver:

$$\vec{S} = \vec{E} \times \vec{H} \quad (9.36)$$

Den engelske fysikeren John Henry Poynting (1852-1914) kom fram til dette uttrykket i 1884, tyve år etter at Maxwell skrev sitt mest berømte verk.

Igjen minner vi om at Poynting vektor bare kan anvendes problemfritt i tilfeller der vi har en enkel plan elektromagnetisk bølge langt fra kilden og langt vekk fra forstyrrende elementer. Sagt på en annen måte: Poynting vektor angir bare en momentan intensitet (momentan energitetthetsstrøm) når det er en perfekt kobling mellom elektrisk og magnetisk felt, uten at feltet fra ladninger i nærheten spiller noen rolle (ren elektrodynamikk).

9.8 Strålingstrykk

Det elektriske og magnetiske feltet vil resultere i en kraft på partikler/gjenstander som elektromagnetiske bølger treffer. Det går an å argumentere for at det elektrisk feltet i bølgen medfører "tvungne svingninger" av ladninger, og at ladninger i bevegelse i sin tur blir påvirket av en kraft $\vec{F} = q\vec{v} \times \vec{B}$. Denne kraften virker i samme retning som den elektromagnetiske bølgen beveger seg.

Det kan vises at en elektromagnetisk bølge medfører et strålingstrykk gitt ved:

$$p_{straling} = S_{midlere}/c = I/c$$

dersom bølgen blir helt absorbert av legemet som blir truffet. Dersom legemet reflekterer bølgene fullstendig, blir strålingstrykket dobbelt så stort, dvs

$$p_{straling} = 2S_{midlere}/c = 2I/c$$

I begge disse uttrykkene er $S_{midlere}$ den tidsmidlerte Poynting vektor.

Det er strålingstrykket som fører med seg at støv i en komethale alltid vender vekk fra Sola. Gravitasjonen som trekker støvet mot Sola er proporsjonal med massen, som igjen er proporsjonal med radien i tredje potens. Kraften som skyldes strålingstrykket er proporsjonal med *flaten* (tverrsnittet) som kan absorbere eller reflektere bølgen, og tverrsnittet går som radien i annen potens. Dette fører med seg at gravitasjon dominerer over strålingstrykk for store partikler, mens det blir motsatt for små partikler.

Det er mulig å betrakte strålingstrykk som en strømningsrate av elektromagnetisk bevegelsesmengde. I et slikt bilde kan det sies at bevegelsesmengde per tid og flate som forflytter seg med bølgen er lik

$$S_{midlere}/c$$

som er samme uttrykk som for strålingstrykk når legemet absorberer bølgen fullstendig.

[♣ ⇒ Beskrivelsen ovenfor gjelder i det tilfellet at lys enten blir absorbert eller totalt reflektert på overflaten til et materiale. Situasjonen er annerledes for lys som går gjennom et gjennomsiktig medium. Det finnes to ulike beskrivelser av hvordan bevegelsesmengden til lys endres når lys går inn i et gjennomsiktig medium. I én beskrivelse hevdes det at bevegelsesmengden øker, i en annen beskrivelse hevdes det motsatte. Dette er et optisk dilemma som henger delvis sammen med hvorvidt lys betraktes som bølger eller som partikler. I så måte er det en klar parallel mellom det dilemmaet vi har i dag og dilemmaet som eksisterte fra 1600-tallet til ca 1850 nevnt i forrige kapittel, der vi lurte på om gruppehastigheten til lys i glass var større eller mindre enn fasehastigheten.]

Har du lyst å lære litt mere om dagens dilemma, kan du starte med å lese en artikkel i Physics World (<http://physicsworld.com/cws/article/news/41873>). ← ♠]

9.9 Feiloppfatninger

9.9.1 Nærfelt og fjernfelt

Gjentatte ganger har vi tidligere i dette kapitlet minnet om at de elektromagnetiske bølgene vi har utledet i ligningene (9.29) og illustrert i figur 9.4 er de *enkleste* bølgeløsningene som finnes av Maxwell's ligninger. Disse relasjonene gjelder *normalt ikke* for tidsavhengige elektromagnetiske fenomen generelt! For å forstå dette, må vi se nøyere på detaljer i utledningen vår.

For det første endte vi opp med ikke-homogene differensielligninger i ligning (9.20) og (9.24) etter å ha kombinert Maxwell's ligninger. Først da vi så bort fra kildeleddene oppnådde vi en enkel homogen bølgeligning.

Videre fant vi en tilfeldig løsning av bølgeligningen, den enkleste løsningen som vi kan tenke oss. Ved innsetting så vi at den passet med bølgeligningen, men det er ikke derved sagt at den passer inn i den fysiske sammenhengen vi betrakter!

I tidligere kapitler har vi drøftet prosedyrer når vi løser en svingeligning eller en bølgeligning. Så lenge vi jobber rent analytisk med en svingeligning, kan vi finne en generell løsning, og siden sette inn initialbetingelser for å bestemme en konkret løsning. Bruker vi numeriske metoder for å finne en løsning av en svingeligning, *starte* vi med med initialbetingelsene, og beregner løsningen trinn for trinn etterpå.

For bølgeligningen er situasjonen betydelig mer kompleks. Det er umulig å gå fra bølgeligningen til en konkret bølge, f.eks. en lydbølge i luft, uten at initialbetingelser og randbetingelser er gitt. Det betyr at vi må kjenne f.eks. lokalt lufttrykk og den tidsdververte av lokalt lufttrykk i alle punkter i rommet hvor bølgen er definert. Vi må også kjenne alle randbetingelser og lokale materialegenskaper i ethvert punkt, dersom vi skal kunne følge bølgens videre utvikling.

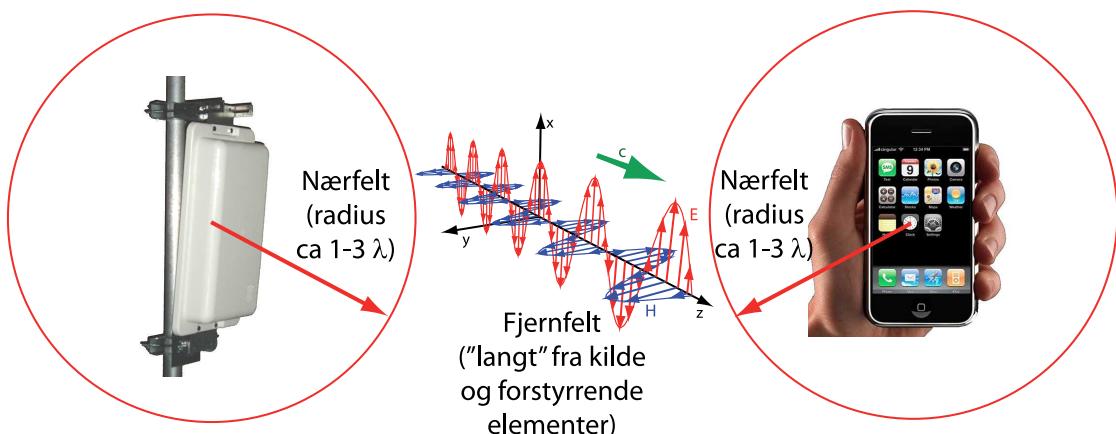
En tommelfingerregel i denne sammenheng er at selv om vi er i vakuum, vil bølgen bli påvirket til dels enormt av "nærliggende" strukturer hvor det er frie elektriske ladninger eller elektriske strømmer. Med "nærliggende" mener vi avstander som er minst flere ganger beregnet bølgelengde. I andre tilfeller betyr "nærliggende" i praksis at vi ikke er flere ganger vekk fra den forstyrrende gjenstandens fysiske utstrekning i rommet (i den retningen vi betrakter). I områder som blir kraftig påvirket av randbetingelser, sier vi at vi finner "nærfelt", i motsetning til "fjernfelt", som vi finner i områder hvor randbetingelser nesten

ikke har noen innflytelse.

Det vil si at selv om vi er i vakuum, er det generelt sett *ikke* slik at plane elektromagnetiske bølger er løsninger av bølgeligningen for vakuum. Randbetingelsene vil påvirke løsningen til dels betydelig. For eksempel, står ikke det elektriske og det magnetiske feltet fra en 50 Hz kraftledning vinkelrett på hverandre før vi er flere jordradier unna!

Dette er en bevissthet som har kommet klarere fram etter hvert som fysikere har brukt numeriske løsningsmetoder. Siden det er en relativt ny erfaring, har den ikke fått så stor plass i lærebøker hittil som den burde ha hatt. Jeg tror det er en viktig grunn til at så mange fysikere lever med en feiloppfatning på dette området.

Faktum er at i de fleste konkrete situasjoner vi kommer borti i praksis, er ikke plane elektromagnetiske bølger noe løsning av problemet. Dette gjelder spesielt i nærfeltområdene.



Figur 9.6: Alle gjenstander med bevegelige ladninger vil påvirke et elektromagnetisk felt ut i en avstand i størrelsesorden en beregnet bølgelengde $\lambda = c/f$. I nærsone er løsningen av Maxwells ligninger ofte svært forskjellig fra løsningen i fjernsonen (langt fra kilden til feltene og langt fra forstyrrende elementer). Det er randbetingelsene som fører til disse forskjellene.

Et hvilket som helst fysisk problem/fenomen har en innebygget tids- og lengdeskala. Slik er det også med elektromagnetisme og elektromagnetiske bølger. For elektromagnetiske bølger er tidsskalaen ofte knyttet opp til periodetiden for bølgene. Lengdeskalaen er ofte knyttet opp til bølgelengden, men den er minst like ofte knyttet opp til utstrekningen på kilden til bølgene og/eller utstrekningen til materialer som påvirker elektriske og/eller magnetiske felt. Det er imidlertid vanskelig å gi en helt generell regel fordi mange faktorer spiller inn. Vi må løse Maxwells ligninger i det aktuelle området med korrekte initialbetingelser og randbetingelser for å virkelig kunne vurdere hvor kraftig det forstyrrende materialet innvirker på løsningen.

Overalt i rommet, også nær en kilde til elektriske og magnetiske felt, gjelder Maxwells ligninger (så lenge vi ikke går helt ned til atomære størrelser). Imidlertid er plane bølger generelt sett *ikke* noe løsning av Maxwells ligninger nær kilden eller forstyrrende elementer. Vi skiller mellom "nærsonen" og "fjernsonen" (se figur 9.6). Det er ingen skarp grense mellom disse. Vi sier at nærsone strekker seg typisk noen få bølgelengder ut fra kilden (når vi med "bølgelengde" mener lyshastigheten dividert på frekvensen til feltvariasjonen, dvs "bølgelengde" = c/f). Fjernsonen overtar ved større avstander.

De plane elektromagnetiske bølgene *kan* likevel være en god beskrivelse for feltene i *fjernfeltområdet* i mange tilfeller. Vi må bare huske på at ingen bølger har uendelig utstrekning, slik at vi i prinsippet *aldri* vil ha en perfekt plan bølge.

Det ligger mye fysikk i å forstå hvorfor vi får en forskjell i nærfelt og fjernfelt. La oss forsøke å skisser denne:

Maxwells ligninger sier at et tidsvariabelt elektrisk felt er en kilde til magnetfelt, og at et tidsvariabelt magnetfelt er en kilde til elektrisk felt. Vi kan godt starte med et område i rommet med temmelig rent tidsvariabelt elektrisk felt. Da vil det elektriske feltet generere et magnetfelt, *også i naboområdene*. Det genererte magnetfeltet vil i sin tur generere et elektrisk felt som modifiserer det opprinnelige, og feltene brer seg lenger og lenger utover for hver runde. *Når vi kommer tilstrekkelig langt unna kilden og forstyrrende randbetingelser*, vil den gjensidige koblingen føre til at det elektriske feltet og magnetfeltet har fått et fast forhold til hverandre lokalt. Geometrien til kilden og randbetingelsene langt borte har da minimalt å si.

Størrelsen på magnetfelt indusert på grunn av et tidsvariabelt elektrisk felt er proporsjonalt med den tidsderiverte til det elektriske feltet. Det betyr at et elektrisk felt på 10 V/m vil generere ti ganger så kraftig magnetfelt dersom frekvensen er 10 MHz enn om den var 1 MHz. Det betyr at vi må forvente at bølgen bruker kortere tid ved 10 MHz enn ved 1 MHz fra vi har f.eks. et nesten rent elektrisk felt til vi har fått en situasjon der forholdstallet mellom elektrisk og magnetisk felt tilnærmet har nådd en grenseverdi. Men bølgelengden ved 10 MHz er bare 1/10 av bølgelengden ved 1 MHz. Følgelig vil avstanden fra kilden til det området hvor feltene omtrent har nådd sin grenseverdi være omtrent like lang *målt i antall bølgelengder* for den frekvensen vi opererer med.

Etter denne argumentasjonen håper jeg at du i det minste har en vag forståelse av at det er meningsfylt å prate om en nærfeltszone som er noen få *bølgelengder* vekk fra kilden.

Det kan være nyttig å tenke på hvor langt ut nærfeltområdet strekker seg fra ulike kilder. For en lyskilde er bølgelengden om lag 500 nm. Nærfeltområdet strekker seg noen få ganger denne avstanden vekk fra kilden, dvs. i størrelsesorden noen mikrometer (tusendedels millimetre) vekk fra kilden.

For en mobiltelefon som fungerer ved 900 MHz, er beregnet bølgelengde om lag 33 cm. Noen få ganger denne avstanden er vi over i fjernfeltszone.

For en kraftledning med frekvensen 50 Hz er beregnet bølgelengde om lag 6000 km. Først når vi er flere ganger denne avstanden vekk fra kraftledningen, er vi i fjernfeltszone, men da er vi ikke lenger på Jorden!

For *fjernfeltområdet* gjelder de relasjonene vi har vist for enkle, plane elektromagnetiske bølger, dvs.

1. Elektrisk og magnetisk felt står vinkelrett på hverandre.
2. Det er et fast forholdstall mellom elektrisk og magnetisk felt.
3. Poynting vektor gir et mål for transport av elektromagnetisk energi.
4. Energien som passerer et tverrsnitt har forlatt kilden en gang for alle og kommer (normalt) ikke tilbake igjen.
5. Det kan derfor være naturlig å bruke ordet “stråling” om energitransporten.

For *nærfeltsonen* gjelder derimot:

1. Elektrisk og magnetisk felt står normalt *ikke* vinkelrett på hverandre.
2. Det er *ikke* et fast forholdstall mellom elektrisk og magnetisk felt.
3. Poynting vektor gir *ikke* et mål for transport av elektromagnetisk energi.
4. Energi kan bygge seg opp i nærområdet til kilden i enkelte tidsperioder, men trekkes tilbake igjen i andre tidsperioder. Bare en liten del av energien som går fram og tilbake til nærområdet vil forlate kilden som bølger (og denne energitransporten blir stort sett ikke synlig før vi kommer i fjernfeltsonen).
5. Det er derfor ikke naturlig å bruke ordet “stråling”. Vi beskriver situasjonen mer som “felt”.

En liten ekstra kommentar bør også tas med. Vi er vant til i mange andre sammenhenger at vi kan beskrive en løsning som en lineær kombinasjon av “basisvektorer” som spenner ut løsningsrommet. For eksempel antar vi at svingningene på en gitarstreng kan beskrives som en lineær kombinasjon av stående bølger der strengen er en halv bølgelengde, en hel bølgelengde, $3/2$ bølgelengder osv.

For elektromagnetiske bølger (og også for akustiske bølger) er dette generelt sett langt vanskeligere. Det skyldes at utbredelsen er et tredimensjonalt problem med randbetingelser som kan være ganske komplekse. For forenklede geometrier lar det seg gjøre, men altså (i praksis) ikke generelt.

9.9.2 Fotonbegrepet

Jeg ønsker her å knytte noen kommentarer til begrepet “foton”. Et foton oppfattes ofte som en “udelelig bølgepakke eller energipakke” som har en begrenset utstrekning i tid og rom. Ordet foton ble opprinnelig tatt i bruk for synlig lys hvor bølgelengden er i størrelsesorden 500 nm (det greske ordet “phos” betyr “lys”). Det vil si at selv en bølgepakke som inneholder ganske mange bølgelengder, vil være liten i forhold til makroskopiske størrelser. I *det* tilfellet er det kanskje ikke så rart at noen kan oppfatte dette som en “partikkel”. Den tenkte udelelige energipakken tilordnes energien $E = h\nu$ hvor h er Plancks konstant og ν er frekvensen.

Hva så dersom vi betrakter “fotoner” ved mobiltelefoni? I så fall vil en bølgepakke som

består av en del bølgelengder fort få en romlig utstrekning på flere meter. Er det naturlig å anse en slik pakke som “udelelig” og at energien utveksles momentant fra pakken til og fra antennen og rommet rundt?

For kraftledninger og 50 Hz felter ville en bølgepakke på flere ganger bølgelengden fort få like stor utstrekning i rommet som omkretsen til Jorda! Vi innser da at vi har alvorlige problemer med å forestille oss et foton som har en utstrekning på flere ganger bølgelengden.

Vi kan selvfølgelig tenke oss et foton som en nærmest punktformig partikkel også når bølgelengden blir lang. Vi får da imidlertid problemer med å forklare små linjebredder samtidig som vi har lange koherenslengder (koherens omtaler vi i kapittel 13). Det er derfor meget problematisk å bruke fotonbegrepet for lange bølgelengder uansett hvordan vi snur og vender oss.

Mange anvender formelen for fotonenergi ukritisk uansett frekvensområde. Det er gjerne de samme folkene som bruker bildet på elektromagnetiske bølger i enhver sammenheng uten å skille mellom nærfelt og fjernfelt. Som nevnt tror jeg personlig at denne kortslutningen skyldes at vi i fysikkundervisningen tradisjonelt bare har anvendt analytiske metoder der vi så lett kan overse poenget at randbetingelser faktisk er avgjørende for hvordan en løsning av en bølgeligning ser ut! Når vi bruker numeriske metoder, blir vi *nødt* til å ta hensyn til randbetingelser, og da gjør vi ikke like mange dumheter som tidligere!

La oss gå tilbake til den uvettige bruken av fotonenergi. Ved 50 Hz er fotonenergien i størrelsesordenen 10^{-13} eV. Når vi trenger flere elektronvolt for å få en ionisering av et atom eller molekyl, mener mange at det teoretisk sett er totalt utenkelig å få ioniseringer ut fra elektromagnetiske felt med frekvensen 50 Hz. Dersom vi går under en kraftig kraftledning, spesielt dersom det er snøkrystaller i luften, hører vi en tydelig knitring fra ledningene. Er det en mørk natt, kan vi faktisk se at det kommer et svakt lys fra overflaten på kraftledningen. Dette fenomenet, som kalles ”corona”, skyldes at det elektriske feltet lokalt er større en ca $3 \cdot 10^6$ V/m. Da får vi dielektrisk gjennombrudd i luften helt lokalt, og vi eksiterer og ioniserer luftmolekyler. Dette faktum sier i alle fall meg at fotoner og fotonenergi bare er nyttige begreper i visse fysiske sammenhenger, spesielt når bølgelengden er svært liten. Fotoner bør ikke brukes for lange bølgelengder fordi de fører til en langt mer komplisert og ”kunstig” beskrivelse enn den som bygger på Maxwells ligninger!!!

♣ ⇒ Fysikernes oppfatning av lys har endret seg opp gjennom århundrene. Huygen og Newton debatterte emnet på 1600-tallet. Huygen mente at lys best kunne beskrives som bølger, mens Newton anså lyset som røde, grønne og blå partikler. Newtons fargeblandingsteori og hans anseelse generelt førte til at partikkelloppfatningen ble den dominerende i litt over 100 år.

Gjennom Thomas Youngs berømte dobbeltspalteksperiment publisert i 1803, og Fresnels arbeider noe senere, ble fysikerne etter hvert overbevist om at Newton hadde tatt feil på dette området, og at lyset best kunne beskrives som bølger.

Lysets partikkelnatur kom tilbake i fysikken gjennom Einsteins forklaring av fotoelektrisk effekt, og ved forklaring av Compton-effekten. Det er imidlertid siden vist at begge disse fenomenene kan beskrives også ved hjelp av bølger. I dag er det derfor stort sett bare ett eksperiment vi hittil ikke har klart å forklare ved hjelp av bølger. Det gjelder et eksperiment publisert i 1986 av P. Grangier, G. Roger og A. Aspect i Europhys. Letters (Vol. 1, s 173). Det er imidlertid også kommet innvendinger mot dette eksperimentet, så kanskje vi om få år kan forklare alle aktuelle eksperiment ved hjelp av bølger. En god del eksperimenter kan *bare* forklares ved hjelp av bølger, mens noen kan beskrives såvel med bølger som med partikler.

W.E. Lamb drøfter fotonbegrepet på en systematisk måte i artikkelen ”Anti-photon” (Appl. Phys. B 60 (1995) 77-84). Lamb mener at fotoner slik de er beskrevet overfor ikke eksisterer, og mener at det er en rekke feil og historiske ulykker som førte til at fotonbegrepet fikk plass i fysikken for om lag hundre

år siden. (Lamb er en av nobelpristakerne i fysikk.)

I lys av eksperimentelle fakta vi har tilgjengelig i dag, og forsøk på å øke presisjonsnivået for hvordan vi bruker de ulike begrepene, er det en økende mengde fysikere som tror at bølge-partikkeldualismen, snart vil falte. Vi begynner å se detaljer i hvordan den kommende modellen av lys vil se ut, rent konkret, men det er fortsatt en betydelig vei å gå.

Det er derfor kanskje ikke så rart at det fortsatt er en majoritet av dagens fysikere som anser lysstråle som en skur med partikler, der hver partikel er en "elementærpartikel: fotonet". Eller kanskje det er mer riktig å si at majoriteten av dagens fysikere skyver problemet foran seg. De opererer med en svært upresis forestilling av hva de mener et foton er. Så lenge de ikke utfordrer skjebnen ved å bli mer presise i sine forestillinger, kan en vag og uklar forestilling overlevere år etter år. Men fysikk er en vitenskap hvor fremgang ofte henger sammen med at vi blir mer presise. Det er først da vi kan teste ulike oppfatninger og velge det som passer godt og forkaste det som duger dårlig. Jeg utfordrer derfor deg som leser, å forsøke å bli mer presis på dette området, for å faktisk kunne se hva som duger og hva som ikke duger!

Det er interessant å se noen trender i dagens forskning relevant for denne problemstillingen:

I en artikkel "Fundamental limit of nanophotonic light trapping in solar cells" av Z. Yu, A. Raman og A. Fan ved Standford University (Proc. Natl. Acad. Sci. Vol 107, no 41 (12. oktober 2010) s 17491-17496) er tema hvordan vi skal kunne fange mest mulig lys som når en solcelle, for å kunne øke den hittil lave effektiviteten vi hittil har hatt på solceller. I deres system bruker de strukturer som er mindre enn bølgelengden. I slike tilfeller duger ikke teorier basert på partikler, og forskerne ved Standford University gikk over til en ren bølgebeskrivelse med bra resultat. Også ved Institutt for energiteknikk på Kjeller utenfor Oslo, brukes lignende strukturer (såkalte "fotoniske krystaller") for å øke effektiviteten til solceller. Fotoniske krystaller er et meget spennende nytt forskningsfelt som antakelig vil prege den teknologiske utviklingen i årene som kommer.

Dette er ett eksempel av flere der forskere vender tilbake fra partikelbeskrivelse til en bølgebeskrivelse av lys for å få resultater som matcher de eksperimentelle.

Det bør forøvrig nevnes at det innen kvantefysikk *ikke* finnes noen "posisjonsoperator" som kan brukes for å fortelle hvor et foton er.

Videre, dersom vi tenker oss lys som fotoner, er det noen som sier: "Dersom vi rir på et foton, vil tiden stå stille!". Denne pussigheten synes jeg blir enklere å håndtere tankemessig når lys behandles som bølger, der bølgen bare representerer en endring av det elektromagnetiske feltet i rommet i seg selv. Forøvrig har jeg en vag tanke om at det nettopp er gjennom elektromagnetismen at koblingen mellom tid og rom finner sted. Det er ikke uten grunn at lyshastigheten har en så sentral plass i relativitetsteorien. Den som kommer fram til en dypere forståelse av dette samspillet, får opplagt en Nobelpris! Dersom du er ung og mestrer både matematikk, modellering og har sans for å se analogier i ulike deler av fysikken (og ikke er så inderlig redd for å tenke utradisjonelt), er dette kanskje et tema du kan sysle litt med i ledige stunder...? Husk at Einstein ikke var gamle karen da han gjorde sine store bidrag til fysikken! ← ♠]

9.9.3 Plan og plan fru Blom

I figur 9.4 tegnet vi en elektromagnetisk bølge på den måten de fleste lærebøker illustrerer en plan bølge. Slike figurer gir imidlertid erfaringsmessig en rekke feiloppfatninger. Figuren blander nemlig tre ulike diagrammer i en og samme figur, og da blir figuren mye mer abstrakt enn vi først innser.

De tre diagrammene er:

1. En linje i det vanlige Euklidske rommet vi alle befinner oss i. I figur 9.4 er dette rett og slett z-aksen. Det er i punkter langs denne linjen vi ønsker å angi elektrisk og magnetisk felt.
2. Siden elektrisk felt i en planpolarisert bølge har samme (eller motsatt samme) ret-

ning i rommet, kan vi tegne inn vektorer som har den korrekte retningen i forhold til aksekorset vi la inn i beskrivelsen av det Euklidske rommet i punkt 1. De røde pilene angir disse vektorene, og lengden er proporsjonal med feltstyrken. Merk at lengdene har absolutt ingenting å gjøre med avstanden i det Euklidske rommet å gjøre! Elektrisk felt er en langt mer abstrakt størrelse enn tid og rom.

3. Det magnetiskefeltet har en veldefinert retning i rommet, slik at vi også kan tegne inn magnetfeltvektorer i det Euklidske rommet i punkt 1 på en lignende måte vi gjorde for det elektriskefeltet. Det blir alle de blå pilene i figuren, og lengden er også i dette tilfellet proporsjonal med feltstyrken.

I figuren blir det gjerne også tegnet inn omhyllingskurver som beskriver enden på vektorpilene for elektrisk og magnetisk felt.

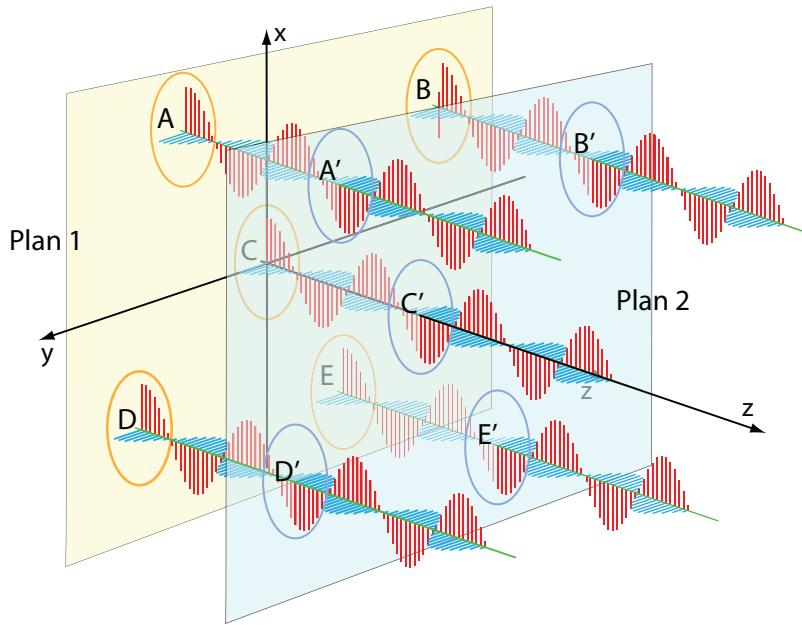
Figur 9.4 alene gir bare indirekte uttrykk for at bølgen er plan. Uten en del ekstra kunnskap som vi ikke har kommet inn på til nå, vil vi ikke kunne si at bølgen er plan ut fra denne figuren alene!

I figur 9.7 har vi forsøkt å få fram det som faktisk er de viktigste trekkene med en plan elektromagnetisk bølge. Disse er:

- Dersom bølgen på ett sted i rommet beveger seg i z-retning, vil den overalt i rommet bevege seg i nøyaktig samme retning. Det vil si at dersom vi ønsker å finne ut hvordan elektrisk og magnetisk felt endrer seg f.eks. i et punkt et annet sted i rommet enn langs z-aksen (som i figur 9.4), kan vi legge en linje parallellt med z-aksen gjennom punktet, og får samme variasjon langs denne linjen som langs z-aksen.
- Fasen i feltvariasjonene langs den nye linjen er identisk med fasen i feltvariasjonene langs z-aksen såfremt vi f.eks. bruker skjæringspunktet mellom linjene og xy-planet som utgangspunkt for målingene. Med andre ord, når elektrisk felt har sin maksimalverdi i punktet der z-aksen skjærer xy-planet, vil også det elektriskefeltet samtidig ha en maksimalverdi i punktet der den andre linjen skjærer xy-planet.
- Velger vi et annet plan som også står vinkelrett på bølgensbevegelsesretning, vil momentanverdien av det elektriskefeltet være identisk overalt i dette planet.

Merk at det ikke er noe som helst som stikker ut av en elektromagnetisk bølge. For et vilkårlig valgt punkt i rommet er det feltet i seg selv som endrer verdi. Feltet har en retning i rommet, men ingen piler skyter ut til siden og ingen sinusbuer finnes langs bølgen. Det er derfor en totalt annen situasjon enn når vi f.eks. klimprer på en gitarstreng der strengen faktisk beveger seg på tvers av lengderetningen.

Det var av denne grunn at vi tidligere laget en definisjon på en transversal bølge som kunne brukes også når det ikke finnes noe forflytning av noe som helst i en retning normalt på bølgebevegelsesretningen. I kapittel 5 sa vi at en transversal bølge er karakterisert ved at den ikke har lokal rotasjonssymmetri rundt vektoren som angir retningen bølgen brer seg i. Når f.eks. et elektrisk felt et sted i rommet er rettet i x-retning og bølgen går i z-retning, får vi ikke samme matematiske beskrivelse av bølgen dersom vi dreier aksesystemet en vilkårlig vinkel omkring z-aksen. Med andre ord har vi ikke lokal rotasjonssymmetri, og bølgen er transversal.



Figur 9.7: Et øyeblikksbilde av en plan elektromagnetisk bølge beskrevet langs fem ulike parallelle linjer i rommet, alle rettet i samme retning som bølgen beveger seg i. Fasen er identisk i et plan vinkelrett på utbredelsesretningen. Fasen varierer med hvilket plan vi velger og hvilket tidspunkt vi betrakter feltet, men innenfor samme plan er altså momen-tanverdien av det elektriske (og magnetiske) feltet identisk. Det er dette som karakteriserer en plan bølge. En annen måte å si dette på er at bølgefronten er plan.

9.10 Hjelpestoff

9.10.1 Nyttige matematiske relasjoner

Vi lister her opp noen nyttige matematiske relasjoner fra matematikken du forhåpentligvis har møtt tidligere:

Felles for alle uttrykk er at vi opererer med et skalarfelt:

$$\phi = \phi(x, y, z)$$

og et vektorfelt

$$\vec{a} = a_x \vec{i} + a_y \vec{j} + a_z \vec{k}$$

En gradient er da definert som:

$$\text{grad } \phi \equiv \nabla \phi \equiv \frac{\partial \phi}{\partial x} \vec{i} + \frac{\partial \phi}{\partial y} \vec{j} + \frac{\partial \phi}{\partial z} \vec{k}$$

Divergensen er definert som:

$$\text{div } \vec{a} \equiv \nabla \cdot \vec{a} \equiv \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}$$

Divergensen til en gradient er definert som:

$$\text{div grad } \phi \equiv \nabla \cdot (\nabla \phi) \equiv \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \equiv \Delta \phi$$

Rotasjon (engelsk: curl) er definert som:

$$\text{rot } \vec{a} \equiv \nabla \times \vec{a} \equiv$$

$$\left| \begin{array}{ccc} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ a_x & a_y & a_z \end{array} \right| =$$

$$\left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z} \right) \vec{i} + \left(\frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x} \right) \vec{j} + \left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y} \right) \vec{k}$$

Merk deg hva som er vektorer og hva som er skalarfelt. Generelt gjelder:

- En gradient omdanner et skalarfelt til et vektorfelt.
- En divergens går motsatt vei.
- Div-grad starter med et skalarfelt, går via et vektorfelt og ender til slutt med et skalarfelt igjen.
- En rotasjon derimot starter med et vektorfelt og ender med et vektorfelt.

Symbolet ∇ symboliserer ulike operasjoner alt etter om den virker på et skalarfelt eller et vektorfelt, og spesielt blir det ekstra utfordrende å anvende ∇^2 på en vektor, siden vi da må anvende Laplace-operatoren på hver av komponentene i vektoren hver for seg:

$$\begin{aligned} \nabla^2 \vec{a} = & \left(\frac{\partial^2 a_x}{\partial x^2} + \frac{\partial^2 a_x}{\partial y^2} + \frac{\partial^2 a_x}{\partial z^2} \right) \vec{i} + \\ & \left(\frac{\partial^2 a_y}{\partial x^2} + \frac{\partial^2 a_y}{\partial y^2} + \frac{\partial^2 a_y}{\partial z^2} \right) \vec{j} + \\ & \left(\frac{\partial^2 a_z}{\partial x^2} + \frac{\partial^2 a_z}{\partial y^2} + \frac{\partial^2 a_z}{\partial z^2} \right) \vec{k} \end{aligned}$$

Noen nytte relasjoner ellers er som følger:

$$\text{rot grad } \phi = \nabla \times (\nabla \phi) = 0$$

$$\text{div rot } \vec{a} = \nabla \cdot (\nabla \times \vec{a}) = 0$$

$$\text{rot}(\text{rot } \vec{a}) = \text{grad}(\text{div } \vec{a}) - \Delta \vec{a} = \nabla \times (\nabla \times \vec{a}) = \nabla(\nabla \cdot \vec{a}) - \nabla^2 \vec{a}$$

9.10.2 Nyttige relasjoner og størrelser fra elektromagnetismen

Her er noen relasjoner fra elektromagnetismen som en oppfriskning av tidligere kunnskap:

Elektrisk feltstyrke \vec{E} måles i V/m.

Elektrisk fluksstetthet \vec{D} måles i C/m².

Magnetisk feltstyrke \vec{H} måles i A/m.

Magnetisk fluksstetthet \vec{B} måles i T.

Elektrisk fluksstetthet betegnes også ofte som "forskyvningsvektor".

Elektrisk tomromspermittivitet ϵ_0 måles i F/m = (As)/(Vm). Den er definert eksakt som

$$\epsilon_0 \equiv \frac{1}{\mu_0 c_0^2} \approx 8.854188 \cdot 10^{-12} \text{ F/m}$$

Relativ permittivitet ϵ_r er normalt et tall større enn 1.0.

Magnetisk tomromspermeabilitet μ_0 måles i H/m, og er gitt eksakt:

$$\mu_0 \equiv 4\pi \cdot 10^{-7} \text{ H/m} \approx 1.256637 \cdot 10^{-6} \text{ H/m}$$

Relativ permeabilitet μ_r er oftest meget nær lik 1.0 for de fleste materialer. Unntak er ferromagnetiske materialer.

Lyshastigheten i vakuum er gitt eksakt som:

$$c_0 \equiv 299\,792\,458 \text{ m/s}$$

SI-grunnenhetene er nå lyshastigheten i vakuum og sekundet. Lengden 1 meter er ikke lenger en av grunnenhetene!

Sammenhengen mellom feltstyrker og fluksstettheter er som følger:

$$\vec{D} = \epsilon_r \epsilon_0 \vec{E}$$

$$\vec{B} = \mu_r \mu_0 \vec{H}$$

9.11 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Omdanne Maxwells ligninger fra integral- til differensiell form (forutsatt at Stokes teorem og divergensteoremet er oppgitt).
- Utlede bølgeligningen for elektromagnetiske felt i vakum forutsatt at ligning (9.16) er oppgitt.
- Gjøre nøyne rede for forskjellen mellom "plan bølge" og polarisasjon.
- Angi hvor stor energitransport det er i en plan elektromagnetisk bølge.
- Anvende Poynting vektor og kjenne begrensinger i betraktninger som ligger bak denne størrelsen.
- Angi og anvende uttrykk for strålingstrykk i et elektromagnetisk felt i en plan bølge.
- Forklare hva vi mener med nærfelt og fjernfelt og hvorfor disse iblant er svært forskjellige.
- Gjøre rede for hvilke egenskaper til elektromagnetiske felt som er forskjellige i de to sonene.
- Gjøre rede for flere problemer med å anvende fotonbegrepet for alle elektromagnetiske felt/bølger.

9.12 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Fortell kort hvordan du kan karakterisere et sted i rommet hvor divergensen av det elektriske feltet er forskjellig fra null. Tilsvarende, fortell kort hvordan du vil karakterisere et sted i rommet hvor rotasjonen til det elektriske feltet er forskjellig fra null.
2. Ved overgang fra Maxwells ligninger på integralform til differensialform bruker vi en argumentasjon som baserer seg på “en midlere” lengde- eller volum-skala. Hva mener vi med dette?
3. Anta at vi måler elektrisk og magnetisk felt i en elektromagnetisk bølge i fjernsonen. Kan vi ut fra målingene bestemme hvilken retning bølgene kom fra?
4. Vi setter en vekselspenning over en kondensator, eller vi sender en vekselstrøm gjennom en solenoide. Forsök å finne retningen for elektrisk og magnetisk felt og relativt størrelsesforhold. Vil disse feltene følge de velkjente lovemessighetene som gjelder for elektriske og magnetiske felt for elektromagnetiske bølger?
5. Det sies iblant at for en elektromagnetisk bølge i vakuum er elektrisk og magnetisk felt vinkelrette på hverandre. Magnetfelt og elektrisk felt har ikke denne relasjonen til hverandre et lite stykke fra en solenoide (“spole”), selv om den er i vakuum og det er et høyfrekvent elektrisk og magnetisk felt til stede. Hva skyldes dette?
6. Er polarisering en egenskap til alle elektromagnetiske bølger, ikke bare med lys? Kan lydbølger ha en polarisering?
7. Tenk deg at du holder et sugerør opp i sollyset slik at sollyset går gjennom røret. Beskriv den elektromagnetiske bølgen i volumelement etter volumelement inne i (langs) sugerøret.
8. En elektromagnetisk bølge (f.eks. kraftig lys) kan ha et elektrisk felt på om lag 1000 V/m. Kan det føre til elektrisk sjokk dersom vi går inn i dette kraftige lyset?
9. Det magnetiske feltet i kraftig laserlys kan være opp til 100 ganger så kraftig som jordmagnetfeltet. Hva vil skje dersom vi lyser med dette laserlyset på nåla i et kompass?
10. Poynting vektor angir effekt som brer seg med en elektromagnetisk bølge. Kan vi bruke Poynting vektor for å beregne effekt som brer seg ut fra en kraftledning til beboere i nærheten? Begrunn svaret.
11. Dersom du blinker med lyset fra en lommelykt, vil du da oppleve en rekyl lignende det vi får når vi skyter med et gevær? Drøft svaret.
12. I ethvert fysisk system/fenomen ligger det innebygget en lengdeskala og en tidsskala. Hva menes med et slikt utsagn når vi betrakter elektromagnetiske bølger?
13. Det finnes mange ulike løsninger av Maxwells ligninger. Kan en av løsningene være elektromagnetiske bølger hvor vi praktisk talt bare har et elektrisk felt (og magnetfeltet er mye lavere enn E_0/c)?

14. I flere ligninger i dette kapitlet inngår den relative elektriske permittiviteten ϵ_r .
- Lyshastigheten er knyttet opp til denne størrelsen. Hvordan?
 - Den relative permittiviteten forteller oss litt om hvilke fysiske prosesser som foregår når lys passerer glass. Hvilke prosesser er det vi da tenker på?
 - Mange synes det er greit nok å forstå at lyset får redusert hastighet når det går fra luft eller vakuum til glass, men de synes det er vanskelig å forstå at lys kommer opp igjen til den opprinnelige hastigheten når lyset går ut av glasset igjen. Hva tror du er grunnen til at mange synes dette er rart?

Regneoppgaver

15. Vis at en plan elektromagnetisk bølge i vakuum tilfredsstiller alle fire Maxwells ligninger.
16. Skriv opp Maxwells ligninger på integralform og angi riktige navn på dem. Utled i detalj Ampères lov på differensiell form.
17. Utledning av bølgeligningen fra Maxwells ligninger følger omtrent de samme tristene enten vi gjennomfører prosedyren for å komme fram til bølgeligningen for det elektriske feltet eller for magnetfeltet. Lag en liste som viser hvilke trinn/triks som benyttes (ønsker bare en relativt kort, punktvis/summarisk liste uten at man går i full detalj).
18. Finn frekvensen til gult lys med bølgelengde 580 nm. Gjør det samme med røntgenstråling med bølgelengde ca 1 nm. De raskeste oscilloskopene vi har tilgjengelig har en samplingsfrekvens i størrelsesorden 10-100 GHz. Kan vi med et slikt oscilloskop se oscillasjonene i elektrisk felt i røntgenbølgene? Hva med gult lys?
19. En elektromagnetisk bølge har et elektrisk felt gitt ved $\vec{E}(y, t) = E_0 \cos(ky - \omega t) \hat{k}$. $E_0 = 6.3e4$ V/m, og $\omega = 4.33e13$ rad/sek. Bestem bølgelengden for bølgen. Hvilken retning beveger bølgen seg? Bestem \vec{B} (vektor). Gjør du noen spesielle antakelser ved beregningene, må disse angis.
20. En elektromagnetisk bølge med frekvensen 65.0 Hz går gjennom et isolerende materiale med relativ permittivitet på 3.64 og relativ permeabilitet på 5.18 for denne frekvensen. Elektrisk felt har en amplitude på 7.20e-3 V/m. Hvor stor er bølgehastigheten i dette mediet? Hva er bølgelengden i mediet? Hvor stor er amplituden for det magnetiske feltet? Hva er intensiteten til bølgen? Er beregningene du har gjort egentlig gyldige? Begrunn svaret!
21. En intens lyskilde stråler ut lys likt i alle retninger. I avstanden 5.0 m unna kilden er strålingstrykket på en flate som absorberer lyset perfekt lik 9.0e-9 Pa. Hvor stor effekt stråler lyskilden ut?
22. En måling ved jordoverflaten viser at lysintensiteten i sollyset er 0.78 kW/m². Hvor stor kraft vil strålingstrykket bli på et 1 m² solpanel? Angi de antakelsene du gjør.
23. For en elektromagnetisk bølge er det gitt at elektrisk felt ved ett tidspunkt er rettet i x-retning og magnetfelt i -z-retning. Hvilken retning brer bølgen seg? Hva dersom retningene var hhv -z og y retning? Gjør vi en antakelse når vi angir svarene?

24. En vanlig lab-helium-neon laser har en effekt på 12 mW og strålen har en diameter på 2.0 mm. Anta at intensiteten er den samme over hele tverrsnittet (hvilket er helt feil, men det kan forenkle beregningene). Hva er amplituden til det elektriske og magnetiske feltet i strålen? Hva er gjennomsnittlig energitethet i elektrisk felt i strålen? Hva med energitetheten i magnetfeltet? Hvor mye energi har vi i en 1.0 m lang bit av strålen?
25. Noen hundre meter unna en basestasjon ble det elektriske feltet målt til 1.9 V/m og magnetfeltet 1.2 mA/m (begge ved om lag 900 MHz). En kyndig person konkluderte at målingene ikke var i overensstemmelse med hverandre. Hva tror du var grunnen til denne konklusjonen?
26. Ved bakken bare et par titalls meter fra en kraftledning ble det målt et elektrisk felt på 1.2 kV/m og "magnetfelt" på $2.6 \mu\text{T}$ (mikrotesla) (begge ved 50 Hz). Det er i praksis ofte magnetisk fluksdækkethet som oppgis ved lave frekvenser, men vi kan gjøre om fra B til H og får da at $2.6 \mu\text{T}$ svarer til magnetfeltverdien 2.1 A/m . Er det samsvar mellom elektrisk felt og magnetfelt i dette tilfellet? Kommenter likheter/-forskjeller mellom situasjonen i forrige oppgave og situasjonen i denne oppgaven.
27. En dag det gjøres målinger av elektrisk felt og magnetfelt samme sted nær kraftledningen som i forrige oppgave, er verdiene 1.2 kV/m og 0.04 A/m . Kan vi konkludere at det er noe feil med et av måleinstrumentene i dette tilfellet?
28. Ifølge StrålevernRapport 2011:6: Radiofrekvente felt i våre omgivelser (www.nrpa.no/dav/388936eccd.pdf, tilgjengelig 15. januar 2011) er "strålingen" fra basestasjoner, trådløse nett, radio m.m. som oftest mindre enn 0.01 W/m^2 rundt omkring i landet vårt. Beregn elektrisk felt og magnetfelt som svarer til 0.01 W/m^2 dersom vi tenker oss at strålingen domineres av mobiltelefonkommunikasjon fra en basestasjon ved 1800 MHz.
29. Når vi bruker en mobiltelefon et sted hvor dekningen er dårlig slik at mobiltelefonen yter maksimal effekt, gir mobiltelefonen om lag $0.7 - 1.0 \text{ W}$ effekt mens kommunikasjonen foregår. Anslå effekttettheten 5 cm fra mobiltelefonen dersom du antar en isotrop effektfordeling omkring mobiltelefonen. Sammenlign verdien med målte effekttettheter fra basestasjoner, trådløse nett osv. gitt i forrige oppgave.
30. Vanligvis oppgis ikke "strålingen" fra en mobiltelefon i form av effekttetthet målt i W/m^2 , men i SAR (Specific Absorption Rate).
- Søk litt på web for å finne litt om SAR. Angi url for den kilden du bruker.
 - Forklar hva SAR innebærer, og hva er enheten for SAR?
 - Hva tror du er grunnen til at man har valgt en slik enhet i dette tilfellet selv om vi bruker effekttetthet fra basestasjoner og denslags, med omtrent samme frekvens som mobiltelefonen?
31. La oss betrakte interplanetarisk støv i vårt solsystem. Anta at støvet er kuleformet og har en radius r og en tetthet ρ . Anta at all stråling som treffer støvkornet blir absorbert. Sola har en total utstrålt effekt P_0 og masse M . Gravitasjonskonstanten er G . Avstanden fra Sola er R . Sett opp et uttrykk som angir forholdet mellom kraften som skyldes strålingstrykket fra solstrålene mot støvkornet, og gravitasjonskraften mellom Sola og støvkornet. Bestem radien i støvkornet når de to kreftene er like store når vi setter inn realistiske verdier for de størrelsene som inngår. ($\rho = 2.5\text{e}3 \text{ kg/m}^3$, $P_0 = 3.9\text{e}26 \text{ W}$, $M = 1.99\text{e}30 \text{ kg}$, $G = 6.67\text{e}-11 \text{ Nm}^2/\text{kg}^2$).

32. Finn forholdet mellom gravitasjonskraften som virker på Jorda fra Sola, og kraften på Jorda som skyldes strålingstrykket fra Sola. Jordas masse er 5.98×10^{24} kg. Jordas radius kan du estimere dersom du husker at avstanden mellom en pol og ekvator er ca. 10 000 km.

Kapittel 10

Refleksjon og transmisjon, polarisasjon, dobbeltbrytning



Lysspillet i edelstener og diamanter fascinerer oss. Også fysikken bak lysspillet er fascinerende. Det er denne beskrivelsen som er tema for dette kapitlet.

Lys som spiller i iskristaller eller i edelsteiner fascinerer oss. Lyset kan reflekteres og transmitteres både på ytre og innvendige flater, og totaliteten kan bli ganske komplisert.

Vi har i tidligere kapitler sett at mekaniske bølger delvis reflekteres og delvis transmitteres ved overgang fra et medium til et annet. Det samme skjer når elektromagnetiske bølger så som lys f.eks. går fra luft til glass. Siden elektromagnetiske bølger er transversale og følger Maxwells ligninger, blir likevel matematikken en del mer komplisert for lys sammenlignet med f.eks. lydbølger. Polarising spiller en betydelig rolle.

Mange materialer kan påvirke polariseringen av elektromagnetiske bølger. Det gir opphav til effekter som kan brukes i alt fra studiet av enkle materialer i laboratoriet til fysiske forhold ved supernovaeksplosjoner i verdensrommet. Polarising kan også brukes for noe så hverdaglig som visning av en 3D film på kino eller TV.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 10032013.

10.1 Innledning

I forrige kapittel fant vi at en plan elektromagnetisk bølge med (fase)hastigheten

$$\begin{aligned} c &= \frac{1}{\sqrt{\epsilon_0 \epsilon_r \mu_0 \mu_r}} \\ &= \frac{1}{\sqrt{\epsilon_0 \mu_0}} \frac{1}{\sqrt{\epsilon_r \mu_r}} = \frac{c_0}{\sqrt{\epsilon_r \mu_r}} \end{aligned}$$

er én mulig løsning av Maxwells ligninger i et uendelig stort homogent medium uten "fri ladninger". Symbolene har vanlig betydning.

Lyshastigheten i et medium (uten frie ladninger) er lyshastigheten i vakuum c_0 dividert på brytningsindeksen n for mediet.

$$c \equiv \frac{c_0}{n}$$

De aller fleste medier vi skal jobbe med er diamagnetiske eller paramagnetiske. Dette gjelder for eksempel for glass som brukes i optikk for lys. Da er $\mu_r \approx 1.00$. Følgelig får vi at:

$$n \approx \sqrt{\epsilon_r}$$

Brytningsindeksen er med andre ord direkte relatert til "polarisasjons-susceptibiliteten" til mediet dersom vi skal forsøke oss på en spesiell beskrivelse. Den relative permittiviteten er et mål for dette. Jo lettere vi kan forskyve elektronskyen rundt atomene vekk fra likevektpunktet sentrert på atomene, desto saktere går lyset gjennom mediet.

I dette kapitlet går vi et skritt videre ved å se hvordan bølgebeskrivelsen blir modifisert når bølgen treffer et grensesjikt mellom to ulike medier i kontakt med hverandre. Igjen står Maxwells ligninger sentralt i beregningene.

Et gjennomgangstema i kapitlet er polarisering, men polarisering dukker opp i to nokså forskjellige sammenhenger. Vær oppmerksom slik at du ikke blander dem sammen!

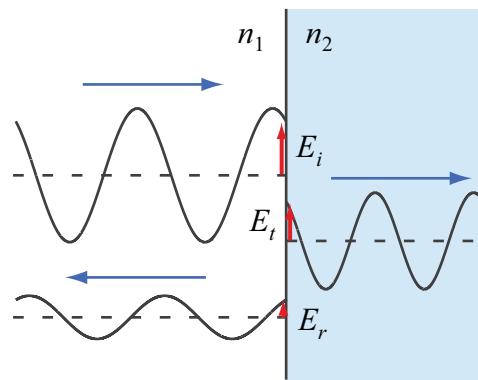
10.2 Elektromagnetisk bølge vinkelrett inn mot et grensesjikt mellom to medier

Generelt sett er det uendelig mange ulike geometrier og uendelig mange forskjellige løsninger av Maxwells ligninger når en elektromagnetisk bølge kommer inn mot et grensesjikt mellom to medier. Vi må forenkle enormt for å komme fram til lovmessigheter som lar seg beskrive på sluttet form.

I dette delkapitlet vil vi bruke Faraday-Henrys lov pluss et energiregnskap for å finne ut hvor mye av en elektromagnetisk bølge som blir reflektert og transmittert når bølgen f.eks. går fra luft inn i glass. Vi antar at den elektromagnetiske bølgen er en tilnærmet plan bølge som treffer normalt på en grenseflate mellom to ulike, homogene medier uten frie ladninger. Vi gjør følgende antakelser for mediet og grenseflaten:

1. Antar at mediet i seg selv er homogent innenfor et volum λ^3 hvor λ er bølgelengden.
2. Antar at grenseflaten er plan over et areal som er mye større enn λ^2 .
3. Antar at tykkelsen på grenseflaten er mye mindre enn bølgelengden λ .

10.2. ELEKTROMAGNETISK BØLGE VINKELRETT INN MOT ET GRENSESJIKT MELLOM TO MEDIUMER



Figur 10.1: En elektromagnetisk bølge som kommer vinkelrett inn mot et annet medium, blir delvis reflektert og delvis transmittert. Bølgene er tegnet adskilt for å indikere momentant elektrisk felt for hver av dem.

Så lenge vi betrakter lys med bølgelengde i området 400 - 800 nm som går gjennom glass der atomene ligger noen få tiendedels nanometer fra hverandre, er disse tre antakelsene rimelig godt oppfylt. Men betingelsene er slett ikke oppfylt i alle vanlige tilfeller. Når lys går gjennom regndråper er dråpene ofte såpass store at vi tilnæmet kan bruke den formalismen vi straks skal utlede. Men når dråpene blir så små at betingelsene ovenfor ikke tilfredsstilles, må Maxwells ligninger brukes direkte. For små dråper får vi såkalt Mie-spredning som ikke gir en vanlig regnbue, men en nærmest fargeløs bue.

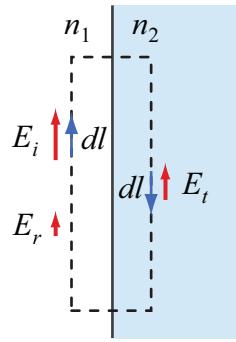
Også for elektromagnetiske bølger i helt andre bølgelengdeområder enn lys, er det viktig å tilfredsstille de tre antakelsene. Ta for eksempel røntgen med bølgelengde omkring 0.1 nm. Da er bølgelengden omtrent like stor som avstanden mellom atomer. For radiobølger blir det problemer å tilfredsstille antakelsene. Det betyr at de lovmessighetene vi nå skal utlede i dette kapitlet i praksis ofte er begrenset til elektromagnetiske bølger i form av synlig lys, eller i alle fall nærliggende bølgelengder.

La oss nå anta at antakelsene ovenfor er tilfredsstilt og at vi sender elektromagnetiske bølger normalt inn på grensesjiktet. Noe av bølgen vil da bli reflektert i grensesjiktet og vandrer bakover i det opprinnelige mediet, mens resten av bølgen transmitteres inn i neste medium og fortsetter der. I figur 13.12 er de tre bølgene tegnet hver for seg for å få frem hovedtrekkene best mulig. Bølgene som er tegnet inn kan anses f.eks. som en komponent av det elektriskefeltet (i en gitt retning vinkelrett på normalen til grensesjiktflaten). Brytningsindeksen på venstre side i figuren er n_1 og på høyre side n_2 , og vi har ikke foreløpig sagt noe om hvilken av disse som er størst. Av samme grunn har vi ikke tatt stilling til om den reflekterte bølgen vil ha motsatt fortegn som den innkommende i selve grensesjiktet. Foreløpig regner vi fortegnene ut fra det som er tegnet inn i figuren, så skal vi diskutere detaljer siden.

Vi lager en integrasjonsvei som indikert i figur 10.2. Integrasjonsveien er orientert slik at de lange strekkene har nøyaktig samme retning som det elektriskefeltet i den elektromagnetiske bølgen. Vi anvender så Faradays lov:

$$\oint \vec{E} \cdot d\vec{l} = -\left(\frac{d\Phi_B}{dt}\right)_{innenfor} \quad (10.1)$$

$$\int_{venstre-side} (\vec{E}_i \cdot d\vec{l} + \vec{E}_r \cdot d\vec{l}) + \int_{tvers-oppe} + \\ \int_{hoyre-side} \vec{E}_t \cdot d\vec{l} + \int_{tvers-nede}$$



Figur 10.2: Integrasjonsvei som benyttes ved bruk av Faradays lov for å finne sammenhenger mellom elektrisk felt fra ulike komponenter. Se teksten for detaljer.

$$= -\frac{d}{dt} \int_A \vec{B} \cdot d\vec{A} \approx 0$$

Elektrisk felt står vinkelrett på integrasjonsveien i tvers-delene slik at vi ikke får noe bidrag her. Grunnen til at fluksen til magnetfelt er omtrent null skyldes ikke at $\vec{B} \cdot d\vec{A}$ er omtrent null, men at vi kan gjøre arealet vilkårlig lite. Vi har jo i starten antatt at grensesjiktet er "uendelig tynt".

Dersom vi antar at positiv retning for det elektriske feltet er slik som vist i figur 10.2, får vi:

$$E_i L + E_r L - E_t L = 0$$

$$E_i + E_r = E_t \quad (10.2)$$

Vi kan bruke Ampère-Maxwells lov på lignende måte og få:

$$H_i + H_r = H_t$$

I tillegg kan vi sette opp et *energiregnskap*: All energi inn per tid må være lik energi som går ut fra grensesjiktet per tid. Vi kunne brukt Poynting vektor, men velger heller varianten som går på elektrisk felt alene, eller nærmere bestemt at intensiteten i en elektromagnetisk bølge er gitt ved:

$$I = c u_E = \frac{1}{2} c \vec{E} \cdot \vec{D} = \frac{1}{2} c \epsilon_0 \epsilon_r E^2$$

hvor u_E er energitettheten i bølgen og c er bølgehastigheten i mediet vi er i. Energiregnskapet gir oss da:

$$\frac{1}{2} c_1 \epsilon_0 \epsilon_{r1} E_i^2 = \frac{1}{2} c_1 \epsilon_0 \epsilon_{r1} E_r^2 + \frac{1}{2} c_2 \epsilon_0 \epsilon_{r2} E_t^2$$

$$c_1 \epsilon_{r1} (E_i^2 - E_r^2) = c_2 \epsilon_{r2} E_t^2$$

$$c_1 \epsilon_{r1} (E_i + E_r)(E_i - E_r) = c_2 \epsilon_{r2} E_t^2$$

Men $E_i + E_r = E_t$, følgelig:

$$c_1 \epsilon_{r1} (E_i - E_r) = c_2 \epsilon_{r2} E_t$$

La oss se nærmere på konstantleddene. Vi benytter da uttrykkene for lyshastigheten i et medium gitt innledningsvis i dette kapitlet:

$$c_1 = \frac{c_0}{n_1} \approx \frac{c_0}{\sqrt{\epsilon_{r1}}}$$

Herav:

$$c_1 \epsilon_{r1} = \frac{c_0}{\sqrt{\epsilon_{r1}}} \epsilon_{r1}$$

$$c_1 \epsilon_{r1} = c_0 \sqrt{\epsilon_{r1}} = c_0 n_1$$

Setter vi dette uttrykket (og tilsvarende for medium 2) inn i uttrykket ovenfor, følger:

$$n_1(E_i - E_r) = n_2 E_t \quad (10.3)$$

Vi kombinerer nå ligning (10.2) og (10.3) og eliminerer i første omgang E_t for å finne en sammenheng mellom E_i og E_r :

$$\begin{aligned} n_1 E_i - n_1 E_r &= n_2 E_i + n_2 E_r \\ (n_1 - n_2) E_i &= (n_1 + n_2) E_r \end{aligned}$$

Forholdet mellom amplituden for reflektert bølge i forhold til innkommende bølge er:

$$\frac{E_r}{E_i} = \frac{n_1 - n_2}{n_1 + n_2} \quad (10.4)$$

Vi ser at høyresiden kan være både positiv og negativ (og lik null dersom $n_1 = n_2$). For $n_2 > n_1$ er uttrykket negativt, det vil si at E_r har motsatt fortegn av E_i (dvs E_r har motsatt retning av det som er angitt i figur 13.12).

♠ ⇒ Noen side-bemerkninger:

Vi sier ofte at "medium 2 er optisk tettere enn medium 1" når $n_2 > n_1$. Dette er imidlertid en uheldig uttrykksmåte som nylig er blitt kraftig kritisert i American Journal of Physics, fordi det gir helt gale assosiasjoner. Uttrykket har antakelig sitt opphav i svingninger på en streng der vi går fra en streng med liten masse per lengde til en streng med større masse per lengde. I en slik situasjon har vi nemlig også at den reflekterte bølgen har motsatt fortegn av den innkommende (i ekstremt tilfelle når strengen er festet til et fast punkt, blir bølgen i sin helhet reflektert).

For lys som kommer fra luft mot glass, er situasjonen en annen. Elektromagnetiske bølger forplanter seg lett i et medium hvor mediet (nesten) ikke kan polariseres. (Nesten) ingen materielle partikler (elektroner) må forflyttes, og bølgen forplanter seg med nær maksimal hastighet som er lyshastigheten i vakuum. Bølgen kommer så til glass hvor elektroner forflytter seg i det elektriske feltet og polariserer glasset i takt med tidsendringene i det elektriske feltet. Elektronene kan imidlertid ikke forflyttes momentant! Det følger av Newtons lover. Følgelig får vi et innkommende elektrisk felt (fra den innkommende bølgen) som vil kombineres med feltet fra de polariserte ladningene. Men polariseringen kommer litt etter i tid sammenlignet med det innkommende feltet. Resultatet er at det kombinerte feltet hele tiden ligger litt på etterskudd i forhold til hvordan det ville vært uten polariseringen av mediet. Det er dette som fører til at bølgen (summen av elektrisk felt fra den innkommende bølgen OG feltet fra de polariserte ladningene) vil forplante seg saktere i mediet enn i vakuum. Det er fasehastigheten vi her omtaler. For pulset felt kommer gruppehastigheten inn, men vi skal ikke gå inn på dette her. ← ♠]

For $n_2 < n_1$ er uttrykkene i ligning (10.4) positive, det vil si at E_r har samme fortegn som E_i (dvs E_r har samme retning av det som er angitt i figur 13.12).

La oss til slutt kombinere ligning (10.2) og (10.3) ved å eliminere E_r for å finne en sammenheng mellom E_i og E_t . Det gir:

$$n_1 E_i - n_1 E_t + n_1 E_i = n_2 E_t$$

Forholdet mellom amplituden for transmittert bølge i forhold til innkommende bølge er:

$$\frac{E_t}{E_i} = \frac{2n_1}{n_1 + n_2} \quad (10.5)$$

Vi ser at den transmitterte bølgen alltid har samme fortegn i det elektriske feltet som den innkommende bølgen (like ved grensesjiktet).

Ligningene (10.4) og (10.5) gir forholdene mellom elektrisk felt på begge sider av grenseflaten. Når vi skal bedømme hvor stor del av lyset som reflekteres og transmitteres, ønsker vi å se på intensitetene. Vi har allerede sett at intensitetene er gitt ved uttrykk av typen:

$$I_i = \frac{1}{2} c_1 \epsilon_0 \epsilon_{r,1} E_i^2 \approx \frac{1}{2} c_0 \epsilon_0 n_1 E_i^2$$

Vi finner da følgende sammenhenger mellom intensitetene:

$$\frac{I_r}{I_i} = \frac{n_1 E_r^2}{n_1 E_i^2} = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \quad (10.6)$$

og

$$\frac{I_t}{I_i} = \frac{n_2 E_t^2}{n_1 E_i^2} = \left(\frac{2n_1}{n_1 + n_2} \right)^2 \cdot \frac{n_2}{n_1} \quad (10.7)$$

Velger vi å se på hva som skjer ved grenseflaten mellom luft og glass (brytingsindeks hhv 1.00 og 1.54), og får:

Reflektert:

$$\frac{I_r}{I_i} = \left(\frac{0.54}{2.54} \right)^2 \approx 0.0452$$

Transmittert:

$$\frac{I_t}{I_i} = \left(\frac{2}{2.54} \right)^2 \cdot 1.54 \approx 0.9548$$

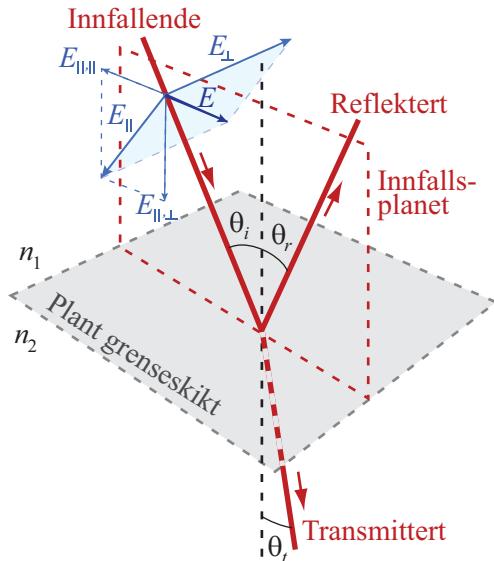
Vi ser altså at om lag 4.5 prosent av lysintensiteten som sendes vinkelrett inn mot en luft-glass flate blir reflektert, mens om lag 95.5 prosent transmitteres. Dette er situasjonen når det ikke er noe form for overflatebehandling ("antirefleksbehandling") av glassoverflaten.

Det kan til slutt bemerkes at refleksjonen i overflaten fører med seg at vi får delvis stående bølger i området foran grensesjiktet.

10.3 Refleksjon og transmisjon når en bølge kommer på skrå inn mot grenseflaten

Vi skal nå se på refleksjon og transmisjon når en (tilnærmet) plan elektromagnetisk bølge kommer på skrå inn mot en grenseflate mellom to medier. Vi har de samme antakelsene som nevnt i begynnelsen av kapitlet om at grenseflaten er plan og "uendelig stor og uendelig tynn".

10.3. REFLEKSJON OG TRANSMISJON NÅR EN BØLGE KOMMER PÅ SKRÅ INN MOT GRENSESJIKTET



Figur 10.3: Geometri når en stråle elektromagnetiske bølger kommer på skrå inn mot et grensesjikt mellom to medier. Elektrisk feltvektor dekomponeres i en komponent normalt på og parallellt med innfallsplanet. Den sistnevnte komponenten dekomponeres så videre i en komponent som er normalt på og parallellt med grensesjiktet. Se teksten for detaljer.

En vesentlig utfordring i utledningene som nå kommer, består i å holde rede på geometri. Bølger kommer på skrå inn mot grenseflaten, og fysikken blir forskjellig alt etter om det elektriske feltet treffer grenseflaten parallellt med eller på skrå i forhold til flaten. Det kan være lurt å bruke tilstrekkelig tid på å forstå dekomponeringen av den elektriske feltvektoren E i figur 10.3 før du leser videre.

Vi tegner inn en “stråle” som kommer på skrå inn mot grenseflaten. Der strålen treffer grenseflaten, tegner vi en normal til grenseflaten og kaller denne innfallsloddet. Strålen og innfallsloddet utspenner da *innfallsplanet*. Vinkelen mellom den innfallende strålen og innfallsloddet er θ_i . Se figur 10.3.

Den reflekterte strålen vil ligge i innfallsplanet og ha samme vinkel med innfallsloddet som den innfallende strålen, dvs $\theta_i = \theta_r$. Den transmitterte strålen vil også ligge i samme plan som de andre strålene, men den har en vinkel θ_t med innfallsloddet (forlengelsen inn i medium 2).

Vi skal ikke gå inn på noe detaljert bevis for at de tre strålene ligger i samme plan, men Maxwells ligninger er symmetriske med hensyn på tid. Med det menes at dersom én løsning av Maxwells ligninger er en innfallende stråle som deler seg i en reflektert og en transmittert stråle, så er en annen løsning at den reflekterte og transmitterte bølgen kan anses som to innfallende stråler som kommer *inn mot* grenseflaten og kombineres til én utgående stråle (som er lik den opprinnelige innfallende strålen, men med motsatt bevegelsesretning).

Siden vi på sett og vis kan snu tidsforløpet for hva som skjer, betyr det at løsningen må ha en viss grad av symmetri. En følge er at de tre strålene må ligge i innfallsplanet.

Vi starter med å anta at alle tre stråler ligger i innfallsplanet og $\theta_i = \theta_r$ i figur 10.3, og vil så bruke Maxwells ligninger for å få fram hvor mye som reflekteres og transmitteres i grensesjiktet.

Bølgen har imidlertid en vilkårlig polarisering. Det betyr at det elektriske feltet E , som står vinkelrett på den innfallende strålen, kan ha hvilken som helst vinkel i forhold til innfallsplanet. Fysikken blir litt forskjellig for komponenten av elektrisk felt som ligger i

innfallsplanet E_{\parallel} sammenlignet med den komponenten som er vinkelrett på innfallsplanet E_{\perp} .

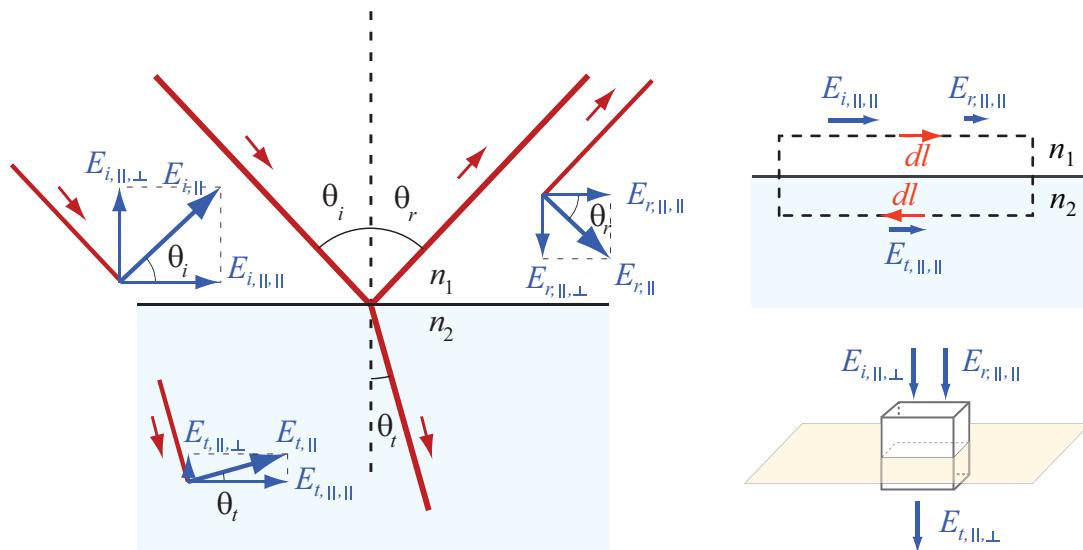
Vi starter med å behandle komponenten av elektrisk felt vinkelrett på innfallsplanet. Denne komponenten vil samtidig være parallel med grenseflaten. Dette svarer til det tilfellet vi hadde for bølgen normalt inn mot grenseflaten (behandlet i forrige delkapittel). Faradays lov anvendt som i figur 13.12 gir som før:

$$E_{i,\perp} + E_{r,\perp} = E_{t,\perp}$$

hvor i , r og t igjen står for innfallende, reflektert og transmittert. \perp indikerer komponenten som er vinkelrett på innfallsloddet, som i sin tur er parallel med grenseflaten. Vi forfølger imidlertid ikke her denne komponenten i detalj.

Det er mer interessant å se på komponenten som er parallel med innfallsplanet, men denne er noe mer komplisert å behandle. Komponenten *parallel med innfallsplanet* kan dekomponeres i en retning som er *normal på grenseflaten* og en som er *parallel med grenseflaten*.

I figur 10.3 har vi forsøkt å angi hvordan det elektriskefeltet i den innkommende bølgen har komponenter både normalt og parallelt med *innfallsplanet*, og at komponenten i innfallsplanet E_{\parallel} igjen kan dekomponeres i en retning parallel $E_{\parallel,\parallel}$ eller vinkelrett $E_{\parallel,\perp}$ på *grensesjiktet/grenseflaten*.



Figur 10.4: Komponenter av elektrisk felt i innfallsplanet for innkommende, reflektert og transmittert stråle. Feltdekomponeringen til venstre er tegnet separat for den innkommende, reflekterte og transmitterte strålen for at ikke for mange detaljer skulle overlappe hverandre. Til høyre er det angitt hvilke retninger komponentene må ha for å regnes som positive i den valgte matematiske utledningen. Se forøvrig teksten for detaljer.

I figur 10.4 er det *bare* komponenten av det elektriske feltet som er parallel med innfallsplanet som er tegnet inn. Dekomponering av denne komponenten er hhv $E_{\parallel,\parallel}$ og $E_{\parallel,\perp}$. Den første delen av indeksen angir komponent i forhold til innfallsplanet, den siste delen av indeksen angir komponent med hensyn til grensesjiktet.

Fra figur 10.3 ser vi at $E_{\parallel,\parallel}$ er vinkelrett på E_{\perp} (komponenten av elektrisk felt normalt på innfallsplanet), selv om begge de to er parallelle med grensesjiktet. Legg også merke til at $E_{\parallel,\perp}$ er vinkelrett på grensesjiktet og dermed parallel med innfallsloddet.

10.3. REFLEKSJON OG TRANSMISJON NÅR EN BØLGE KOMMER PÅ SKRÅ INN MOT GRENSEFLETA

Vi kan anvende Faradays lov på $E_{\parallel,\parallel}$ komponentene av innfallende, reflektert og transmittert bølge, og på samme måte som for bølger normalt inn mot grenseflaten finner vi:

$$E_{i,\parallel,\parallel} + E_{r,\parallel,\parallel} = E_{t,\parallel,\parallel}$$

Positiv retning er definert i høyre del av figuren. Herav følger:

$$E_{i,\parallel} \cos \theta_i + E_{r,\parallel} \cos \theta_r = E_{t,\parallel} \cos \theta_t$$

Siden $\theta_i = \theta_r$, kan vi endelig skrive:

$$E_{i,\parallel} + E_{r,\parallel} = \frac{\cos \theta_t}{\cos \theta_i} E_{t,\parallel} \quad (10.8)$$

Vi trenger enda en ligning for å eliminere en av de tre størrelsene for å finne en sammenheng mellom de to øvrige. I stad brukte vi energiregnskap for å få en ligning til. Det er ikke så enkelt i vårt tilfelle siden vi må ta hensyn til mange komponenter samtidig i det skrå tilfellet. Vi velger i stedet å bruke Gauss lov for elektrisk felt på en liten lukket terningflate med flater parallele med grensesjikt og innfallsplan. Terningen har sideflater med areal A og normal til flaten $d\vec{A}$, og vi har:

$$\oint \vec{D} \cdot d\vec{A} = Q_{fri,innenfor}$$

Det smarte ved dette valget er at alle komponenter av det elektriskefeltet som er parallelle med grensesjiktet vil gi netto null bidrag til integralet. De går inn og ut av sideflatene i samme medium, og disse feltkomponentene er tilnærmet konstant langs flaten så lenge vi lar terningen ha liten sidelengde sammenlignet med bølgelengden. Derimot får vi bidrag fra komponenten som er normalt på endeflaten i sylinderen (og normalt på grensesjiktet). Ved å angi hvordan vi definerer positive feltretninger i høyre del av figur 10.4, følger:

$$D_{i,\parallel,\perp} + D_{r,\parallel,\perp} = D_{t,\parallel,\perp}$$

$$\epsilon_0 \epsilon_{r1} E_{i,\parallel,\perp} + \epsilon_0 \epsilon_{r1} E_{r,\parallel,\perp} = \epsilon_0 \epsilon_{r2} E_{t,\parallel,\perp}$$

Vi bruker nå relasjonen $n \approx \sqrt{\epsilon_r}$, og får:

$$n_1^2 E_{i,\parallel,\perp} + n_1^2 E_{r,\parallel,\perp} = n_2^2 E_{t,\parallel,\perp}$$

Ut fra valgte positive retninger for vektorene i høyre del av figur 10.4, følger da:

$$-n_1^2 E_{i,\parallel} \sin \theta_i + n_1^2 E_{r,\parallel} \sin \theta_r = -n_2^2 E_{t,\parallel} \sin \theta_t$$

Vi bruker så Snels² brytningslov (utledes nedenfor):

$$n_1 \sin \theta_i = n_2 \sin \theta_t$$

og dessuten $\theta_i = \theta_r$. Eliminerer θ_t og får:

$$-n_1^2 E_{i,\parallel} \sin \theta_i + n_1^2 E_{r,\parallel} \sin \theta_i = -n_2 E_{t,\parallel} n_1 \sin \theta_i$$

Forkorter med $n_1 \sin \theta_i$ og får:

²Vi er vant til å skrive "Snell" med to l-er. I en artikkel i American Journal of Physics for få år siden ble vi imidlertid minnet om at Snel egentlig bare skal ha én l, og vi har fulgt oppfordringen om å endre praksis.

$$E_{i,\parallel} - E_{r,\parallel} = \frac{n_2}{n_1} E_{t,\parallel} \quad (10.9)$$

Vi har nå to ligninger som forbinder E_{\parallel} for innkommende, reflektert og transmittert bølge. Vi kan bruke en av disse ligningene for å eliminere ett av de tre, og få sammenhengen mellom de to øvrige. Trekker vi for eksempel ligning (10.8) fra ligning (10.9), får vi:

$$2E_{r,\parallel} = \left(\frac{\cos \theta_t}{\cos \theta_i} - \frac{n_2}{n_1} \right) E_{t,\parallel} \quad (10.10)$$

Ligning (10.10) er interessant i seg selv fordi det synes som det er mulig å få innholdet i parantesen til å bli null. I så fall vil ingenting av den innfallende bølgen reflekteres dersom E ligger i innfallsplanet (for da er jo $E_{\perp} = 0$! Betingelsen er at:

$$\frac{\cos \theta_t}{\cos \theta_i} = \frac{n_2}{n_1}$$

Vi bruker på ny Snel's brytningslov og får:

$$\frac{\cos \theta_t}{\cos \theta_i} = \frac{\sin \theta_i}{\sin \theta_t}$$

$$\sin \theta_i \cos \theta_i = \sin \theta_t \cos \theta_t$$

Vi vet at $\sin(2x) = 2 \sin x \cos x$, følgelig:

$$\sin(2\theta_i) = \sin(2\theta_t)$$

Vi vet videre at $\sin x = \sin(\pi - x)$, som i vårt tilfelle gir:

$$\sin(2\theta_i) = \sin(\pi - 2\theta_t)$$

Denne relasjonen tilfredsstilles dersom:

$$2\theta_i = \pi - 2\theta_t$$

eller

$$\theta_i = \pi/2 - \theta_t$$

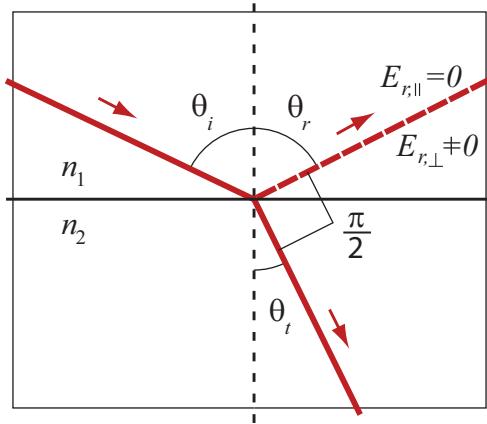
Siden $\theta_i = \theta_r$ kan vi endelig si:

Dersom

$$\theta_r + \theta_t = \pi/2 \quad (10.11)$$

vil det ikke bli reflektert noe lys med polarisering parallellt med innfallsplanet. Da er også vinkelen mellom reflektert og transmittert stråle lik $\pi/2$ som indikert i figur 10.5.

10.3. REFLEKSJON OG TRANSMISJON NÅR EN BØLGE KOMMER PÅ SKRÅ INN MOT GRE



Figur 10.5: Når vinkelen mellom reflektert og transmittert stråle er 90 grader, blir det ikke noe elektrisk felt parallelt med innfallsplanet i den reflekterte strålen.

Siden innfallsvinkel og refleksjonsvinkel er like, er det lett å vise at vinkelen hvor vi ikke har noe reflektert lys med polarisering i innfallsplanet, er karakterisert ved at vinkelen mellom reflektert og transmittert stråle er 90 grader.

Vi ønsker å finne et uttrykk for en vinkel der dette skjer, og starter da med:

$$\frac{n_2}{n_1} = \frac{\cos \theta_t}{\cos \theta_i}$$

og kombinere dette med $\cos \theta_t = \sin \theta_i$ og får:

$$\tan \theta_i = \frac{n_2}{n_1} \equiv \tan \theta_B \quad (10.12)$$

Vinkelen θ_B kalles Brewster-vinkelen. På grenseflaten mellom luft og glass med brytningsindeks 1.54 får vi:

$$\begin{aligned} \tan \theta_B &= \frac{1.54}{1.00} \\ \theta_B &\approx 57^\circ \end{aligned}$$

Ved å sette denne vinkelen inn i Snels brytningslov, kan vi også bestemme θ_t . Resultatet er ca 33° og vi ser at summen av innfallsvinkel og transmisjonsvinkel ("utfallsvinkel") er 90 grader, som forventet.

Det kan være verdt å merke seg at vi også kan få null refleksjon (for lys med elektrisk vektor parallelt med innfallsplanet) dersom lyset går fra glass mot luft. Da får vi:

$$\begin{aligned} \tan \theta_B &= \frac{1.00}{1.54} \\ \theta_B &\approx 33^\circ \end{aligned}$$

Brewster-vinkelen-fenomenet kan med andre ord forekomme når lys går inn i et nytt medium uansett om brytningsindeksen blir høyere eller lavere! Til sammenligning kan totalrefleksjon (som vi kommer tilbake til om litt) bare forekomme når lyset treffer et medium med lavere brytningsindeks.



Figur 10.6: *Upolarisert lys som reflekteres i en luft-glass-grenseflate kan bli fullstendig polarisert når innfallsvinkel er lik Brewstervinkelen. Disse fotografiene viser dette. Venstre bilde er tatt uten polarisasjonsfilter. Høyre bildet er tatt med et polarisasjonsfilter dreid slik at det bare slipper gjennom lys med polarisering parallelt med innfallsplanet. All refleksjon ved Brewstervinkelen fjernes da, og vi ser direkte inn mot gardinene på innsiden av glassruten praktisk talt uten noe reflekser. Det betyr at alt reflektert lys ved Brewstervinkelen er fullstendig polarisert i en retning vinkelrett på innfallsplanet (parallelt med grenseflaten luft - glass). Merk forvrig at reflekser på malingoverflaten påvirkes på lignende måte som reflekser fra glasset.*

10.3.1 Brewster-vinkel-fenomenet i praksis

Det er faktisk relativt enkelt å observere at lys som reflekteres fra en flate ved enkelte vinkler er fullstendig polarisert.

Det essensielle er at vanlig upolarisert lys kan dekomponeres i lys med polarisering parallelt med innfallsplanet og vinkelrett på. For komponenten parallelt med innfallsplanet kan vi oppnå null refleksjon dersom lyset kommer inn med Brewstervinkelen. I så fall vil reflektert lys være fullstendig polarisert normalt på innfallsplanet. Dette kan vi observere ved å bruke et polarisasjonsfilter som bare slipper gjennom lys med polarisering i en bestemt retning. Figur 10.6 viser et eksempel på denne effekten.

10.3.2 Fresnels ligninger

For å komme fram til lovmessigheten mellom refleksjon og transmisjon brukte vi Maxwells ligninger, men lovmessigheten ble utledet lenge før Maxwell hadde systematisert elektromagnetiske fenomen i sine ligninger. Fresnel utledet ligningene allerede i første halvdel av 1800-tallet. Du kan lese mer om dette f.eks. på Wikipedia under stikkordet "Fresnel equations". Her skal vi bare gjengi resultatet i to formler og en graf. I ligningene (10.13) og (10.14), og i figur 10.7 er refleksjonskoeffisienten gitt for lys som er fullstendig polarisert vinkelrett på innfallsplanet (R_s) og fullstendig polarisert parallelt med innfallsplanet (R_p). [Indeksene s og p stammer fra tysk: *Senkrecht* (loddrett) og *parallel*, henholdsvis.] Refleksjonskoeffisienten refererer seg til intensiteter, så i vår språkbruk ville f.eks.

$$R_s = \left(\frac{E_{r,\perp}}{E_{i,\perp}} \right)^2$$

10.3. REFLEKSJON OG TRANSMISJON NÅR EN BØLGE KOMMER PÅ SKRÅ INN MOT GRENSESJIKTET

De fullstendige uttrykkene kan skrives slik:

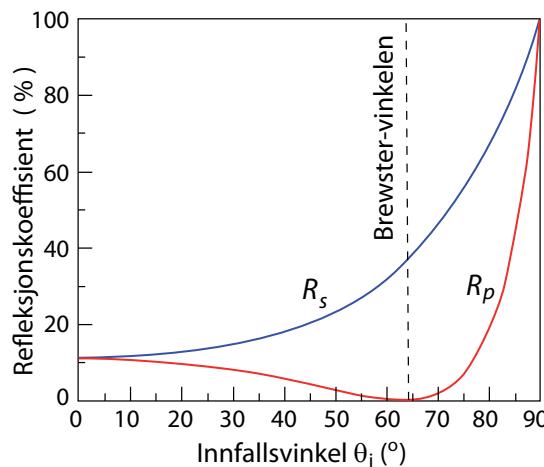
$$R_s = \left(\frac{n_1 \cos \theta_i - n_2 \sqrt{1 - (\frac{n_1}{n_2} \sin \theta_i)^2}}{n_1 \cos \theta_i + n_2 \sqrt{1 - (\frac{n_1}{n_2} \sin \theta_i)^2}} \right)^2 \quad (10.13)$$

og

$$R_p = \left(\frac{n_1 \sqrt{1 - (\frac{n_1}{n_2} \sin \theta_i)^2} - n_2 \cos \theta_i}{n_1 \sqrt{1 - (\frac{n_1}{n_2} \sin \theta_i)^2} + n_2 \cos \theta_i} \right)^2 \quad (10.14)$$

Transmisjonen er da gitt ved $T_s = 1 - R_s$ og $T_p = 1 - R_p$.

Dersom lyset som faller inn på flaten er totalt upolarisert (alle polariseringer forekommer), er total refleksjon gitt ved $R = (R_s + R_p)/2$.



Figur 10.7: Refleksjons- og transmisjonskoeffisient for elektromagnetiske bølger som sendes skrått inn mot en grenseflate mellom to medier med brytningsindeks $n_1 = 1.0$ og $n_2 = 2.0$. Indeksen s betyr at elektrisk felt-komponenten av bølgen er normalt på innfallsplanet, og indeksen p at komponenten er parallell med innfallsplanet. (Figuren er laget med basis i en figur fra Wikipedia under stikkordet "Fresnel equation".)

Figur 10.7 gir refleksjonen i prosent for ulike innfallsvinkler. Figuren gjelder for $n_1 = 1.0$ og $n_2 = 2.0$. For en bølge som da vender normalt inn mot grensesjiktet, er refleksjonen om lag 11 % og selvfølgelig uavhengig av polarisasjonsretning. Brewstervinkelen for disse brytningsindeksene er om lag 63° , og for denne vinkelen er refleksjonen om lag 36 % for bølger polarisert normalt på innfallsplanet.

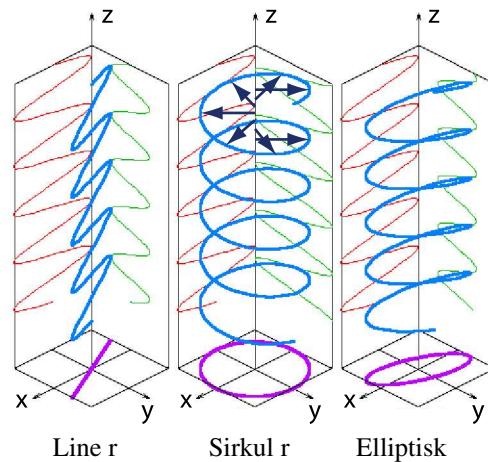
Merk forøvrig at refleksjonskoeffisienten går mot 1.0 (100 %) når innfallsvinkelen går mot 90 grader. Dette gjelder begge komponenter av det elektriske feltet.

10.4 Polarisasjon

Vi har allerede omtalt polarisasjon en god del i dette kapitlet, nemlig som retningen den elektriske feltvektoren har når en elektromagnetisk bølge beveger seg gjennom rommet.

Det er imidlertid ikke slik at polarisering alltid ligger i et bestemt *plan*. For en elektromagnetisk bølge kan elektrisk felt gjerne endre retning på en systematisk måte ettersom bølgen beveger seg. Tegner vi inn elektrisk feltvektor i ethvert punkt langs en linje som beskriver den retningen bølgen beveger seg, kan tuppen på de mange feltvektorene f.eks. beskrive en skrulinje (helix) med én omdreining per bølgelengde. Vi sier i så fall at bølgen er sirkulært polarisert.

Figur 10.8 viser tre ulike varianter for polarisering der elliptisk polarisering er en mellomting mellom lineær polarisering (polarisering i et plan) og sirkulær polarisering.



Figur 10.8: Tre ulike polariseringer av en plan elektromagnetisk bølge. De røde kurvene markerer tuppen på elektrisk feltvektor tegnet ut fra alle punkter langs en linje med retning lik bølgens bevegelsesretning. Noen få eksempler på slike vektorer er vist øverst i midtre del av figuren. Utgangspunktet for figuren er hentet fra Wikipedia under stikkordet "polarization" 12. april 2009, men er noe modifisert.

Det kan virke som om lineær polarisering er svært forskjellig fra sirkulær, men faktum er at det er ganske enkelt å skifte fra den ene til den andre. Ta utgangspunkt i en plan lineær polarisert elektromagnetisk bølge som beveger seg i z-retning. Polarisingen ligger i et plan mellom xz -planet og yz -planet (lignende orientering som venstre del av figur 10.8). Vi kan da si at $E_x(t)$ og $E_y(t)$ varierer i takt, eller "i fase" sagt med andre ord.

Matematisk kunne vi beskrive bølgen til venstre i figur 10.8 slik:

$$\vec{E} = E_x \cos(kz - \omega t) \vec{i} + E_y \cos(kz - \omega t) \vec{j}$$

hvor $E_x > E_y$.

Dersom vi kan forsinke $E_x(t)$ med en kvart periodetid i forhold til $E_y(t)$, og amplitudene like store, er polarisingen sirkulær (lignende som midtre del av figur 10.8), og polarisingen følger en skrulinje som på en vanlig skrue. Vi sier da at vi har en høyredreid sirkulær polarisering fordi polarisingsskrulinjen følger fingrene på høyre hånd når vi griper om aksen som angir bølgens bevegelsesretning med tommelen i denne retningen.

Dersom vi derimot fremskynder $E_x(t)$ med en kvart periodetid i forhold til $E_y(t)$, er polariseringen venstredreid sirkulær (da blir det akkurat som i midtre del av figur 10.8).

Matematisk kan vi beskrive en venstredreid sirkulært polarisert bølge (som midt i figur 10.8) slik:

$$\vec{E} = E_x \cos(kz - \omega t) \vec{i} + E_y \sin(kz - \omega t) \vec{j}$$

hvor $E_x = E_y$. Det elektriske feltet i x-retning er som vi ser forskjøvet en kvart periodetid (eller en kvart bølgelengde) i forhold til det elektriske feltet i y-retningen.

Polarisering til en plan elektromagnetisk bølge kan angis enten med to planpolariserte bølger med polarisering normalt på hverandre som basisvektorer, eller med en høredreid og en venstredreid sirkulært polarisert bølge som basisvektorer.

Vær *helt sikker* på at du skjønner hva som menes med en “plan, elektromagnetisk bølge med (f.eks. høyredreid) sirkulær polarisering”.

10.4.1 Dobbeltbrytning

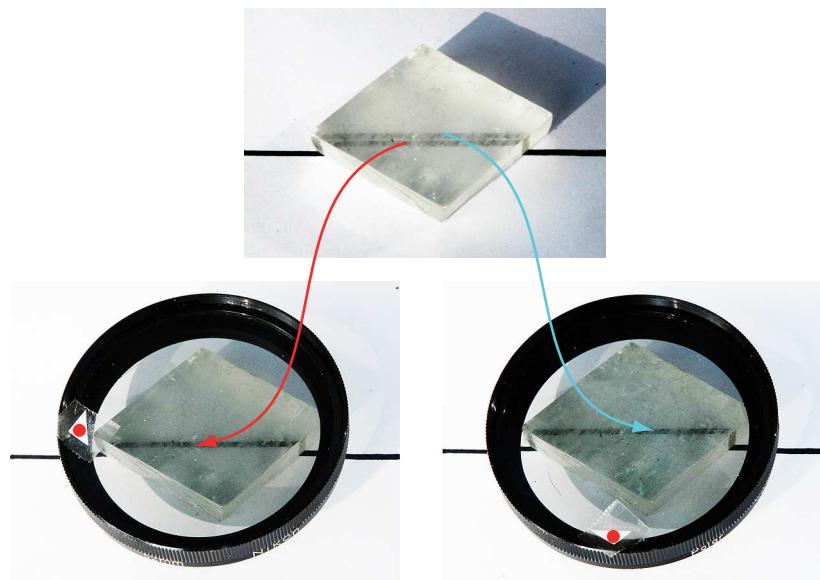
I forrige avsnitt hevdet vi at det er lett å endre fra lineær polarisering til sirkulær eller motsatt. Det eneste som må til er å endre fasen i tidsvariasjonen til én komponent av det elektriske feltet sammenlignet med den andre. Men hvordan skal vi i praksis oppnå en slik endring i fase? Endring i fase svarer til en tidsforsinkelse, og en forsinkelse kan oppnås dersom bølgen vandrer saktere når elektrisk feltvektor har én retning i rommet sammenlignet med om feltvektoren har en retning vinkelrett på den første.

Det finnes materialer som har den egenskapen at bølger med polarisering i én retning har en annen hastighet enn bølger med polarisering vinkelrett på den første. Det betyr at brytningsindeksen er forskjellig for de to polariseringene. Slike stoffer kalles dobbeltbrytende.

Et glass kan ikke være dobbeltbrytende fordi et glass er en uordnet materie der bindinger mellom atomer har alle mulige retninger i rommet. For å få et dobbeltbrytende materiale, må det være en systematisk forskjell mellom én retning og en annen retning, og denne forskjellen må være konstant innenfor makroskopiske deler av materialet (helst en hel bit av materialet). Et dobbeltbrytende materiale er derfor oftest en krystall. Kalsitt-krystaller er et velkjent dobbeltbrytende materiale. Ved 590 nm er brytningsindeksen 1.658 for lys med polarisering orientert i såkalt “ordinær” retning, og 1.486 for lys med polarisering i såkalt “ekstraordinær” retning i kalsitt-krystallen.

Det er morsomt å vite at dobbeltbrytning først ble beskrevet av den danske vitenskapsmannen Rasmus Barholin i 1669.

Det går an å lage en tynn skive av en kalsitt-krystall som har akkurat den tykkelsen som skal til for å forsinke tidsvariasjonen med en kvart periodetid i én komponent av elektrisk feltvektor sammenlignet med komponenten vinkelrett på. En slik skive kalles en “kvart-bølge-plate”. En kvart-bølge-plate vil sørge for at lineært polarisert lys blir transformert til sirkulært polarisert eller omvendt. En kvart-bølge-plate vil bare fungere optimalt for et relativt snevert bølgelengdeområde slik at dersom vi skal kjøpe en slik plate, må bølgelengden den skal brukes ved spesifiseres.



Figur 10.9: Øvre del av figuren viser en rett linje betraktet gjennom et dobbeltbrytende stoff (orientert på velvalgt måte). Vi ser to linjer! Disse skyldes at lys med ulik polarisering har ulik brytningsindeks gjennom krystallen. Dette kan demonstreres ved å holde et lineært polarisasjonsfilter foran krystallen. Orienterer vi polarisasjonsfilteret på én måte, ser vi bare én av de to linjene, men dreier vi polarisasjonsfilteret 90 grader, ser vi bare den andre linjen. En markering er satt på filteret for å vise dreiningen som er foretatt mellom de to nederste bildene.

To ulike brytningsindekser i ett og samme materiale gir opphav til et artig fenomen. Øverste del av figur 10.9 viser hvordan en rett linje ser ut når vi ser den gjennom en kalsitt-krystall orientert på en spesiell måte. Orienteringen er slik at vi ser *to* linjer i stedet for én. Det er lett å forstå ordet "dobbeltbrytende materiale" når vi ser en slik splitting av et bilde.

Vi kan tenke oss at lyset fra linjen (området rundt) har alle mulige lineære polarisasjonsretninger. Lys med en bestemt polarisering går med forskjellig hastighet sammenlignet med lys med polarisering vinkelrett på den første. Det vil si at lys med disse to polariseringene har forskjellig brytningsindeks, - det er derfor vi ser *to* linjer gjennom krystallen.

De siste to bildene i figuren viser hvordan linjen ser ut når vi holder et polarisasjonsfilter mellom krystallen og oss. For en bestemt orientering av filteret slipper vi gjennom bare lys med én polariseringsretning. Ved å rotere filteret i én retning, ser vi bare den ene linjen gjennom krystallen. Dreier vi filteret 90 grader, ser vi bare den andre linjen gjennom krystallen. Dette er en fin indikasjon på at de to brytningsindeksene er knyttet til polariseringen til lyset gjennom krystallen.

[♠ ⇒ Kommentar:

Hittil har vi angitt sammenhengen mellom elektrisk feltstyrke \vec{E} og elektrisk fluksstetthet (eller forskyvningsvektor) \vec{D} slik:

$$\vec{D} = \epsilon_0 \epsilon_r \vec{E}$$

hvor ϵ_0 er permittiviteten i det tomme rom og ϵ_r er den relative permittiviteten (også kalt dielektrisk konstanten). Begge disse størrelsene har vært enkle skalarer, og derfor har \vec{D} og \vec{E} vært parallelle vektorer.

Komponentvis kan ligningen skrives som:

$$D_i = \epsilon_0 \epsilon_r E_i \quad (10.15)$$

hvor i f.eks. kan være x, y eller z.

Ved dobbeltbrytning duger ikke denne enkle beskrivelsen lenger. Elektrisk felt rettet i én retning, vil kunne gi en polarisering av et materiale (f.eks. kalsitt) også i en annen retning. For å innlemme denne oppførselen i den matematiske formalismen, må skalaren ϵ_r erstattes med en tensor med elementer $\epsilon_{r,i,j}$ hvor i og j svarer til x, y og z. Da blir ligning (10.15) erstattet med:

$$D_j = \epsilon_0 \epsilon_{r,i,j} E_i \quad (10.16)$$

Dette er bare ett eksempel på hvordan en enkel beskrivelse må kompletteres med flere detaljer når et fysisk system framviser egenskaper ut over de aller mest elementære.

Vi nevner disse detaljene for å minne om at en av fysikkens oppgaver er å gi en matematisk modellering av prosessene vi observerer. Når prosessene i naturen er kompliserte, trengs tilsvarende komplisert matematisk formalisme.

For en del år siden var det viktig å bygge inn komplisert matematikk, men i dag kan dette gjøres langt enklere såfremt vi bruker numeriske metoder i beregningene. .← ♠]

10.4.2 Lysets vekselvirkning med materie

Alt lys har opphav i en eller annen prosess hvor materie er involvert. Da lyset ble skapt, fikk det en polarisering ut fra de geometriske føringene som lå i det lokale området hvor lyset ble til. Når lys går gjennom vakuum, endres ikke polariseringen, men så snart lyset vekselvirker med materie igjen, kan polariseringen endres. Det er mange ulike mekanismer som påvirker polariseringen til lyset. Det betyr at vi ved å studere endring av polarisering når lys passerer materie, kan oppnå mer kunnskap om det materialet vi betrakter. Et samlenavn på alle slike studier er "polarimetri".

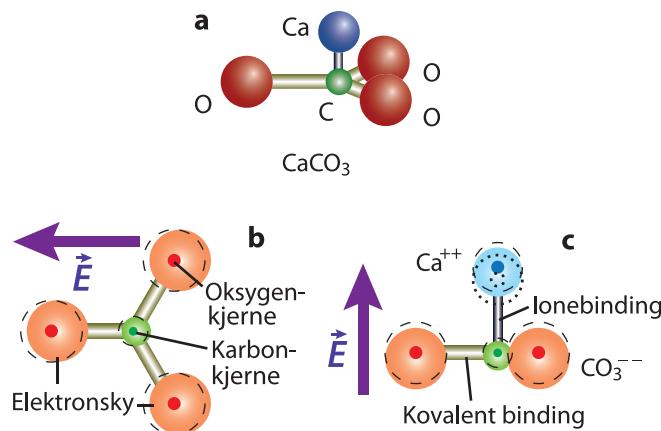
La oss diskutere hva som skjer når lys sendes gjennom et stykke kalsitt for å få en idé om aktuelle virkningsmekanismer. Kalsitt har kjemisk formel CaCO_3 , og vi vil betrakte kalsittkrystaller. Disse er "dobbeltbrytende", hvilket betyr at brytningsindeksen er forskjellig alt etter hvilken retning lysets polarisering har i forhold til krystallaksen. Enhetscellen i en kalsittkrystall er relativt komplisert.³ I hver CaCO_3 -gruppe ligger alle atomene i CO_3^{--} -delen omrent i samme plan, og alle CO_3^{--} planene har samme retning i krystallen.

I figur 10.10 er det vist en perspektivisk skisse av CaCO_3 . Det er kovalente bindinger mellom karbon og oksygenatomene, mens bindingen mellom kalsiumatomet og resten har en betydelig karakter av ionebinding. En ionebinding er oftest mye mer tøyelig enn en kovalent binding.

Nederst til venstre i figuren har vi indikert hva som skjer når lys passerer krystallen dersom polariseringen er slik at det elektriske feltet i lysbølgen har en retning parallelt med CO_3^{--} -planet. Når det elektriske feltet er rettet som vist, vil elektronskyene rundt hver av atomkjernene få en ørliten forskyvning i forhold til kjernen. Følgen er at hvert atom får en indusert polarisering. Energi stjeles fra det elektromagnetiske feltet i lyset og "lagres" midlertidig i polariseringen i krystallen. Når så det elektriske feltet i løpet av en periodetid går mot null og så øker igjen med motsatt retning av den som er vist på figuren, vil det induseres polarisering av krystallen på ny, men nå med motsatt forflytninger av elektronskyene relativt til atomkjernene.

Vi bygger imidlertid ikke opp mer og mer polarisering etter som tiden går. Den lagrede energien i polariseringen av materialet vil på en måte virke som "antenner" og ge-

³Se f.eks. Wikipedia (engelsk utgave) med oppslagsord "calcite".



Figur 10.10: Kalsitt er bygget opp av atomgruppene CaCO_3 . En perspektivisk tegning er gitt i a. I b og c er det vist et øyeblikksbilde på hvordan et ytre elektrisk felt fra lys som passerer forbi vil polarisere atomene. I b er feltet rettet i CO_3^{2-} planets retning (figuren viser CaCO_3 "ovenfra", og Ca-atomet er ikke tatt med). Stiplede sirkler indikerer elektronskyenes plassering når det ikke er et ytre elektrisk felt til stede. I c ser vi inn langs CO_3^{2-} planet når det elektriske feltet i lyset er rettet vinkelrett på dette planet. I dette tilfellet får vi en endring i lengden på ionebindingen i tillegg til polariseringene av hvert enkelt atom.

nerere elektromagnetiske bølger. Disse bølgene har samme frekvens som de som skapte polariseringen opprinnelig. Det er denne polariseringen av materialet og reemittering av elektromagnetiske bølger fra de små induserte dipolene i materialet som fører til at lys går med lavere hastighet i krystallen sammenlignet med vakuum. Så snart bølgen går ut av krystallen, er det ikke noe materie å polarisere (når vi ser bort fra luft), og lyshastigheten blir selvfølgelig den samme som i vakuum.

Så kommer det spennende inn! Dersom vi sender lys inn i kalsittkrystallen slik at det elektriske feltet i lysbølgen har en retning vinkelrett på CO_3^{2-} -planene, vil vi akkurat som tidligere få forskyvninger av elektronskyene relativt til atomkjernene. Men i tillegg får vi en endring av bindingsavstanden mellom CO_3^{2-} -planet og Ca^{++} siden en ionebinding er mindre stiv enn en kovalent binding. Det betyr at vi får en større grad av polarisering av materialet ved denne orienteringen av det elektriske feltet enn i stad. Følgen er at brytningsindeksen for polariseringen av lyset vist i nedre høyre del av figuren blir større enn brytningsindeksen for en polarisering vinkelrett på denne.

Det ligger i sakens natur at effekter lignende den vi hadde i kalsitt, får vi bare ved krystallinske materialer, eller i det minste materialer med forskjellige egenskaper i en retning sammenlignet med en annen (anisotrop). Vi kan imidlertid få tilsvarende effekter også for et i utgangspunktet isotrop materiale dersom det har blitt utsatt for stress i en bestemt retning slik at det ikke lenger er isotrop. Et isotrop plastmateriale kan gjøres svakt anisotrop ved f.eks. å bøye det eller strekke det. Forøvrig er ofte enkelte typer plast litt anisotrope allerede i utgangspunktet dersom de er laget ved støping der molekylene har fått en viss ensretting lokalt idet plasten ble presset inn i formen fra et bestemt matingspunkt.

UTFORDRING

I dette kapitlet har vi brukt ordet "polarisering" om to vidt forskjellige forhold. Vi brukte ordet da vi omtalte ulike elektriske permittiviteter (som har med forskjell mellom elektrisk felt \vec{E} og elektrisk feltstyrke \vec{D} å gjøre), og vi brukte ordet da vi skilte mellom

f.eks. lineær og sirkulær polarisering. Vær sikker på at du skjønner fullt ut forskjellen på disse to ulike (men likevel relaterte) begrepene med samme navn. I motsatt fall bør du diskutere med medstudenter og/eller gruppelærer/foreleser.

10.4.3 Polarisasjonsfiltre

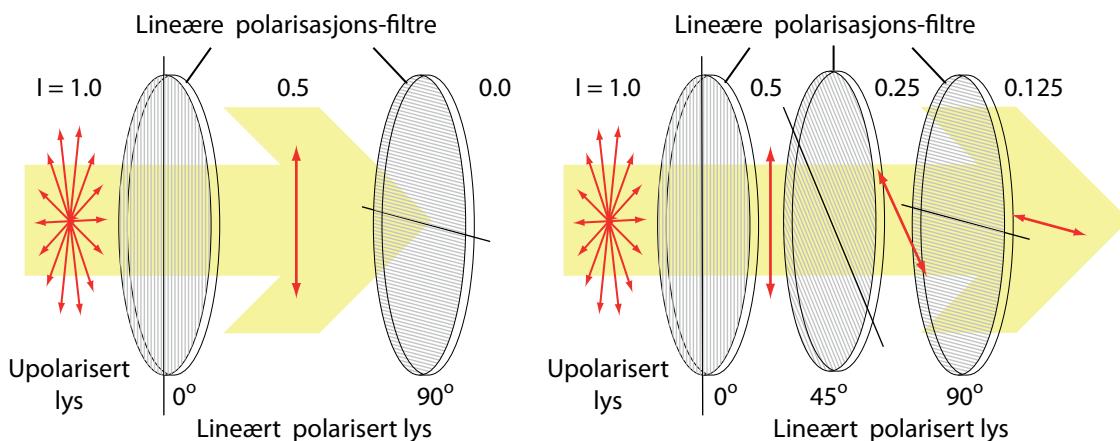
Lineære polarisasjonsfiltre

Da vi diskuterte Brewstervinkel-fenomenet, så vi et eksempel på et lineært polarisasjonsfilter. Grovt sett kan vi si at et slikt filter (dersom det er tykt nok) skræller av én komponent av den elektriske feltvektoren i de elektromagnetiske bølgene (synlig lys). Dersom lyset var fullstendig upolarisert i utgangspunktet, vil intensiteten avta til det halve når lyset passerer et lineært polarisasjonsfilter.

Anta at vi har en horisontal lysstråle. Ved hjelp av ett lineært filter kan vi da sørge for at alt lys som passerer har elektrisk felt som er rettet horisontalt.

Dersom vi setter inn enda et slikt filter, og orienterer det akkurat som det forrige, vil alt lys som har passert filter 1 også passere filter 2.

Dersom filter 2 dreies 90 grader slik at det bare kan slippe gjennom lys med vertikal polarisering, finnes det ikke noe slikt lys etter filter 1. Da vil *ikke noe* lys passere filter 2 (venstre del av figur 10.11).



Figur 10.11: *To etterfølgende lineære polarisasjonsfiltre rettet 90 grader på hverandre, slipper ikke gjennom noe lys (venstre del). Plasseres et tredje filter mellom de to første, med polarisasjonsretning mellom de to andre, vil det likevel slippe lys gjennom filtrene (høyre del).*

Dersom vi derimot f.eks. dreier filter 2 45 grader relativt til filter 1, vil lys med horisontal polarisering etter filter 1 faktisk ha en komponent også i retningen filter 2 er rettet. Lys som nå passerer filter 2, får en polarisering 45 grader i forhold til polariseringen det hadde etter filter 1. Vi *endrer* med andre ord polariseringen, men amplituden til det elektriske feltet er nå mindre enn hva det var før filter 2 (bare E -feltskomponenten i filter 2's retning slippes gjennom).

Intensiteten til lyset som går gjennom filter 2 er gitt ved Malus' lov:

$$I = I_0 \cos^2(\theta_2 - \theta_1) \quad (10.17)$$

Her er I_0 intensiteten til lyset etter at det har passert filter 1. Argumentet for cosinusfunksjonen er forskjellen i dreievinkel mellom filter 1 og 2.

La oss nå starte med to polarisasjonsfiltre med polarisasjonsaksen vinkelrett på hverandre, og plasserer et tredje polarisasjonsfilter mellom de to første. Velger vi en annen orientering enn 90 grader i forhold til det første, får vi lys gjennom alle tre filtrene (høyre del av figur 10.11). Dette skyldes at det midtre filteret har endret polarisasjonen før det treffer siste filteret.

Det er viktig å merke seg at polarisasjonsfiltre av denne typen faktisk har en aktiv rolle idet det *endrer* polarisasjonen til lys som slipper gjennom.

[♠ ⇒ Kommentar:

Vi vil nå presentere et bilde som kan være nyttig analogi til hva som skjer i et lineært polarisasjonsfilter: Tenk deg at filteret består av pendler som bare kan svinge i ett plan. Dersom pendlene blir forsøkt dyttet på i den retningen de faktisk kan svinge, vil pendlene svinge. En svingende pendel kan forplante sin bevegelse til en nabopendel av samme type osv og slik kan en bølge forplante seg gjennom materialet.

Dersom vi derimot forsøker å dytte på pendlene i en retning de ikke *kan* svinge, blir det ingen svingninger. Da kan bølgen ikke forplante seg gjennom mediet. Dersom vi dytter på skrå, vil pendlene kunne svinge, men bare i den retningen de faktisk kan svinge. Det betyr at svingeretningen i bølgen vil endre seg når bølgen forplanter seg gjennom mediet, men vi får en reduksjon i bølgen fordi bare den komponenten av vår dytting som er langs pendlenes svingeplanet vil bli utnyttet i svingningene. .← ♠]

Sirkulære polarisasjonsfiltre

Et sirkulært polarisasjonsfilter er i utgangspunktet et filter som bare slipper gjennom sirkulært polarisert lys. Det er to varianter av slike filtre, én type som slipper gjennom høyredreid-sirkulært polarisert lys, og en annen type som slipper gjennom venstredreid sirkulært polarisert lys.

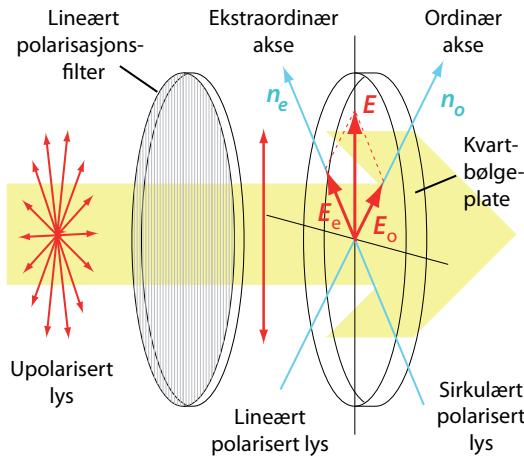
I dag er det imidlertid en helt annen type filter som går under navnet sirkulært polarisasjonsfilter. Vi tenker da på polarisasjonsfiltre som brukes i fotografien. I mange fotoapparater bygger autofokus på sirkulært polarisert lys. Dersom vi vil ha et polarisasjonsfilter foran objektivet, må filteret lages slik at sirkulært polarisert lys når en detektor inne i apparatet.

Et slike sirkulært polarisasjonsfilter er satt sammen på en ganske spesiell måte. Når lyset går inn i filteret møter det først et ordinært lineært polarisasjonsfilter. Like bak dette filteret er det satt inn en såkalt kvartbølgplate med en spesiell orientering. Resultatet er at lyset først blir omdannet til rent lineært polariser lys, dernest omgjort til nær fullstendig sirkulært polarisert lys. Lyset som kommer inn i fotoapparatet er derfor sirkulært polarisert, og autofokusen fungerer.

Vi skal se litt nærmere på detaljer i denne sammenheng.

En kvartbølgplate er laget av et dobbeltbrytende stoff, f.eks. kalsitt. Vi har allerede sett at i et dobbeltbrytende stoff er fasehastigheten for lys med en viss orientering forskjellig fra fasehastigheten for lys med polarisering vinkelrett på den føstnevnte orienteringen.

I foto-polarisasjonsfiltrene er orienteringen til det dobbeltbrytende stoffet valgt slik at elektrisk vektor etter det lineære polarisasjonsfilteret danner 45 grader med hver av de to



Figur 10.12: Prinsippskisse for et sirkulært polarisasjonsfilter brukt i fotografi. Lyset går først gjennom et ordinært lineært polarisasjonsfilter og dernest gjennom en kvartbølgeplate. De to delene ligger i virkeligheten tett inntil hverandre. Orienteringen av halvbølgplatene er valgt slik at midlere bølgelengder omdannes fra lineært til sirkulært polarisert lys.

spesielle retningene i det dobbeltbrytende stoffet. Vi dekomponerer elektrisk vektor som vist i figur 10.12. Komponenten E_o vil gå gjennom stoffet med en viss fasehastighet (dvs en viss bølgelengde), mens komponenten E_e går gjennom stoffet med en annen fasehastighet (og bølgelengde).

Ved å velge en bestemt tykkelse på det dobbeltbrytende stoffet, kan vi oppnå at E_o har akkurat en kvart bølgelengdes forskjell fra det E_e har når det forlater filteret. I så fall oppnår vi akkurat det vi ønsker, nemlig at lineært polarisert lys er omdannet til sirkulært polarisert lys.

Ser vi nøyere på denne argumentasjonen, oppdager vi at vi ikke kan få en perfekt transformasjon fra lineært til sirkulært lys for alle bølgelengder i det synlige spekteret samtidig. I praksis vil derfor tykkelsen på det dobbeltbrytende stoffet velges slik at de midtre bølgelengder (omlag spektral grønt) får en optimal omforming, mens andre bølgelengder får en mindre perfekt transformasjon. Det spiller liten rolle, for autofokus må ha noe lys som er sirkulært polarisert, og behøver ikke ha perfekt sirkulær polarisering for alle bølgelengder.

[♠ ⇒ Kommentar:

En lineært polarisert bølge kan betraktes som en sum av en høyredreid og en venstredreid sirkulært polarisert bølge, og en sirkulært polarisert bølge kan betraktes som en sum av to lineært polariserte bølger med polarisering vinkelrett på hverandre (og faseforskjøvet). Det betyr at vi kan mikse sirkulære polarisasjonsfiltre og lineære polarisasjonsfiltre på ulikt vis.

Filterkombinasjoner hvor fotografiens sirkulære polarisasjonsfiltre inngår, gir imidlertid en del overraskelser nettopp fordi disse filtrene er sammensatt av to elementer.

Plasserer vi to foto-sirk-pola-filtre med innerflatene mot hverandre, vil lyset gå gjennom begge filtre med omtrent samme intensitet omtrent som etter første filter. Intensiteten er tilnærmet uavhengig av hvilken dreievinkel det ene filteret har i forhold til det andre. Dette skyldes jo at lyset etter at det har passert første filter er tilnærmet sirkulært polarisert, og har derved en sirkulær symmetrisk E -feltfordeling (når vi betrakter intensitet).

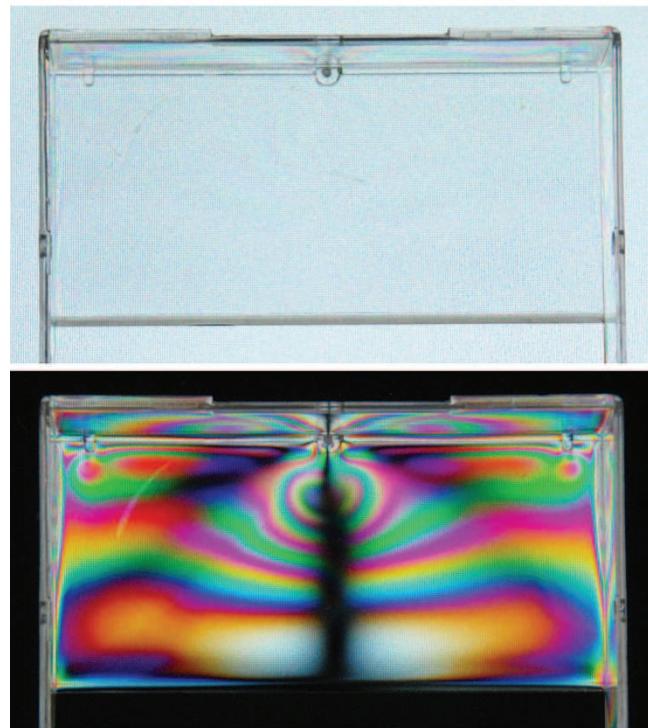
Plasserer vi derimot to foto-sirk-pola-filtre med ytterflatene mot hverandre, vil to lineært polariseringsfiltre følge like etter hverandre i lysveien. Filterparet vil da oppføre seg omtrent som to vanlige lineært

polarisjonsfiltre, og intensiteten som slipper gjennom er gitt ved Malus' lov (ligning (10.17)).

Den spesielle konstruksjonen av foto-sirk-pola-filtre gjør at vi i fotografien oppnår samme effekt som ved lineære polarisasjonsfiltre fotografisk sett. Pola-filte blir brukt for å fjerne reflekser (som vist i figur 10.6) og bl.a. å fjerne virkningen av dis i atmosfæren (siden lys fra dis er delvis lineært polarisert). Vi kan ved et hjelp av polafiltre få fram en flott kontrast mellom blå himmel og hvite skyer, noe som gir ekstra liv i bildene. .=< ♠]

10.4.4 Polariometri

Vi har tidligere sett at et polarisasjonsfilter stukket inn mellom to kryssepolareerde polarisasjonsfiltre fører til at lys slipper gjennom kombinasjonen av tre filtre. Dette gir oss et utmerket utgangspunkt for å studere enkelte materialegenskaper. Ethvert materiale som endrer polarisasjonen til lyset, vil sørge for at noe lys slipper gjennom kryssepolareerde filtre med dette materialet mellom. Eksempelvis vil mange plastgjenstander ha forskjeller i optiske egenskaper ulike steder i gjenstanden alt etter hvordan plastmaterialet strømmet inn i en støpeform før og under herding. Anisotropi ulike steder i materialet fører til at polarisasjonsretningen til lys vil dreie seg litt eller at vi får en viss overgang mellom planpolarisert og sirkulært polarisert bølge. Effekten er ofte avhengig av bølgelengden. Resultatet er at vi kan få nydelige fargebilder gjennom de kryssepolareerde polarisasjonsfiltrene dersom vi sender hvitt lys inn mot oppsettet.



Figur 10.13: Fotografi av en plastikkboks for små videokassetter i polarisert lys (øverst), og i polarisert lys pluss et lineært polarisasjonsfilter i såkalt krysset orientering (nederst). Lys slipper bare gjennom dersom polariseringen på lyset endres når det passerer plasten.

Figur 10.13 viser hvordan bildet ser ut av en plastikkboks for små videokassetter i krysset konfigurasjon. Jeg kunne ha brukt hvitt lys (f.eks. fra sola eller fra en glødelampe), to kryssepolareerde polarisasjonsfiltre og denne plastikkboksen mellom filtrene. Men siden filtene jeg hadde tilgjengelig ikke var like store som boksen, valgte jeg i stedet å bruke en dataskjerm som lyskilde for planpolarisert lys. Dataskjermer, mobiltelefonskjermer

og en del andre display som bygger på flytende krystaller teknologi, gir nemlig fra seg planpolarisert lys. Plastboksen plasserte jeg da direkte inn mot dataskjermen og jeg kunne nøye meg med ett polarisasjonsfilter like foran kameraobjektivet.

Som det går fram av figur 10.13 kommer anisotropier i plasten godt fram ved polarimetri. Varianter av denne metoden brukes for mange ulike materialer og i mange ulike sammenhenger i industri og forskning. Man kan kjøpe spesialisert utstyr for denne type analyser.

10.5 Polarisasjon i astronomien

I de senere år er det gjennomført flere studier av polariseringen av lys fra Sola og fra fjerne lyskilder i universet. Riktignok er det ikke akkurat det astronomer først tenker på. Vanligvis er utfordringen å samle nok lys for å få gode bilder eller spektroskopiske data. Dersom vi setter inn et polarisasjonsfilter, mister vi halve lysintensiteten. Og dersom vi vil ha informasjon om polarisasjonen til lyset, trenger vi gjerne minst to fotografier med polarisering vinkelrett på hverandre. Det betyr at studiet ved en rett fram prosedyre vil ta minst fire ganger så lang tid som ett bilde uten å bry seg om polarisering.

Grunnen til at polarisering likevel er blitt interessant i astronomien er omtrent den samme som for polarimetri av ulike materialer. La oss for eksempel betrakte lys fra Sola. Lyset kan gjerne bli sendt ut som upolarisert lys i prosesser vi kjenner som "sort ståling" (stråling fra et varmt legeme). Lyset vil imidlertid vekselvirke med plasma og atomer på veien til oss. Dersom elektronene i et plasma blir påvirket av et sterkt tilnærmet statisk magnetfelt, vil bevegelsen til elektronene ikke foregå like lett i alle retninger (husk kryssproduktet i uttrykket for Lorentz-kraften).

Når så lyset fra f.eks. en del av Sola passerer elektronene i et plasma, vil det elektriske feltet i vår elektromagnetiske bølge sette elektronene i plasmaet i bevegelse. Det relativt sett statiske magnetfeltet fra solar aktivitet som finnes i regioner på soloverflaten, vil da føre til at elektronbevegelsen som følge av det elektromagnetiske feltet (lyset) ikke foregår like lett i alle retninger på tvers av lysets retning.

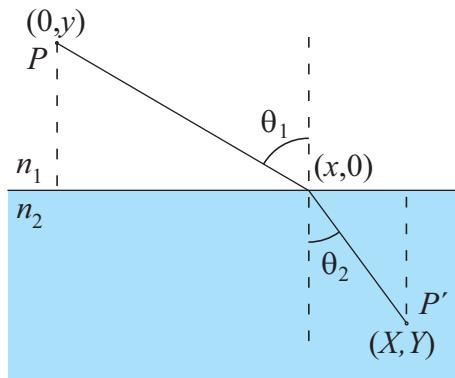
Resultatet er at vi får generert en polarisering av lyset som sendes videre mot oss. Polarisingens retning vil fortelle oss noe om det "kvasistatiske" magnetfeltets størrelse og retning i den delen av sola lyset kom fra.

Det er også en rekke andre faktorer som kan påvirke polariseringen av lys fra astronomiske objekter, så her er det mye å ta fatt i! Polarimetrien vil utvilsomt i årene som kommer kunne gi oss opplysninger om astronomiske prosesser som inntil nylig var utilgjengelig på annet vis. En kort populærvitenskapelig artikkel om emnet kan du finne på www.lbl.gov/today/2006/Jul/18-Tue/AstronomyGetsPolarized.pdf (tilgjengelig 10. mars 2013).

10.6 Snels brytningslov

Snels brytningslov gir oss sammenhengen mellom en lysstråles retning inn mot et grensesjikt mellom to materialer og strålens retning etter grensesjiktet.

Brytningsloven kan utledes på flere måter. Vi vil her velge å bruke "Fermats prinsipp" som også kalles *prinsippet om minste tid*. Fermats prinsipp uttrykkes i vår moderne tid ved å si at *optisk veilengde må være stasjonær*. Litt upresist betyr dette at for den veien lyset transporterer energi ("der lyset faktisk går") er optisk veilengden samme (i første tilnærming) for en mengde optiske veier som ligger nær hverandre. Det betyr at optisk veilengde må være maksimal, minimal eller ha et sadelpunkt for små variasjoner i veivalg. Når vi skal utlede Snels brytningslov, bruker vi minimumspunktet.



Figur 10.14: Ved utledning av Snels brytningslov bruker vi de koordinatene som er gitt i denne figuren. Se forøvrig teksten.

Vi viser til figur 10.14. En lysstråle sendes fra punktet P i et medium med brytningsindeks n_1 til P' i medium med n_2 . Vi antar i figuren at $n_2 > n_1$. Lyset går raskere i medium 1 enn i medium 2, og kortest tid vil lyset bruke ved å gå litt lenger i medium 1 enn medium 2 i forhold til den rette linjen. Bruker vi symbolene i figuren, følger at medgått tid er:

$$t = \frac{\sqrt{x^2 + y^2}}{c_0/n_1} + \frac{\sqrt{(X-x)^2 + Y^2}}{c_0/n_2}$$

$$t = \frac{1}{c_0} \left(n_1 \sqrt{x^2 + y^2} + n_2 \sqrt{(X-x)^2 + Y^2} \right)$$

Variabelen er x og minimum tid kan vi finne ved å sette $\frac{dt}{dx} = 0$:

$$\frac{dt}{dx} = \frac{1}{c_0} \left(n_1 \frac{\frac{1}{2} \cdot 2x}{\sqrt{x^2 + y^2}} + n_2 \frac{\frac{1}{2}(X-x) \cdot 2 \cdot (-1)}{\sqrt{(X-x)^2 + Y^2}} \right) = 0$$

$$n_1 x \sqrt{(X-x)^2 + Y^2} - n_2 (X-x) \sqrt{x^2 + y^2} = 0$$

$$\frac{n_1}{n_2} = \frac{(X-x) \sqrt{x^2 + y^2}}{x \sqrt{(X-x)^2 + Y^2}} = \frac{\frac{X-x}{\sqrt{(X-x)^2 + Y^2}}}{\frac{x}{\sqrt{x^2 + y^2}}}$$

Vi ender da opp med Snels brytningslov:

$$\frac{n_1}{n_2} = \frac{\sin \theta_2}{\sin \theta_1}$$

eller

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (10.18)$$

Fermats prinsipp har klare koblinger til Huygens prinsipp og også tankegangen bak kvante-elektrodynamikk (CED). Bølgene følger alle mulige veier, men i enkelte retninger vil bølgene virke forsterkende på hverandre, andre steder vil de ødelegge hverandre. Det er med andre ord interferens som spørker i bakgrunnen, og helt essensielt for tankegangen som ligger bak er at vi må holde orden på relativ fase for de ulike bidragene for å komme riktig ut. Ved “minimum tid” oppnår vi dette automatisk siden minimum tid betyr at mange bølger, som vi kan tenke oss er blitt sendt ut fra P , vil ha nær minimum tid, og alle disse bølgene vil da automatisk ha samme fase og virke sammen med konstruktiv interferens.

10.6.1 Totalrefleksjon

Totalrefleksjon er selvfølgelig en viktig effekt, men vi kommer ikke til å bruke mye tid på fenomenet i dette kapitlet siden stoffet antas å være godt kjent fra før. Poenget er at dersom lys går fra et medium med brytningsindeks n_1 til et medium med indeks n_2 og $n_1 > n_2$, vil “innfallsvinkel” være mindre enn “utfallsvinkel” for den transmetterte strålen. Vi kan først sende strålen normalt inn på grensesjiktet og så øke innfallsvinkel gradvis. Utfallsvinkelen vil da øke gradvis den med, men alltid være større enn innfallsvinkelen.

Før eller senere vil vi få en innfallsvinkel som fører til at utfallsvinkelen blir (nesten) 90 grader. Øker vi innfallsvinkelen ytterligere, kan vi ikke få tilfredsstilt Snels brytningslov for sinus til en vinkel kan aldri bli større enn 1.0.

Grensevinkelen, der utfallsvinkelen er 90 grader, er gitt ved å sette $\theta_t = 90^\circ$ i Snels brytningslov:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 = n_2 \sin 90^\circ = n_2$$

Innfallsvinkelen som svarer til at utfallsvinkelen er 90 grader, kalles iblant “kritisk vinkel”, og er gitt ved:

$$\sin \theta_1 = \frac{n_2}{n_1} \quad (10.19)$$

Øker vi innfallsvinkelen ut over denne grensevinkelen får vi ikke noe transmettert stråle. Alt vil bli reflektert fra grensesjiktet tilbake i det opprinnelige mediet igjen, noe vi kaller *totalrefleksjon*.

Er vi under vannet og kikker opp mot overflaten, vil kritisk vinkel være gitt ved:

$$\sin \theta_1 = \frac{1.00}{1.33}$$

$$\theta_1 = 48.8^\circ$$

Forsøker vi å se opp mot overflaten med en større vinkel enn dette (relativt til loddlinjen), vil vannoverflaten bare virke som et speil.

Totalrefleksjon benyttes i stor utstrekning i dagens samfunn. Signalkabler for internett og telefoni og nærmest all informasjonsoverføring skjer nå i stor grad via optiske fibre. For optiske fibre som har en diameter som er mange ganger bølgelengden (såkalte "multimode" fibre), er det greit å si at det er totalrefleksjon som gjør seg gjeldende.

En optisk fiber består av en tynn kjerne med superrent glass. Utenfor denne kjernen legges det et lag med glass som har nesten identisk brytningsindeks som kjernen, men likevel ørlite mindre enn kjernen. Forskjellen er godt under én prosent! Følgen er at kritisk vinkel blir meget nær 90 grader. Det betyr at bare lys som beveger seg meget nær parallelt med fiberaksen blir reflektert i overgangen mellom indre kjerne og neste lag glass utenfor. Det er viktig at bølgene er mest mulig parallelle med aksen for at pulser som sendes inn i fiberen skal beholde sin form over så mange kilometre som mulig før pulsene må gjenskapes av elektronikk før videresending.

I mange optiske fibre er diameteren på den indre glasskjernen bare noen ganske få ganger bølgelengden. Slike fibre kalles "single mode" fibre, og det er mest slike som brukes i telekommunikasjon og liknende. For single mode fibre er det egentlig misvisende å forklare bølgebildet i fiberen med totalrefleksjon. Vi må i stedet bruke Maxwell's ligninger direkte med den aktuelle geometrien. Bølgebildet inne i fiberen kan ikke lenger anses som en plan bølge slik vi finner den i vakuums langt fra kilden og fra forstyrrende randbetingelser. Randbetingelsene framtvinger en helt annen løsning. Vi kommer tilbake til dette når vi siden behandler bølgeledere.

Single mode fibre er utfordrende å arbeide med fordi tverrsnittet av fiberen er svært liten og lyset som skal inn i fiberen må ha en retning meget nær fiberens retning. Det er derfor vanskelig å få lys *inn* i fiberen uten for mye tap. Standardisering av koblingsenheter gjør at det likevel går greit i telekommunikasjonsutstyr, men når vi i en laboratoriesammenheng skal koble lys inn i en fiber fra en fri ståle i luft, er det en utfordring!

Det er atskillig lettere å få lys inn i multimode fibre på grunn av større tverrsnitt og at det ikke er like kritisk med lysretningen inn i fiberen. Multimode fibre egner seg imidlertid ikke så godt for langdistansekommunikasjon siden pulser "flyter ut" etter relativt korte overføringsavstander.

10.7 Flyktige bølger (Evanescent waves)

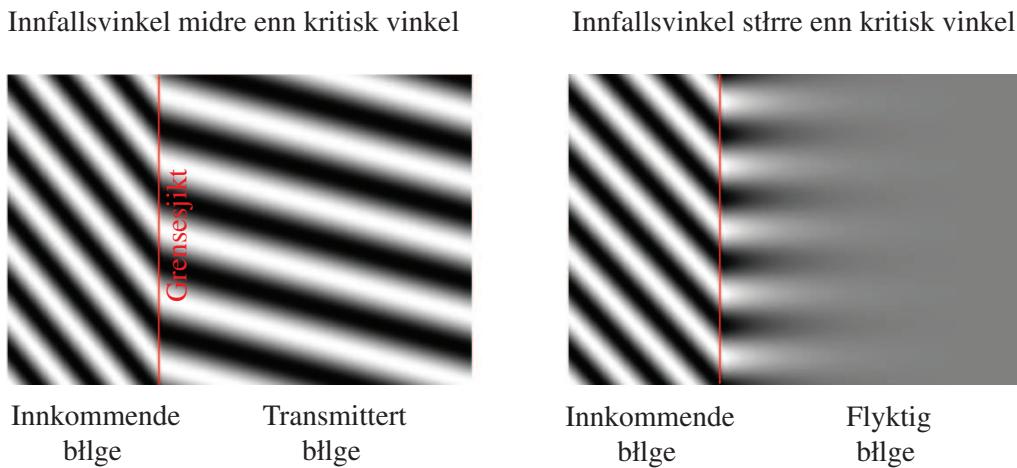
I forrige kapittel skilte vi mellom nærfelt og fjernfelt og pekte på at mange kjente relasjoner mellom elektrisk og magnetisk felt for elektromagnetiske bølger bare gjelder i fjernfeltet. Vi nevnte at nærfeltet strekker seg i størrelsesordenen noen få beregnede bølgelengder vekk fra kilden eller de strukturene som fører til at vi får nærfelt.

Denne erkjennelsen har det siste tiåret og vel så det, slått gjennom med stor tyngde innen f.eks. optikk. Da vi ovenfor utledet uttrykket for totalrefleksjon, gjorde vi det ut fra Snel's brytningslov, rett og slett ved å manipulere et matematiske uttrykk alene.

Dersom vi imidlertid anvender Maxwells ligninger i en grundigere analysere av totalrefleksjon, vil vi innse at *et elektrisk felt på innsiden av glasset ved totalrefleksjon ikke kan ende brått ved grenseflaten mellom glass og luft. Det elektriske feltet må avta gradvis.*

En detaljert løsning av Maxwells ligninger for regionen nær grenseflaten viser en slags stående bølge der amplituden (og intensiteten) avtar eksponensielt når vi fjerner oss fra grensesjiktet (se figur 10.15). Denne stående bølgen kalles på engelsk for en “evanescent wave”, noe som betyr en bølge som liksom forsvinner raskt når vi fjerner oss fra grensesjiktet; en bølge som “fordamper” som dugg for solen når vi går bort fra grensesjiktet.

Det er ikke ennå etablert et norsk navn på fenomenet, og fornorskingen jeg forsøker meg på her er ikke helt god siden “flyktig” ofte betyr noe som endrer seg i tid, mens det her er snakk om bølger som forsvinner raskt når vi fjerner oss fra grensesjiktet. Et annet mulig navn på fenomenet kunne være “usynligsbølge” eller “forsvinningsbølge”, fordi bølgen avtar så raskt fra grensesjiktet at den er vanskelig å få øye på.



Figur 10.15: Flyktige bølger når vi har totalrefleksjon (til høyre). Grenselinjen mellom to medier med ulik brytningsindeks er markert med rød strek. Bare innkommende og transmittert/flyktig bølge er tegnet inn, det vil si at reflektert bølge er ikke tatt med. Illustrasjonen er basert på http://en.wikipedia.org/wiki/File:evanescent_wave.jpg per 10. mars 2012.

Flyktige bølger finner vi i mange sammenhenger, ikke bare ved totalrefleksjon. Et meget viktig eksempel er grenseflaten mellom metall og luft eller metall og et annet dielektrikum. I metallet vil imidlertid elektroner bevege seg langs grensesjiktet på en spesiell måte. Vi kaller dette fenomenet for “plasmoner” (“surface plasmon-polariton waves”). Plasmonene (kollektiv elektronbevegelse) er kraftig medvirkende til hvordan det elektromagnetiske feltet vil endre seg i området nær grensesjiktet mellom de to materialene, og følgelig også de flyktige bølgene utenfor metallet.

Flyktige bølger er nå svært populære innen fysikk-forskning, ikke minst fordi vi i det siste også har hatt en betydelig utvikling innen nanoteknologi. Vi kan i dag lage strukturer mye mindre enn bølgelengden av lys. Resultatet er blant annet at man har funnet smarte måter å forbedre opplosning f.eks. i mikroskop i. Vi skal i et senere kapittel beskrive diffraksjon, og ifølge de klassiske analysene av diffraksjon kunne vi aldri oppnå en bedre opplosning enn såkalt “diffraksjons-begrenset opplosning”. I dag kan vi imidlertid for spesielle geometrier sprengre denne grensen. Hemmeligheten er at en del strukturer i naturen, f.eks. en celle, har langt mindre utstrekning enn lysets bølgelengde. De flyktige bølgene (evanescent waves) som danner seg i grensesjiktet mellom f.eks. cellen og omgivelsene, vil ha en utstrekning som er betydelig mindre enn bølgelengden. Dersom man kan oppfange

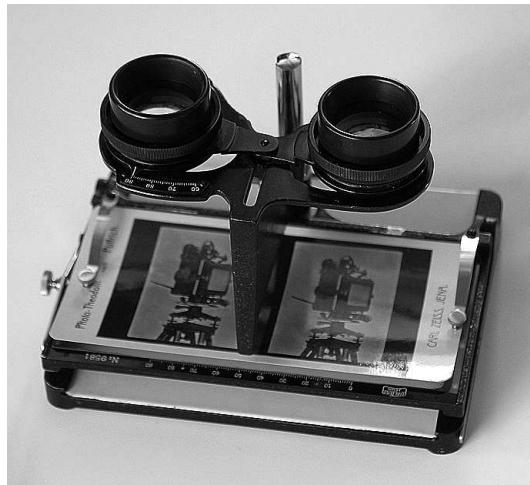
de flyktige bølgene på en eller annen måte, vil vi kunne få en betydelig forbedret oppløsning. Dette er imidlertid ikke lett, siden de flyktige bølgene avtar eksponensielt i styrke når vi kommer bort fra grensesjiktet. De flyktige bølgene er først og fremst betydelige i avstander mindre enn ca. $\lambda/3$ vekk fra grensesjiktet. Det er rom for mye kreativitet i årene som kommer innen forskning på flyktige bølger og utnyttelse av disse!

[♣ ⇒ Kommentar: Vi har i flere kapitler i denne boka sett at det finnes klassiske analogier til Heisenbergs uskarphetsrelasjon i kvantefysikken. Disse klassiske analogiene er alle koblet til bølger, og diffraksjonsbegreset opplösning er et eksempel i så måte (mer om dette i et senere kapittel). Nå når flyktige bølger gjør sitt inntog i fysikken, kan det få følger for Heisenbergs uskarphetsrelasjon. I tilfeller der vi unngår begrensingen som ligger i tradisjonell ”diffraksjonsbegrenset optikk”, regner noen allerede dette som brudd på Heisenbergs uskarphetsrelasjon. Tiden vil vise hvordan vi innen fysikken etter hvert vil forstå uskarphetsrelasjonen når vi beskriver disse fenomenene. .← ♣]

10.8 Orienteringsstoff: 3D Stereoskop

Mennesker har et vel utviklet ”dybdesyn”. De to bildene vi fanger opp med våre to øyne er litt forskjellige fordi det er 6–7 cm avstand mellom øynene. Det betyr at synsvinkelen til nærliggende objekter er forskjellig for de to øynene, mens synsvinkelen er omtrent identisk for fjerntliggende objekter.

Helt siden fotografiens barndom har det vært eksperimentert med å ta bildepar som svarer til bildene vi får på netthinnene våre, det vil si såkalte ”stereoskopiske par”. Vi kunne betrakte bildene hver for seg gjennom spesielle stereoskop (se figur 10.16). Denne teknikken ble også brukt i kommersielle ”ViewMaster” kikkerter som var mektig populære på 1970-80-tallet.



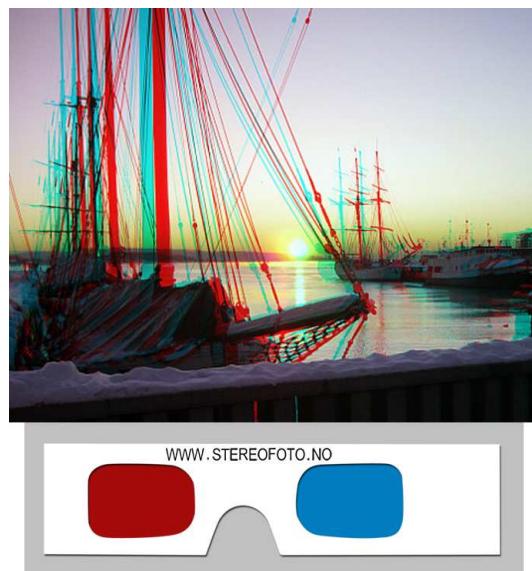
Figur 10.16: *Stereoskopiske bildepar og tilsvarende luper ble utviklet allerede for over 100 år siden. Bildet viser et stereoskop fra 1908 som ble brukt for å betrakte par av stereoskopiske fotografier.*

Stereoskopiske bilder kan også lages ved å legge to stereoskopiske bilder oppå hverandre, der bildet som skal til venstre øye pålegges blågrønt fargeskjær mens bildet som skal til høyre øye pålegges et rødlig fargeskjær. Det resulterende bildet ser derfor noe merkelig ut, med rødfargede og blågrønnfargede objekter side om side (se figur 10.17). Når et slikt bilde betraktes gjennom såkalte ”amaglyfbryllar” som er røde på venstre side og blågrønne på høyre, vil de blå-grønne delene av bildet bli synlige gjennom det røde filteret (lite farge

slipper gjennom, og objektet ser mørkt ut). De røde delene av bildet vil gå gjennom det røde filteret like godt som hvitt lys, og vil bare se hvitt ut og blir “usynlig”.

Bruk av fargekoding og amaglyfbriller er fint til sitt bruk, men metoden kan ikke brukes når vi ønsker å benytte oss av alle farger en fotografisk reproduksjon kan gi.

Da kommer polarisasjonsfiltre inn, og polarisasjon av lys er som skapt for dette formålet. Vi har to øyne, og trenger å “kode” to bilder slik at ett bilde når det ene øyet, og et annet bilde det andre øyet. Lar vi lyset fra ett bilde være horisontalt polarisert og lyset fra det andre bildet være vertikalt polarisert (vi antar at vi nå ser horisontalt mot bildene), vil vi ved hjelp av en brille med polarisasjonsfiltre med horisontal akse på ene glasset og vertikal akse på det andre glasset, oppnå akkurat det vi ønsker.



Figur 10.17: *Stereoskopiske bilder kan lages ved å legge to fargekodete bilder oppå hverandre. Bildene betraktes gjennom fargede briller for å sikre at et stereoskopisk par fotografier oppfattes bare av det øyet hvert enkelt bilde er laget for. Illustrasjon er hentet fra en ordinær medlemsutsendelser fra foreningen Stereofoto Norge.*

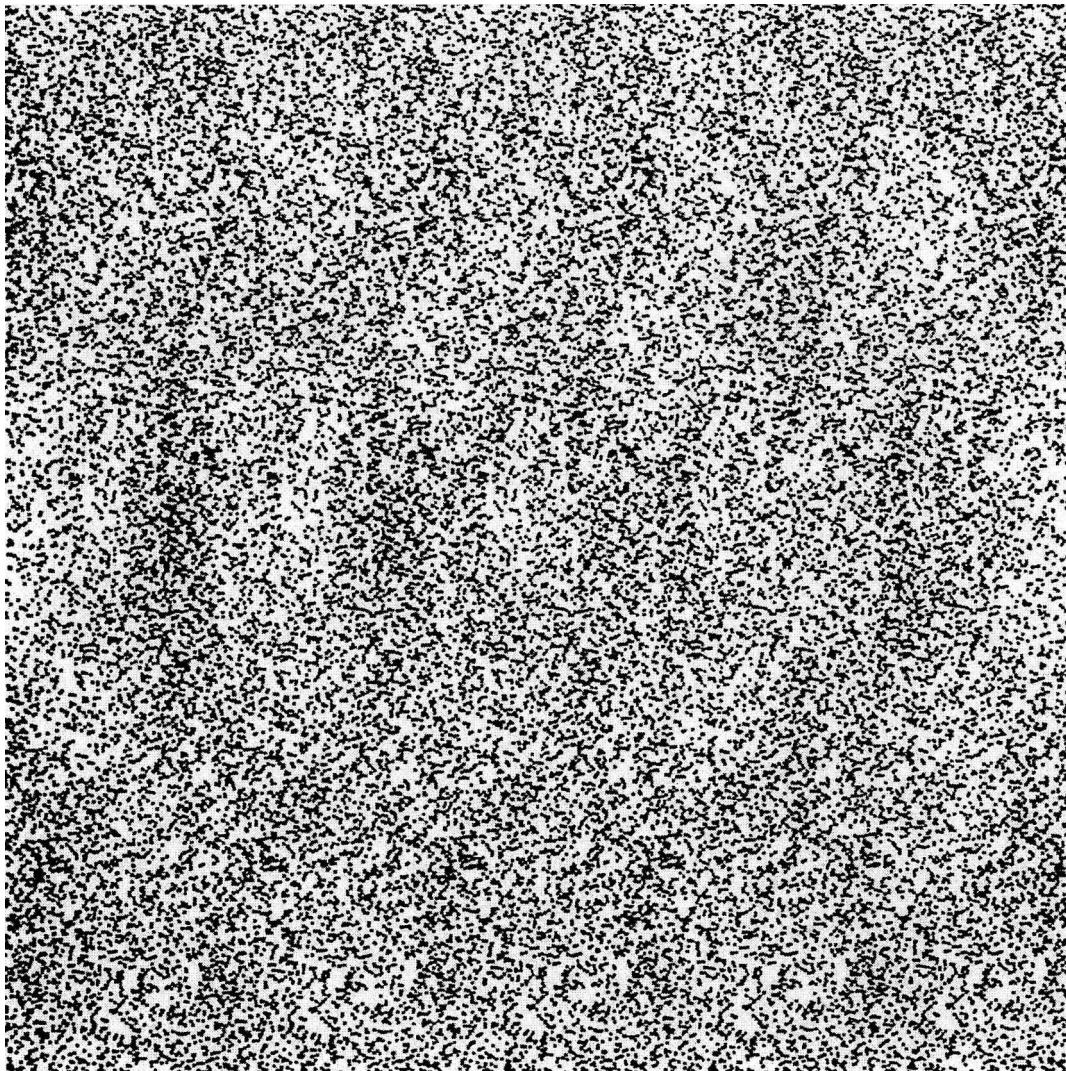
Bruk av lineært polarisert lys fungerer utmerket så lenge vi ser en film og har hodet rett opp og ned. Men legger vi hodet 45 grader på skakke, vil hvert av øynene få inn like mye lys fra hvert av de to bildene. Vi vil i så fall se dobbeltbilder på begge øynene.

Bruker vi derimot sirkulært lys, slipper vi denne ulempen. De to prosjektorer som trengs for en stereoskopivisning må da gi fra seg henholdsvis høyredreid og venstredreid sirkulært polarisert lys. Brillene må ha tilsvarende polarisasjonsfiltre (sirk-pola-filtre fra fotografien vil *ikke* kunne brukes!).

Flere hundre filmer er spilt inn med stereoskopisk teknikk til nå. For eksempel var filmen “Avatar” populær på Oslo-kinoene i 2009/10, og i 2012 hadde filmen “Historien om Pi” betydelig suksess. Stadig lages nye filmer basert på stereoskop. Mange fjernsynsaparatet for stereoskop er kommet på markedet. De baserer seg på en brille som sørger for at annethvert bilde kommer til venstre og annethvert bilde til høyre øye. Kommersielle fotoapparater og videokameraer ment for konsum-markedet finnes allerede. Bare framtiden vil vise hvor stort omfang stereoskopiske bilder/filmer vil få.

For kuriositetens skyld tar vi til slutt med et stereoskopisk bilde som kan betraktes uten hjelpemedler (se neste side). Det er dannet ved mange punktpar som er plassert slik at det ene punktet i hvert par passer for det ene øyet og det andre punktet i paret passer

for det andre øyet. Det er laget en rekke bøker basert på dette prinsippet i mange ulike varianter av virkemidler.



Figur 10.18: Stereoskopisk bilde laget ved hjelp av prikker. Hold boka (eller skjermen) helt opp mot nasen og skyv boka (skjermen) langsomt, langsomt bort fra ansiktet. Ikke forsøk å fokusere på prikkene i bildet, men la selve det stereoskopiske totalbildet etter hvert komme i fokus (kan gjerne fokusere på "uendelig" i starten). Dette stereoskopiske bildet vil, når du endelig får øye på det, synes å ligge nesten like bak papiret som øynene dine er foran papiret. Illustrasjonen er visstnok laget av Lars Olof Björn og ble visstnok publisert i det svenske tidsskriftet *Forskning og framsteg*, nr 4, 1992.

10.9 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Bruke Maxwells ligninger for å på egen hånd utlede relasjonene mellom innfallende, reflektert og transmittert bølge når en plan elektromagnetisk bølge kommer normalt inn mot et plant grensesjikt mellom to ulike dielektriske materialer.
- Gjøre rede for utregningen av refleksjon og transmisjon når en plan elektromagnetisk bølge kommer på skrå inn mot et plant grensesjikt (spesielt holde orden på de to komponentene av det elektriske feltet som inngår i beregningene).
- Gjøre rede for fenomenet knyttet til Brewstervinkelen, og sette opp et matematisk uttrykk for denne vinkelen.
- Angi definisjonene for refleksjonskoeffisient og transmisjonskoeffisient.
- Forklare forskjellen mellom en lineært og en sirkulært polarisert plan elektromagnetisk bølge, og sette opp matematiske uttrykk for de to eksemplene.
- Forklare hva som karakteriserer et dobbeltbrytende materiale, og forklare hvordan vi kan benytte et slikt materiale for å omdanne en lineærpolarisert bølge til en sirkulært polarisert bølge.
- Forklare hva som skjer når lys sendes gjennom flere etterfølgende polarisasjonsfiltre, herunder å kunne angi Malus' lov.
- Gjøre rede for "Fermats prinsipp" (også kalt prinsippet om at optisk veilengde må være stasjonær). Kunne anvende dette prinsippet for å utlede Snels brytningslov og loven om at "innfallsvinkel er lik utfallsvinkel" ved refleksjon av lys mot en plan flate.
- Gjøre rede for fenomenet "totalrefleksjon" og kunne gi et eksempel på bruk av totalrefleksjon i moderne teknologi.

10.10 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Kan vannbølger og/eller lydbølger i luft bli reflektert og transmittert (slik vi har sett at transversale bølger kan) ?
2. Når vi ser en refleks i et vindu, ser vi ofte to bilder litt forskjøvet i forhold til hverandre. Hva skyldes dette? Har du sett mer enn to bilder av og til?
3. Du sender en laserstråle mot en glassplate. Kan du oppnå totalrefleksjon? Forklar.
4. Hvordan kan du avgjøre om solbriller er av polaroid-typen eller ikke?
5. Hvordan kan du bestemme polariseringsaksen til et enkelt lineært polarisasjonsfilter?
6. Nevn to vesentlige forskjeller mellom totalrefleksjon og Brewstervinkel-fenomenet.
7. Hastigheten til lydbølger i luft øker med temperaturen, og lufttemperaturen i luft kan variere mye med høyden. Om dagen varmes bakken ofte opp mer enn lufta slik at temperaturen i lufta nær bakken er varmere enn litt lenger opp. Om natta avkjøles bakken (ved utstråling) og vi kan ende opp med at temperaturen i lufta etter hvert blir lavest nær bakken og stiger litt (før den igjen blir kjøligere enda lenger oppe). Kan du bruke Fermats prinsipp for å forklare at vi ofte hører lyder fra fjerne lydkilder bedre om natten enn om dagen?
8. Hvorfor ser gjerne havet lyst og blankt når vi betrakter en solnedgang i havet?
9. Er det mulig å lage en *plan* elektromagnetisk bølge som samtidig er *sirkulært* polarisert. Som vanlig: Begrunn svaret!
10. Ved omtalen av figur 10.3 ble det sagt at Maxwells ligninger er symmetriske med hensyn til tid. Dersom én løsning er som gitt i figur 10.3, vil en annen løsning være den hvor alle lysstrålene går i motsatt retning. Kan det være tilfelle? (Se spesielt på lyset som da kommer nedenfra inn mot et grensesjikt. Skulle det ikke vært en reflektert stråle nedover i dette tilfellet?) Har du noen eksempler på eksperimentelle situasjoner som ligner på dette tilfellet?

Regneoppgaver

11. En lyskilde har bølgelengde 650 nm i vakuum. Hva er lyshastigheten i en væske med brytningsindeks 1.47? Hvor stor er bølgelengden i væsken?
12. Lys går gjennom en glassterning som er helt nedsenket i vann. Innfallsvinkelen for en lysstråle som går mot glass-vann grenseflaten er 48.7 grader. Dette svarer til kritisk vinkel der vi går over fra å ha noe transmisjon til ren totalrefleksjon. Bestem brytningsindeksen for glasset. Brytningsindeksen for vann ved 20 °C ved 582 nm er 1.333.

13. Anta at en (multimode) optisk fiber har en forskjell i brytningsindeks på 1 % mellom glasset i den indre kjernen der lyset skal gå og den omliggende laget med glass. Bestem maksimal vinkel (i forhold til fiberaksen) lyset kan ha og likevel få totalrefleksjon. Hvor mye vil en kort lyspuls (digitale signaler) flyte ut etter å ha gått 1.0 km langs fiberen? Hva blir da største bitrate (pulser pr sekund) som kan sendes over fiberen før signalet renskes opp for neste etappe med kommunikasjon? (Vi forutsetter at utflyting i signal er hovedårsak til begrenset bitrate i vårt tilfelle.)
14. Når en parallel upolarisert lysbunt treffer en glassflate med innfallsvinkel 54.5 grader, er den reflekterte strålen fullstendig polarisert. Hvor stor er brytningsindeksen for glasset? Hvilken vinkel har den transmitterte lysstrålen?
15. En horisontal upolarisert lysstråle går gjennom et lineært polarisasjonsfilter med polarisasjonsaksen dreid 25.0 grader fra vertikalen. Lysstrålen fortsetter gjennom et nytt, makent polarisasjonsfilter der aksen er dreid 62.0 grader fra vertikalen. Hvor stor intensitet har lyset etter det har gått gjennom begge filtrene sammenlignet med intensiteten før første filterert?
16. En horisontal upolarisert lysstråle går gjennom et lineært polarisasjonsfilter med polarisasjonsaksen dreid +15.0 grader fra vertikalen. Lysstrålen fortsetter gjennom et nytt, makent polarisasjonsfilter der aksen er dreid -70.0 grader fra vertikalen.
- Hvor stor intensitet har lyset etter det har gått gjennom begge filtrene sammenlignet med intensiteten før første filterert?
 - Det settes så inn et tredje polarisasjonsfilter makent til det to andre, men nå med aksen dreid -32.0 grader fra vertikalen. Det tredje filteret plasseres *mellan* de to andre. Hvor stor intensitet har lyset nå som går gjennom alle tre filtrene?
 - Ville det blitt et annet resultat dersom det tredje filteret var plassert *etter* de to andre i stedet for mellom dem?

17. Vis at dersom vi sender en tynn lysstråle på skrå gjennom en plan glassplate med jevn tykkelse, vil strålen som går gjennom glasset ha samme retning som den innkommende strålen, men være parallellforskjøvet fra denne. Vis at parallelforskyvningen utgjør en avstand d gitt ved:

$$d = t \sin(\theta_a - \theta_b) / \cos(\theta_b)$$

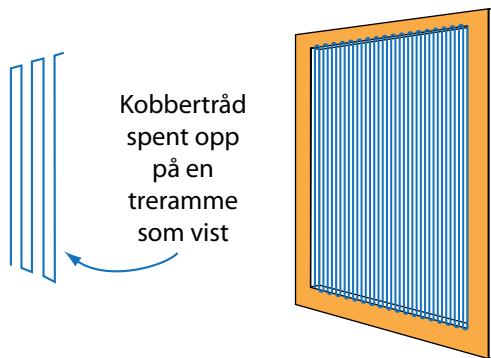
hvor t er tykkelsen på glassplaten og θ_a er innfallsvinkelen og θ_b er vinkelen mellom innfallsloddet og strålen inne i glasset. Finn forskyvningen d dersom glassplaten er 2.40 cm tykk, har brytningsindeks 1.80 og innfallsvinkelen for den innkommende strålen er 66.0 grader.

18. Vis matematisk at innfallsvinkel er lik "utfallsvinkel" (vinkel mellom reflektert stråle og innfallsloddet) ved å bruke Fermats prinsipp.
19. Et dobbeltbrytende materiale har en brytningsindeks n_1 for lys med en viss lineær polariseringsretning og n_2 for lys med polariseringsretning vinkelrett på den første. Dersom dette materialet skal kunne brukes som en kvart-bølgelengde-plate, må lys med den ene polariseringen ha en kvart bølgelengde mer innenfor platen enn lys med vinkelrett polarisering. Vis at platen da må ha (minimum) en tykkelse gitt ved:

$$d = \lambda_0 / (4(n_1 - n_2))$$

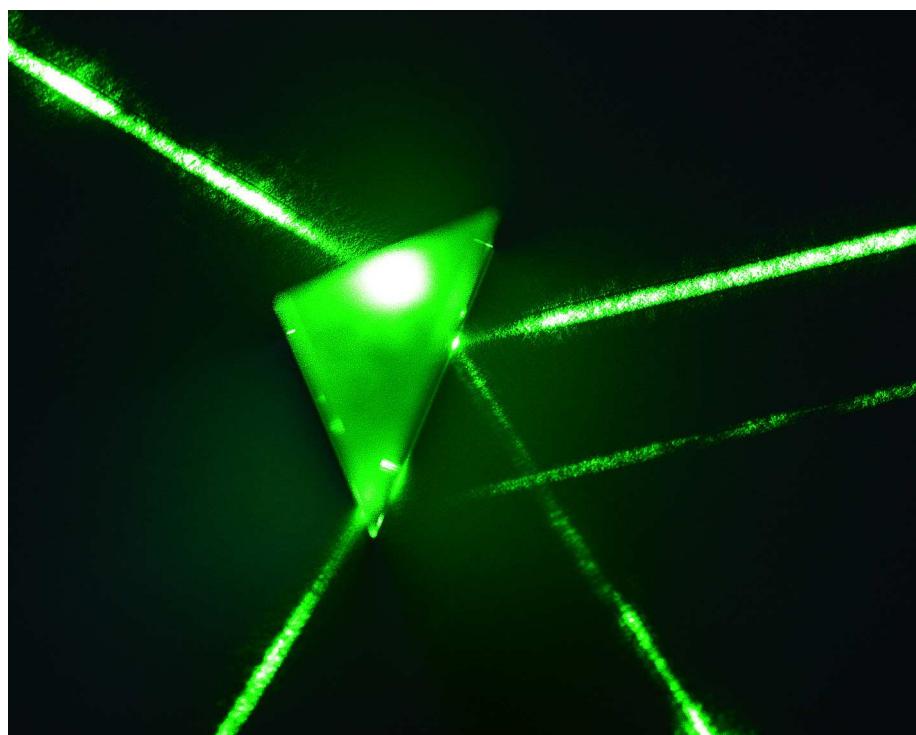
hvor λ_0 er bølgelengden i vakuum (luft). Finn minste tykkelse for en kvart-bølgelengde-plate som er laget av kalsitt ($n_o = 1.658$ og $n_e = 1.486$ ved 590 nm i kalsitt.). Hva er neste tykkelse som vil gi kvart-bølgelengde-plate-funksjon? Hvilken funksjon har forresten en kvart-bølgelengde-plate?

20. Bestem hvor mye en lysstråle blir avbøyd dersom den går gjennom et likesidet trekantet glassprisme på en slik måte at lysstrålen inne i prismet er parallel med en sideflate. Glasset har brytningsindeks n .
21. Vis at ligning (10.6) kan utledes av ligningene (10.13) og (10.14) i det tilfellet at alle er gyldige samtidig.
22. Forsøk å analysere hvordan lyset går gjennom en kuleformet vanndråpe. Definer de vinklene du trenger og benytt deg av symmetrier. Finn ut hvilke parametere som er viktige for å forutsi hvor regnbuen viser seg. For å få fram selve regnbuen må vi enten bruke numeriske løsningsmetoder (noe dere ikke skal gjøre) eller Fermats prinsipp i moderne form. Forsøk uten å gjennomføre regningen å vise hvorfor dette prinsippet kommer inn i bildet. (Ikke bruk mer enn ca 15-20 min på oppgaven.)
23. Vi kan lage et polarisasjonsfilter for radiobølger ved hjelp av en kobbertråd strukket over en ramme som vist i figur 10.19. Forklar virkemåten, og forklar hvilken polarisasjonsretning radiobølgene må ha for å bli stoppet av filteret og hvilken retning som slipper gjennom. Det er mulig at filteret ville vært enda litt mer effektivt om filteret var laget *litt* annerledes. Har du noen gode idéer i så måte?



Figur 10.19: Et “polarisasjonsfilter” for radiobølger.

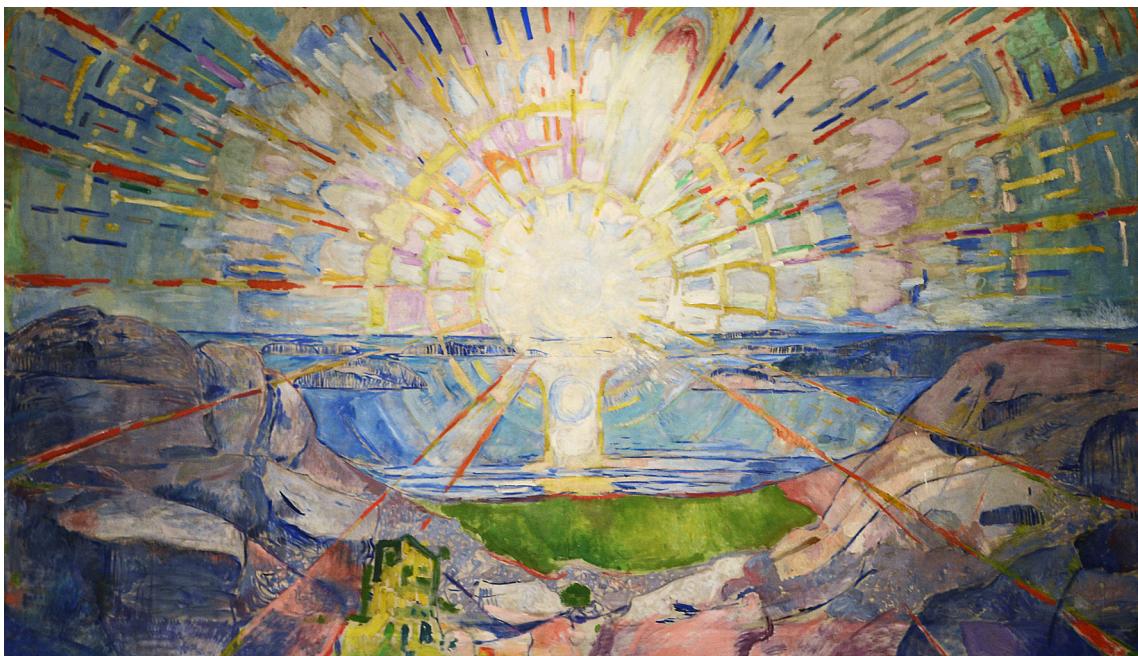
24. Figur 10.20 viser lysstrålen fra en laserpeker inn mot et rettvinklet glassprisme, sammen med lysstråler som på ulikt vis kommer fra lysstrålens gang i og utenfor prismet.
- Finn ut hvilken stråle som er den innkommende strålen, det vil si strålen mellom laseren og prismet. (Det er ikke nødvendigvis den kraftigste strålen!)
 - Identifiser alle fem strålene og finn hvordan lysveien i prismet forbinder lysstrålene med hverandre.
 - Bestem brytningsindeksen til glasset i prismet.
 - Lysstrålen som er brukt har bølgelengden 532 nm. Kan du ved å bruke en figur fra et tidligere kapittel finne ut hvilken type glass prismet synes å være laget av?
 - To stråler synes å være parallelle. Kan du finne ut om dette bare er en tilfeldighet, eller om det faktisk kan bevises at de er parallele? (Litt vanskelig).



Figur 10.20: Lysstrålen fra en laser treffer et rettvinklet prisme og fører til at fire stråler kommer ut fra prismet.

Kapittel 11

Lysmåling, dispersjon av lys, farger



Lys og farger spiller en stor rolle i vårt liv. Munch's "Solen" i Universitetets Aula boltrer seg i lys og farger.

I dette kapitlet gir vi diverse måleenheter for lys. Det er faktisk ganske komplisert når det kommer til stykket. Vi skiller mellom rent fysiske (radiometriske) mål og mål knyttet til følsomhetskurvene til et menneskeøye (fotometriske mål). I tillegg er det et vell av detaljer alt etter om vi mäter lys fra en kilde eller lys som faller inn på et objekt.

Det er store forskjeller mellom hørselsansen og synssansen, noe som blant annet henger sammen med frekvensforskjellen mellom hørbar lyd og lys. Men det er også andre forskjeller. Vi kan ikke se flere farger samme sted til samme tid, mens vi kan høre lyd med flere frekvenser samtidig. Den mest sentrale delen av kapitlet er knyttet opp til fargesyn. Mange kjenner ikke sammenhengen mellom spektralfarger og fargeopplevelser når mange spektralfarger finnes samtidig. Digitale bilder utnytter disse sammenhengene maksimalt.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 22032013.

11.1 Lysmåling

Da vi omtalte lyd, kom vi inn på dB-skalaene som brukes for å angi lydintensitet og lignende. Den gang så vi at vi kunne angi lydstyrke i et rent fysisk mål, f.eks. som amplituden i den gjennomsnittlige romlige oscillasjoner til luftmolekylene, eller som lokal trykkamplitude mens lydbølgjen passerer. Et slikt mål har en begrenset verdi siden svingninger med frekvens lavere enn ca 20 Hz eller høyere enn ca 20 kHz hører vi ikke uansett hvor kraftig en slik lyd er. Vi måtte derfor innføre et eget lydstyrkemål som var knyttet opp til hvor følsomt menneskeøret er for lyd med ulike frekvenser.

På tilsvarende måte finnes det to helt parallelle målesystem når vi skal angi lysmengder/lysintensiteter. Et "radiometrisk" målesystem er basert på fysiske energimål hvor enheter er knyttet til watt på en eller annen måte. På den annen side finnes et "fotometrisk" målesystem som er bygget på synsinntrykk, det vil si menneskeøyets følsomhetskurve. Her er den grunnleggende måleenheten lumen (som igjen er basert på SI-enheten candela).

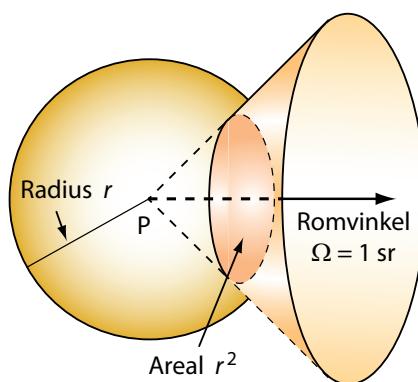
La oss først se på radiometri. Det er først og fremst fire ulike mål som da er aktuelle:

- Energi eller effekt som forlater en kilde totalt.
- Effekt som forlater kilden i en begrenset romvinkel.
- Effekt per flate som faller inn mot en overflate.
- Effekt per flate som stråler ut fra en overflate.

Det kan også være aktuelt å angi effekt per bølgelengde (eller frekvens) for mer eller mindre alle størrelsene ovenfor. Vi får da såkalte spektrale varianter av grunnenhetene.

I flere enheter inngår størrelsen "romvinkel". SI-enheten for romvinkel er en *steradian* (forkortes *sr*). En romvinkel med toppunkt P er 1 sr dersom et kuleskall sentrert i P med radius r kutter av en flate med areal r^2 av romvinkelens kjegleflate, som vist i figur 11.1.

Målinger basert på romvinkel er som oftest bare interessante når målingene foregår langt fra kilden sammenlignet med kildens utstrekning.



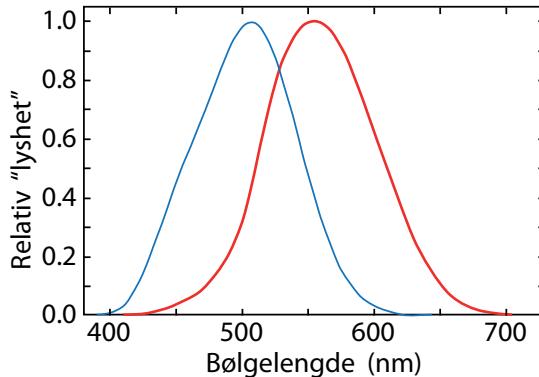
Figur 11.1: En romvinkel på 1 steradian (sr) kutter av et areal r^2 fra en kuleflate med radius r . Toppunktet til romvinkelen og sentrum for kuleskallet må sammenfalle.

De mest vanlige størrelser som brukes innen radiometri er som følger:

- *Strålingsenergi* (radiant energy, eng), måles i J. Karakteriserer kilden.
- *Strålingseffekt / strålingsfluks* (radian flux), måles i W. Karakteriserer kilden.
- *Strålingsintensitet* (radian intensity), måles i W/sr. Karakteriserer stråling per romvinkel i en gitt retning.
- *Radians* (radiance), måles i W/(sr m²). Karakteriserer utstrålt effekt per kvadratmeter projisert flate per steradian romvinkel i en gitt retning.
- *Irradians / innstrålingstetthet* (irradiance), måles i W/m². Karakteriserer strålingsintensiteten inn mot en overflate.
- *Utstrålingstetthet / strålingseksitans* (radian exitance), måles i W/m². Karakteriserer strålingsintensiteten ut fra en overflate.

Legg merke til at det ikke er noen enkel proporsjonalitet mellom radiometriske verdier og lysstyrke slik øyet oppfatter det. En rent infrarød kilde kan ha en høy strålingsfluks, likevel vil ikke øyet oppfatte nærmest noe som helst lys fra en slik kilde. Det er derfor et klart behov også for måleenheter som er knyttet til menneskets oppfatning av lysmengder.

Figur 11.2 viser øyets følsomhetskurve, både for fargesyn (tappene) og nattsyn (stavene). Det er bare kurvens form som er vist, ikke absolutt sensitivitet (toppunktene er normert til 1.0).



Figur 11.2: Kurver som viser hvor "lys" menneskeøyet oppfatter ulike spektralfarger når radiansen inn mot øyet holdes konstant. Den røde kurven (til høyre) gir "lyshetskurven" for fargesyn (tapper), mens den blå kurven gir tilsvarende kurve for mørkesyn (staver). Den første har toppunkt ved 555 nm (grønt), mens den andre har toppunkt ved ca 505 nm. Kurver av denne typen varierer noe fra person til person, slik at en standardkurve må lages ut fra mange målinger. Figuren er basert på Wikipedia (eng) med oppslagsordene "scotopic vision" og "photopic vision" per 22. 3. 2013, hvor CIE's luminosity function er inkludert blant flere. Om CIE: Se referanselisten.

Overgangen mellom radiometri og fotometri er helt analog til overgangen mellom lydintensitet i W/m² og dB(A) (eller dB(C)) i kapittel 6 om lyd og hørsel. Vi må integrere opp fysiske mål i W/m², men vekte bidragene for ulike bølgelengder med øyets lyshetsfølsomhet (figur 11.2) for samme bølgelengde. Det betyr for eksempel for fargesyn at det må dobbelt så kraftig radians i W/(sr m²) for lys ved 510 nm (eller 620 nm) for at det skal bidra like mye til lyshetsopplevelsen som lys ved 555 nm (se figur 11.2).

Absolutt skala i fotometrisammenheng bestemmes ved å angi en sammenheng mellom radiometri og fotometri for toppunktet i kurven. Sammenhengen finnes i definisjonen for grunnenheten for fotometri i SI-systemet, nemlig *candela* (forkortet *cd*):

Monokromatisk lys med bølgelengde 555 nm og strålingsintensitet (radianc intensity) 1/683 W/sr har per definisjon en lysstyrke på 1.0 candela (cd).

Tallet 1/683 virker merkelig, men har sammenheng med at måleenheten tidligere var "normallys", som svarte omtrent til lyset fra et stearinlys. Candela er valgt slik at den omtrent svarer til den gamle enheten.

Det kan bemerkes at i SI-systemet er det bare syv grunnenheter, slik at candela virkelig har en sentral rolle. Det er i grunnen nokså pussig når vi vet at enheten er så nøye knyttet opp til menneskets synssans.

Lumen (lm) er en avledet enhet, lik candela multiplisert med romvinkelen. Lumen angir hvor stor synlig lysstyrke en kilde gir fra seg (integrert over alle retninger). En kilde som gir mange lumen vil gi kraftigere synlig lys enn en kilde med få lumen (se figur 11.4).

Her følger da størrelser som brukes innen fotometri:

- *Lysmengde* (luminous energy, eng), måles i cd sr s = lm s (candela steradian sekund = lumen sekund). Karakteriserer kilden.
- *Lysfluks / lysstrøm* (luminous flux), måles i lumen: $lm \equiv cd \cdot sr$. Karakteriserer strøm av synlig lys som kommer fra kilden.
- *Lysstyrke / lysintensitet* (luminous intensity), måles i candela: $cd = lm/sr$. Karakteriserer synlig lysintensitet fra en lyskilde (per romvinkel) i en gitt retning.
- *Luminans* (luminance), måles i cd/m^2 . Karakteriserer utstrålt lysintensitet fra hver (projisert) kvadratmeter av kilden.
- *Illuminans / belysningstetthet* (illuminance), måles i $lm/m^2 \equiv lux$. Karakteriserer lysfluks inn mot en overflate.
- *Lysekstans / lysutstrålingstetthet* (luminous emittance), måles også i $lm/m^2 \equiv lux$. Karakteriserer lysfluks som stråler ut fra en overflate.
- *Lysutbytte* (luminous efficacy), måles i lm/W . Karakteriserer hvor effektiv en lyskilde er å omsette fysisk effekt til synlig lysstyrke.

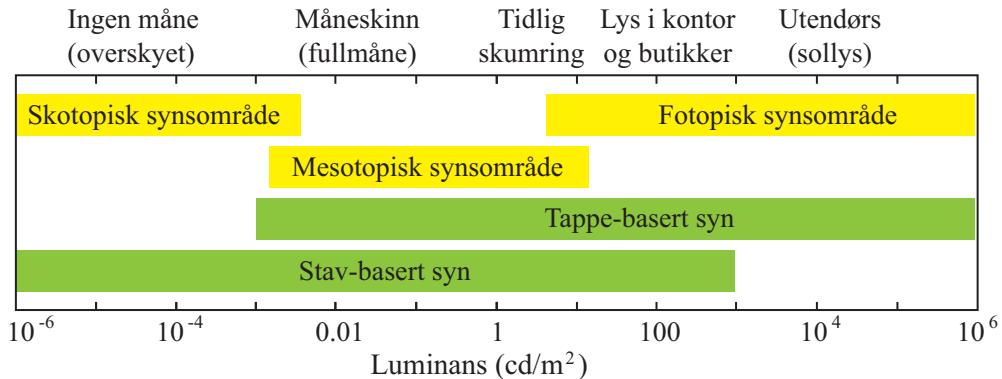
La oss forsøke å sette noen av størrelsene i sammenheng:

Anta at vi har en punkt-lyskilde som har en lysfluks/lysstrøm på 4π lumen totalt. Dette er en oppintegrert lysstyrke for kilden (tar med alle retninger.)

Lysintensiteten fra denne kilden er 1 candela i alle retninger (antar en punktformig lyskilde). Lysstyrken er 1 candela uansett hvor langt unna lyskilden vi er, for lysstyrken karakteriserer kilden og ikke lyset på et gitt sted.

Setter vi imidlertid opp en skjerm på tvers av lysretningen, vil belysningstettheten/illuminansen på skjermen avta med avstanden fra kilden. Vår punktlyskilde med lysfluks 4π lumen vil ha en illuminans på 1 lux på innsiden av et kuleskall med radius 1 m sentrert

i lyskilden. Øker vi radien i kuleskallet til det doble, avtar illuminansen til $1/4$ lux.



Figur 11.3: Omrentilg lysnivå (luminans) i omgivelsene hvor synssansen vår fungerer. Figuren er basert på www.ecse.rpi.edu/~schubert/Light-Emitting-Diodes-dot-org/Sample-Chapter.pdf per 22. mars 2013.

Øyet vårt kan tilpasse seg et enormt spenn med lysintensiteter. Figur 11.3 viser en oversikt over luminans i omgivelsene våre under ulike forhold. I mange år har det vært hevdet at det svakeste lyset menneskeøyet kan oppfatte er noen få fotoner per akkumuleringstid i stav-synscellene (det vil si fem til ti fotoner per sekund innenfor et litet område på netthinnen). Dette er imidlertid en oppfatning som bygger på fotoner som udelelige partikler, og mer moderne forskning stiller spørsmålstege med om dette er en god beskrivelse eller ikke (se referanselisten til slutt i kapitlet). Uansett er det en utrolig lav lysintensitet sammenlignet med fullt sollys.

11.1.1 Lumen vs watt

Det har de siste årene vært mye oppmerksomhet på lyskilders effektivitet. En lyspære av Edison-typen, for å si det slik, omsetter om lag 80 % av energien til utstrålt energi. Resten blir til varme i ledninger, sokkel osv. Av den utstrålte energien går det meste til infrarød stråling som ikke er synlig lys. Når slike lamper brukes i omgivelser hvor vi har bruk for oppvarming, er den lave effektiviteten i å lage synlig lys ikke noe problem. I omgivelser der vi tvert om har bruk for å fjerne varme (i varme strøk), er slike lamper ugunstige.

I figur 11.4 er det vist fire ulike lyspærer. Den klassiske 40 W pæra gir 300 lumen, dvs 7.5 lm/W. For halogenpæra er lysutbyttet $130 \text{ lm} / 10 \text{ W} = 13 \text{ lm/W}$. Fra lysstofferøret får vi $600 \text{ lm} / 9 \text{ W} = 67 \text{ lm/W}$, og endelig, for sparepæra er utbyttet $650 \text{ lm} / 11 \text{ W} = 59 \text{ lm/W}$. Det er med andre ord om lag en faktor 9 mer lysutbytte for lysstofferør sammenlignet med den tradisjonelle lyspæra.

I dag er det stor interesse for neste generasjons lyskilder. Mest kjent er hvitlys LED (Light Emitting Diode). I lommelykter osv. har slike lysdioder vært i bruk i flere år allerede, og det gir mye lengre levetid for batteriene. Det er imidlertid bare en moderat forbedring i lysutbyttet for lysdioder sammenlignet med lysstofferør og sparepærer foreløpig. Vi oppnår om lag 90 lm/W. Fordelen med lysdioder er imidlertid at de er små i utstrekning, at de kan kjøres med lavere spenning enn sparepærer, og at levetiden hevdes å være om lag ti ganger lengre (estimert 100 000 timer, dvs 11 år).

En nyere og mindre kjent teknologi er såkalte organisk lysemitterende dioder (oled). De lages gjerne som lysende flater, se høyre del av figur 11.4. Hittil er maksimalt oppnådd lysutbytte på oled på vel 100 lm/W, men målet er å nå 150 lm/W i 2015.



Figur 11.4: Eksempler på dagens lyspærer (til venstre). Effekt, lysstrøm, levetid i timer og spenning er angitt: 10 W halogenpærer (130 lumen, 3000 t, 12 V), 40 W Classic (300 lumen, 2500 t, 240 V), 11 W sparepære (“= 60 W”) (650 lumen, 10000 t, 240 V), og 9 W lysstoffer (=” 60W”) (600 lumen, ukjent levetid, 240 V anlegg). Til høyre: Bilde av organisk lys-emitterende dioder (oled-panel) sammen med noen vanligere lyskilder. Bildet er hentet fra en artikkel om oled skrevet av Gary Boas i Photonics spectra februar 2010.

Det foregår for tiden intens forskning og utvikling på nye former for lyskilder. Det er mange penger i sikte for dem som får patent på lyskilder som vil dominere markedet i årene som kommer. Det blir i alle fall ikke dagens sparepærer!

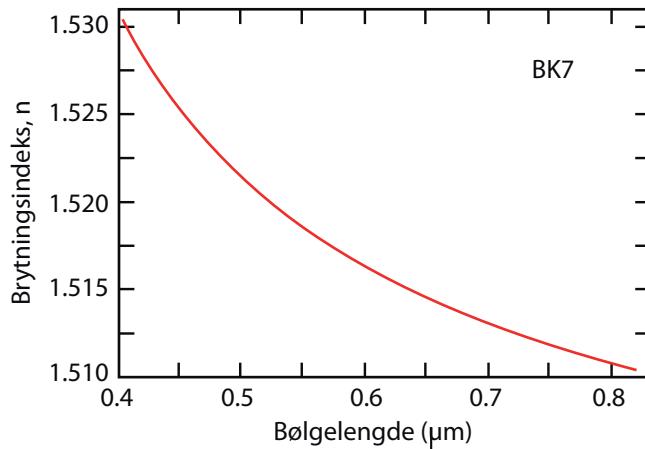
11.2 Dispersjon

Vi har tidligere sett at dispersjon fører til at elektromagnetiske bølger med forskjellig bølgelengde har ulik hastighet gjennom glass. Det er ensbetydende med at brytningsindeksen varierer med bølgelengden.

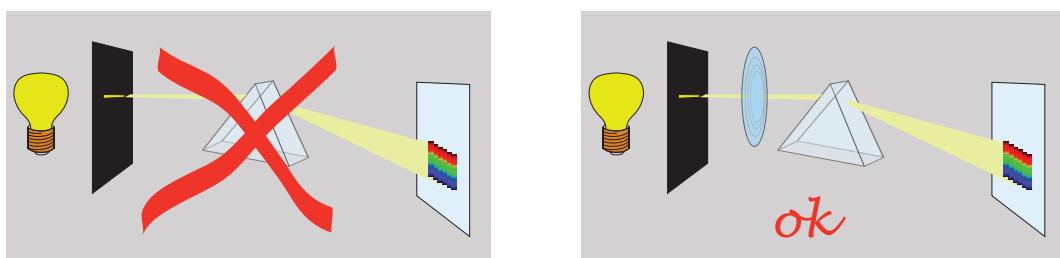
I figur 11.5 gjengis et diagram som viser hvordan brytningsindeksen til lys endrer seg med bølgelengden for en vanlig type optisk glass (Schott BK7). Kurven varierer til dels betydelig for ulike typer glass, så dersom vi ønsker å demonstrere de ulike farge-fenomenene som tas opp i dette kapitlet, bør det brukes en type glass som gir stor dispersjon (ofte knyttet til høy brytningsindeks). I kikkerter søker vi helst materialer med minst mulig dispersjon for at det såkalte kromatiske avviket skal bli så lite som mulig (mer om kromatisk avvik i kapittel 11). Dette gjelder først og fremst for bølgelengder i det synlige området.

Newton s klassiske farge-eksperiment blir vanligvis beskrevet omrent slik: “Når lys sendes gjennom et glassprisme, får vi et spekter”. I praksis er det mer som skal til for at spekteret skal ha den kvaliteten vi forventer, og figur 11.6 indikerer dette. Vi må sende lys gjennom en *smal* spalt, og når lyset gjennom spalten treffer en skjerm bakenfor, må vi ha en avbildning av spalten. Med det mener vi at vi må se spalten som en relativt vel avgrenset lysende, smal flate på skjermen. Dette kan vi oppnå ved å bruke f.eks. sollys (fjern lyskilde) gjennom en egnet spalt (ganske smal). Enda bedre er det å bruke en linse for å få en skarp avbildning av spalten på skjermen.

Først når disse forholdene er tilfredsstilt, kan vi sette inn prismet i lysveien med en



Figur 11.5: Eksempel på dispersjon for en type optisk glass (BK7). Figuren er basert på en figur fra <http://mintaka.sdsu.edu/GF/explain/optics/disp.html> 22. mars 2013.



Figur 11.6: Newton fikk et fargespekter da han avbildet en spalt på en skjerm og lot lyset underveis passere et glassprisme. Geometrien i oppsettet er avgjørende for resultatet.

sidekant parallel med spalten. Lysbunten vil da avbøyes, men vil danne et forskjøvet bilde av spalten på skjermen. Vi må eventuelt etterjustere linsens plassering slik at avbildningen av spalten på skjermen blir skarpest mulig.

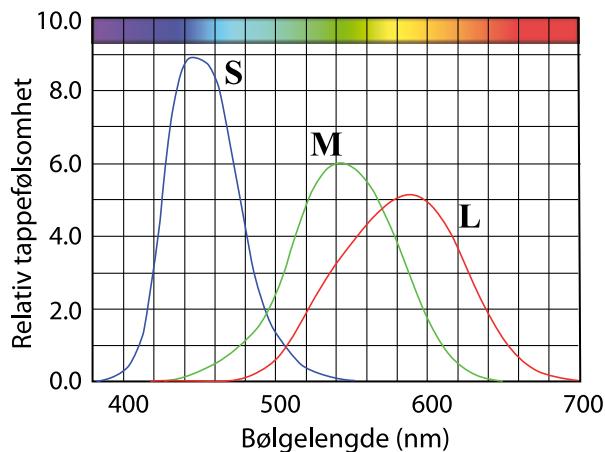
Det resulterende spekteret kan beskrives som *mange* avbildninger av spalten, litt forskjøvet i forhold til hverandre. Dersom lyskilden inneholdt et kontinuerlig spekter med bølgelengdekomponenter i hele det synlige området, vil f.eks. det røde lyset avbildes på ett sted, det grønne et annet sted, og det blå på et tredje. Summen av alle disse avbildningene gir et synlig “spekter” på skjermen.

11.3 “Farge” hva er det?

En rekke detaljer kommer ikke fram i en så enkel beskrivelse av Newtons spekter som den vi ga ovenfor. For det første: Hva mener vi med “farge”? Mange har et svært mangelfullt bilde av farger.

Farge er noe vi *opplever*, et sanseintrykk. Fargefornemmelsen har en komplisert sammenheng med de fysiske stimuli som kommer inn i øyet vårt. Lyset blir delvis absorbert i spesielle proteiner i netthinnens staver og tapper, celler som går under betegnelsen “fotoreseptorer”. Stavene er de mest lysfølsomme reseptorene og er ansvarlig for syn i mørke. Stavene kan ikke gi fargeinformasjon, så vi ser bort fra deres funksjon her. Tapene derimot gir fargeinformasjon. Det finnes tre typer tapper som i første omgang kan kalles blå-følsomme, grønn-følsomme og rød-følsomme. Disse betegnes også som S, M og L-tapper der bokstavene står for “short”, “medium” og “long” bølgelengde for toppen i

følsomhetskurvene deres.



Figur 11.7: *Relative følsomhetskurver for de tre typer tapper i øyet vårt. Figuren er en litt omarbeidet versjon av en figur på www.fho-emden.de/~hoffmann/ciexyz29082000.pdf per 22. mars 2013. Se forøvrig Wikipedia ved oppslagsordet CIE 1931 color space.*

Kort fortalt er følsomhetsområdet og det mest følsomme området for de tre tappene som følger:

- S-tapper, 380 - 560 nm, topp 440 nm
- M-tapper, 420 - 660 nm, topp 540 nm
- L-tapper, 460 - 700 nm, topp 580 nm

Det finnes til dels ganske ulike tall i ulike kilder fordi det er individuelle forskjeller fra person til person og delvis fordi måling av følsomhetskurver ikke er en triviell oppgave slik at verdiene er noe avhengig av målemetoden som brukes. CIE (se referanselisten) har vedtatt en standard som gir gjennomsnittsverdien for hvert av toppunktene.

Figur 11.7 viser følsomhetskurvene for de tre tappetyper. Figuren må forstås slik at dersom vi sender inn *monokromatisk lys* (lys med bare én bølgelengde), viser kurvene følsomheten til hver av de tre typer tapper. Ved ca 444 nm er de "blå-følsomme" tappene (S-tappene i figuren) omrent dobbelt så følsomme som ved ca 480 nm (hhv 0.88 og 0.42 i y-retning i diagrammet). Det betyr at det må dobbelt så intenst lys til ved 480 nm for å gi samme respons fra denne tappen som lys ved 444 nm. Sagt på en annen måte: Dersom en person er født med defekte grønnfølsomme (M) og rødfølsomme (L) tapper, vil vedkommende ha nøyaktig samme synsopplevelse for monokromatisk lys ved 480 nm som for monokromatisk lys ved 444 nm, men med halve intensiteten. For monokromatisk lys med bølgelengde mindre enn 380 nm og større enn 560 nm gir ikke denne type tapper noen nevneverdig respons.

Merk altså at de blåfølsomme tappene gir samme type signal fra seg uansett hvilket lysstimuli som eksiterer tappene. Signalet er et tog av nervepulser. Den eneste variasjonen i signalet er intensitet (antall nervepulser per sekund). På samme måte er det for de to øvrige typer tapper.

[♣] ⇒ En liten digresjon:

Når du ser kurver som i figur 11.7 får du forhåpentligvis assosiasjoner til et grunnleggende fenomen fra bokas første kapitler. Du møtte kurver som lignet svært mye på hver enkelt av kurvene i figur 11.7 da vi

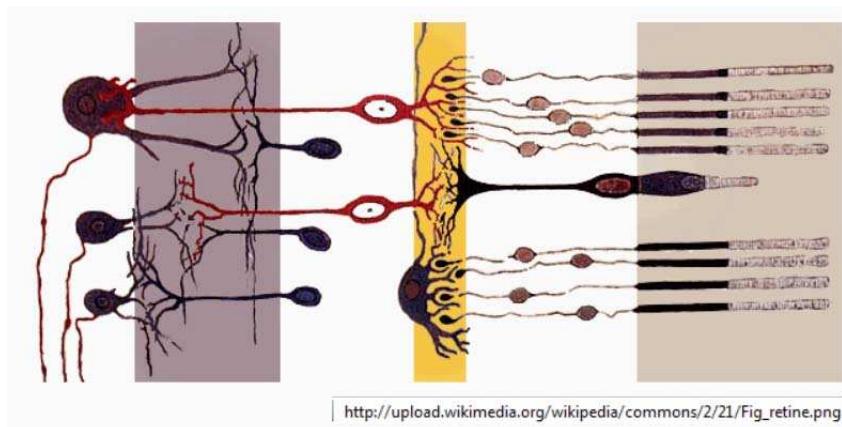
diskuterte tvungne svingninger og resonans. Resonanskurveene tegnet vi riktig nok med frekvens langs x-aksen og et eller annet utslag langs y-aksen, mens vi i figur 11.7 har bølgelengde langs x-aksen. Imidlertid, siden frekvens og bølgelengde er knyttet sammen via $f\lambda = c$, kan bølgelengdeaksen lett gjøres om til en frekvensakse.

Ifølge semiklassisk teori er det en nøye sammenheng mellom tvungne svingninger og absorpsjon i synsreceptorene. Proteinene som endrer form og som i neste omgang endrer på ionetransport og sørger for at synscellene sender en elektrisk puls inn i syns-nerve-banene, har elektrisk ladde partier som vil vibrere i takt med det elektriske feltet i lysbølgene. Vi har da en situasjon helt analogt med tvungne svingninger av et lodd som henger i en fjær, bare at tidsskalaen er helt forskjellig siden både masser og avstander er totalt forskjellig for et protein sammenlignet med et lodd i en fjær.

Kvantemekanisk anses lyset som en strøm av partikler (fotoner), uten utstrekning. En slik beskrivelse gir ikke den samme mekanistisk forklaring på bredden på absorpsjonskurvene. Mange tror at kvantemekanikken er DEN ultimate teori som alltid må brukes for å beskrive mikroskopiske fenomen. Jeg selv mener at semiklassiske betraktningsmåter ofte er langt bedre egnet enn kvantefysikk for å få "visuelle bilder i sitt eget sinn" for de prosessene som foregår. Og ut fra slike bilder kan vi bruke analogier fra andre deler av fysikken for å forutsi nye aspekter ved et fenomen. Jeg selv har derfor stor glede av semiklassiske forklaringer ved siden av kvantemekaniske beregninger når vi forsøker å finne ut hvordan lys og mikroskopiske system vekselvirker med hverandre. – Men nå tilbake til forklaringen av fargesyn hos mennesker! ⇐ ♠]

Det spesielle er at de tre kurvene i figur 11.7 overlapper hverandre, til dels meget sterkt! Det betyr at monokromatisk lys med bølgelengde ca 570 nm vil stimulere (eksitere) både de rød-følsomme og de grønn-følsomme tappene omtrent like mye! Vi skjønner da at uttrykkene "rød-følsom" og "grønn-følsom" egentlig er nokså villedende, og vi går derfor heretter over til bare å omtale tapper etter type S, M og L (forkortinger for "short", "medium" og "long" mhp bølgelengde).

Hvordan kan vi få fargeinformasjon når monokromatisk lys kan eksitere både M og L-tappene like mye?



http://upload.wikimedia.org/wikipedia/commons/2/21/Fig_retine.png

Figur 11.8: *Signalene fra tappene (helt til høyre) blir prosessert av mange typer celler i øyet vårt og på vei til og i selve hjernen. Synsprosessen er derfor svært komplisert. (Venstre del av figuren svarer til den siden av netthinnen der lyset kommer inn. Figuren er hentet fra Wikipedia, oppslagsord "Rod cell" 22. mars 2013.)*

Signalene fra tappene blir kraftig bearbeidet. Allerede i netthinnen har vi fire typer celler som bearbeider responsen fra fotorezeptorene. Disse er såkalte horisontalceller, bipolare celler, amakrinceller og ganglionceller. De ulike cellene har hver sin funksjon, blant annet å forsterke kontraster eller å reagere spesielt på tidsmessige endringer i lysstyrke. Cellene er også involvert i signalbehandlingen relatert til opplevelsen av farger. Det er også

en utstrakt bearbeiding av signalene fra netthinnen i visse relé-knutepunkt i synsbanen, og enda mer i hjernens synssenter. Det er et imponerende maskineri som ligger bak våre synsinntrykk!

Vi skal nøye oss med noen av de enkleste prinsippene for fargeopplevelse, og *hovedregelen* i den sammenheng er at fargen bestemmes av det innbyrdes forholdet mellom hvor mye lys som absorberes i de tre tappetypene.

11.3.1 Fargemetri

Holder vi oss bare til *monokromatiske* bølger i det synlige området, ser vi at lysabsorbsjonen i de tre tappetypene vil endres entydig når bølgelengden varieres. Monokromatisk lys gir synsformemmelser vi kaller "spektralfarger". Disse fargene er i en særstilling, og de oppleves som "mettede" farger. Vi kan ikke gjøre en spektral rød mer rød enn den allerede er (i alle fall ikke med den fargevaløren den representerer).

Dersom vi slipper til lys med *flere* bølgelengder, vil responsen fra tappene være temmelig lik *summen* av responsen fra de monokromatiske bidragene hver for seg. Dette er en summasjonsregel som svarer til superposisjonsprinsippet. Selvfølgelig gjelder denne summasjonen bare innenfor et begrenset intensitetsområde, men vi holder oss til det enkle bildet her.

Tappe-energiabsorbsjonen i M-tapper kan angis matematisk som følger:

$$M = \int \phi(\lambda)M(\lambda)d\lambda$$

der $\phi(\lambda)$ er den spektrale intensitetsfordelingen i innkommende lys (formelt kalt fargestimulus-funksjonen). $M(\lambda)$ er den spektrale energifølsomheten for M-tappene svarende til den midterste kurven i figur 11.7.

Tappe-energiabsorbsjonen i de andre to tappene kan angis på tilsvarende vis. De tre uttrykkene vi ender opp med gir bare relativ absorbsjon (det er ingen absolutt kalibrering involvert i uttrykkene).

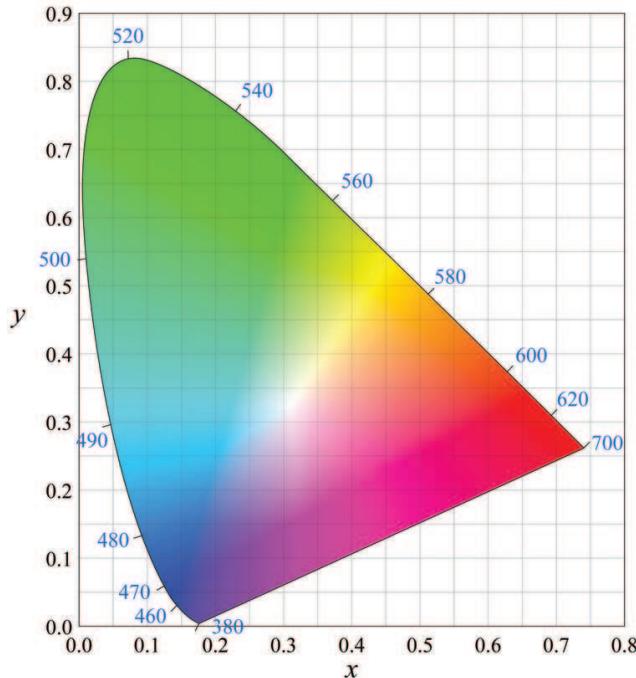
Det er relativt enkelt å innse at monokromatisk lys ved ca 570 nm pluss monokromatisk lys ved ca 420 nm vil gi omtrent samme stimulering av de tre tappetypene som en blanding av monokromatisk lys ved 660, 500 og 410 nm. Det eneste som må sørges for er at:

$$\begin{aligned} M &= \phi_1(570)M(570) + \phi_1(420)M(420) = \\ &\phi_2(660)M(660) + \phi_2(500)M(500) + \phi_2(410)M(410) \end{aligned}$$

og tilsvarende for L og S . Vi får tre ligninger med tre ukjente (når vi antar at de to ϕ_1 -verdiene er kjent).

Poenget med denne analysen er å påpeke at *vi kan få samme fargefornemmelse fra temmelig vidt forskjellige fysiske stimuli*. Med "stimuli" mener vi da spesifikke fordelinger av intensitet mhp bølgelengde, det vil si lysets spektralfordeling. Spektralfordelingen for lys som vi mener har en spesiell grønnfarge kan altså være svært forskjellig fra spektralfordelingen til en annen lyskilde selv om vi vil si den har nøyaktig samme grønnfarge som den første (betegnes "metomeri"). Det er altså *ikke* slik at "farge" er ekvivalent med spektralfarge, definert ovenfor!

Det er faktisk ganske heldig for oss at det er slik! Vi benytter oss av dette i stort monn i dag, ikke minst når vi har med fotografering og farger på en TV-skjerm eller dataskjerm å gjøre. I alle disse tilfellene starter vi vanligvis ut med tre farger og blander dem med hverandre i ulike mengdeforhold for å danne “alle andre farger”. Men det er noen begrensninger ...:



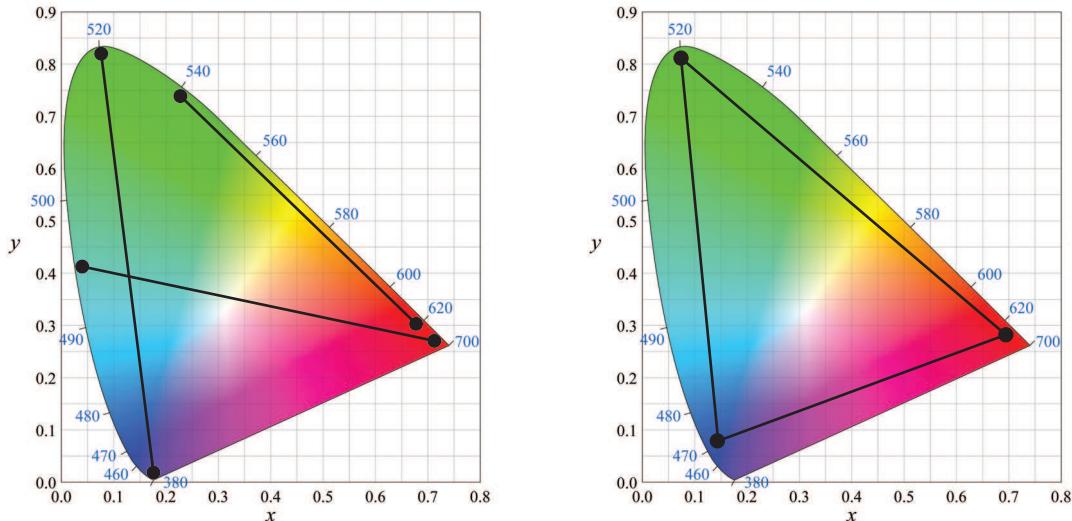
Figur 11.9: “Fargehesteskoen” definert av CIE i 1931. Nærmere omtale i teksten. Figuren er hentet fra Wikipedia, oppslagsord “CIE 1931 color space” 22. mars 2013.

Figur 11.9 viser en såkalt “fargehesteskoen” som er konstruert på en spesiell måte. Langs den krumme randen ligger spektralfargene fra rødt til fiolett. På den rette linjen mellom rødt og fiolett ligger de såkalte “purpurlargene” (purpurlinjen). Midt i hesteskoen er “hvittpunktet”.

Langs aksene er det angitt såkalte x og y-koordinater. Hvordan kan det ha seg at et fargestimulus bestemmes av tre parametre: (S, M, L) mens fargehesteskoen gjengir farger i et todimensjonalt plot?

De tre stimuliene angir både informasjon om farge og om lysintensitet. For en gitt lysintensitet (eller rettere sagt *luminans*) vil de tre parametrene ikke være uavhengig av hverandre. Bare to kan velges fritt. Ved å anvende en passende transformasjon av tappeabsorbsjonene kan vi transformere til to nær uavhengige parametre x og y som angir fargen uavhengig av lysintensiteten (luminansen). Den omvendte transformasjonen er ikke entydig! Fargehesteskoen angir i prinsippet alle farger vi kan oppleve ved en gitt luminans, og kan derfor betraktes som et generelt “fagekart”. Den kalles derfor for et *kromatisitetsdiagram*.

Matematikken bak de aktuelle transformasjonene er utviklet over mange år. Fargehesteskoen ble vedtatt allerede i 1931 som en standard for fargemåling av CIE (Commission Internationale de l’Eclairage på fransk, The International Commission on Illumination på engelsk). Transformasjonene som brukes diskuteres fortsatt, og flere norske fysikere (for eksempel Arne Valberg, Knut Kvaal og Jan Henrik Wold) har arbeidet med denne problemstillingen i mange år.



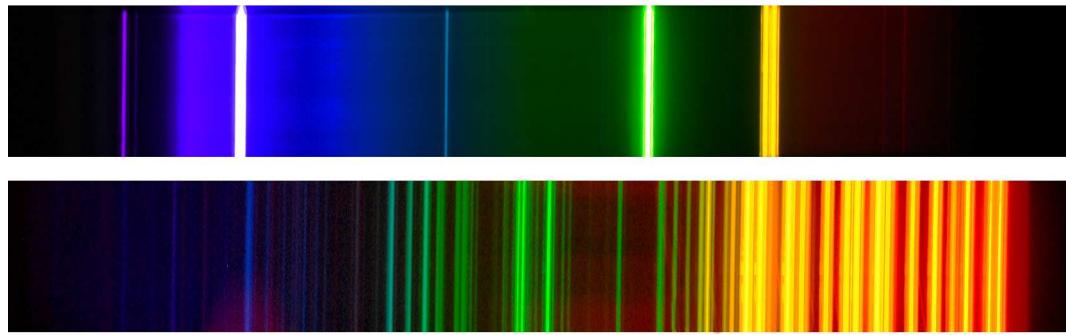
Figur 11.10: Å finne fargen etter additiv fargeblanding svarer til å finne “tyngdepunktet” for de fargekoordinatene som inngår i blandingen. Til venstre er det gitt tre eksempler på fargene som kan oppnås ved blanding av to farger, markert med rette linjer. Nærme- re omtale i teksten. Høyre del: Fargeomfanget til en dataskjerm som benytter tre typer fargepixler ligger innenfor trekanten utspent av fargekoordinatene til pixlene. Fargeomfanget innen den resulterende trekanten er betydelig mindre enn fargeomfanget som hele fargehesteskoen representerer.

Fargehesteskoen er nyttig på mange måter. Starter vi med to farger (to koordinatpunkter inne i fargehesteskoen) og blander disse i samme intensitsforhold (definert på egnet vis), vil fargeoppfatningen vår svare til punktet i fargehesteskoen som er midt mellom de to punktene vi startet ut med. Dette er indikert i venstre del av figur 11.10. Starter vi med like mengder av nærspektrale stimuli ved hhv 540 nm og 620 nm, vil fargen vi oppfatter være temmelig lik fargen til et spektralstimulus med bølgelengde 572 nm (de fleste ville betegne den som gul). Blander vi derimot i omtrent lik mengde nærspektralt stimuli ved 495 nm med nærspektralt stimuli ved 680 nm, vil vi oppfatte fargeblandinga som tilnærmet ”hvit” (en lys gråfarge uten kulør).

Når vi betrakter en dataskjerm, en mobiltelefonskjerm, en iPad, en TV-skjerm, eller lignende, er det tre typer lys som bygger opp bildet: ”Rødt”, ”grønt” og ”blått”. Disse stimuliene har hver sine koordinatpunkter (kromatisitespunkter) (x, y) i fargehesteskoen. *Fargene vi kan danne med disse tre primærfargene ligger innenfor trekanten som de tre koordinatpunktene danner i fargehesteskoen.* Mengden av alle farger vi kan danne med de tre primærfargene kalles *fargeomfanget* til f.eks. dataskjermen.

Vi kan forsøke å velge tre punkter og trekke linjer mellom dem for å få fram hvilke farger som kan fremstilles ved tre primærfarger, og vi vil da oppdage at en rekke farger ligger *utenfor* trekanten som punktene utspenner. Et eksempel på en slik trekant er angitt i høyre del av figur 11.10. Siden en innvendig trekant aldri kan dekke hele fargehesteskoen, betyr det at fargene vi kan få fram på en dataskjerm osv er en ganske blek avbildning av det fargeomfanget vi kan oppleve i naturen. En rekke farger på blomster for eksempel, er langt mer mettede når du ser blomsten i virkeligheten enn det vi kan gjengi på en dataskjerm (eller foto for den saks skyld).

Et eksempel på det manglende fargeomfanget som er oppnåelig med tre-fargers re-



Figur 11.11: Spektrallinjer viser vakre, mettede farger når de betraktes direkte i laboratoriet. Etter fotografering og reproduksjon (som her) blir fargeomfanget langt mindre.

produksjon er vist i figur 11.11. I figuren er det gjengitt to spektre av gasser, et med få spektrallinjer og et med en god del flere. Spektrallinjene er i virkeligheten de mest mettede fargene vi kan få, og når linjene betraktes direkte i et laboratorium, innser vi dette. De samme spektrallinjene gjengitt i et fotografi er bare en blek kopi av virkeligheten (slik figuren viser).

I industriell sammenheng er det utviklet andre fargesystemer enn CIE. Farger inngår i langt flere deler av et moderne samfunn enn det vi vanligvis tenker over. For eksempel bedømmes matkvalitet ved hjelp av farger. To fargesystemer som brukes i industriell sammenheng er NCS og Munsell.

11.3.2 Farger på en mobiltelefon- eller dataskjerm

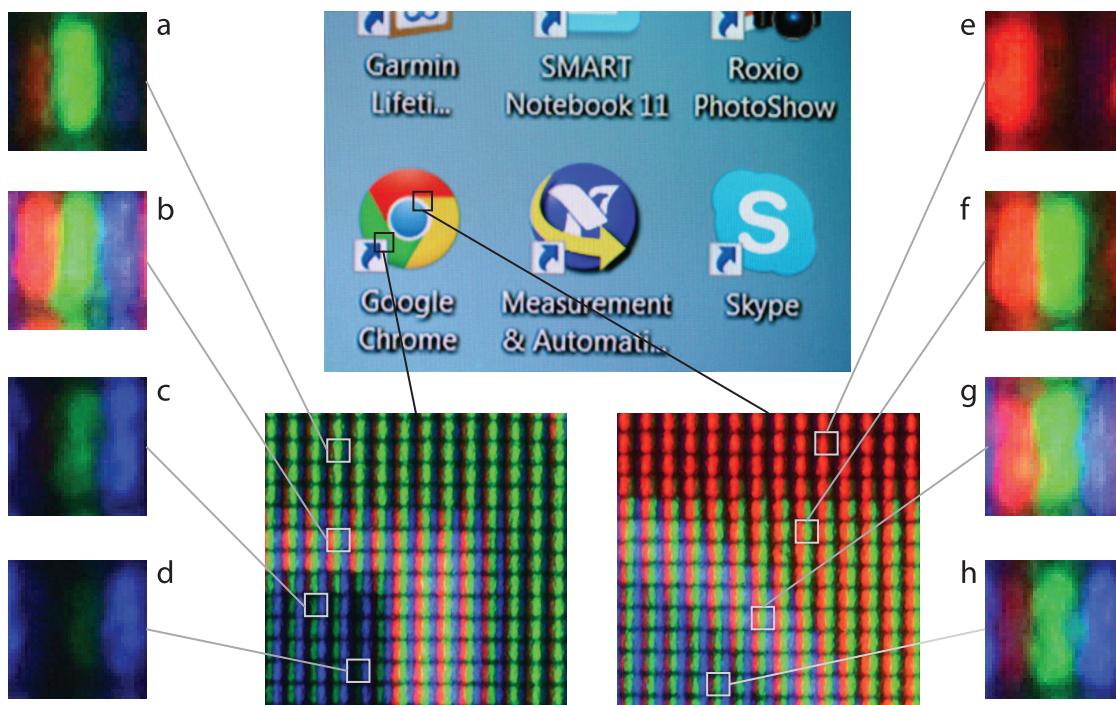
La oss nå sjekke i praksis hvordan farger genereres på en TV, en mobiltelefon eller en dataskjerm. Figur 11.12 viser i midten øverst en liten del av en dataskjerm med Windows-ikoner. Vi har tatt et bilde tetttere innpå skjermen for å se detaljer. Et utsnitt fra Google ikonet har fargene grønt, hvitt og mørk blå. Et annet utsnitt har fargene rød, gul, hvitt og lys blå.

Helt til høyre og helt til venstre er det plukket ut representative “piksler” som bildet er bygget opp med. Hvert piksel på denne skjermen har tre loddrette felt. Disse feltene kan gi henholdsvis fargene rødt, grønt og blått, og kun disse. Disse tre fargene svarer til punktene vist i høyre del av figur 11.10. Vi ser her ganske tydelig at f.eks. fargen gul på dataskjermen egentlig genereres bare ved hjelp av rødt og grønt lys. Pikslene er så små at lyset fra det røde og det grønne feltet i en piksel, treffer de samme synscellene i øyet.

Legg forøvrig merke til at mørk blå eller blåsvart genereres praktisk talt ved å bruke null rødt og grønt lys, og bare svakt blått lys. Lys blå (litt lyst blågrønt) genereres imidlertid med nær maksimalt med blått, en del grønt og litt rødt. Hvitt genereres med kraftig rødt, kraftig grønt og kraftig blått samtidig. Det er fascinerende at vi kan generere så mange farger som vi faktisk kan ved hjelp av bare tre primærfarger!

11.3.3 Additiv versus subtraktiv fargeblanding

I kunstplakater brukes f.eks. syv-fargetrykk, ni-fargetrykk, 13-fargetrykk osv. En av grunnene til dette er at fargeomfanget i det endelige bildet skal bli så stort som mulig. Det er naturlig å trekke paralleller til trekanten til høyre i figur 11.10 i denne sammenheng.



Figur 11.12: Fotografier av en dataskjerm. Et utsnitt av ikoner på ”skrivebordet” på en Windows-maskin er vist øverst i midten. To små utsnitt fra Google-ikonet er vist nedenfor. Piksler fra ulike fargede områder er vist ytterst til venstre og høyre i figuren. Ethvert piksel kan bare gi fra seg rødt, grønt eller blått lys (i hvert sitt område av pikselen). Farger pikslene gir oss er: a) grønt, b) hvitt, c) mørk blå (blå-grønn), d) blåsvart, e) rød, f) gul, g) hvitt, og h) lys blå (blågrønn).

Imidlertid må vi huske at når farger blandes ved hjelp av pigmenter som blyses av en ytre lyskilde, er all fargeblanding langt mer komplisert enn den vi har gjengitt ovenfor. Vi har hittil bare omtalt *additiv* fargeblanding som oppstår ved blanding (overlagring) av lys. I en kunstplakat (eller i et maleri eller fotografi) har vi med *subtraktiv* fargeblanding å gjøre. Pigmenter absorberer noe av det lyset som faller inn på dem, og lyset som sendes tilbake til oss vil gjøre at den pigmenterte overflaten fremstår med en bestemt farge når den blyses f.eks. med sollys. Legger vi flere pigmenter sammen, f.eks. blander gule og blå pigmenter, vil flaten se grønn ut. I alle fall ofte. Men dersom pigmentene blyses av lys med bare noen få bølgelengder (f.eks. lys fra enkelte diodelys (LED) eller lysstoffrør), er det slett ikke sikker at blandingen av gule og blå pigmenter vil se grønn ut!

Ordet subtraktiv fargeblanding er forresten litt misvisende. For å finne tappeabsorbsjonen når stimulus svarer til lys reflektert fra en blanding av to pigmener, må pigmentenes spektrale refleksjonskoeffisienter *multipliseres* med hverandre.

Det var forresten Helmholtz som første gang beskrev forskjellen mellom additiv og subtraktiv fargeblanding. Dette skjedde om lag 200 år etter Newtons fargeblandingsmodell basert på overlagring av lys (additiv fargeblanding).

♣ Det er ikke trivielt å lage fargepigmenter fra scratch. Ofte brukes pigmenter fra naturen, f.eks. fra planter eller mineraler. Det er et begrenset antall pigmenter tilgjengelig, og når vi skal trykke en kunstplakat kan det iblant være nyttig å bruke mer enn tre ”farger” (pigmenter) for å gjengi et bilde best mulig, selv om originalen bare finnes som RGB (tre måltall) fra et digitalt kamera. Vi kan ikke utvide fargeomfanget i forhold til bildeopptaket (fargeomfanget utsspent av RGB-verdiene), men vi kan *reprodusere på papir* fargeomfanget bedre enn om vi hadde brukt færre pigmenter.

Skal vi oppnå et større fargeomfang, må vi allerede i dataopptaket starte ut med flere enn tre stimuli.

Det hjelper lite å starte med et digitalt kamera med kun tre detektorer per pixel og tro at hvis vi bare har en god printer, så skal totalresultatet bli bortimot perfekt! Dette har analogier til lydopptak: Skal vi behandle lyd med en samplingsfrekvens flere ganger den vi bruker i CD-lyd, så nytter det ikke å begynne med denne oppløsningen i behandlingen av lyd og siden utvide. Vi må ha den høyeste samplingsfrekvensen allerede ved den aller første digitaliseringen av lyd. I studioopptak av lyd er det nå temmelig vanlig å bruke høyere samplingsfrekvens enn CD-standarden. For opptak av bilder er det såvidt begynt å eksperimenteres med kameraer med flere enn tre detektorer per pixel, og likeså er det såvidt begynt å produseres skjermer med flere enn røde, grønne og blå lysende punkter. Det er slett ikke utenkelig at fremtidens fotografiapparat og dataskjermer vil bygge på teknologi med flere enn tre basisstimuli. ← ♠]

11.3.4 Andre kommentarer.

Opplevelse av fargene til en blomst avhenger ikke bare av responsene i de tre typene synstapper på det stedet blomsten danner et bilde på netthinnen. Fargeopplevelsen avhenger også av lysintensiteten, samt av tappeabsorbsjonene andre steder på netthinnen. Vi sier at øyet “adapterer”. Vi vet at øyet adapterer når det gjelder intensitet. I sollys reflekterer en “grå” flate mye mer lys enn en “hvit” flate vil gjøre i skumringen. Likevel kaller vi den første for grå og den andre for hvit. Hva vi kaller hvitt, grått og sort flate er altså ikke så mye avhengig av lysintensiteten fra flaten som den relative intensiteten fra flaten i forhold til omgivelsene. Slik er det også til en viss grad for kulørte farger. Hva vi kaller en rød, grønn og blå flate avhenger ikke bare av tappeabsorbsjonene (S, M, L) fra flaten, men også i høy grad av omgivelsene. Fotograferer vi i lampelys vil ofte bildet se ganske rød-gult ut sammenlignet med et bilde tatt i sollys (såfremt ikke kameraet selv korrigerer på en lignende måte som øyet vårt gjør). Når vi er til stede i lampelyset, adapterer øyet slik at vi oppfatter fargene omtrent som om vi hadde sollys til stede.

Dersom vi driver fargekorreksjon av digitale bilder, f.eks. i Photoshop eller tilsvarende programvare, er det viktig å ha en grå flate sammen med bildet der fargene skal vurderes. Sjekker vi stadig at den grå flaten ser grå ut, har vi en rimelig god garanti for at øynene våre ikke har adaptert seg til selve bildet som skal korrigeres. Dersom vi ikke sjekker øyets adaptasjonstilstand i forhold til en grå flate, kan vi komme til å lure oss selv slik at det endelige resultatet blir uheldig.

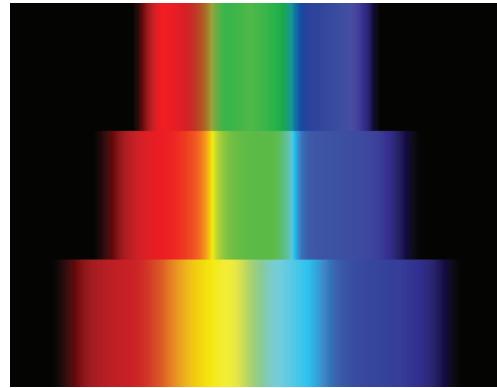
Det er mange andre finurligheter knyttet til øyet og de øvrige deler av synssystemet, ikke minst knyttet til kontraster, men vi kan ikke ta oss tid til å gå inn i denne materien mer enn vi allerede har gjort.

En liten kommentar til slutt om fargehesteskoen: Dersom vi betrakter figur 11.9 på flere ulike datamaskiner, vil vi oppdage at fargene ser nokså forskjellig ut fra skjerm til skjerm. Delvis skyldes dette at de tre pixelfargene er noe forskjellige fra skjerstype til skjerstype. Grafikere utfører ofte en kalibrering av skjerm og transformerer fargeinformasjon ved hjelp av matriser før de betrakter bilder på skjerm eller før bildene trykkes. En slik transformasjon kalles gjerne en “fargeprofil”. I arbeidet med å komme fram til en god fargeprofil, benyttes ofte en standard-plate (bilde) som legges inn f.eks. i motivet ved fotografering. Fargeprofilen kan da utformes slik at sluttresultatet blir så nær opp til den opprinnelige standardplaten som mulig. Gretag Macbeth platen er et eksempel på en slik plate. Den kan bl.a. skaffes fra Edmund Optics (se referanselisten).

Fargehåndtering er et av problemene vi har å hanskes med når vi skal forholde oss til dagens teknologi. Fargekorrigering er en profesjon!

11.4 Spekter fra et prisme

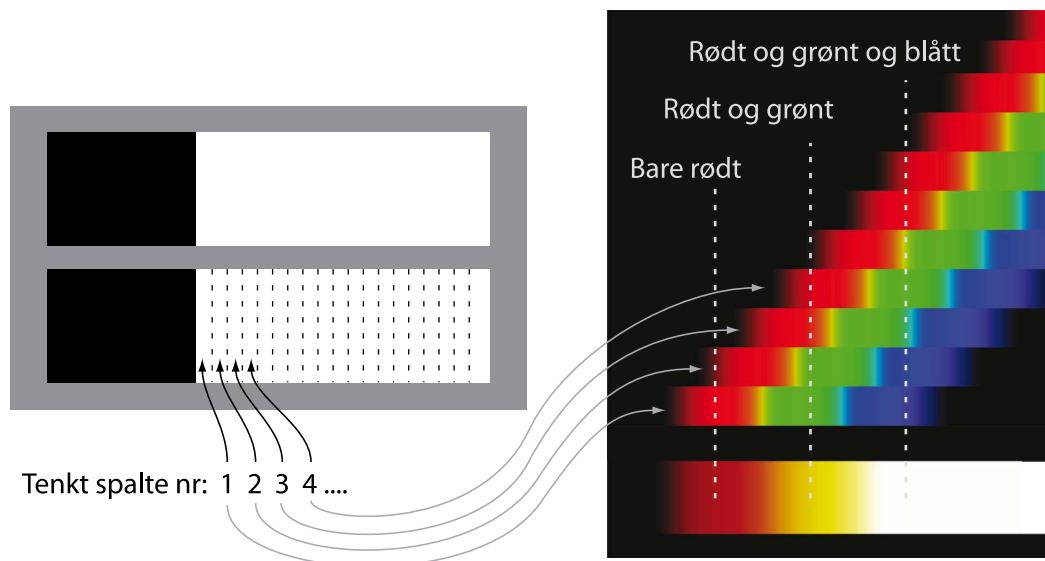
Nå når vi vet litt mer om hvordan vi oppfatter farger, er vi klar til å gå tilbake til Newtons fargespekter fra et prisme. Mange tenker på spektralfarger som de fargene vi ser i regnbuen: ROGGBIF: rødt, orange, gult, grønt, blått, indigo og fiolett. Men hva ser vi egentlig når vi betrakter et Newton-spekter fra en smal spalt? Jo, spekteret ser da omtrent ut som øverst i figur 11.13. Det spesielle er at vi faktisk stort sett bare ser rødt, grønt, blått og til dels fiolett. Det er svært lite gult og orange! Og det er klart mindre gult enn i regnbuen! Hvordan kan det forklares?



Figur 11.13: Spekter fra en smal spalt (øverst) og fra økende spaltbredde nedenfor.

Forklaringen finner vi ved å øke spaltebredden noe. I de neste to eksemplene i figur 11.13 har vi simulert spektere fra spalter med økende bredde. Nå får vi inn gult! Hva skyldes det?

Det skyldes at når øyets betrakter spektralfarger, er det bare et meget snevert bølgelengdeområde som gir oss fargeinntrykket ”gult”. Det meste gule vi oppfatter skyldes blanding av røde og grønne spektralfarger (additiv fargeblanding). Vi får da et koordinatpunkt i fargehesteskogen som ligger litt innenfor randen.



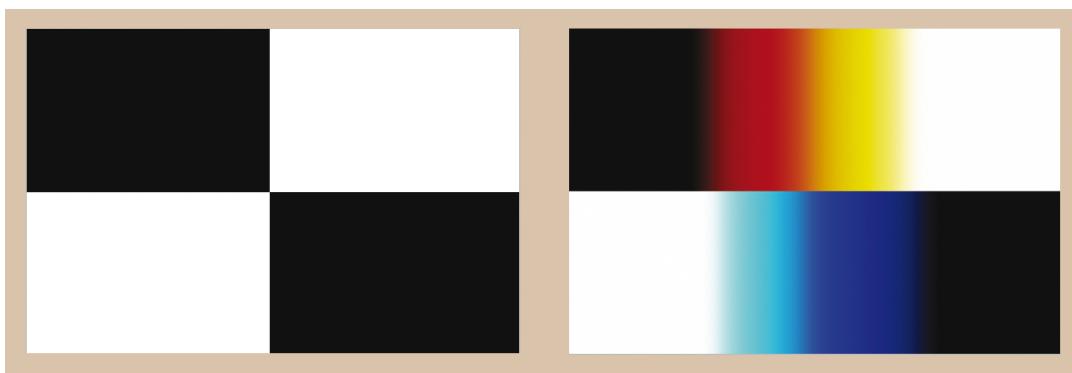
Figur 11.14: Avbildning av en kant kan betraktes som sum av avbildninger av mange spalter ved siden av hverandre. Se teksten.

For å forstå hvordan vi tenker oss fargeblandingen henviser vi til figur 11.14 som viser hvordan bildet ville se ut dersom vi ikke avbildet en spalt på skjermen, men i stedet

en “kant” mellom en flate uten lys og en flate med homogent “hvitt” lys. Lyset passerer også her et glassprisme. Vi kan da tenke oss at det lyse området er en sum av mange enkeltpalster som ligger tett i tett (inntil hverandre). Hver av spaltene (dersom de er smale) vil gi et spekter som ser rødt, grønt og blått ut. Hver spalte er litt forskjøvet i forhold til hverandre, slik at spektrene også blir litt forskjøvet i forhold til hverandre.

Summerer vi lys som kommer inn i forskjellige posisjoner på skjermen, ser vi at helt yttert til venstre kommer det bare rødt lys inn. Summen er da rød. Like til høyre for denne posisjonen får vi blanding av rødt og grønt lys, men ikke blått. Denne summen vil vi oppfatte som gul. Like til høyre for dette følter igjen, vil vi få en blanding av rødt, grønt og blått. Summen oppfatter vi som hvit. Slik fortsetter det utover, og resultatet blir som nederst i figuren.

Vi ser da at når vi avbilder en kant på en skjerm, men lar lyset gå gjennom et prisme, vil kanten bli farget med en rød og gul stripe.



Figur 11.15: Det finnes to typer randfarger, avhengig av hvilken side som er sort og hvilken er hvit i forhold til prismets orientering. Venstre del av figuren viser sort-hvitt-fordelingen av lys vi starter ut med. Høyre del viser hvordan avbildningen av den opprinnelige lysfordelingen ville se ut dersom lyset gikk gjennom et prisme. Fargeeffekten er en simulering tilpasset synsinntrykk fra en dataskjerm. Virkelige randfarger ser penere ut, men de må oppleves *in vivo!*

Avbilder vi en kant hvor det lyse og mørke har byttet plass, vil det fiolette / blå området ikke blandes med andre farger. Ved siden av dette får vi et område med cyan (blanding mellom grønt og fiolett for å si det litt omtrentlig). Ved siden av dette får vi igjen blanding av alle farger, og vi opplever dette som hvitt.

En kant av denne typen vil ha to fargede stripene i overgangen mellom hvitt og sort: Blå-fiolett og cyan. De rød-gule og de fiolett-cyan stripene kaller vi *randspektre* (iblant også kalt randfarger). Et eksempel på randfargene er vist i figur 11.15. Når du ser randfarger i praksis, er stripene ganske mye smalere enn vi kan få inntrykk av i denne figuren, men det avhenger selvfølgelig av avstander og mange andre detaljer.

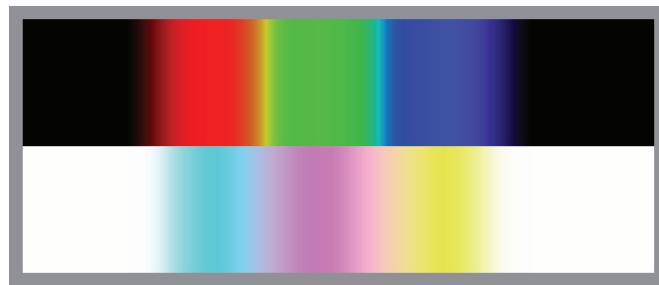
Dersom du ser gjennom en kikkert og velger å betrakte en hvit-sort overgang ute i periferien av synsfeltet, vil du nesten bestandig se randfarger i overgangen. På gode kikkerter, hvor det er forsøkt korrigert for dispersjonen til lys gjennom glass (ved å bruke kombinasjoner av ulike glasstyper i linsene), er randfargene ikke særlig tydelige. I billigere kikkerter er randfargene betydelige og ødelegger kontrasten og skarpheten til bildet vi kan se.

Så tilbake til regnbuen: Hvorfor ser vi gult mye tydeligere i regnbuen enn i et New-

tonspekter hvor vi bruker en smal spalt? I regnbuen er “spalten” i praksis regndråper, og vinkelutstrekningen av hver enkelt regndråpe er kompatibelt med en smal spalt. MEN sola har selv en utstrekning på om lag en halv grad (vinkeldiameter) på himmelen! Regnbuen blir derfor i praksis en summasjon av mange regnbuer som ligger litt utenfor hverandre (som stammer fra ulike soner på soloverflaten). Det er *denne* summasjonen (svarende til at vi bruker en bred spalt) som gir oss tydelig gult i regnbuen!

11.4.1 En digresjon: Goethes fargelære

I Newtons spekter har vi en svært spesialisert geometri som gir oss det vanlige spekteret. Historisk sett reagerte Goethe på Newtons forklaring, ikke minst fordi Newton ikke klart betonte at hans spekter bare fremkom ved avbildning av en spalt (gjennom et prisme). Goethe viste at andre geometrier ga helt andre farger. Blant annet har “spekteret” fra den omvendte geometrien til Newton, nemlig en smal sort stripe på hvit bakgrunn, et totalt annet fargeforløp enn Newton-spekteret (som indikert i figur 11.16). Goethe mente at Newtons forklaring var alt for enkel, og at vi må trekke inn de kromatiske randbetingelsene for å kunne forstå de fargene som oppleves i ulike geometrier.



Figur 11.16: *Fargespekteret vi får fra en Newtonsk spalt og fargespekteret vi får fra en “omvendt spalt”, det vil si en sort smal stripe på lys bakgrunn. Også dette bildet er resultat av en simulering. Virkelige spektere (uten å gå via fotografi eller dataskjermer) er langt vakkre å betrakte!*

Goethe utforsket mange ulike geometrier og fant mange symmetrier i fenomenene og innførte visse “fargeharmonier”, men vi skal ikke gå i detalj.

Her i landet var dikteren André Bjerke en viktig disippel av Goethe. Han ledet en diskusjonsgruppe over en del år, der blant annet fysikerne Torger Holtsmark og Sven Oluf Sørensen var ivrige deltakere. En bok om emnet er: “Goethes fargelære. Utvalg og kommentarer ved Torger Holtsmark”, utkommet på Ad Notam Gyldendals forlag (1994).

Personlig har jeg ikke oppdaget at Goethes fargelære har en større forklaringsevne enn vår vanlige fysikk-modell (basert på lys som bølger) når det gjelder fargefenomener. Figur 11.14 viser etter min mening hovedprinsippet for hvordan vi kan gå fram for å bygge opp hvordan fargene vil komme ut ved en rekke ulike geometrier.

På den annen side har Goethes fargelære hatt en viktig historisk funksjon fordi den fokuserte på at Newtons spekter ikke bare var “lys som gikk gjennom et prisme”. Den fokuserte på symmetrier og geometrier på en flott måte som inntil da ikke var så velkjent som nå. I min språkdrakt vil jeg si at Goethe-tilhengernes poenger minner oss om at beregninger basert på Maxwells ligninger avhenger i høy grad av randbetingelsene! Iblast er vi fysikere alt for slumsete når vi beskriver fenomener og når vi gir forklaringer. Da mister vi fort verdifulle detaljer, og kan komme til å sitte igjen med oppfatninger som ikke duger i andre sammenhenger enn dem som det opprinnelige eksemplet er hentet fra.

Du har forhåpentlig sett at jeg i denne boka har lagt mye vekt på at vi lett tror at spesielle løsninger kan brukes også i andre sammenhenger enn de ble utledet for. Det er en en tvilsom sport å slumse slik, og det gagner ikke fysikkfaget!

11.5 Referanser

I Norge er det kanskje Høgskolen i Gjøvik som har best kompetanse om farger lokalisert på ett sted. De er samlet under paraplyen Det Norske fargeforskningslaboratorium (<http://www.colorlab.no>).

Norsk Lysteknisk Komite er det nasjonale organ for den globale belysningsorganisasjonen CIE. Mer informasjon på websidene til "Lykskultur, Norsk kunnskapssenter for lys" på www.lyskultur.no.

Her er et par andre lenker som kan være av interesse dersom du er interessert i å lese mer om farger:

<http://www.brucelindbloom.com/> og <http://www.efg2.com/>

International Commission of Illumination, Commission Internationale de L'Eclairage (CIE) er en organisasjon som tar seg av belysning og synsoppfatning av lys. Deres hjemmeside er <http://www.cie.co.at/>. Spesielt gis det detaljer angående overgang mellom fysikk og persepsjon i rapporten Photometry - The CIE System of Physical Photometry ISO 23539:2005(E)/CIE S 010/E:2004 beskrevet delvis på http://cms.cie.co.at/Publications/index.php?i_ca_id=475.

Gretag Macbeth fargeplate for å lette fargekorrigering kan kjøpes bl.a. hos Edmund Optics:

<http://www.edmundoptics.com/test-targets/7color-grey-level-test-targets/x-rite-colorchecker/1815> (per. 22. mars 2013).

En bok om emnet er: Richard J.D. Tilley: Colour and optical properties of materials. John Wiley 2000.

Om øyets absolutte sensitivitetsgrense:

Se G.D.Field, A.P.Sampath, F.Rieke. Annu.Rev.Physiol. 67 (2005) 491-514.

Tilgjengelig på web fra www.cns.nyu.edu/~david/perceptionGrad/Readings/FieldRieke-AnnuRevPhysiol2005.pdf per 24. 2. 2012.

11.5.1 Takk!

Jeg vil gjerne rette en hjertelig takk til Jan Henrik Wold (nå Høgskolen i Buskerud, Drammen), og Knut Kvaal (nå Universitetet for molekylær og biovitenskap, Ås) for nyttige kommentarer til dette kapitlet. Eventuelle feil og mangler i denne utgaven av kapitlet er likevel helt og holdent mitt ansvar (AIV), og ikke deres.

11.6 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Forklare behov for ulike størrelser som eksempelvis strålingseffekt, strålingsintensitet, radians, irradians, lysstyrke, lysutbytte m.fl. innen radiometri og fotometri, og kunne gjøre beregninger der du går fra én slik størrelse til en annen.
- Gjøre rede for sammenhengen mellom fargeoppfatning (kulør) og fysisk spektralfordeling (som vi kan få ut fra et spektroskop f.eks. ved å sende lys mot et optisk gitter eller sende lys gjennom et prisme). Eksempelvis skal du kunne forklare at “gult lys” faktisk ikke behøver ha noe spektralt gult i seg overhodet, og likevel bli oppfattet som “gult”.
- Gjøre rede for CIE fagehesteskogen og teori for additiv fargeblanding, og forklare hvilke farger vi kan og ikke kan gjengi fullt ut ved hjelp av f.eks. digitale bilder på en TV- eller dataskjerm.
- Gi relativt detaljerte kvalitative forklaringer på hvordan vi kan oppnå et “Newton”-spekter, randfarger og omvendt spekter, og peke på randbetingelsenes betydning.
- Reflektere litt over at en detektor, f.eks. øyets synsreceptor (tappe-cellene i netthinnen) bare har et begrenset følsomhetsområde, og gjerne knytte dette opp mot såkalte tvungne svingninger tidligere i boka.

11.7 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Fra figur 11.7 kan du finne et par spekralområder der bare én av tappene absorberer lys. Hva betyr det for fargeopplevelsen vi kan ha for ulike spektralfarger innenfor hvert av disse intervallene?
2. Betrakt fagehesteskogen i figur 11.9, og spesielt bølgelengdetallene som står langs randen av hesteskogen. Forsök å estimere hvor lang rand som svarer til spektralfarger hhv i intervallet 400 – 500 nm, 500 – 600 nm, og 600 – 700 nm. Grunnen har noe med hvor lett vi kan oppdage *endringer* i kulør når vi vurderer farger. Hvordan vil du tro sammenhengen er (grovtt regnet)?
3. Finner du holdepunkter i figur 11.7 for sammenhengen du fant i forrige oppgave? Forsök å skrive ned ditt argument på en så presis og lett forståelig måte som mulig! (Dette er en øvelse i å kunne argumentere klart innen fysikk.)
4. Hvorfor er det ikke oppført bølgelengder langs den rette randen av fagehesteskogen?

5. Forsök å se randfarger ved å betrakte en skarp kant mellom et lyst og et mørkt område gjennom en kikkert eller en linse. Lag en skisse som viser omtrentlig det du ser. Påpek hvordan kanten må ligge for at du skal se henholdsvis rød-gul og fiolett-cyan randfarger. Alternativt kan du påpeke randfarger i figur 11.17



Figur 11.17: *Fotografi av diverse strukturer sett gjennom Fresnell-linser i glass (gjenstannede er ut av fokus). Randfarger er synlige.*

6. Digitaliseringskretser som f.eks. brukes ved digitalisering av lyd, har vanligvis 12 - 24 bits oppløsning. Hvor mange dekaders lydintensitet samtidig kan vi dekke med slikt utstyr? For fotoapparater er det vanlig med 8-16 bits oppløsning for å angi lysintensitet. Hvor mange dekaders lysintensitet kan vi dekke samtidig med et slikt kamera? Sammenlign med det intensitetsspennet som gjelder for menneskets hørsel og syn. Hvorfor fungerer lydopptak og fotografier likevel tilfredsstillende til tross for begrenset antall bits oppløsning?
7. Hvor mange ti-er-potensers variasjon i intensiteter fungerer vår hørselsans og vår synssans (se figur 11.3)? Begrunn at sanser som har så stort "dynamisk område" faktisk må være basert på å gi en logaritmisk respons (i alle fall ikke en lineær respons).
8. I forsök hvor vi ønsker å bestemme hvor følsomt øyet er for lys (absolutt terskel-verdi), har det vist seg at responsen fra mange synsceller innenfor et område på netthinnen summeres. Disse synscellene anta å være koblet til samme nervefiber ("område for romlig summasjon"). Et slikt summasjonsområde svarer til at lyskilden har en utstrekning på om lag 10 bueminutter ($1/6$ grad) sett fra øyets posisjon. Drøft hvilke(n) av de fotometriske størrelsene som er relevant å ha kontroll over i denne type forsök. Drøft også hvorfor en del av de andre størrelsene ikke er relevante.
9. Hvide glasskupler blir ofte brukt i lamper for å få en jevn og fin belysning uten markante skygger. Til tross for at lampekuppelen er kuleformet, synes lysintensiteten fra randen av kuppelen å være omtrent lik med lysintensiteten fra de sentrale områdene (se figur 11.18). Drøft hva grunnen til dette er, og drøft hvilket radiometrisk mål som er aktuelt å bruke for å få fram denne egenskapen. Kan du se for deg andre lampekuppelkonstruksjoner som *ikke* ville gitt samme resultat?



Figur 11.18: Fotografi av lyset fra en kuleformet lampekuppel av hvitt glass.

10. Anta at vi for enkelhets skyld kan tenke oss at lys kan beskrives som fotoner med en energi lik $E = hf$, der h er Plancks konstant og f er frekvensen til fotonene. En vanlig laserpeker har gjerne en effekt på noen få milliwatt. Hvor mye må vi dempe en slik laserstråle for å komme ned på et nivå som svarer til grensen for hva øyet vårt kan oppfatte (anta grensen svarer til ca “500 fotoner per sekund”) når laseren har bølgelengden 532 nm?
11. I en tabell på Wikipedia med fysiske data om Sola står det at solas “Luminosity” er $3.846\text{e}26$ W.
 - a) “Luminosity” (innen astronomi) kan ut fra navnet virke som en fotopisk enhet. Hvilken størrelse er det *egentlig* snakk om ut fra tabellene over radiometriske og fotometriske størrelser gitt i teksten?
 - b) I samme tabell finnes følgende opplysninger: “Mean intensity” = $2.009\text{e}7 \text{ W m}^{-2} \text{ sr}^{-1}$. Sola har en diameter $1.392\text{e}9$ m. Skriv med ord hva “mean intensity” forteller oss og hvilken norsk betegnelse vi har for denne størrelsen. Vis ved utregning at det faktisk er den forventede sammenhengen mellom verdiene for “mean intensity” og “luminosity” gitt her i oppgaveteksten.
 - c) Vi skal beregne hvor stor effekt som teoretisk sett kan fanges opp fra sollys ved bakken på Jorden, f.eks. i en solfanger eller solcellepanel. Hvilken radiometrisk eller fotometriske størrelse er vi da interessert i å bestemme? Finn verdien av denne størrelsen når det er oppgitt at avstanden mellom Jorda og Sola er $1.496\text{e}11$ m og at om lag 30 % av solstrålene som kommer inn mot den ytre atmosfæren blir reflektert eller absorbert der.
12. Eigerøy fyr er et av kystens stolteste fyr (se figur 11.19). Fyret ble bygget i 1854 og var det første fyr med støpjernskall. Fyret er 33 m høyt, og lyset kommer ut 46.5 m over havet. I høyre del av figur 11.19 er det vist en bit av det imponerende linsesystemet sammen med lyspæra (pluss en reservepære). Fyret gir fra seg lys med lysstyrke $3.9\text{e}6$ cd i tre lysstråler 90 grader på hverandre (tre linsesett som står 90 grader på hverandre). Fyret er et av de sterkeste langs kyst og når 18.8 nautiske mil utover havet.
 - a) Hvordan kan det ha seg at en lyspære med effekt noen få hundre watt kan sees 18.8 nautiske mil unna?
 - b) Er forresten 18.8 nautiske mil begrenset av lysstyrke eller av jordkrumming?
 - c) Anta at pæren er på 500 W og har et lysutbytte om lag som for glødelamper flest. Estimer romvinkelen til strålene.



Figur 11.19: *Eigerøy fyr er et mektig monument fra en tid før GPS gjorde sitt inntog. Til høyre: Linsesystemet som omkranser lyspæra er imponerende. Det er tre ekvivalente linsesystem som er plassert 90 grader på hverandre. Siste “veggen” er åpen slik at vi slipper inn til pæra. Linsesystemene er over to meter høye.*

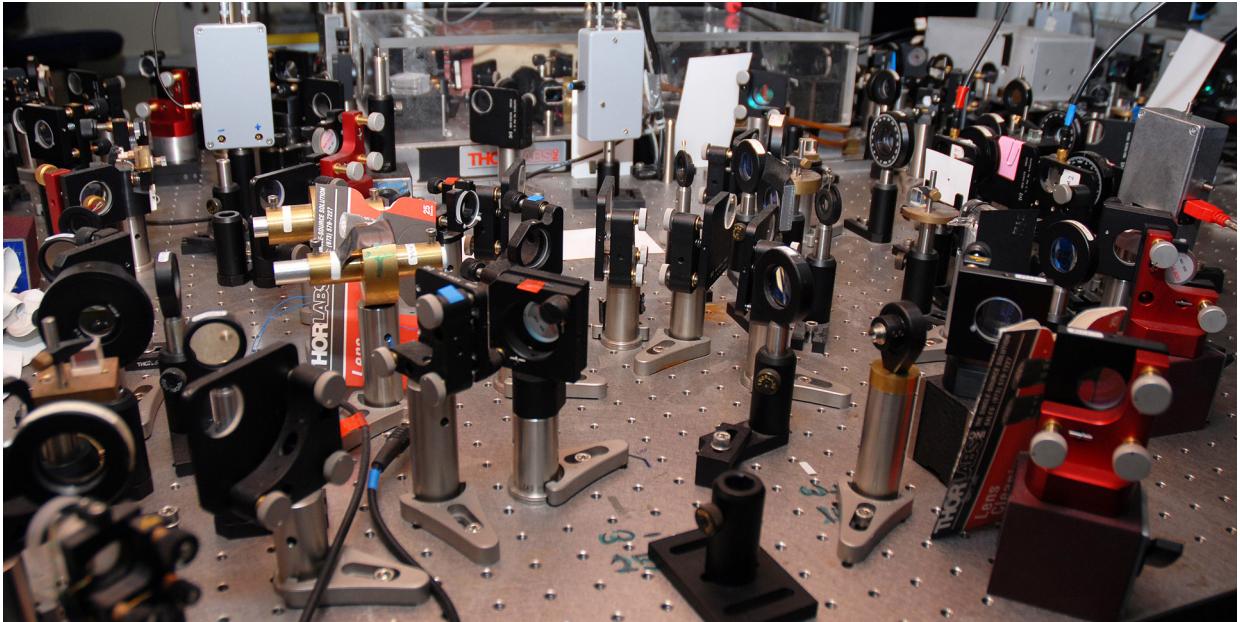
13. Finn ut omtrent hvor bred regnbuen er (vinkelbredde fra rødt til fiolett). Sammenlign dette med vinkeldiameteren til Sola. Synes det å være hold i påstanden om at Solas utstrekning kan ha noe med fargene vi observerer i regnbuen (sammenlignet med et vanlig Newton-spekter når vi bruker en meget smal spalt)?
14. Plukk med deg et par fargeprøver fra en malerbutikk, og forsøk å finne ut hvilket fargesystem fargene er angitt i.
15. Forsøk å analysere hvordan lyset går gjennom en kuleformet vanndråpe. Definer de vinklene du trenger og benytt deg av symmetrier. Finn ut hvilke parametere som er viktige for å forutsi hvor regnbuen viser seg (les gjerne artikkelen som er omtalt helt til slutt i dette kapitlet). Skarpheten og fargene i regnbuen kan anses som et resultat av Fermats prinsipp i moderne form. Forsøk uten å gjennomføre regningen å vise hvorfor dette prinsippet kommer inn i bildet. (Ikke bruk mer enn ca 15-20 min på oppgaven.)

[♣ ⇒ En morsom fortelling i denne sammenheng:

En gruppe studenter gjorde numeriske beregninger av regnbuen som en prosjektoppgave i FYS2130 (UiO) i 2008. Arbeidet ble så veldig godt at de fikk inn en artikkel i tidsskriftet American Journal of Physics: David S. Amundsen, Camilla N. Kirkemo, Andreas Nakkerud, Jørgen Trømborg og Arnt Inge Vistnes: *The rainbow as a student project involving numerical calculations.* Am.J.Phys. 77 (2009) 795-798. Du finner artikkelen selv ved å gå inn på <http://scitation.aip.org/ajp/> og søke f.eks. på "Amundsen". For å hente opp artikkelen i pdf-format, må du benytte deg av en PC knyttet til universitetets nett. ← ♣]

Kapittel 12

Geometrisk optikk



Utsnitt fra et velutstyrt optisk bord i Quantop-laboratoriet på Niels Bohr Instituttet i København 2007.

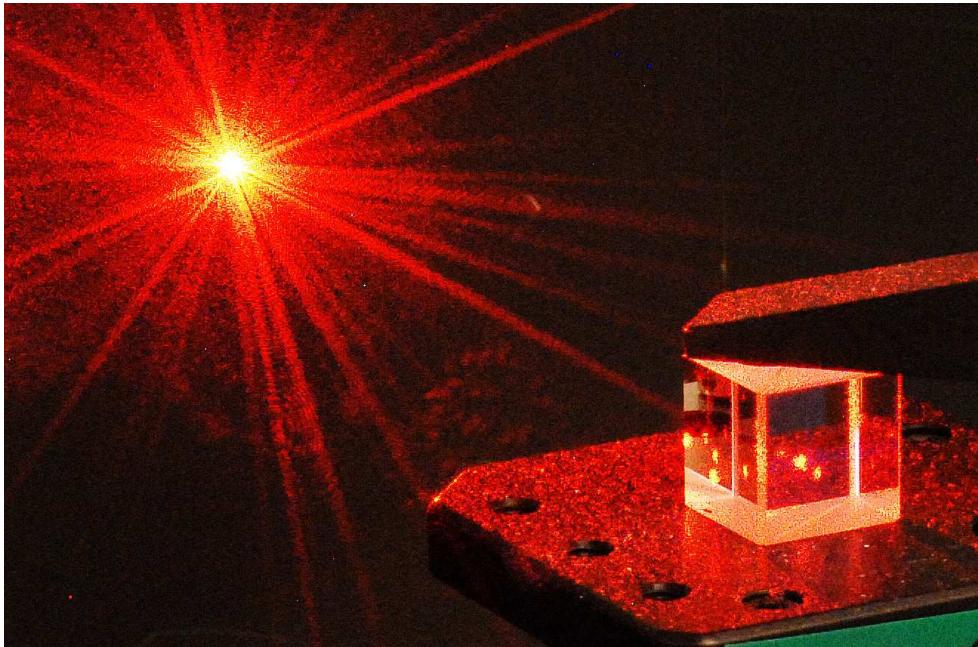
Optikk har i lang tid vært en meget viktig del av fysikken. Kikkerten hjalp Galilei til å innse at månene kretset rundt Jupiter, og det ga støtet til et nytt verdensbilde. I kvantefysikk brukes ofte lys og optikk for å utforske kvantefenomener. Optikk er avgjørende for moderne kommunikasjon siden alle viktige internettforbindelser foregår vha lys. Også fremtidens datamaskiner kan muligens bygges opp av optiske brytere og andre elementer.

Også på et mer personlig plan spiller optikk en betydelig rolle. Mange av oss trenger briller eller linser for å se ordentlig. Vi bruker kameraer for å forevige viktige hendelser i livet vårt. Kopimaskiner eller scannere foreviger skriftlig materiale, og vi bruker kikkerter for å utforske sjeldne fugler eller verdensrommet, og mikroskop for å identifisere bakterier eller pollen. I dette kapitlet ønsker vi å belyse noen generelle trekk innen optikk vi håper du vil ha glede av senere.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 06042013.

12.1 Lysstråler

Vi har tidligere sett at Maxwells ligninger sammen med energikonservering gir størrelse og retning på reflektert og transmittert elektrisk felt etter at plane elektromagnetiske bølger kommer inn mot en plan grenseflate mellom to ulike dielektriske medier. Fra Maxwells ligninger følger både refleksjonslov og Snels brytningslov. Vi har også sett at både refleksjonslov og Snels brytningslov også kan utledes ut fra prinsippet om minste tid, eller mer korrekt, prinsippet om at tiden lyset bruker på veien har en ekstremalverdi. Vi har med andre ord to ulike forklaringer. Hvilken skal vi regne som den mest fundamentale? Det er ikke godt å si, men her er et innspill du kan reflektere over.



Figur 12.1: “Lysstråler” er et mye mer komplisert begrep enn det vi ofte tenker over.

Dersom vi har en lyskilde som sender ut lys i alle mulige retninger, kan vi se for oss at lyset “velger” å følge veien som gir kortest tid dersom vi på forhånd har valgt ut lyskildens posisjon og endepunktets posisjon. Dersom vi derimot sender en vel avgrenset laserstråle i en gitt retning inn mot grenseflaten, vil strålen bli brutt i en bestem retning gitt av Snels lov (utledet av Maxwells ligninger). Har vi valgt et endepunkt som ikke ligger langs den brutte strålen, vil lyset faktisk ikke nå endepunktet. Vi må selv endre på den innfallende strålen inntil den brutte strålen når endepunktet. Med en slik beskrivelse er kriteriet om kortest mulig tid fra startpunkt til sluttpunkt nokså meningsløs. Retningen på den brutte strålen er fullt ut bestemt av retningen til den innfallende strålen. Likevel er det slik at dersom vi velger et sluttpunkt et eller annet sted langs den brutte strålen, så representerer lysveien den veien hvor lyset bruker kortest mulig tid, men det er på en måte en ekstra gevinst, ikke det primære. [Vi kommer tilbake til denne type refleksjoner når vi omtaler diffraksjon, for da får vi inn også Richard Feynman’s tenkning som ligger til grunn for kvanteelektrodynamikk (QED).]

Går vi tilbake til lyskilden som sender ut lys i alle mulige retninger, kan vi i avstander større enn noen centimeter fra kilden anse at lyset tilnærmet kan betraktes som plane elektromagnetiske bølger så lenge vi betrakter en smal “bunt” i en gitt retning. Når denne bunten kommer f.eks. skrått inn mot en plan grenseflate mellom luft og glass, vil den oppføre seg omtrent som laserstrålen vi beskrev i stad. Det vil si at lysbunten vil bli brutt på en entydig måte ut fra hvilken innfallsvinkel strålen har mot grenseflaten.

I geometrisk optikk vil vi operere også med krumme overflater, f.eks. overflaten til en linse eller et krumt speil. Kan vi da bruke lovene vi har utledet for plane grenseflater? Vel, det kommer an på krumningen! En lysbunt hvor vi kan betrakte bølgen som tilnærmet plan, må være “mange” bølgelengder vid for at diffraksjon ikke skal ødelegge “lysbunten” i vesentlig grad. Bølgelengden for lys er i størrelsesordenen $500 \text{ nm} = 0.5 \mu\text{m} = 0.0005 \text{ mm}$. En “lysbunt” med diameter i størrelsesordenen minst 0.01-1 mm (avhengig av ulike detaljer), vil som oftest være vid nok til at den kan betraktes som en bunt med tilnærmet plane bølger.

Dersom denne bunten møter en overflate som er nokså plan innenfor denne tenkte bunten på 0.01-1 mm, vil vi kunne anvende refleksjonslov og brytningslov *lokalt* uten store problemer. Krumme overflater er derfor ikke et nevneverdig problem når vi analyserer lys gjennom linser, nettopp fordi bølgelengden er så liten sammenlignet med krumningsradien.

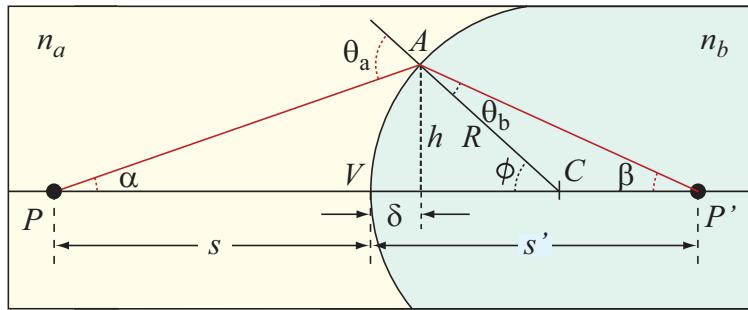
For vanndråper blir det fort annerledes. For vanlige store vanndråper kan vi tilnærmet bruke refleksjonslover og brytningslover for å f.eks. beregne regnbuens utseende, men for svært små vanndråper bryter det hele sammen. Da må vi tilbake til Maxwells ligninger med krumme grensesjikt, og beregningene blir da ekstremt omfattende. Lysspredning fra slike små dråper går under navnet Mie-spredning. Først etter at vi fikk slagkraftige datamaskiner kunne vi gjøre gode beregninger for Mie-spredning. Tidligere ble Mie-spredning bare sett på som en akademisk kuriositet.

Du lurer kanskje på hvilken hensikt jeg har med denne innledningen til geometrisk optikk, og den skal jeg røpe nå. I geometrisk optikk snakker vi om *lysstråler* som går i rette linjer i homogene medier. Men er det egentlig forenlig med moderne fysikk å snakke om lysstråler som går i rette linjer fra f.eks. en lampe, eller fra en blomst på bordet? Vel, så lenge vi er mange bølgelengder vekk fra kilden og fra kanter som begrenser lysets utstrekning, er vi i fjernfeltområdet for de elektromagnetiske bølgene. Da kan lyset lokalt betraktes som tilnærmet plane bølger, og lyset kan tenkes sammensatt av små lysbunter med diameter minst 0.01-1 mm. En såpass bred lysbunt vil kunne fortsette med tilnærmet konstant diameter innenfor lengdeskalaen som er aktuell for klassiske linser og speil. Det betyr at “lysstråler” er et greit nok begrep i klassisk geometrisk optikk. Samtidig må vi ikke glemme at i andre sammenhenger (om f.eks. i Mie-spredning) er begrepet helt ubrukelig.

12.2 Lys gjennom en krum grenseflate

Tenk deg en glasskule i luft og et lite lysende punkt et stykke fra kula som sender ut lys i alle mulige retninger (treffer i alle fall kula). Vi skal nå undersøke hvordan ulike tenkte lysstråler fra det lysende punktet vil gå når de treffer ulike områder på overflaten til glasskula.

I figur 12.2 er det lysende punktet i P , og vi har valgt et snitt hvor både dette punktet og sentrum av kula ligger. En lysstråle fra P som følger linjen mellom P og C vil treffe kuleoverflaten vinkelrett på. Den delen av lyset som transmitteres vil da fortsette rett fram og fortsette i forlengelsen av linjen PC .



Figur 12.2: *Lysstråler fra et lysende punkt (objekt) i P vil danne et “bilde” i punktet P' . Se teksten for detaljer.*

Vi velger så en lysstråle som treffer kuleoverflaten i et punkt A i det planet vi betrakter. Linjen CA og forlengelsen av denne blir da innfallsloddet, og innfallsplanet og utfallsplanet ligger i det planet vi betrakter. Strålen vil *lokalt* synes å treffe en plan flate, og vanlig Snels brytningslov gjelder. Den brutte strålen får en bestemt retning, og lysstrålen vil i punktet P' krysse den første lysstrålen (som gikk gjennom kulesentrum).

Vi skal nå gjennomføre en del geometri for å finne ut hvor skjæringspunktet P' er plassert. Kulas radius er R . Linjen som går gjennom lyskilden P , kulas sentrum C og krysningspunktet P' , kaller vi optisk akse. Punktet der optisk akse skjærer kuleoverflaten kaller vi verteks, og er markert med V på figuren. Avstanden fra lyskilden til verteks kaller vi s og avstanden fra verteks til krysningspunktet P' kaller vi s' . Den loddrette avstanden fra punktet A inn mot den optiske aksen kaller vi h . Avstanden mellom verteks V og punktet der normalen fra A inn på den optiske aksen treffer aksen, kaller vi δ . Ellers er en del vinkler angitt med sine symboler på figuren.

For å gjøre beregningene så generelle som mulig, sier vi at brytningsindeksen for lys er n_a i mediet hvor lyskilden ligger (til venstre i figuren) og n_b i kula (til høyre i figuren). Vi antar også at $n_b > n_a$.

Snels lov gir:

$$n_a \sin \alpha = n_b \sin \beta$$

Vi har også:

$$\tan \alpha = \frac{h}{s + \delta}, \quad \tan \beta = \frac{h}{s' - \delta}, \quad \tan \phi = \frac{h}{R - \delta}$$

Videre vet vi at en utenforliggende vinkel i en trekant er lik summen av de motstående vinklene:

$$\theta_a = \alpha + \phi, \quad \phi = \beta + \theta_b \tag{12.1}$$

Vi gjør nå en meget vanlig forenkling når vi jobber med enkel optikk, nemlig såkal *paraksial* forenkling. Med det menes at vi begrenser oss til forhold der vinklene α og β er så små at både sinuser og tangenser kan tilnærmedes med vinkelen selv (i radianer). Under den samme forutsetningen vil δ være liten sammenlignet med s , s' og R . Ligningene ovenfor kan da tilnærmet skrives:

$$n_a \theta_a = n_b \theta_b \quad (12.2)$$

og

$$\alpha = \frac{h}{s}, \quad \beta = \frac{h}{s'}, \quad \phi = \frac{h}{R} \quad (12.3)$$

Kombineres den første ligningen i (12.1) med ligning (12.2), får vi:

$$n_a \alpha + n_a \phi = n_b \theta_b$$

Setter vi inn også for andre del av ligning (12.1), får vi:

$$n_a \alpha + n_a \phi = n_b \phi - n_b \beta$$

og videre:

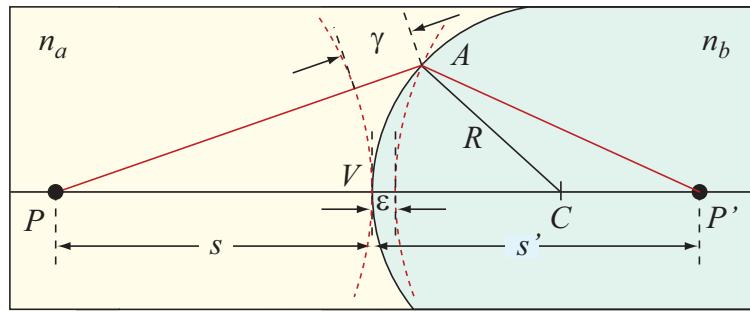
$$n_a \alpha + n_b \beta = \phi(n_b - n_a)$$

Setter vi nå inn uttrykkene (12.3) og forkorter med h , får vi:

$$\frac{n_a}{s} + \frac{n_b}{s'} = \frac{n_b - n_a}{R} \quad (12.4)$$

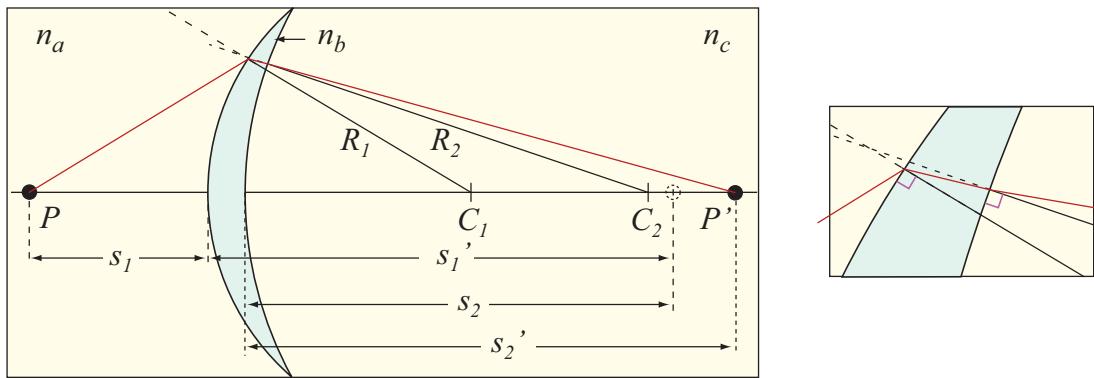
Denne formelen er ganske viktig. Den viser at relasjonen gjelder uavhengig av vinkelen α såfremt vi jobber med paraksial approksimasjon (små vinkler). *Alle* lysstråler fra lyskilden som har liten vinkel relativt til optisk akse, vil krysse optisk akse i punktet P' . Vi kaller lyskilden for *objektpunkt* og krysningspunktet for *bildepunkt*.

Så langt så godt. Men hva så? Er det noe spesielt at lysbunten krysser hverandre? Vil de ulike lysbuntene tilsammen gi et spesielt resultat, eller kan det hende at en lysbunt vil slokke ut en annen, og at det ikke blir noe spesielt ved krysningspunktet uansett?



Figur 12.3: Hvor lang tid vil to lysstråler bruke fra et lysende punkt i P til et “bilde” i P' ? Se teksten for detaljer.

Figur 12.3 viser de samme to lysbuntene som på forrige figur, men vi har nå fokusert på noe annet enn tidligere, nemlig *tiden* lyset bruker fra objektpunktet til bildepunktet. For den rette lysstrålen som følger optisk akse, vil lyshastigheten være c/n_a fram til verteks og c/n_b resten. Hastigheten i glasset er minst. På en liknende måte er det for lyset som følger den andre retningen (via A). Vi ser at avstanden fra objektpunkt til A er γ lengre enn fra objektpunkt til verteks. Lyset vil altså bruke lengre tid på PA enn på PV . På den



Figur 12.4: En linse kan tenkes sammensatt av to krumme grenseflater mellom luft og glass. Bildet av P , dersom vi bare hadde første grenseflate, er markert med en stiplet sirkel. Til høyre viser detaljert strålegang gjennom linsen med brytning først mot innfallsloddet (luft til glass) og dernest fra innfallsloddet (glass til luft). Se teksten for detaljer.

annen side ser vi at lyset vil bruke mindre tid inne i glasset når lyset følger den brutte linjen, for avstanden AP' er ϵ kortere enn VP' . Vi ser at ϵ er kortere enn γ , men dersom vi gjør en nøye geometrisk analyse (noe vi ikke skal gjøre), kan vi vise at *tiden* lyset bruker på avstanden γ i luft er identisk med tiden lyset bruker på avstanden ϵ i glass (gjelder bare for paraksial-tilnærmingen).

Med andre ord: Lyset bruker samme tid fra lyskilden (objektet) til skjæringspunktet (bildet) uansett hvilken retning lysstrålen går (innenefor paraksial-tilnærmingen). Siden lyset har samme frekvens uansett om det går gjennom luft eller glass, betyr dette at det er nøyaktig like mange bølgelengder langs den brutte lysstrålen som den rette. Følgelig vil lyset som kommer til krysningspunktet alltid være i fase med hverandre. Følgelig får vi en addering av amplituder. Kunne vi satt inn en skjerm på tvers av den optiske aksen i punktet P' , ville vi kunne bekrefte dette ved at vi ville se en lysende flekk akkurat der. Ordet "bildepunkt" kan brukes, fordi vi kan danne oss et virkelig bilde av lyskilden på dette stedet.

12.3 Linsemakerformelen

I forrige underkapittel så vi hvordan lysstråler fra en lyskilde (objektpunkt) utenfor en glasskule samlet seg inne i kula i et bildepunkt. Et slikt system er imidlertid av ganske begrenset interesse. Vi skal nå se hvordan vi kan sette sammen to krumme flater, f.eks. fra luft til glass, og deretter fra glass tilbake til luft, for å få fram lover som gjelder for linser. Anta at vi har et opplegg som indikert i figur 12.4. For å få fram hvordan ligning (12.4) benyttes, velger vi å operere med tre ulike brytningsindeks, og vi lar linsen være "tynn", det vil si at linsens tykkelse er liten sammenlignet med både objektavstand, bildeavstand og radiene for både den ene grenseflaten og den andre. Under disse betingelsene (samtidig at vi fortsatt arbeider bare innenfor paraksial-tilnærmingen) får vi:

$$\frac{n_a}{s_1} + \frac{n_b}{s'_1} = \frac{n_b - n_a}{R_1}$$

$$\frac{n_b}{s_2} + \frac{n_c}{s'_2} = \frac{n_c - n_b}{R_2}$$

For en glasslinse i luft er $n_a = n_c = 1$ og $n_b = n$. Videre vil bildepunktet for andre

grenseflate ligge på motsatt side av det vi brukte da vi utledet ligning (12.4). Det kan vises at vi kan implementere dette i ligningene våre ved å sette $s_2 = -s'_1$. Vi gjør da en tilnærming idet vi ser bort fra linsetykkelsen, dvs betrakter linsen som "tynn". Følgelig kan ligningsparet ovenfor skrives som:

$$\frac{1}{s_1} + \frac{n}{s'_1} = \frac{n-1}{R_1}$$

$$-\frac{n}{s'_1} + \frac{1}{s'_2} = \frac{1-n}{R_2}$$

Adderer vi ligningene, følger:

$$\frac{1}{s_1} + \frac{1}{s'_2} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

Dersom linsen betraktes som ett element, som attpå til et tynt, er det naturlig å snakke om objektavstand og bildeavstand relativt til linsen som sådan (midt i linsen), i stedet for å operere med avstander til overflatene. Da ender vi opp med ligningen som går under navnet "*linsemakerformelen*":

$$\frac{1}{s} + \frac{1}{s'} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (12.5)$$

Et spesialtilfelle er når objektpunktet er "uendelig langt borte" (relativt til radiene R_1 og R_2). Da vil $\frac{1}{s} \approx 0$ og

$$\frac{1}{s'} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

Bildeavstanden for dette spesialtilfellet at objektpunktet er "uendelig langt borte", kaller vi "*brennvidden*" til linsen, og betegner den med f (fokal lengde). Bildepunktet ligger da i "*brennpunktet*" for linsen (en brennvidde fra midtpunktet av linsen). Med den gitte definisjonen av brennvidden f ender vi opp med:

$$\frac{1}{s} + \frac{1}{f} = \frac{1}{f} \quad (12.6)$$

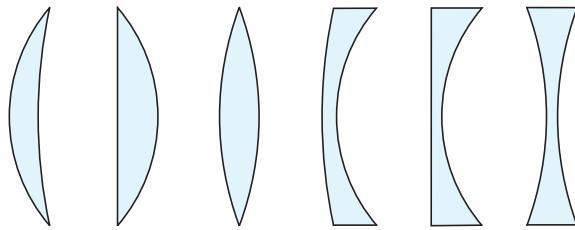
hvor brennvidden f er definert ved:

$$\frac{1}{f} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (12.7)$$

Den første av disse formlene kalles "*linseformelen*", og vi skal benytte oss av den i resten av kapitlet.

Før vi går videre skal vi se hvordan linser vil se ut for ulike valg av R_1 og R_2 i linsemakerformelen. Disse radiene kan være positive og negative, endelige og uendelige. Ulike varianter er gitt i figur 12.5. Linser med størst tykkelse på optisk akse kalles *konvekse* (vokser i midten), men linser som er tynnest ved den optiske aksen kalles *konkave* (de er nesten "av" på midten).

Vi skal også ta med en liten påminning før vi går videre:



Figur 12.5: Snitt gjennom en rekke ulike linseformer. Fra venstre mot høyre: Meniskformet konveks, plankonveks og bikonveks linser. Dernest: Meniskformet konkav, plankonkav og bikonkav linser.

Utledningene ovenfor innebar en rekke tilnæringer, og resultatene vi kom fram til er bare omtrentlige. Dette er typisk for geometrisk optikk. De enkle formlene gjelder bare tilnærmet, og alle beregninger med disse er så enkle at de godt kunne vært gjort i den videregående skolen. På universitetsnivå skulle vi tro at vi kunne gå inn på mer kompliserte og mer nøyaktige beskrivelser, men disse er faktisk så kompliserte at de ikke egner seg i en såpass generell bok som denne. I dag brukes numeriske metoder for de mer avanserte beregningene. Det har da vist seg at det ikke er mulig å lage perfekte linser. Vi må gjøre avveininger, og en linse som skal brukes stort sett ved korte avstander, vil måtte utformes på en annen måte enn en linse som hovedsakelig skal brukes ved lange avstander.

Vi har tatt utgangspunkt i sfæriske grenseflater. Dette skyldes at det inntil ganske nylig var mye lettere å fabrikere linser med sfæriske overflater enn andre former. I de siste årene er det blitt mer vanlig å fabrikere linser med noe annen form, og da reduseres problemet som kom til syne ved paraksial-tilnærmingen. Vi kan redusere såkalte sfæriske feil ved å forme overflatene ikke-sfærisk.



Figur 12.6: Eksempel på et moderne objektiv for fotografi: AF Nikkor 28 mm f/1.4 GED objektiv. I stedet for én enkel tynn linse, slik vi tenker oss et objektiv i vår gjennomgang av geometrisk optikk, har Nikkon-objektivet 12 linser som spiller sammen som én. De fleste enkeltlinsene har sfæriske overflater, men to (de blå) har ikke-sfæriske overflater. To linser (de gule) er laget av ekstra dispersivt glass, det vil si at brytningsindeksen har en annerledes variasjon med bølgelengden enn det som er mest vanlig i andre glasstyper. Illustrasjonene er hentet fra websidene til Nikon 6. april 2013.

Vi ser forresten av ligning (12.4) at brytningsindeksene inngår. Når vi vet at brytningsindeksen avhenger av bølgelengden, vil det si at bildepunktet P' vil ha en annen posisjon for rødt lys enn for blått lys. Ved hjelp av flere linser med ulikt glass (ulike brytningsindeks) kan vi kompensere delvis for denne type feil (kromatisk avvik). Totalt sett

er det likevel en meget krevende oppgave å lage en god linse. Det er derfor ikke uten grunn at entusiaster gransker nye objektiver fra Nikon, Canon, Leitz osv. med stor interesse like etter at de er kommet på markedet. Har spesialistene klart å lage noe spesielt godt denne gangen, og i så fall, i hvilket henseende? Den helt perfekte lensen finnes ikke!

12.4 Lysstråleoptikk

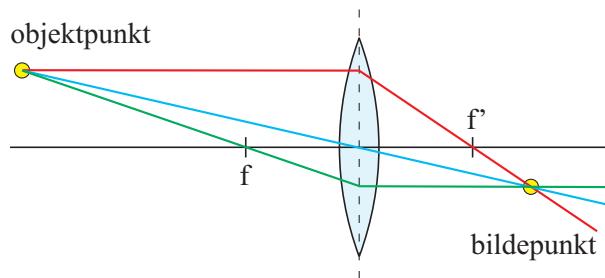
Vi skal nå gå løs på den delen av optikken som omhandler briller, fotoapparater, luper, kikkerter, mikroskop osv. På engelsk går den under navnet “Ray optics” i motsetning til “Beam optics” som mer koncentrerer seg om hvordan en laserstråle utvikler seg (der diffraksjon er helt essensiell).

Når vi tar et fotografi, avbildes et objekt på f.eks. en CMOS-brikke. Situasjonen er da en del forskjellig fra det vi hittil har omtalt. Hittil har vi latt objektet være et lysende punkt plassert på den optiske akse. Nå må vi kunne behandle også objektpunkter som ikke ligger på den optiske aksen.

Det er da tre hovedregler vi forholder oss til om igjen og om igjen.

1. Innkommende lys parallellt med optisk akse, vil gå gjennom brennpunktet.
2. Lys som går gjennom linsens midtpunkt (der optisk akse skjærer gjennom linsen), vil gå videre i samme retning som det kom inn.
3. Lys som passerer linsens brennpunkt *foran* linsen, vil gå parallellt med optisk akse *etter* linsen.

Reglene kommer delvis fra linseformelen. Vi så at dersom objektavstanden s ble gjort uendelig stor, vil bildepunktet ligge i en avstand lik brennvidden etter linsen. Trekkes flere ulike lysstråler fra objektet i dette tilfellet, vil strålene komme inn (tilnærmet) parallellt med den optiske aksen, og alle slike stråler skal gå gjennom brennpunktet. Herav den første regelen.



Figur 12.7: Et lysende objektpunkt som ikke ligger på den optiske aksen, vil avbildes i et punkt på motsatt side av en konveks linse. Bildepunktet ligger ikke på den optiske aksen. Tre hjelpeelinjer brukes for å finne plasseringen til bildepunktet.

Linseformelen kan imidlertid kjøres både forlengs og baklengs for å si det litt upresist. Dersom vi plasserer en svært liten lyskilde på den optiske aksen i en avstand lik brennvidden foran linsen, vil lyset fra kilden gå til linsen i mange ulike retninger, men bildepunktet vil da ligge i en avstand $s' = \infty$. Det betyr at strålene, uansett hvor de går gjennom linsen, vil fortsette tilnærmet parallellt med optisk akse etter linsen.

Den midterste regelen er kanskje enda lettere å skjonne. Midt på linsen (der optisk akse skjærer gjennom linsen) er de to overflatene tilnærmet parallelle. Dersom en lysstråle sendes gjennom et stykke planparallelt glass, vil lysstrålen bli brutt ved første grenseflate, men brutt tilbake til opprinnelige retning når den går gjennom andre grenseflate. Utgående lysstråle vil bli litt parallelfforskjøvet i forhold til innkommende lysstråle, men dersom vinkelen ikke er for stor, og linsen tynn, vil parallelfforskyvningen bli såpass liten at vi kan se bort fra den i våre beregninger.

Objekt utenfor brennpunktet.

Anvendes disse reglene for et lysende objektpunkt som ikke ligger på optisk akse, får vi resultatet gitt i figur 12.7. De tre spesielle lysstrålene, angitt ved våre generelle regler, møtes eksakt i ett bildepunkt. Bildepunktet ligger på motsatt side av optisk akse sammenlignet med objektpunktet.

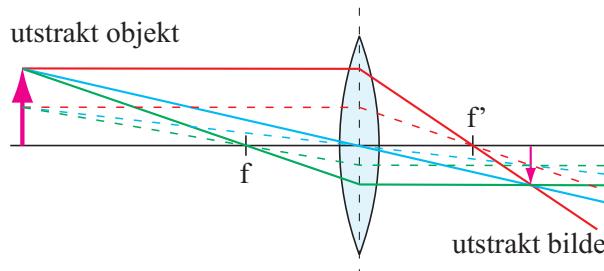
Dersom objektet ikke lenger er ett lyspunkt, men et utstrakt legeme, som for eksempel en pil, finner vi noe interessant (se figur 12.8). Fra hvert punkt i legemet sendes det ut lys, og for hvert punkt i objektet blir det et tilsvarende punkt i bildet. For våre forenkede regneregler er det slik at alle punkter i objektet som ligger i et plan vinkelrett på optisk akse, vil de tilsvarende bildepunktene ligge i et plan vinkelrett på optisk akse på motsatt side av linsen (under betingelser som angitt i figuren). Det vil si at vi kan avbilde et objekt (f.eks. forsiden av en avis) til et bilde, som kan fanges opp på en skjerm. Bildet vil da være en tro kopi av objektet (avissiden), bare at det vil ha en forstørrelse eller forminskning sammenlignet med originalen, og bildet vil være opp ned (men ikke speilvendt).

Forstørrelsen er rett og slett avhengig av s og s' . Er $s = s'$ vil objekt og bilde være like store. Er $s' > s$ er bildet større enn objektet (originalen), og visa versa. Lineær forstørrelse er rett og slett gitt ved:

$$M = -\frac{s'}{s}$$

Minustegnet er tatt med bare for å markere at bildet er opp ned sammenlignet med objektet.

Det er også mulig å definere en forstørrelse i areal. I så fall blir den kvadratet av uttrykket angitt her.



Figur 12.8: Et utstrakt objekt kan tenkes å bestå av en mengde objektpunkter, og hvert punkt avbildes i et tilsvarende punkt på motsatt side av en konveks linse. Resultatet er at objektet som sådan avbildes som et bilde. Bildet er opp-ned og har en annen størrelse enn objektet.

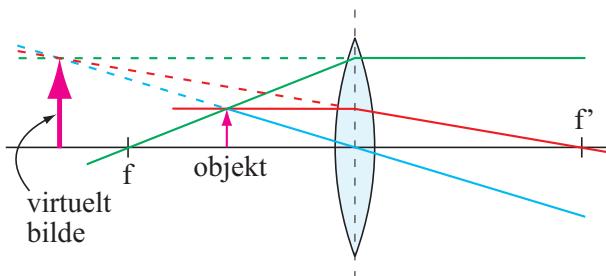
Objekt innenfor brennpunktet.

Hittil har det vært relativt lett forståelige sammenhenger mellom objekt og bilde, og

vi har kunnet fange opp bildet på en skjerm og se at det er der. Men hva så dersom vi plasserer objektet nærmere linsen enn brennvidden? Figur 12.9 viser hvordan de tre referansestrålene nå går. De divergerer etter at de har passert linsen! Det finnes ikke noe punkt hvor lysstrålene møtes og hvor vi kan samle opp lyset og se på det. Derimot synes lysstrålene å komme fra ett og samme punkt, et punkt *på samme side av linsen som objektet*, men på et annet sted.

I tilfeller som dette snakker vi også om et bilde, men omtaler det som “virtuelt bilde” i motsetning til “reelt bilde” som vi hittil har omtalt. Et virtuelt bilde kan ikke samles opp på en skjerm. Derimot kan vi betrakte det virtuelle bildet dersom vi bringer inn enda en linse på en slik måte at objektet alt i alt, etter å ha gått gjennom den nye linsen, danner et reelt bilde.

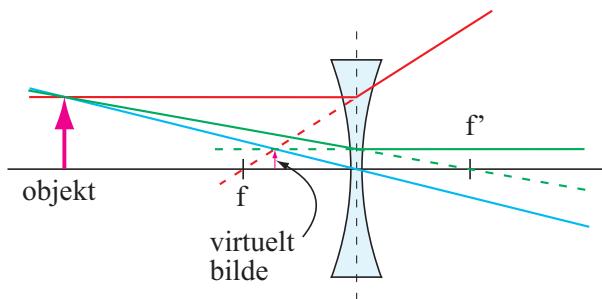
Dersom vi f.eks. betrakter lyset som kommer gjennom linsen i figur 12.9 ved hjelp av øynene våre, vil øyelinsen vår kunne samle lyset slik at det danner seg et reelt bilde på netthinnen. Da ser vi bildet. Bildet på netthinnen er et resultat av lys fra objektet går gjennom den frittstående linsen og deretter gjennom øyelinsen vår. Avstandene mellom objekt, linse og øyet er som gitt. Vi kunne imidlertid fått nøyaktig samme bilde på netthinnen dersom vi erstattet det virkelige objektet, med et forstørret objekt med størrelse og plassering som angitt ved “virtuelt bilde” i figuren, men nå uten noe ytre linse. For oss ser det altså ut som om vi (når vi ser gjennom den ytre linsen) ser på en forstørret gjenstand i en annen avstand enn den virkelige. Dette er grunnen til at vi snakker om “virtuelt” bilde.



Figur 12.9: Når et utstrakt objekt plasseres innenfor brennvidden til en konveks linse, dannes det ikke noe bilde på motsatt side av linsen. Tvert om indikerer hjelpe linjene at objektet og linsen synes å kunne erstattes av et forstørret objekt på samme side av linsen som det virkelige objektet. Dette tilsvarende, forstørrede objektet kalles et virtuelt bilde.

Konkav linse.

En konkav linse alene kan vi ikke danne noe reelt bilde for noen som helst posisjon av objektet (se figur 12.10). Konkave linser alene gir bestandig virtuelle bilder, og det er ofte litt uvant og krevende å jobbe med strålegang for konkave linser. Dersom vi skal bruke linseformelen, sier vi at brennvidden er negativ for konkave linser. Vi må også operere med negative objektavstander og negative bildeavstander alt etter om objekt og/eller bilde er på “vanlig” side av linsen eller ikke. Det finnes et sett fortegnsregler for hvordan vi skal behandle s , s' og f i linseformelen for alle kombinasjoner av tilfeller.



Figur 12.10: En konkav linse vil aldri kunne produsere et reelt bilde alene. Dersom vi betrakter et objekt gjennom en konkav linse, vil det imaginære bildet se mindre ut enn det virkelige objektet.

12.4.1 Fortegnsregler for linseformelen

Linsemakerformelen og linseformelen kan brukes for både konvekse og konkave linser og speil, men iblant må vi operere med negative verdier for posisjoner, krumningsradier og brennvidder for at formlene skal fungere.

Fortegnsreglene for lys som kommer inn mot linser eller speil er som følger:

- Objektavstanden $s > 0$ dersom objektet er et reelt objekt, $s < 0$ ellers.
- Bildeavstanden $s' > 0$ dersom bildet er reelt (virkelige lysstråler møtes i bildet), $s' < 0$ ellers.
- Brennvidden $f > 0$ for konvekse linser, $f < 0$ for konkave linser.
- Brennvidden $f > 0$ for konkave speil (hulspeil), $f < 0$ for konvekse speil.

I tillegg gjelder konvensjonen:

- Forstørrelsen m regnes som positiv når bildet har samme retning som objektet, $m < 0$ når bildet er opp-ned.

Fortegnsreglene er ofte greie å ha, men erfaring viser at de iblant er til mer forvirring enn til nytte. Av den grunn velger noen å bestemme hvilket fortogn størrelsene må ha ved å tegne opp lysgangen i lysstråleoptikk, få et omtrentlig mål for bildeavstand i forhold til objektavstand, og sjekke om bildet blir reelt eller imaginært. Derved gir fortegnet seg av seg selv. Framgangsmåten betinger likevel at vi kjenner fortegnsreglene for brennvidden til konvekse og konkave linser og speil.

Slavisk bruk av linseformelen og fortegnsregler uten samtidige tegninger basert på lysstråleoptikk, vil nesten garantert medføre dumme feil før eller siden!

12.5 Optiske instrumenter

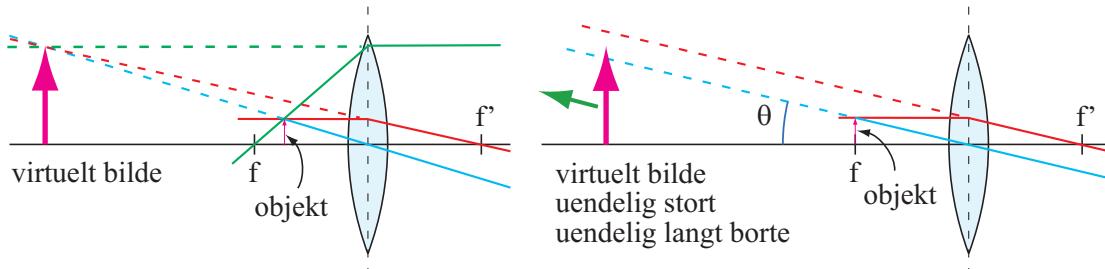
Flere linser ble satt sammen til optiske instrumenter på begynnelsen av 1600-tallet. Teleskopet åpnet opp for Galilei slik at han fikk sett månene rundt Jupiter, noe som fikk avgjørende betydning for utviklingen av vårt verdensbilde. Mikroskopet åpnet opp for studier av bakterier og celler, og åpnet opp for vår forståelse av biologiske systemer. Op-

tiske instrumenter har spilt og spiller fortsatt en enorm betydning for vår utforskning av naturen og også som et meget nyttig teknologisk hjelpemiddel.

Vi skal straks se på hvordan vi kan bygge opp et teleskop og mikroskop ved hjelp av to linser. Aller først skal vi imidlertid ta for oss en enkel linse brukt som lupe, siden denne konstruksjonen også inngår i såvel teleskop som mikroskop.

12.5.1 Lupen

En luke består i sin enkleste variant av en enkel konveks linse. Strålegangen ved en luke er noe annerledes enn vi har angitt i figurene hittil.



Figur 12.11: Når et objekt plasseres litt innenfor brennvidden, får vi et virtuelt oppreist bilde på samme side av linsen som objektet. Lar vi objektet nærme seg brennpunktet, flytter det virtuelle bildet lenger og lenger vekk fra linsen, samtidig som størrelsen av det virtuelle bildet øker. Vinkelen som toppen av objektet danner med optisk akse vil derimot gå mot en bestemt grense θ . Når objektet ligger i brennplanet, vil lysstrålene som stammer fra toppen av pilen og som virkelig passerer linsen, alle være parallelle (høyre del av figuren). De vil derfor se ut som om de kommer fra et objekt som er plassert uendelig langt borte.

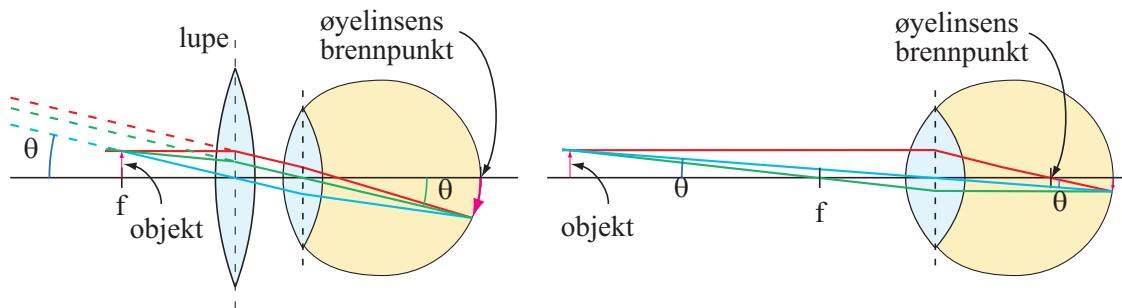
Figur 12.11 viser hvordan en linse ofte er plassert når den brukes som en luke. Det essensielle er at objektet plasseres på eller såvidt innenfor brennpunktet for linsen. De utvalgte lysstrålene vil da sprike svakt eller gå omtrent parallelt ut på høyre side av linsen. Vi får ikke dannet noe reelt bilde, men et imaginært bilde langt vekk fra linsen, på samme side som objektet. Til venstre i figur 12.11 har vi valgt å plassere objektet en del innenfor brennpunktet for at vi skal få plass til det virtuelle bildet innenfor figuren.

I høyre del av figuren er gjenstanden plassert på en mer vanlig måte, nemlig i brennpunktet (egentlig brennplanet). Det virtuelle bildet ligger da uendelig langt til venstre i figuren og lysstrålene fra et punkt på objektet er da parallelle med hverandre på høyre side av linsen.

En luke fungerer bare sammen med øyet vårt. Hva skjer når vi setter inn øyelinsen etter lupen? Jo, lyset fra toppen av pilen (i vår figur) vil komme inn til øyelinsen som parallele lysstråler. Øyet vil danne et reelt bilde på netthinnen (se figur 12.12), forutsatt at øyet innstiller seg på å betrakte ting som er langt unna (fokuserer på uendelig).

Bildets størrelse på netthinnen er proporsjonal med *vinkelutstrekningen* til de innfallende lysstrålene.

Vi har bare tegnet inn lysstrålene fra spissen av pilen. Lysstrålene fra bunnen av pilen vil fokusere på netthinnen der optisk akse skjærer netthinnen. Det synes altså som om pilen har en annen størrelse på netthinnen enn om vi hadde sett på pilen direkte.



Figur 12.12: Her vises lysstrålene som går gjennom lupen og videre inn gjennom øyelinsen og danner et reelt bilde på netthinnen. Når vi bruker en lupe slik at objektet er plassert i brennplanet (venstre side av figuren), vil alle lysstråler fra et vilkårlig valgt punkt (f.eks. toppen av pilen i figuren) komme ut parallelle etter lupen. Øyet vil da fokusere på uendelig, hvilket vil si at brennpunktet for øyelinsen ligger på netthinnen. Bildet av objektet spenner seg ut på netthinnen (her markert ved en vinkel θ). Dersom vi skulle betrakte objektet uten lupe, måtte objektet flyttes vekk en avstand $d \approx 25$ cm fra øyet for at vi skulle se et skarpt bilde (høyre del av figuren). Øyelinsen vil da krumme seg så kraftig den kan, og brennpunktet flytter seg inn mot øyets sentrum. Vi får da en vanlig bildedannelse hvor bare en konveks linse inngår, og resultatet blir igjen et reelt bilde på netthinnen. Bildets størrelse blir likevel mindre enn det var da lupen var i bruk. Legg merke til hvordan vinklene θ kommer inn i de to tilfellene.

Forstørrelsen.

Hvordan ville pilen sett ut dersom vi betraktet den direkte? Vel, det kommer an på hvor nært vi kunne ha pilen til øyet og fortsatt kunne fokusere skarpt på den. Et “normaløye” (se siden) kan ikke fokusere på objekter nærmere enn ca 25 cm. Størrelsen på netthinnen for pilen plassert 25 cm fra øyet blir da det største bildet vi kan få på netthinnen uten hjelpe midler.

Forstørrelsen lupen gir blir forholdet mellom tangens til vinklene lyset har når det kommer inn mot øyet via en lupe, og vinkelen lyset har når det kommer direkte fra objektet i $d = 25$ cm avstand uten lupe. For en lupe med brennvidde f blir forstørrelsen:

$$M = \frac{h/f}{h/d} = \frac{d}{f}$$

Her er h høyden på objektet. En lupe som har en brennvidde på 5 cm vil da ha en forstørrelse på $25 \text{ cm} / 5 \text{ cm} = 5$. Vi skriver gjerne 5 X (fem ganger forstørrelse). Merk forøvrig at forstørrelsen her er positiv, fordi bildet vi ser gjennom lupen har riktig retning i forhold til objektet.

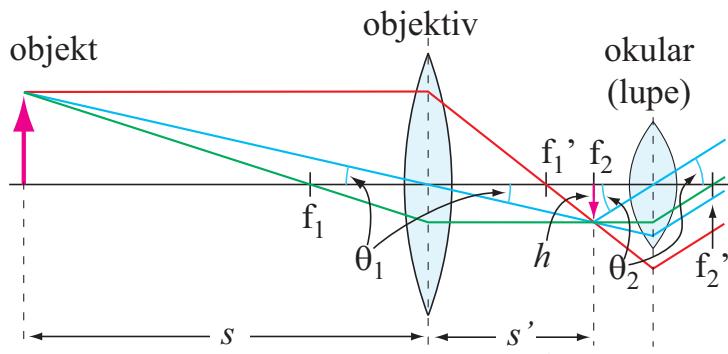
Kort sagt kan vi si at lupen bare har den funksjonen at objektet kan flyttes nærmere øyet vårt enn om vi ikke hadde lupen der. Den effektive avstanden er rett og slett lupens brennvidde. Har vi en lupe med brennvidde 2.5 cm, vil vi kunne betrakte en sommerfuglvinge i en effektiv avstand 2.5 cm i stedet for å måtte flytte sommerfuglvingen 25 cm vekk fra øyet for å få et skarpt bilde av den. Effekten er et ca ti ganger så stort reelt bilde på netthinnen når vi bruker lupen sammenlignet med uten.

I et mikroskop eller teleskop brukes en lupe sammen med enda en linse. Lupen kan ha

brennvidder på ned til ca 3 mm. Det gir automatisk en bortimot 100 ganger forstørrelse sammenlignet med om vi ikke hadde benyttet oss av lupen.

12.5.2 Teleskopet

Et teleskop består av minimum to linser (eventuelt minst ett krumt speil og en linse). Linsen (eller speilet) som er nærmest objektet kalles et *objektiv*, mens linsen som er nærmest øyet kalles et *okular*. Objektivets rolle er å lage en lokal avbildning av objektet (på en måte flytte objektet mye nærmere oss enn det egentlig er). Okularet brukes som en lupe for å betrakte den lokale avbildningen. Selv om den lokale avbildningen nesten alltid er *mye* mindre i utstrekning enn objektet, er det også mye nærmere øyet enn objektet selv. Når vi at på til kan bruke en lupe når den lokale avbildningen betraktes, kan vi få en (angulær) forstørrelse på opp til flere hundre ganger. En vanlig prismekikkert har dog en begrenset forstørrelse på om lag 5 - 10 X. Kikkerter med større forstørrelser krever at vi har et stødig stativ for at bildet ikke skal hoppe og sprette sjenerende mye i synsfeltet.



Figur 12.13: For et teleskop er objektet langt borte sammenlignet med brennvidden. Objektivet lager et reelt, forminsket "lokalt" bilde litt bakenfor brennvidden. Dette bildet betraktes så med en lupe. Forstørrelse totalt måles som vinkelforstørrelse til objektet sett gjennom kikkerten sammenlignet med uten kikkert.

I figur 12.13 er det tegnet en prinsippskisse for et teleskop. Vi bruker standard valg av lysstråler fra objektets største vinkelavstand fra optisk akse (fra toppen av pilen). Punkter i objektet som ligger på optisk akse vil bli avbildet på optisk akse, og vi tegner vanligvis ikke inn disse linjene.

Vi merker oss at objektivet gir et reelt opp-ned bilde litt lengre vekk fra linsen enn brennplanet. Objekter som er meget langt unna, vil avbildes temmelig nær brennplanet. Objekter som er nærmere faller lengre og lengre utenfor brennpunktet til objektivet.

Okularet plasseres slik at bildet fra objektivet faller i okularets brennplan. Da vil alle lysstråler fra et valgt punkt i objektet, etter det har gått gjennom okularet, komme ut parallelle. Øyet vil fokusere på uendelig og det dannes et reelt bilde på netthinnen.

Forstørrelsen.

Forstørrelsen til teleskopet er gitt som forholdet til tangens til vinkler mellom optisk akse og lysstrålene som går gjennom sentrum av linsene.

Fra figur 12.13 ser vi at vinkelforstørrelsen kan defineres som:

$$M = -\frac{\tan \theta_2}{\tan \theta_1} = -\frac{h/f_2}{h/s'}$$

Forstørrelsen varierer med andre ord med bildeavstanden fra objektivet til det reelle bildet som ligger mellom objektiv og okular. Denne avstanden vil variere alt etter hvor nær objektet er objektivet. Det er mer hensiktsmessig å angi forstørrelsen som ett tall. Det oppnås ved å velge forstørrelsen når objektet er uendelig langt unna. Da er s uendelig, og s' blir lik brennvidden til objektivet f_1 . Forstørrelsen kan da skrives:

$$M = -\frac{h/f_2}{h/f_1} = -\frac{f_1}{f_2}$$

Vinkelforstørrelsen er med andre ord lik forholdstallet mellom brennviddene til objektiv og okular.

For et teleskop med brennvidde $f_1 = 820$ mm og okular med brennvidde $f_2 = 15$ mm, blir vinkelforstørrelsen:

$$M = \frac{820}{15} = 54.7 \approx 55X$$

Merk at siden det er så mange tilnærminger som gjøres i den enkle varianten av geometrisk optikk, har det ingen hensikt å angi f.eks. en forstørrelse med mer enn to gjeldende siffer.

Okularprojeksjon.

Før vi forlater teleskopet skal vi nevne en nyttig liten detalj. Det er vel og bra å kikke gjennom et teleskop eller et mikroskop, men i dag ønsker vi ofte å kunne dokumentere detaljer som observeres slik at andre også kan se dem. Siden okularet normalt virker som en lupe, kan vi ikke uten videre fange opp noe reelt bilde ved å plassere f.eks. en CMOS-brikke eller en gammeldags film et eller annet sted bak lupen. En mulighet er å fjerne hele okularet og plassere CMOS-brikken akkurat der det reelle bildet danner seg. Men det vi da egentlig har gjort er å lage et vanlig fotoapparat, med telelinse dersom brennvidden på objektivet er lang nok. For teleskopet vi nevnte ovenfor vil vi da få et kamera med en 820 mm telelinse. CMOS-brikken kan være fra et speilreflekskamera med utskiftbar optikk. Vi kan da fjerne det vanlige objektivet og bruke teleskopets objektiv i stedet.

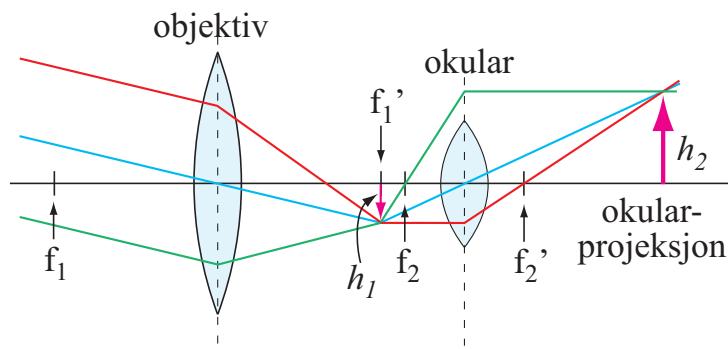
Anta at vi ønsker å ta bilder av månen. Vinkeldiameteren til månen er ca. en halv grad. Størrelsen på det reelle bildet som dannes av teleskopobjektivet med 820 mm brennvidde, vil da bli:

$$h = 820 \text{ mm} \cdot \tan(0.5^\circ) = 7.16 \text{ mm}$$

Dersom CMOS-brikken er 24 mm i minste utstrekning, betyr det at månen vil dekke $7.2/24 = 0.3$ av denne dimensjonen. Bildet av hele månen får en diameter på kun 30% av bildets minste dimensjon ("høyde"). Det er da umulig å få med seg fine detaljer fra måneoverflaten selv om eksponeringen av bildet er optimal.

Finnes det en mulighet for at bildet av månen kan blåses opp slik at vi f.eks. kan ta et utsnitt av måneoverflaten? Ja, det er mulig. Vi bruker da såkalt "okularprojeksjon".

Prinsippet er ganske enkelt. Normalt brukes okularet som en lupe, og da plasseres bildet fra objektivet i brennplanet for okularet. Skyver vi okularet lenger bort fra objektivet,



Figur 12.14: Ved okularprojeksjon brukes okularet ikke som en lupe, men som en avbildende linse nr 2. Se tekst for detaljer.

vil det reelle bildet ligge utenfor brennplanet, og da vil okularet faktisk lage et nytt reelt bilde ved å bruke det første reelle bildet som sitt eget objekt. Figur 12.14 viser prinsippet. For at det nye reelle bildet skal bli større enn det første reelle bildet, må okularet bare skyves *litt* lengre vekk enn normalt fra objektivet. Vi kan da i prinsippet få så stort reelt bilde nummer to vi vil, men avstanden fra okularet til dette siste reelle bildet står i forhold til størrelsen på bildet. Vi må da ha et egnet stativ for å holde CMOS-brikken på plass et stykke bak okularet. Med en slik teknikk kan vi lett ta bilder av utsnitt av måneoverflaten fra f.eks. teleskopet med brennvidde 820 mm.

Det er imidlertid en catch ved metoden. Objektivet fanger opp like mye lys uansett om vi bruker okularprojeksjon eller ikke. Når lyset spres ut over en større flate, betyr det at lysstyrken per pixel på CMOS-brikken avtar til dels betydelig. Eksponeringen må da skje over en lengre tid for å få et brukbart bilde. Det bør legges til at okularer vanligvis optimaliseres for normalt bruk. Linsefeil kan dukke opp ved okularprojeksjon som vi ellers ikke legger merke til.

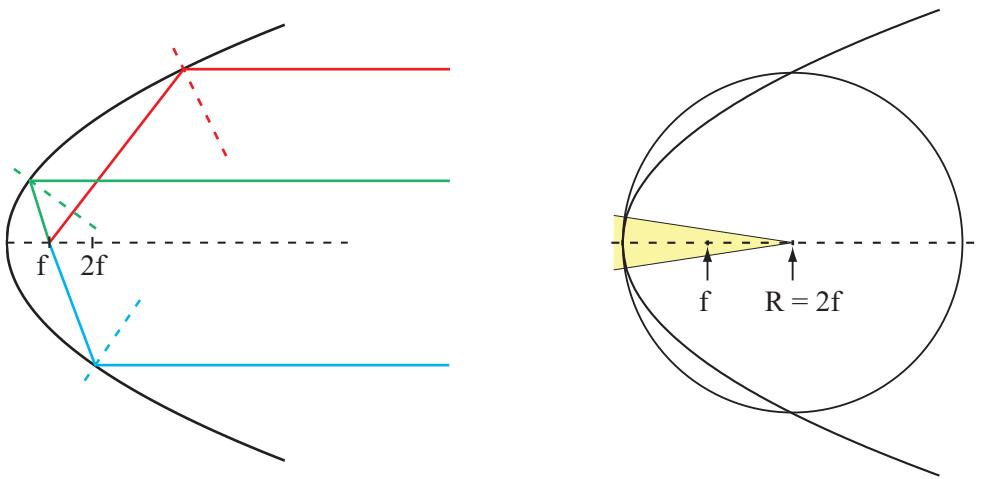
Okularprojeksjon kan brukes også ved mikroskopering.

12.5.3 Speilteleskop

Store astronomiske teleskop benytter som oftest krumme speil som objektiv. Den vesentligste grunnen for dette er at refleksjonslovene for et speil ikke er bølgelengdeavhengige. Lys med lang bølgelengde oppfører seg omtrent likt med lys med kort bølgelengde, og vi slipper da det kromatiske avviket som skyldes at brytningsindeksen for glass er bølgelengdeavhengig.

I likhet med linser er det lettest å lage krumme speil når overflaten er kuleformet. Denne formen er likevel ikke god fordi parallelle lysstråler med optisk akse vil fokuseres på ulikt sted alt etter hvor langt fra aksen lysstrålene kommer inn. Matematisk ville det vært langt bedre å velge en overflate som har form som en paraboloide. I venstre del av figur 12.15 er det vist eksempler på tre ulike lysstråler som kommer inn mot et parabolisk speil parallelt med optisk akse. Strålene blir reflektert ifølge refleksjonslovene, og ender opp i nøyaktig samme punkt (brennpunktet). En kikkert med et slikt parabolisk speil som objektiv, kan få meget skarpe bilder og samtidig svært høy lysstyrke (se siden).

Dessverre er det komplisert å lage kikkerter med parabolisk overflate med den presisjonen som trengs for lysbølger (siden bølgelengden er så liten). I de fleste tilfeller velges derfor speil med sfærisk overflate, men da med en svært liten åpningsvinkel i forhold til



Figur 12.15: Venstre del: Et parabolspeil sikrer at alle lysstråler som kommer inn parallellt med optisk akse bli fokusert i samme punkt, uansett om strålene ligger nær eller lengre vekk fra optisk akse. Høyre del: Et parabolspeil og et sfærisk speil har nært samme form forutsatt at "åpningsvinkelen" er liten, dvs at speilets diameter er liten sammenlignet med brennvidden.

radien (se høyre del av figur 12.15). Da er det nemlig ikke så stor forskjell mellom parabelform og kuleform. Alternativt kan vi kombinere et sfærisk speil med en korreksjonslinse av glass for å få et bra totalresultat til en lavere pris enn om vi skulle laget et nært perfekt parabolsk speil.

Konstruksjonsregler for speil.

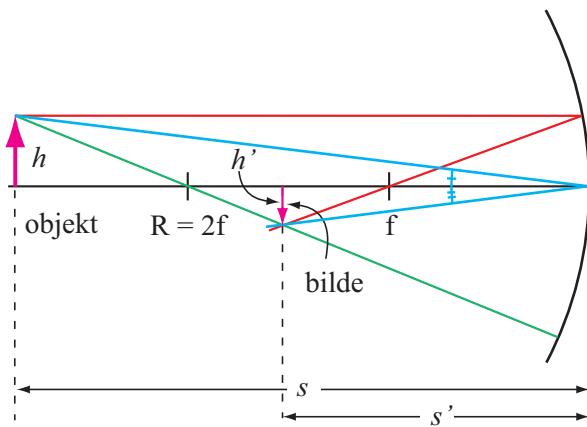
Vi kan konstruere bildedannelsen ved et krumt speil på temmelig lik måte som for tynne linser. Vi kombinerer egenskaper med sfærisk og parabolsk form for å gjøre reglene så enkle som mulig, og får:

1. Lys som kommer inn parallellt med den optiske akse, vil reflekteres gjennom brennpunktet.
2. Lys som treffer speilets midtpunkt (der optisk akse skjærer gjennom speilet), vil reflekteres med samme vinkel til optisk akse som det kom inn.
3. Lys som passerer speilets krummingssentrum (to ganger brennvidden) *foran* linsen, vil reflekteres tilbake samme vei som den innkommende strålen kom fra.

I figur 12.16 er det vist bildedannelsen for et konkavt speil der objektet er litt utenfor to ganger fokallengden. Legg merke til alle detaljer knyttet til hvordan hjelpeelinjene er trukket.

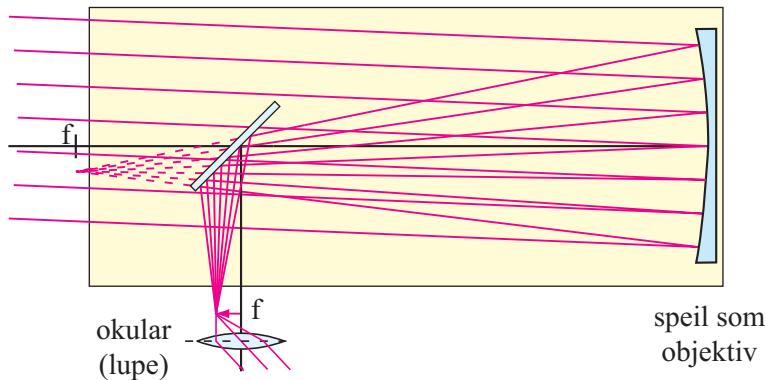
Vi kan bruke linseformelen også for et speil, men må da være ekstra påpasselig med å vurdere fortegn for å komme riktig ut.

Et hulspeil (konkavt speil) vil danne et reelt bilde av objektet forutsatt at objektet plasseres lengre vekk fra speilet enn én brennvidde. Et problem med speil er at bildet danner seg i samme område som det innfallende lyset går gjennom. Setter vi opp en skjerm for å fange opp bildet, vil det for det første fjerne lys som når speilet, og for det andre får vi diffraksjonseffekter på grunn av kanten mellom lys og skygge (se siden). Det finnes flere triks for å minimalisere ulempene så mye som mulig. Et av de klassiske triksene



Figur 12.16: Eksempel på konstruksjon av bildedannelse for et konkavt speil.

er å sette inn et skråspeil som reflekterer strålebunten bort fra det området lyset kommer inn (se figur 12.17). Et teleskop av denne typen kalles en Newton reflektor.

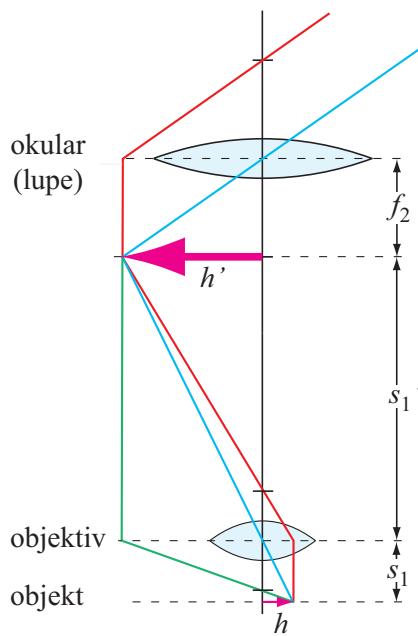


Figur 12.17: I en Newton-reflektor brukes et skråspeil for å bøye av lysbunten fra hovedspeilet slik at vi kan bruke et okular og ikke på stjerner uten å selv komme i veien for innkommende lys. Skråspeilet tar likevel en del av dette lyset.

12.5.4 Mikroskopet

I teleskopet brukte vi objektivet for å lage et lokalt bilde av objektet, og dette bildet ble betraktet gjennom en lupe. Det er en strategi som fungerer bra når objektet er så langt unna at vi ikke kan komme nær det. Det er nettopp i slike situasjoner vi har bruk for et telekop.

Når vi skal betrakte f.eks. cellene i en plantestengel, har vi objektet rett foran oss. Vi trenger ikke å lage noe lokalt bilde, for vi har originalen. Da bruker vi en annen strategi for å kunne se et forstørret bilde. Strategien er egentlig nøyaktig den samme som for okularprojeksjon. Vi plasserer objektet like utenfor brennpunktet til objektivet (som nå har liten brennvidde) for å danne et reelt opp-ned bilde et godt stykke bakenfor objektivet. Dette forstørrede bildet av objektet betraktes så med en lupe. Strålegangen i et mikroskop er illustrert i figur 12.18.



Figur 12.18: Strålegangen i et mikroskop. Objektet kan plasseres vilkårlig nær brennpunktet til objektivet, følgelig gir objektivet et reelt forstørret bilde av objektet. Dette bildet betraktes så med okularret som fungerer som en lupe.

Forstørrelsen som følger av objektivet alene er gitt som:

$$M_1 = \frac{s'_1}{s_1}$$

Denne forstørrelsen kan i prinsippet gjøres vilkårlig stor, men da vil også det reelle bildet forflytte seg langt fra objektivet, og mikroskopet ville bli uhåndterlig stort. Ved å bruke et objektiv med meget kort brennvidde, gjerne bare noen få mm, kan vi oppnå en betydelig forstørrelse selv for en tubelengde (avstand mellom objektivet og okularret) på 20-30 cm. Lupen gir i tillegg en forstørrelse slik luper gjør, nemlig:

$$M_2 = \frac{25 \text{ cm}}{f_2}$$

Den totale forstørrelsen til mikroskopet blir da:

$$M_{tot} = \frac{25 \text{ (cm)} s'_1}{f_2 s_1}$$

For et 8 mm objektiv og en tubelengde på 30 cm, samt et okular med brennvidde 10 mm, blir den totale forstørrelsen (måltall i mm i mellomregning):

$$M = \frac{25 \text{ (cm)} s'_1}{f_2 s_1} \approx \frac{250 \cdot (300 - 10)}{10 \cdot 8} = 906 \approx 900 \times$$

12.5.5 Bildekvalitet

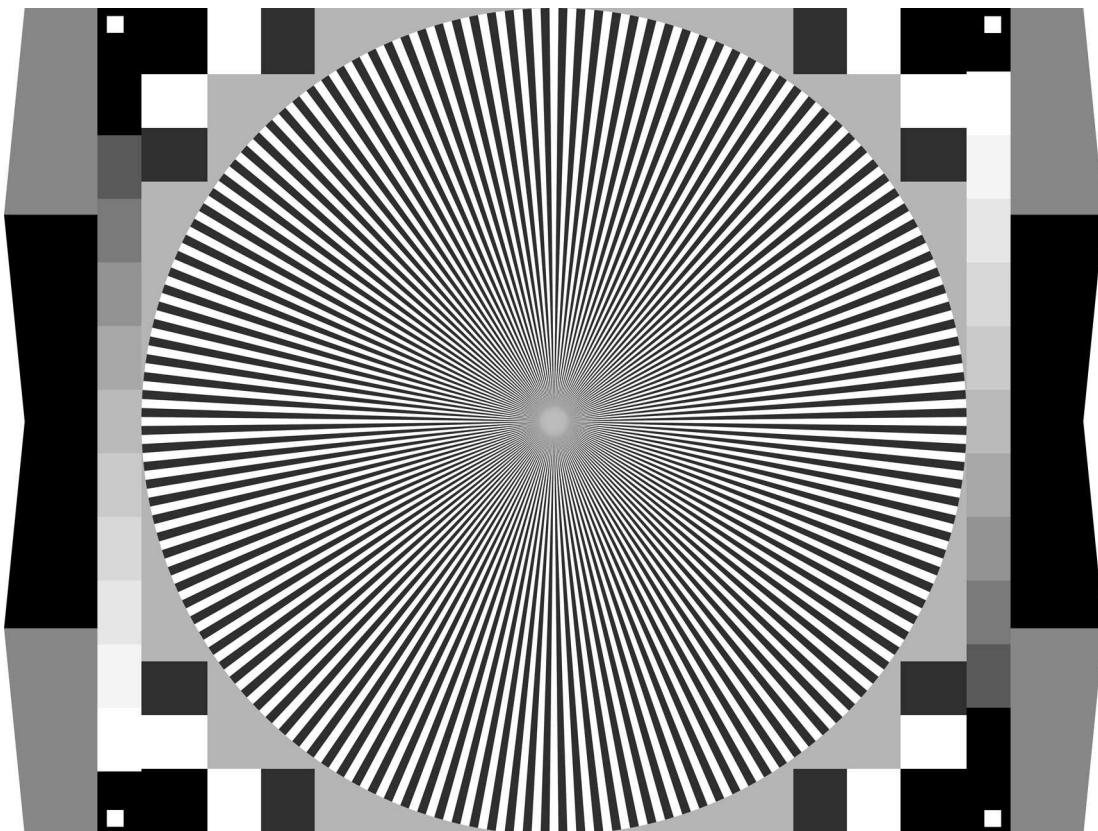
Her må det likevel inn et advarende ord. Kjøper vi et mikroskop (eller teleskop for den saks skyld), kan vi gjerne få billige mikroskop med like stor forstørrelse som dyre instrumenter. Forstørrelsen i seg selv er egentlig langt mindre viktig enn bildekvaliteten. I alt for lang tid

har det vært et problem at bildekvaliteten ikke kunne angis i noe vel etablert og utbredt system. Det var derfor rom for lureri i stor grad, og mange har kjøpt både mikroskop og kikkerter som bare var penger ut av vinduet fordi bildekvaliteten var for dårlig. Enn så lenge er det slik at går du til en optiker og skal kjøpe en kikkert, er det i høy grad bare en vag subjektiv synsing om kvaliteten vi kan støtte oss til. Det er frustrerende!

Heldigvis er dette i ferd med å endre seg. Et mål for optisk kvalitet som nå synes å få fotfeste er å angi måleresultater basert på en såkalt "Modulation Transfer Function" (MTF). Dette er først og fremst en metode for å bestemme hvor skarpe bilder vi kan få. Fargegjengivelse blir ikke vurdert ved denne målemetoden.

Kort fortalt forteller MTF-verdiene oss om hvor tett linjene i et sort-hvitt stripemønster kan ligge, før det sorte og hvite flyter mye over i hverandres områder. Når stripene ligger tett, vil stripemønsteret bare bli mer og mer et gråtone-stripemønster, i stedet for sort og hvitt, og for de tettleste linjene forsvinner stripemønsteret helt.

Det er utviklet flere testbilder som kan brukes for å bestemme MTF-verdier og dermed si litt om det optiske systemets kvalitet med hensyn på kontrast og oppløsning. I figur 12.19 er det gitt ett eksempel hvor stripemønsteret blir tettere og tettere jo nærmere sentrum i sirkelen man kommer. Har vi et ark med stripemønsteret skrevet ut i høy kvalitet (høy oppløsning), kan vi i prinsippet f.eks. betrakte mønsteret gjennom et teleskop og se hvor fine stripedetaljer vi kan oppdage. Vi kommer tilbake til denne problemstillingen siden i kurset, men da med et litt annet testobjekt.



Figur 12.19: Et av flere populære testobjekt for å måle kvaliteten på et optisk system. Stjernen brukes ved test av oppløsning, mens gråtone-trinnene i ytterkantene kan brukes for å teste om optikken gir en god gjengivelse av ulike lyshetsgrader. Testobjektet var gratis tilgjengelig fra <http://www.bealecorner.org/red/test-patterns/star-chart-bars-full-600dpi.png> den 6. april 2013 (originalen har langt bedre oppløsning enn vår gjengivelse her).

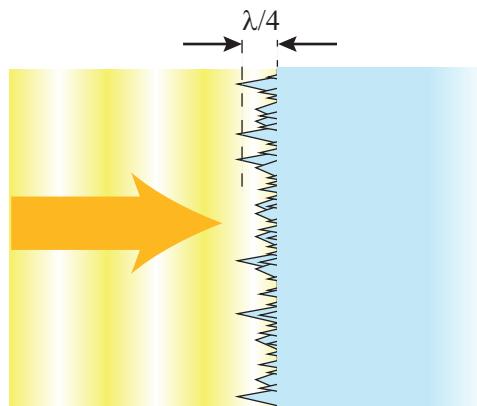
Det er mange grunner til at kvaliteten i et optisk system kan ødelegges. Diffraksjon, som jo skyldes at lyset har en bølgenatur, vil alltid spille inn. Diffraksjon vil imidlertid bare gi en begrensning for svært gode optiske systemer. De fleste systemer har alvorligere kilder til forringelse av optisk kvalitet enn diffraksjon.

For å unngå sfærisk og kromatisk avvik i linser, er objektiver og okularer i dag ofte sammensatt av flere (mange) linselementer (se figur 12.6). Vi vet fra tidligere kapitler, at når lys går fra luft til glass, reflekteres ca 5 % av intensiteten i overflaten. Kommer lyset på skrå inn mot overflaten, kan refleksjonen gjerne være enda større (for en viss polarisering, slik vi så det i Fresnels ligninger).

Reflekteres 5 % i enhver ytre og innvendige glassflate i et objektiv som består av f.ek. åtte elementer, vil ganske mye lys gå fram og tilbake flere ganger mellom elementer og ha en tendens til å ødelegge skarphet og kontrast.

Vi har i mange år redusert dette problemet ved å bruke antirefleksbelegg på overflaten til glasset. Da kan vi få redusert refleksjonen mye. Problemet er imidlertid at slik behandling både er avhengig av bølgelengden og av vinkelen lyset treffer overflaten med. Antirefleksbehandling av denne typen gir en vesentlig forbedring av bildekvaliteten, men behandlingen er ikke så god som vi kunne ønske for systemer så som kameraer og kikkerter hvor lys med mange bølgelengder betraktes samtidig.

Siden om lag 2008 har denne situasjonen endret seg dramatisk til det bedre, og det er morsom fysikk som ligger bak! Nikon kaller deres variant av teknikken for "Nano Crystal Coating", mens konkurrenten Canon kaller den for "Subwavelength Structure Coating". Figur 12.20 viser hovedprinsippet.



Figur 12.20: En skjematiske figur som viser prinsippene for strukturene som ligger bak den nye typen antirefleks-behandling basert på nanoteknologi.

Da vi i et tidligere kapittel beregnet hvor mye lys som ville bli reflektert og transmittert ved en grenseflate mellom luft og glass, var vår bruk av Maxwells ligninger basert på noen antakelser. For å sette ting på spissen, sa vi at grensesjiktet måtte være "uendelig glatt, plant og vidt" og "uendelig tynt" i forhold til bølgelengden. Da ble integrasjonen lett å gjennomføre, og vi fikk de svarene vi fikk. Vi hevdet at betingelsene kunne oppfylles ganske bra f.eks. på en glassoverflate, siden atomene er så små i forhold til bølgelengden.

Det nye konseptet som nå er tatt i bruk baserer seg på "nanoteknologi", som i vår sammeheng betyr at vi lager og bruker strukturer som er litt mindre enn bølgelengden for lys.

På overflaten til glasset legges det på et sjikt som har en uordnet struktur med elementer som i størrelse langs sjiktet er mindre enn bølgelengden, og som bygger en tykkelse på

laget på omtrent en kvart bølgelende (ikke like kritisk som for de tradisjonelle antirefleksbeleggene). Et slikt sjikt virker helt vilt når vi tenker tradisjonelt. Vi ville tro at lyset ville bli brutt i hytt og vær når det går ned gjennom de skråstilte overflatene. Men slik er det ikke! Lyset slik vi behandler det er elektromagnetiske bølger som er utstrakt i tid og rom. På en måte kan vi si at bølgen ser mange av de små strukturene *samtidig*, og detaljer mindre enn bølgelengden vil ikke kunne spores hver for seg når bølgen har forplantet seg flere bølgelengder videre.

En annen måte å beskrive fysikken bak disse nye antirefleks-beleggene på, er å si at overgangen fra luft til glass skjer gradvis over en avstand ca en kvart bølgelengde. Da blir refleksjonen kraftig redusert.

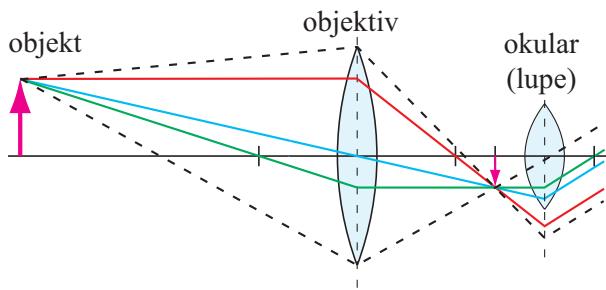
Det bør legges til at siden atomene fortsatt er små i forhold til nanokrystallene som brukes, kan vi fortsatt bruke Maxwells ligninger for å se hva som skjer når en elektromagnetisk bølge treffer en linse med nanokrystaller på overflaten. Er f.eks. strukturene 100 - 200 nm store, er det om lag 1000 atomer i lengderetningen for disse krystallene. Alle beregninger på slike systemer krever bruk av avanserte numeriske metoder.

Nanokrystall-belegget er så vellykket at om lag 99.95 % av lyset slipper gjennom, og bare 0.05 % blir reflektert. Denne nye teknologien brukes i dag på alle dyre objektiver fra Canon og Nikon, og har forbedret bildekvaliteten betraktelig.

12.5.6 Synsfelt

Hittil har vi tegnet de tre hjelpe linjene fra objekt til linseplan til bildeplan uten å bekymre oss om linjene går utenfor eller innenfor selve linseelementene. Det er greit så lenge vi bare er interessert i hvor bildet dannes og hvilken forstørrelse det har. Lyset fra et objekt følger alle mulige vinkler, og så snart vi har etablert hvor bildet er plassert og hvor stort det er, kan vi fylle på med så mange ekstra lysstråler vi måtte ønske. Vi har da nok opplysninger om hvordan linjene må trekkes.

På dette tidspunktet er det meningsfylt å vurdere hvilke lysstråler som faktisk vil bidra til det endelige bildet vi f.eks. ser gjennom et teleskop. Ut fra denne type betraktninger kan vi bestemme hvilken bildevinkel f.eks. et teleskop eller mikroskop vil gi.



Figur 12.21: Når vi først har etablert de tre hjelpe linjene for å vise bildedannelse, kan vi fylle inn med alle tenkelige lysstråler som faktisk går gjennom en linse. Når flere linser kombineres, vil ikke nødvendigvis alt lys som kommer gjennom første linsen gå gjennom neste. Denne type betraktning kan gi et omtrentlig mål for hvilken bildevinkel et teleskop eller mikroskop vil ha.

Figur 12.21 gir en indikasjon på hvordan dette fungerer i praksis. De stiplete linjene indikerer yttergrensene for hvilke lysstråler fra pilens spiss som fanges opp av objektivet,

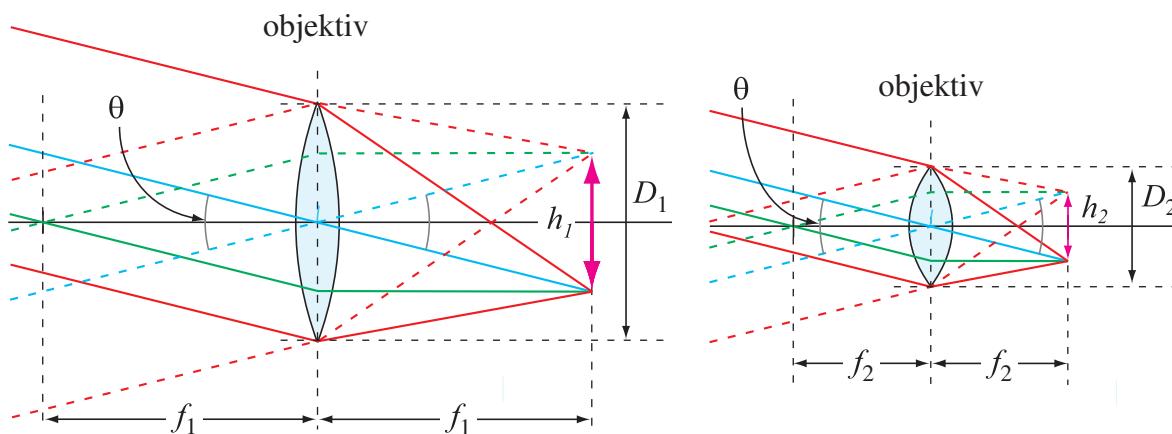
og hvordan disse utvikler seg videre. I dette tilfellet ser vi at bare vel halvparten av lyset som objektivet fanger inn vil gå videre gjennom okularet. Dersom objektet hadde en enda større vinkelutstrekning, ville vi kunne risikere at ikke noe lys fra de ytterste delene av objektet vil nå okularet, selv om det faktisk går en del lys gjennom objektivet. Ved å gjøre denne type analyse, kan den maksimale bildevinkelen til f.eks. et teleskop bestemmes, forutsatt at vi faktisk kjenner diametern til såvel objektiv som okular.

I praksis er det ikke fullt så enkelt, fordi det brukes objektiver og okularer som er sammensatt av mange linser for å redusere sfærisk og kromatisk feil m.m. Likevel kan denne type betraktnign gi et omtrentlig mål for synsfeltet.

Det bør også legges til at ulike konstruksjoner av okularer gir ganske forskjellige opplevelser når vi ser gjennom f.eks. et teleskop. I gamle dager måtte vi holde øyet i en ganske bestemt avstand fra det nærmeste elementet i okularet for å se noe som helst, og det vi så var stort sett sort, bortsett fra et lite rundt felt der motivet befant seg. I dag gir gode okularer mye mer spillerom i hvor vi må holde øyet i forhold til okularet for å se noe bilde ("eye relief" på engelsk, på opp til 10 mm), og bildet vi ser fyller mer eller mindre hele det effektive synsfeltet til øyet. Det er ikke noe sort område utenfor som vi legger merke til uten å kikke aktivt etter det. Når vi ser gjennom slike okularer får vi inntrykk at vi ikke ser gjennom noe teleskop i det hele tatt, men bare rett og slett *er* der bildet viser. I astronomi snakker vi om en følelse av "space walk" når slike okularer benyttes.

12.5.7 Lysstyrke, blendertall

Alle har vel opplevd enkelte kikkerter der bildet er lyst og fint, men andre kikkerter hvor bildet er mye mørkere enn det vi forventet. Hva er det som bestemmer lysstyrken på bildet vi ser gjennom en kikkert?



Figur 12.22: En linse fanger opp en begrenset mengde lys fra et objektet, og denne lysmengden blir spredt ut over arealet til bildet som dannes. I denne figuren antas objektet å være svært langt unna slik at bildet dannes i brennplanet. Sammenligning mellom venstre og høyre del: Dersom linsens diameter reduseres til det halve, samtidig som brennvidden reduseres til det halve, vil lysintensitetstettheten på bildet som dannes være uendret. Se tekst for detaljer.

I figur 12.22 er det tegnet inn en enkel linse med objekt langt unna, sammen med bildet som danner seg tilnærmet i brennplanet. Når den totale bildevinkelen som objektet utspenner er θ , brennvidden for linsen f , og utstrekningen av bildet i bildeplanet er h_1 ,

følger:

$$\tan(\theta/2) = \frac{h_1/2}{2}$$

$$h_1 = 2f \tan(\theta/2) \quad (12.8)$$

Avbilder vi f.eks. månen, er vinkeldiameteren om lag en halv grad. Dersom brennvidden er 1 m, vil bildet linsen danner ha en diameter på 8.7 mm. Hvor mye lys samles i månebildet i brennplanet? Det avhenger av hvor mye lys vi faktisk fanger inn av lyset som sendes ut fra Månen. Når lyset når linsen, har det en innstrålingstetthet (irradians) S gitt f.eks. i antall mikrowatt per kvadratmeter. Total strålingseffekt som samles opp av et objektiv med diameter D er da $S\pi(D/2)^2$ (i antall mikrowatt). Den totale strålingseffekten vil fordele seg utover bildet av månen i brennplanet, slik at:

$$S\pi(D/2)^2 = S_b\pi(h_1/2)^2$$

Irradianse S_b i bildeplanet blir:

$$S_b = \frac{\pi(D/2)^2}{\pi(h_1/2)^2}$$

Bruker vi ligning (12.8) og omgrupperer leddene litt, får vi:

$$S_b = \frac{S}{4 \tan^2(\theta/2)} \left(\frac{D}{f} \right)^2$$

hvor θ er vinkeldiameter for månen og f og D er hhv. brennvidden og diameteren for linsen.

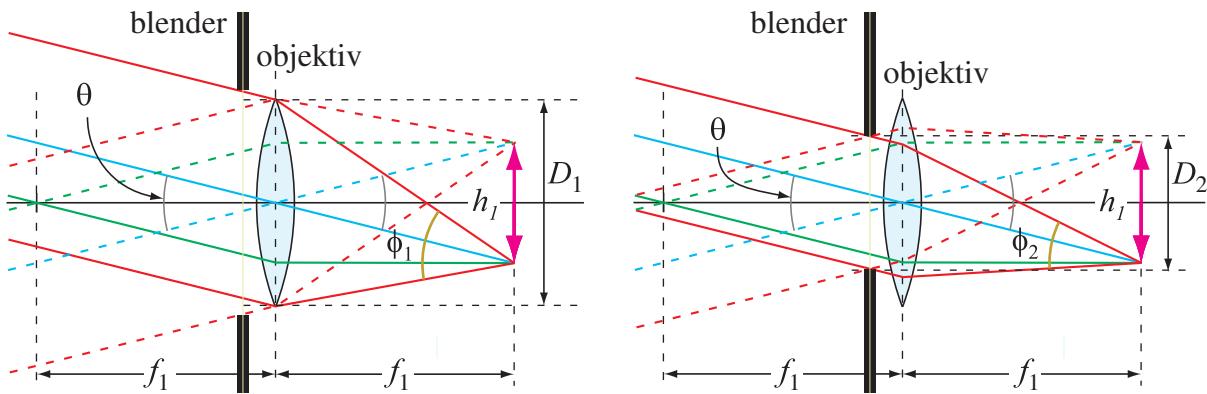
Det første ledet har bare med lyskilden å gjøre, mens det siste ledet har bare med linsen å gjøre. Jo større forholdet D/f er, desto mer intens lysstyrke får bildet som linsen danner (og det er dette bildet som eventuelt siden betraktes ved hjelp av et okular eller samles opp på en CMOS-brikke eller liknende).

I figur 12.22 er det tegnet inn to ulike linser, med ulik radius og ulik brennvidde. Dersom brennvidden går ned til det halve (høyre del av figuren), vil bildets størrelse (diameter) bli halvparten så stor som i venstre del. Men dersom linsens diameter også går ned til det halve, vil arealet som kan fange inn f.eks. lys fra Månen, gå ned til fjerdeparten. Men når *arealet* av bildet også har gått ned til fjerdeparten, betyr det at irradiansen i bildeplanet er identisk med hva vi har i venstre del av figuren. Forholdstallet D/f er det samme i begge tilfeller. Det har derfor vist seg at forholdet D/f er et mål for lysstyrken til bildet en linse danner.

Dersom vi setter inn en film eller en CMOS-brikke i brennplanet og fanger opp bildet av f.eks. Månen, vil vi måtte samle inn lys i en viss tid for å få en passe eksponering. Dersom linsen har stor lysstyrke, vil vi trenge kortere tid enn dersom linsen har liten lysstyrke. Et teleskop med stor diameter på objektivet (eller speilobjektivet), vil kunne fange inn svakt lys fra fjerne galakser mye lettere enn et teleskop med liten diameter. Det er imidlertid ikke diameteren i seg selv som avgjør dette, men lysstyrken angitt i forholdet D/f .

f-tallet.

I fotoapparater brukes en blender, som rett og slett er en nær sirkulær åpning hvor vi kan endre diametern. Ved hjelp av en blender kan vi endre hvor mye lys som skal slippe inn på filmen eller CMOS-brikken. Dette er antydet i figur 12.23. Bildets størrelse



Figur 12.23: Dersom en blender brukes for å redusere lysintensiteten i bildet av et objekt, vil størrelsen av bildet hele tiden være uendret – bare lysstyrken går ned. Forholdstallet mellom brennvidde og diameter for lysbunten som slipper inn i linsen, angir det såkalte blendertallet, eller f-stop på engelsk.

endres ikke om vi blander ned for å få mindre lys inn på CMOS-brikken, men irradiansen i bildeplanet vil gå ned. Lysstyrken på et kameraobjektiv blir gjerne angitt i et såkalt f-tall som er inversverdien av lysstyrken slik vi har definert den. f-tallet er gitt som:

$$f - tall \equiv f/D$$

Typiske f-tall er 1.4, 2, 2.8, 4, 5.6, 8, 11, 16, 22. Dersom vi tar kvadratet av disse tallene får vi tilnærmet: 2, 4, 8, 16, 32, 64, 128, 256, 512. Det vil si at dersom vi endrer f-tallet med ett hakk, svarer det til at irradiansen i bildeplanet endres med en faktor to opp eller ned. Økende f-tall svarer til mindre effektiv diameter for linsen og derved mindre lysintensitet på bildebrikken. For å få samme mengde energi samlet opp per pixel i CMOS-brikken må da eksponeringstiden enten økes til det doble (for økende f-tall) eller reduseres til det halve (for avtakende f-tall).

Egentlig angis f-tallet gjerne som en brøk, slik: $f/4$, $f/5.6$, $f/8$ osv. Dersom vi da oppfatter dette som $1/4$, $1/5.6$ og $1/8$, blir verdien av brøken mindre og mindre for høyere blendertall slik vi brukte tallet ovenfor. Et objektiv som har lavest mulig blendertall, f.eks.



Figur 12.24: Når et lysende punkt i et objekt avbildes av en linse, vil det avbildes omtrent som et punkt i det aktuelle fokalplanet. Dersom vi skal avbilde flere objekter som ikke har samme avstand til linsen, finnes det ikke ett fokalplanet for bildene som dannes. Da vil lysende punkter i objekter som har en annen avstand til linsen enn det vi har fokusert på, bli avbildet som sirkulære skiver, og bildet blir "uskarpt". Ved å redusere åpningen til bladeren vil uskarpheten bli redusert (vinkel ϕ_2 er mindre enn vinkel ϕ_1 i figur 12.23), og vi sier at vi har fått større dybdeskarphet. I venstre fotografi er det brukt et objektiv med blendertall 3.5 og lukkertid 1/20 s, i høyre foto er det brukt samme objektiv og fokusering, men nå med blendertall 22 og lukkertid 1.6 s.

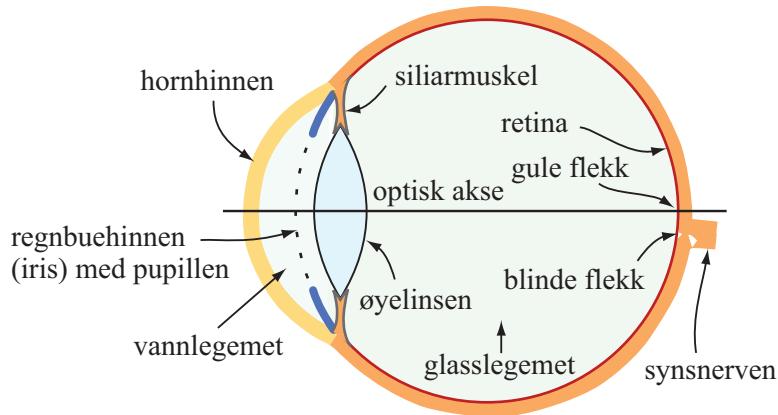
1.4, eller $f/1.4$ dersom vi ønsker å skrive det slik, har en større lysstyrke enn en linse med minste blendertall f.eks. 5.6 ($f/5.6$).

Av figur 12.23 kan vi merke oss en annen detalj. Dersom vi forskyver CMOS-brikken litt fram eller tilbake i forhold til der bildet er, vil et lysende punkt erstattes av en lysende skive. Lysbunten inn mot brennplanet har en romvinkel som avtar når åpningen på linsen blenes ned, antydet som ϕ_1 og ϕ_2 i figuren. Det betyr at skyver vi bildebrikken et lite stykke vekk fra fokalplanet, vil sirklene på bildebrikken bli større når blenderen er åpen (mest mulig lysintensitet inn) enn når blenderen er mindre (mindre lys slipper inn).

Dersom vi tar et bilde av et motiv hvor ikke alle objektene har samme avstand fra objektivet, vil vi ikke kunne få bildet av alle objektene i fokus på samme tid. Har vi åpen blender, vil uskarpheten for objektene som ikke ligger i fokalplanet være større enn dersom blenderen har redusert åpning. Jo mindre åpning (høyere blendertall), desto mindre vil uskarpheten bli. Innen fotografi sier vi at vi har større "dybdeskarphet" ved liten blenderåpning (stort f-tall) sammenlignet med stor blenderåpning (lite f-tall). Figur 12.24 viser et eksempel på denne effekten.

12.5.8 Øyets optikk

Figur 12.25 viser skjematiske detaljer om hvordan et menneskeøye er bygget opp. Optisk sett er det en kraftig sammensatt linse som danner et reelt bilde på netthinnen i bakre del av øyet. Mengde lys som når netthinnen kan reguleres ved å endre størrelsen på hullet (iris) i regnbuehinnen. Netthinnen har meget stor oppløsning (som gjør at vi kan se detaljer) bare i et lite område av retina der øyets optiske akse treffer. Dette området kalles *den gule flekk* (se figur 12.27), og her er synscellene såkalte tapper som kan gi fargeinformasjon (jamfør kapittel 10). I retina forøvrig ligger synscellene ikke så tett, og det er mest staver (mer lysfølsomme enn tappene, men gir ikke fargeinformasjon). Merkelig nok må lyset gå gjennom flere cellelag før det treffer selve synscellene. Dette har kanskje en evolusjonsmessig opprinnelse siden mennesket oppholder seg i dagslys hvor sollyset er ganske kraftig. Det finnes arter som lever på store (mørke) havdyp hvor lyset når synscellene direkte uten å gå gjennom andre cellelag først.



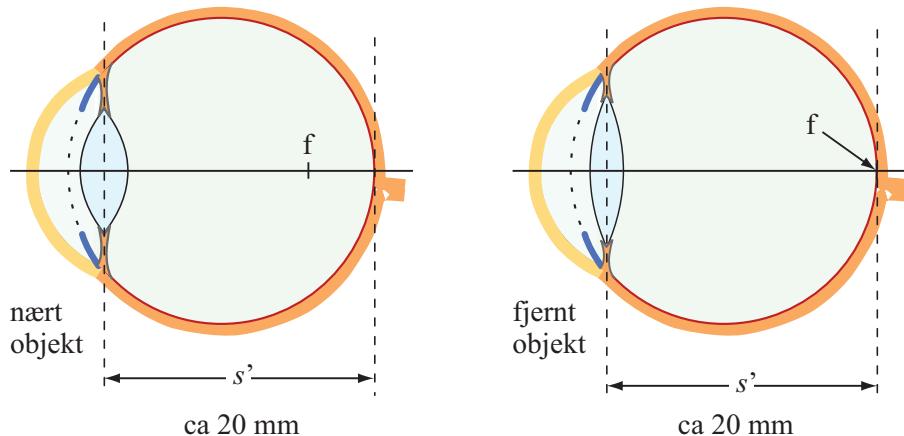
Figur 12.25: Skjematiske oppbygging av et menneskeøye.

Den vesentligste linseeffekten til øyet kommer fra den krumme overflaten mellom luft og hornhinnen. Øyelinsen virker bare lett modifiserende på linsestyrken som hornhinnen selv setter opp. Brytningsindeksen til vannlegemet (aqueous humor) og glasslegemet (vitreous humor) er om lag 1.336 (omtrent som for vann), mens øyelinsen har en brytningsindeks på 1.437. Forskjellen mellom disse brytningsindeksene er nokså liten, og dette er en vesentlig grunn til at total linsestyrke i størst grad skyldes den krumme flaten mellom luft og hornhinna.

Tabell 12.1: Aldersutvikling

Alder (år)	Nærpunkt (cm)
10	7
20	10
30	14
40	22
50	40
60	200

Øyets størrelse er omrent uendret under bruk, slik som indikert i figur 12.26. Når vi skal fokusere på objekter som er nær oss og langt fra oss, justeres øyelinsen. Øyelinsen lar seg forme, og når siliarmusklene tekker hardt i linsen, blir den tynn og får liten krumning. Linsestyrken går da ned. For et normaløye vil da brennvidden for linsene totalt ligge på netthinnen. Objekter som er "uendelig langt borte" vil da avbildes som et reelt opp-ned bilde på netthinnen. Når siliarmusklene slapper mere av, vil linsen krympe sammen og bli mer krum og få større linsestyrke. Da faller brennpunktet inn i glasslegemet et sted, og objekter som ikke er så langt fra øyet vil kunne danne et reelt bilde på netthinnen. Bildeavstanden s' i linseformelen vil altså alltid holde seg konstant, mens brennvidden på den totale linsen vil endre seg. Vi regner ofte at bildeavstanden s' er lik 20 mm.



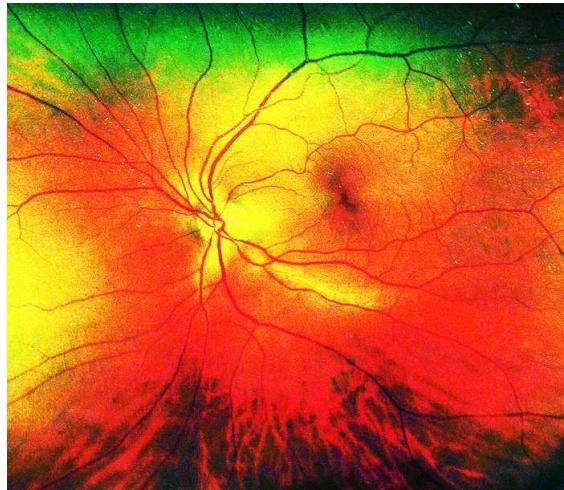
Figur 12.26: Øyets størrelse holder seg nær konstant, men brennvidden kan endres. På den måten kan vi se gjenstander skarpt i et helt intervall av avstander.

Øyelinsen blir stivere med alderen og klarer ikke å gjennomløpe så store endringer i brennvidde som linser til barn. Normalt betyr det at linsen ikke blir så krum som den pleide være når vi slapper av i siliarmusklene. Total linsestyrke går da ned, og følgen er at vi må holde et objekt lengre vekk fra øyet for å kunne se det skarpt.

Den minste avstand et objekt kan ha som samtidig gir skarpt bilde på netthinnen kalles øyets *nærpunkt*. Den største avstanden et objekt kan ha og samtidig gi skarpt bilde kalles øyets *fjernpunkt*. Et normaløye har fjernpunktet i "det uendelige". Nærpunktet endrer seg med alderen, av grunner allerde nevnt, og tabell 11.1 gjengir en oversikt tatt fra en amerikansk lærebok.

Ta disse tallene med en klype salt, men de antyder i det minste en aldersutvikling som mange vil oppleve med årene. Det bør bemerknes at normalavstanden vi bruker for å

beregne en luples fortørrelse er 25 cm, altså nærpunktet for en person i ca 40-årsalderen. En tiåring vil kunne holde en gjenstand ved en tredjedel av denne avstanden og fortsatt se den skarpt. For en tiåring vil derfor reell forstørrelse for en lupe bare være tredjeparten så stor som verdien vi fikk fra den generelle regelen vi ga tidligere.



Figur 12.27: Netthinnen tatt med et kamera gjennom pupillen med en spesiell blitz-anordning. Bildet er tatt ved en rutinekontroll ved Krogh optikk, Strømmen, 2009. Den blinde flekk er det gule området hvor blodårer og nerver går ut fra øyeplet. Den såkalte gule flekk (makula på latin) er et ca. 1.5 mm diameter stort område omtrent midt inni det svakt mørke ovale området litt til høyre fra den blinde flekk. I makula ligger tappene tettest, og området er ansvarlig for vårt skarpsyn. Lyset må gå gjennom flere cellelag før det når de lysomfintlige synscellene. I vårt bilde er ikke den gule flekken gul! (Bildet er bearbeidet for å øke kontrast m.m.)

Øyets linsestyrke, briller.

La oss nå bli litt mer kvantitative ved å bruke linseformelen for øyet.

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

For et normaløye som fokuserer på et gjenstand langt, langt borte, vil s tilnærmet være uendelig, og s' omtrent lik 20 mm = 0.02 m. Da følger:

$$\frac{1}{f} = \frac{1}{0.02 \text{ m}} = 50 \text{ m}^{-1}$$

Størrelsen $1/f$ kalles *linsestyrken* og denne angis i antall *dioptre* som er lik m^{-1} . Et normaløye som fokuserer på uendelig har altså en linsestyrke på 50 dioptre.

Et normaløye som har nærpunkt på 25 cm, har en linsestyrke gitt ved:

$$\frac{1}{f} = \frac{1}{0.25 \text{ m}} + \frac{1}{0.02 \text{ m}} = 54 \text{ dioptre}$$

Øyelinsen klarer med andre ord bare å endre total linsestyrke med om lag ti prosent. Hornhinnen tar seg av resten av linsestyrken.

Ikke alle øyne er "normale". Noen har hornhinner med alt for stor krumming, slik at linsestyrken blir for stor i forhold til avstanden mellom linse og retina. Når et slikt

menneske forsøker å se et objekt langt borte, vil det reelle bildet ikke treffe netthinnen, men ligge et sted inne i glasslegemet. Personen vil derfor ikke se skarpt når hun/han ser på noe langt borte. En slik person kaller vi *närsynt*. Närsynhet kan avhjelpes ved bruk av briller eller kjøpe-øyelinser. Øyets naturlige linsestyrke må da *motvirkes*, siden den er for stor, og brillene eller de ytre øyelinssene må være konkave (negativ brillestyrke).

Noen har hornhinner med alt for lite krumning, og linsestyrken blir da for liten. Når personen forsøker å se på en gjenstand 25 cm fra øyet (nominelt nærpunkt), vil det reelle bildet ikke falle på netthinnen, men vil teoretisk sett falle bak denne. Bildet på netthinnen blir igjen uskarpt. En slik person kalles *langsynt*. Igjen kan vi kompensere feilen ved å sette inn en ekstra konveks linse i form av briller eller eksterne øyelinser inntil vi ser skarpt også for legemer i 25 cm avstand.

Som ung kan en person ved hjelp av én brille få et tilnærmet normalsyn, med nærpunkt i 25 cm og fjernpunkt i det uendelige. Når alderen øker, minker *akkomodasjonsevnen*, og én brille vil ikke lenger både kunne gi nærpunkt ved 25 cm og samtidig fjernpunkt i det uendelige. Det blir nødvendig å stadig ta av seg og på seg briller, kanskje til og med skifte mellom to sett for å få kunne se tilfredsstillende både nært og fjernt. Det finnes også såkalte "progressive briller" hvor øvre del av glasset har én brennvidde og nedre del en annen (med en glidende overgang mellom disse).

Noen har linsefeil som kan karakteriseres ved at hornhinnen ikke har kuleflateform, men mer form som en ellipsoide. Det blir da forskjellig linsestyrke for linjer rettet en retning sammenlignet med linjer rettet vinkelrett på den første. Denne linsefeilen kalles *astigmatisme* og kan rettes opp ved hjelp av linser som har overflater til sylinder heller enn overflater til kuler. Slike linser kalles cylinderlinser. De kombineres i praksis med sfæriske overflater om ønskelig.

I dag er det temmelig vanlig å ta en slags laseroperasjon for å endre overflaten til hornhinnen dersom lensen har betydelige feil fra fødselen av. I så fall kan deler av hornhinnen brennes bort og formas slik at personen får et normalt syn og slipper å gå med briller (inntil aldersfenomener gjør det nødvendig).

Eksempler.

Det er nokså enkelt å selv danne oss et grovt inntrykk for hvilke briller vi trenger i tilfelle vi er litt närsynt eller langsynt. Her er et par eksempler:

Anta at vi bare kan se skarpt ut til 2.0 m, det vil si at fjernpunktet uten briller ligger på 2.0 m. Det betyr at øyets linsestyrke er:

$$\frac{1}{f} = \frac{1}{2.0 \text{ m}} + \frac{1}{0.02 \text{ m}} = 50.5 \text{ dioptre}$$

Linsestyrken er da 0.5 dioptre for stor, siden linsestyrken burde bare vært 50.0 dioptre for fjernpunktet. Løsningen er å bruke en brille på -0.5 dioptre, i alle fall når vi skal betrakte objekter langt borte. Det nydelige med disse beregningene er at vi kan legge sammen linsestyrker på enklest mulig måte, slik at 50.5 dioptre for øyets optikk pluss briller på -0.5 dioptre får en total linsestyrke på 50.0 dioptre.

I neste eksempel tar vi for oss en person som ikke klarer å fokusere på avstander nærmere enn 50 cm. Personens linsestyrke er da:

$$\frac{1}{f} = \frac{1}{0.5 \text{ m}} + \frac{1}{0.02 \text{ m}} = 52.0 \text{ dioptre}$$

I dette tilfellet, når vi betrakter nærpunktet, burde linsestyrken vært 54.0 dioptre. Her mangler det to dioptre. Personen trenger altså en brille med brillestyrke +2.0 dioptre for å kunne flytte nærpunktet fra 50 cm til 25 cm.

12.6 Oppsummering

Geometrisk optikk er basert på tankegangen at lyset fra ulike objekter brer seg som "lysstråler" i ulike retninger, der hver lysstråle oppfører seg tilnærmet som (avgrensede) plane elektromagnetiske bølger. Disse lysstrålene vil reflekteres og brytes ved overganger fra ett medium til et annet, og lovmessigheten er bestemt ved Maxwells ligninger og tilfredsstiller refleksjonslover og Snels brytningslov i vanlige materialer.

Når en grenseflate mellom to medier er krum, vil lysstråler som kommer inn mot grenseflaten ha ulik vinkelfordeling sammenlignet med lysstrålene som går ut.

For tynne linser kan vi definere to brennpunkt, ett på hver side av linsen. Lysstråler vil ha uendelig mange ulike retninger i praksis, men det holder å bruke to eller tre hjelpe-linjer for å konstruere hvordan et objekt blir avbildet av en linse til et bilde. Hjelpe-linjene karakteriseres ved at lys parallelt med optisk akse blir brutt gjennom brennpunktet på motsatt side for en konveks linse, men vekk fra brennpunktet på samme side som innkomende lysstråle for konkave linser. Lysstråler gjennom linsens sentrum blir ikke brutt. Vi tegner normalt bare linjer til linsenes midtpunkt i stedet for å ta med detaljert bryting ved hver overflate. Hjelpe-linjer kan gjerne gå utenfor linsens fysiske utstrekning, men bare lysstråler som faktisk går gjennom linsene bidrar til lysintensiteten i bildet.

Lys som følger ulike lysstråler fra ett objektpunkt til ett bildepunkt bruker alle like lang tid på veien, uansett om de går gjennom sentrale eller perifere deler av linsen. Dette sikrer at lys fra ulike lysstråler har konstruktiv interferens når de møtes. Når virkelige lysstråler møtes på denne måten, kan lys fanges opp på en skjerm, og vi snakker om et reelt bilde. Dersom de virkelige lysstrålene bare fjerner seg fra hverandre etter å ha passert en linse, men synes alle å komme fra et punkt bakenfor linsen, sier vi at vi har et virtuelt bilde der lysstrålene synes å komme fra. Betrakter vi lysstråler som divergerer fra hverandre ved hjelp av øyet vårt, vil lysstrålene igjen samles på netthinnen og danne et reelt bilde der. Vi kan derfor "se" et virtuelt bilde, til tross for at vi ikke kan samle dette bildet opp på en skjerm.

Linseformelen er en forenkling av linsemakerformelen, og kun objektavstand, bildeavstand og brennvidde inngår. I linseformelen regnes brennvidden for positiv for en konveks linse og negativ for en konkav. Fortengnet på objektavstand og bildeavstand varierer med hvordan lysstrålene kommer inn mot linsen i forhold til hvor de går ut. Vi må vurdere fortegn i hvert enkelt tilfelle for å ikke komme galt ut. En tegning som viser strålegangen er helt vesentlig for å ikke gjøre feil i slike tilfeller (må sjekke at resultatet ser rimelig ut).

En linse kan brukes som lupe. Forstørrelse er da en vinkel-forstørrelse, for plasseres objektet i linsens brennpunkt, vil det virtuelle bildet tilsvarende være uendelig langt borte og være uendelig stort. Ulike vinkler som angir maksimal utbredelse av et objekt fører til tilsvarende fysisk utstrekning på det reelle bildet på netthinnen når vi betrakter objektet gjennom lupen. Lupens funksjon er først og fremst at vi effektivt kan holde objektet mye nærmere øyet enn øyets nærpunkt. Det vil si, vi kan effektivt plassere objektet mye nærmere øyet og likevel se skarpt, sammenlignet med å se på objektet nærmest mulig (skarpt) uten hjelpe-middel.

Linser kan settes sammen til optiske instrumenter så som teleskop og mikroskop. For teleskopet brukes et objektiv for å lage et lokalt bilde av objektet som vi så kan betrakte med en lupe. Resultatet kan være en betydelig forstørrelse. For mikroskopet plasseres objektet utenfor, men meget nært objektivets brennpunkt. Det reelle bildet som objektivet da lager, er da betydelig større enn objektet. Igjen brukes en lupe for å betrakte det reelle bildet som objektivet lager.

Et menneskeøye har en fast bildeavstand på ca. 20 mm, men kan endre linsestyrken noe slik at brennpunktet kan flyttes og vi får et helt avstandsområde der vi kan se et objekt skarpt. Normaløyet skal kunne se gjenstander skarpt fra ca. 25 cm til uendelig, hvilket svarer til en linsestyrke som varierer fra 54 til 50 dioptre. Har hornhinnen for stor eller for liten krumming, er linsestyrken for stor eller for liten. Da vil vi ikke kunne se skarpt i hele området 25 cm til uendelig, og vi trenger briller for å kompensere for mangler i øyets linsestyrke.

12.7 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Forklare hvorfor lys fra objekter kan betraktes som “lysstråler” når lyset treffer f.eks. en linse.
- Beregne hvor bildepunktet til en punktformig lyskilde er etter at lyset har møtt overflaten til en glasskule.
- Gjøre rede for størrelsene objekt, bilde, brennpunkt, objektavstand, bildeavstand, brennvidde, krumningsradius, konkav, konveks, reelt og imaginært bilde.
- Utlede (evt. med litt hjelp) linsemakerformelen for en meniskformet konveks linse, og angi hvilke forenklinger som vanligvis blir gjort.
- Utlede linseformelen under samme betingelser som i forrige punkt.
- Angi de tre hovedreglene som benyttes ved konstruksjon av strålegangen gjennom linser og speil (ray optics).
- Gjøre rede for at vi iblant må skifte fortegn på enkelte størrelser når linseformelen benyttes.
- Gjøre rede for to ulike måter å angi forstørrelsen for optiske instrumenter.
- Forklare hvordan en lupe vanligvis brukes og hvilken forstørrelse den har.
- Angi hvordan et teleskop og et mikroskop er bygget opp, og forstørrelse det har.
- Angi hvordan et speilteleskop fungerer og hvordan vi unngår å stå for mye i veien for innkommende lys.
- Beregne hvor stort bilde av et gitt motiv (i en gitt avstand) vi kan oppnå i bildeplanet for ulike kameraobjektiver.
- Beregne omtrentlig synlig bildevinkel gjennom en kikkert når nødvendige geometriske mål er oppgitt.
- Gjøre kort rede for hvordan nanovitenskap har ført til bedre fotografiske linser.
- Gjøre rede for lysstyrke til en linse / objektiv og kjenne til hva blendertallene sier oss.
- Gjøre rede for ulik “dybdeskarphet” og hvordan denne endres med blendertallet.
- Gjøre rede for øyets optikk, og forklare hva uttrykkene nærpunkt, fjernpunkt og akkomodasjon betyr.
- Kjenne til øyets linkestyrke og hvor linkestyrken oppnås.
- Beregne omtrentlig hvilken brillestyrke en person eventuelt trenger ut fra enkle målinger av nærpunkt og fjernpunkt.

12.8 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Refleksjons- og brytningslovene som er omtalt i dette kapitlet gjelder for synlig lys. Lys er elektromagnetiske bølger. Vil de samme lovene gjelde for elektromagnetiske bølger generelt? (Som vanlig, svaret må begrunnes!)
2. En mottakerantenne for satellitt-TV er formet som et krumt speil. Hvor plasseres selve antenneelementet? Er dette analogt med bruk av speil i optikk? Hvilken bølgelengde har forresten satellitt-TV-signaler? Og hvor lang er bølgelengden i forhold til antennediskens størrelse?
3. En mottakerantenne for satellitt-TV med diameter 1 m koster omlag en tusenlapp, mens et speil for et optisk teleskop med diameter 1 m, ville koste anslagsvis en million kroner. Hvorfor er det så stor forskjell i pris?
4. Et "brennglass" er en konveks linse. Sender vi sollys gjennom linsen og holder et stykke papir i brennplanet, kan papiret ta fyr. Dersom linsen er bortimot perfekt, vil vi ut fra geometrisk optikk alene forvente at alt lyset kan sammles i et punkt med nærmest ingen utstrekning?
5. Dersom du har forsøkt å bruke brennglass kan du ha oppdaget at papiret tar lettere fyr dersom solflekken treffer et sort område på papiret sammenlignet med et hvitt. Kan du forklare hvorfor?
6. Det er rapportert tilfeller at fiskeboller med vann og kuleformede vaser med vann har fungert som brennglass slik at ting har tatt fyr. Er det teoretisk mulig ut fra lovene vi har utledet i dette kapitlet?
7. Ut fra linsemakerformelen ser vi at effektiv brennvidde avhenger av bølgelengden siden brytningsindeksen varierer med lysets bølgelengde. Er det mulig for en bikonveks linse å ha positiv brennvidde for én bølgelengde og negativ brennvidde for en annen bølgelengde?
8. Hvordan kan du raskt finne den omtrentlige brennvidden til en samlelinse? Har du også en like rask test for en spredelinse?
9. Blir brennvidden endret når du senker en konveks linse ned i vann?
10. Blir brennvidden endret når du senker et konkavt speil ned i vann?
11. Dersom du ser under vann, ser du uskarpt, men dersom du har dykkerbriller på deg, ser du skarpt. Forklar! Kunne du i stedet klart deg med ekstra briller, uten noe luftlag noe sted? I så fall, måtte brillene være konkave eller konvekse?
12. Et reelt bilde (f.eks. skapt av et objektiv) kan detekteres ved å plassere et papir, film eller CMOS-brikke i bildeplanet. Er det mulig å registrere et virtuelt bilde på et eller annet vis?
13. Refleksjonslover, brytningslover, linsemakerformel og linseformel er alle symmetriske med hensyn til hvilken vei lyset går. Vi kan med andre ord bytte om hva vi anser som objekt og bilde. Kan du påpeke rent matematisk hvordan denne reversibiliteten kommer til uttrykk i de aktuelle lovene? Er det noen unntak fra regelen?

14. a) Vi har et vertikalt speil på en vegg. En lysende glødelampe holdes foran speilet slik at lys som reflekteres av speilet treffer gulvet. Det er likevel ikke mulig å danne noe bilde av glødelampen på gulvet. Hvorfor?
- b) Vi har en laserpeker og lyser på lignende vis som med glødelampen, slik at lys fra laserpekeren som reflekteres av speilet når gulvet. Nå synes det som om vi har fått dannet et bilde av laserpekeren (åpningen av denne) på gulvet. Kan du forklare hva som foregår?
15. Hvor høyt må et speil være, og hvor høyt må det plasseres på en loddrett vegg, for at vi skal se hele oss selv i speilet på en gang? Har avstanden til speilet betydning?

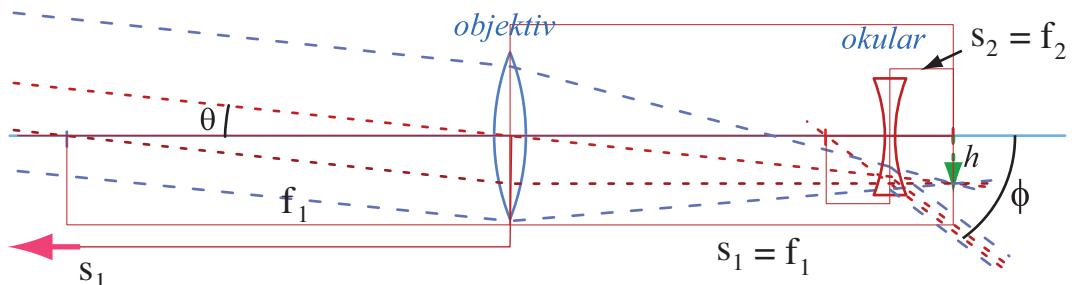
Regneoppgaver

16. Bestem ved å ta utgangspunkt i linseformelen og en av reglene for lysstrålegeometri, minste og største forstørrelse (i tallverdi) en konveks linse kan ha. Bestem betingelsen for at forstørrelsen skal bli 1.0.
17. Gjenta samme beregning som i forrige oppgave, men nå for en konkav linse. Bestem på ny betingelsen for at forstørrelsen skal bli (tilnærmet lik) 1.0.
18. Når vi skal finne bildet som en konveks linse lager av et objekt "uendelig langt borte", kan vi ikke bruke de tre standard lysstrålene for konstruksjon av bildet. Hvordan går vi fram i et slikt tilfelle for å finne bildets plassering i bildeplanet?
19. Vi har en konveks meniskformet linse med overflater som svarer til kuleflater med radius på 5.00 og 3.50 cm. Brytningsindeksen er 1.54. Hvor stor er brennvidden? Hvilken bildeavstand får vi dersom et objekt plasseres 18.0 cm vekk fra linsen?
20. En smal lysstråle fra et fjernt objekt sendes inn i en glasskule med radius 6.00 cm og brytningsindeks 1.54. Hvor vil lysstrålen bli fokusert?
21. Tegn lysstrålediagram for en konveks linse for følgende objektavstander: $3f$, $2f$, $1.5f$, $1.2f$, $1.0f$, $0.8f$, $0.5f$. For en av disse avstandene kan bare to av de vanlige tre standardlysstrålene benyttes i konstruksjon av bildet. Hvilken? Legg merke til hvorvidt vi har forstørrelse eller forminskning av bildet, om bildet er opprett eller opp ned, og om bildet er reelt eller virtuelt.
22. Anta at du har et kamera og skal ta bilde av en 1.75 m høy venn som står oppreist 3.5 m unna. Kameraet har en 85 mm linse (brennvidde). Hvor stor avstand er det mellom linsen og bildeplanet når bildet tas? Får du plass til hele personen innenfor bildet dersom bildet registreres på en gammeldags film eller en "fullformat" CMOS bildebrikke med størrelse 24 x 36 mm? Hvor mye av personen får du plass til på bildet dersom det beyttes en CMOS bildebrikke med størrelse 15.8 x 23.6 mm?
23. Når Mars er nærmest Jorden er avstanden ca $5.58 \cdot 10^7$ km. Diameteren på Mars er 6794 km. Hvor stort bilde får vi fra et konvekst objektiv (eller hulspeil) med brennvidde 1000 mm?
24. Et teleskop har et objektiv med brennvidde 820 mm og diameter 100 mm. Okularet har en brennvidde på 15 mm og en diameter på 6.0 mm. Hvor stor forstørrelse har teleskopet? Hvor stor er bildevinkelen? Kan vi se hele måneskiven på én gang?

25. En lysbildefremviser (eller data-projektor for den saks skyld) har et objektiv med brennvidde 12.0 cm. Lysbildet er 36 mm høyt. Hvor stort blir bildet på en skjerm 6.0 m fra projektoren (objektivet)? Er bildet opprett eller opp ned?
26. Anta at vi har to briller, en med linsestyrke +1.5 dioptre på begge glass og en med linsestyrke +2.5 dioptre på begge glass. Vi finner bare et av brillene og får lyst til å sjekke om dette var de med sterkest eller svakest linsestyrke. Kan du gi en prosedyre på hvordan vi kan bestemme linsestyrken på de brillene vi fant?
27. a) Hvor er nærpunktet til et øye der en optiker foreskriver en brille med linsestyrke 2.75 dioptre? b) Hvor er fjernpunktet til et øye der en optiker foreskriver en brille med linsestyrke -1.30 dioptre (når vi skal se ting på lang avstand)?
28. a) Hvilken brillestyrke trenger en pasient som har nærpunkt i 60 cm avstand. b) Finn brillestyrken en pasient trenger som har fjernpunkt ved 60 cm.
29. Bestem akkomodasjonen (i betydning mulig endring i linsestyrke) hos en person som har nærpunkt ved 75 cm og fjernpunkt ved 3.0 meter.
30. I en forenklet modell av øyet ser vi for oss at hornhinnen, væsken innenfor, linsen og glassvæsken i det indre av øyet alle har brytningsindeks 1.4. Avstanden mellom hornhinnen og retina er 2.60 cm. Hvor stor må krumningsradius for hornhinnen være for at en gjenstand 40.0 cm fra øyet skal bli fokusert på netthinnen?
31. En lupe har brennvidden 4.0 cm. Hvilken forstørrelse vil den gi under "normalt" bruk? Er det mulig å få en forstørrelse på hele 6.5 X ved å bruke lupen på en litt annen måte enn beskrevet som standard (tenker ikke her på okularprojeksjon)? I så fall, fortell hvor objektet vi betrakter må plasseres, og si litt om hvordan vi nå må bruke øyet.
32. Det gamle Yerkes teleskopet ved University of Chicago var verdens største linsekikker. Det hadde et objektiv som var 1.02 m i diameter og et f-tall (blendertall) på 19.0. Hvor lang var brennvidden? Hvor stor diameter har bildet av Månen i fokalplanet til denne linsen? (Vinkeldiameteren til Månen er om lag en halv grad.)
33. Et sfærisk konkavt speil vil ikke samle alle parallelle stråler i ett punkt, fordi effektiv brennvidde vil avhenge av hvor nær speilets optiske akse strålene treffer speilet.
- a) Forsök å sette opp et matematisk uttrykk for effektiv brennvidde for en stråle som kommer inn parallelt med optiske akse en viss avstand fra aksen. Speilets krumningsradius settes lik R . Som parameter kan vi med fordel bruke vinkelen θ mellom innfallende stråle og linjen som går mellom krumningscentrum for speilet og punktet der strålen treffer speiloverflaten.
- b) For hvilken vinkel vil effektiv brennvidde ha endret seg med 2 % i forhold til brennvidden for de stråler som kommer inn meget tett til den optiske akse?
- c) Kan du med dette forklare hvorfor speilkikkerter som er basert på sfærisk speil ofte har høyt f-tall (liten lysstyrke)?
34. En "tynn" linse med brennvidde 5.00 cm blir brukt som en lupe når vi betrakter detaljer i et fotografi. Hva menes med forstørrelse til en lupe? Hvor stor forstørrelse gir vår lupe? Forstørrelsen varierer litt med hvor lang avstand vi bruker mellom linsen og fotografiet. Hvor stor er forstørrelsen dersom vi stiller avstanden s mellom linse/fotografi slik at øyet kan *fokusere som om objektet var uendelig langt borte*? Hvor stor er avstanden s dersom den fører til at øyet må fokusere som om objektet bare var 25 cm unna? Hvor stor er forstørrelsen da? [Hint: Ved sammenligning av

vinkler må du bruke vinkelen mellom optisk akse og lysstrålen som tenkes å gå gjennom linsens sentrum.]

35. Et linseteleskop skal brukes av en amatørastronom. Brennvidden på objektivet er 820 mm, diameteren på er 10.0 cm. Objektivet sitter i en ende av teleskoprøret og en okularholder i motsatt ende. Okularholderen kan justeres slik at vi får et klart bilde av stjernehimmelen og planetene. For å kunne bruke litt ulik forstørrelse på ulike objekter, har amatørastronomen fire forskjellige okularer med brennvidde 30, 15, 7.5 og 3.0 mm. Diameteren på linsen i disse okularene er hhv 48, 20, 11 og 3.7 mm. Vi behandler alle linser som om de var "tynne".
- Hvor langt må teleskoprøret være (avstand mellom objektiv og okular)?
 - Hvor stor endring i posisjon må okularholderen tillate?
 - Hvor mye lengre må okularholderen kunne bevege seg dersom vi også skal kunne bruke kikkerten som landskapskikkert med minste objekt-avstand lik 20 m?
 - Hvor stor lysstyrke (blendertall) har objektivet?
 - Hva mener vi med "forstørrelse" til en kikkert?
 - Anslå omtrentlig hvor stor forstørrelse vi får for de fire ulike okularene.
 - Anslå omtrentlig hvor stor bildevinkel vi får for 30 mm og 3.0 mm okularene.
 - Sammenlign dette med bildevinkelen til månen, som er om lag 0.5 grader.
 - Hvor stor vil Jupiter se ut når forholdene ligger best til rette for observasjoner, når vi betrakter den gjennom vårt teleskop med 3.0 mm okularet? (Tilnærmet radius i Jordens bane er $1.50 \cdot 10^{11}$ m og i Jupiters bane $7.78 \cdot 10^{11}$ m. Jupiters diameter er om lag $1.38 \cdot 10^9$ m.)
36. Teleskopet som Galilei laget bestod av et konvekst objektiv og et konkavt okular. Vi kaller en slik kikkert i dag for en "opera-kikkert", og en litt utsydelig prinsippskisse er vist i figur 12.28. Bildet fra objektivet plasseres da "bak" okularet (okularet er



Figur 12.28: *Galeiskop, forenklet.*

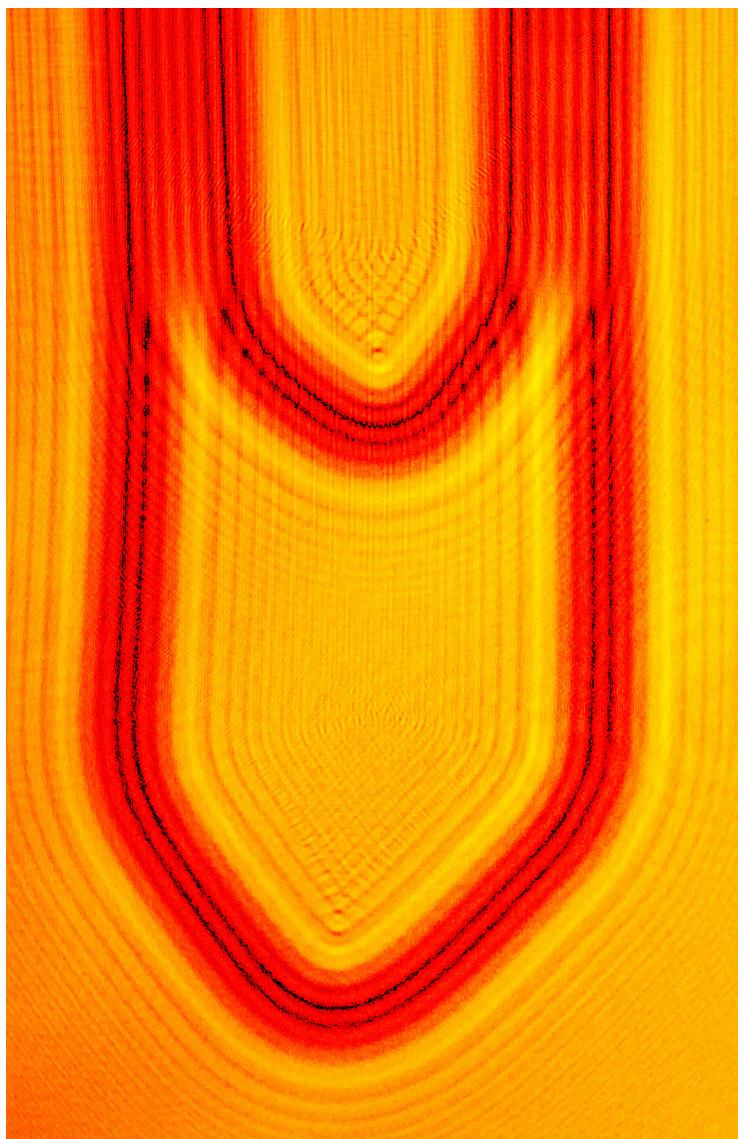
nærmere objektivet enn bildet fra objektivet). Kikkerten blir derfor kortere enn for eks. et teleskop beskrevet ovenfor. Anta at vi starter med samme objektiv som i forrige oppgave. Vi antar videre for enkelhets skyld at objekter vi kikker på er "uen delig langt borte" og at øynene fokuserer som om objektene var plassert uendelig langt borte.

- Tegn på egen hånd opp strålegangen i Galilei-kikkerten, og pass på å få detaljene korrekt nær okularet.
- Vis at den angulære forstørrelsen (tallverdien) for Galilei-kikkerten er gitt ved $M = f_1/f_2$, der f_1 og f_2 er tallverdien av brennvidlene til objektiv og okular.
- Hvilken brennvidde må vi velge på okularet i Galilei-kikkerten dersom vi skal få samme forstørrelse som i kikkerten i forrige oppgave når okularet med nest lengste brennvidde er i bruk?
- Sammenlign lengden på kikkerten i forrige oppgave og lengden på Galilei-kikkerten for dette tilfellet.

- e) Har Galilei-kikkerten en annen fordel sammenlignet med kikkerten i forrige oppgave (og som også er medvirkende til at denne konstruksjonen brukes i “operakikkerter”)?
37. I et mikroskop på labben brukes et objektiv med brennvidde 8.0 mm og et okular med brennvidde 18 mm. Avstanden mellom objektiv og okular er 19.7 cm. Vi bruker mikroskopet slik at øynene fokuserer som om objektet var plassert uendelig langt borte. Vi behandler linsene som om de var “tynne”.
- Hvor stor avstand må det være mellom objektet og objektivet når vi bruker mikroskopet?
 - Hvor stor lineær forstørrelse gir objektivet (alene)?
 - Hvor stor forstørrelse gir okularet alene?
 - Hvordan er forstørrelse definert for et mikroskop?
 - Hvor stor er dette mikroskopets forstørrelse?
38. Vi har to tynne linser med brennvidde 12.0 cm (i tallverdi), den ene konveks, den andre konkav. Linsene plasseres 9.00 cm fra hverandre. Et 2.50 mm høyt objekt plasseres på den optiske aksen 20.0 cm utenfor de to linsene, nærmest den konvekse linsen.
- Hvor langt fra denne første linsen dannes det endelige bildet?
 - Er det endelige bildet reelt eller imaginært, oppned eller rettvend?
 - Hvor stort er det endelige bildet?
39. Vis at når to tynne linser er i kontakt med hverandre, vil brennvidden f til de to linsene tilsammen være gitt ved:
- $$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}$$
- hvor f_1 og f_2 er brennviddene til de to enkeltlinsene. Vi har en konvergerende menisk-formet linse med brytningsindeks 1.55 og krumningsradier på 4.50 og 9.00 cm. Den konkave flaten vendes vertikalt oppover, og vi fyller “gropen” med en væske som har brytningsindeks $n = 1.46$. Hvor stor blir den totale brennvidden for linse pluss væske-linse?
40. En tynn linse er laget som en plankonveks linse med diameter 2.5 cm og radien i den krumme delen er 5.0 cm. Linsen er laget av BK7 glass med brytningsindeks vs bølgelengde som gitt i en figur i forrige kapittel.
- Hvilken tykkelse har linsen på midten dersom tykkelsen i randen er 1.0 mm?
 - Bestem omtrentlig brennvidde for lys med bølgelengde 600 nm når den rette delen av linsen vender mot objektet (som antas å ligge “uendelig” langt borte).
 - Bestem sfærisk avvik ved 600 nm (det vil si variasjonsområdet for brennvidden for lys som kommer inn mot linsen i ulike avstander fra optisk akse). Anta at lyset kommer inn parallellt med optisk akse.
 - Bestem kromatisk avvik for lys i hele bølgelengdeområdet 400 - 800 nm for lys i én valgt avstand fra optisk akse (holder å gjøre beregningen for 400 og 800 nm).
 - Beregn sfærisk avvik for samme betingelser som i oppgave c, men nå med linsen snudd slik at den krumme delen av linsen vender mot objektet. Det er ok å gjøre en tilnærming (“tynn linse”) for at ikke beregningen skal bli for omfattende.
 - Ut fra resultatene i punkt c og e, kan du gi en anbefaling om hvilken vei en plankonveks linse bør stå for å få minst mulig sfærisk avvik ved avbildninger av objekter “uendelig” langt borte?
 - Gjenta samme beregninger som i c og e, men nå med et objekt som er halvannen brennvidde unna linsen (en arbeidskrevende oppgave for de som liker utfordringer!)

Kapittel 13

Interferens - Diffraksjon



I min studietid var et av høydepunktene en lang kveld på et laboratoriekurs hvor Sven Lilledal Andersen ga meg frie tøyler til å leke meg med diffraksjon og interferens. Jeg syntes da som nå at det er fascinerende å se hvordan lysets bølgenatur manifesterer seg med strukturer både i skyggepartier og i partier der lyset tilsynelatende har fri tilgang. Ved koherent lys kan randbetingelser mange tusen bølgelengder borte fortsatt spille inn på bølgeutbredelsen!

Når du leser dette kapitlet anbefaler vi at du merker deg den underliggende modellen både diffraksjon og interferens bygger på. Legg også merke til hovedstrukturer i diffraksjons- og interferensmønstrene. Matematikken er kun tatt med for å vise hvordan vi kan utlede de ulike uttrykkene, og er ellers ikke noe vi behøver kunne utenat! Derimot er det fint om du forstår hvorfor laserstrålen ved Alomarobservatoriet på Andøya utvides til 50 cm diameter før den sendes opp i atmosfæren. Diffraksjonens lover er pussige og vakre!

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes.

Versjon 22042013.

13.1 Superposisjon og linearitet

Når to eller flere bølger virker sammen, kan en rekke spennende fenomener observeres. I dette kapitlet skal vi først og fremst diskutere interferens og diffraksjon. Historisk sett kan vi kanskje si at ordet “interferens” først og fremst ble brukt når to separate bølger samvirket, mens ordet “diffraksjon” oftest ble brukt når noen deler av en bølge samvirket med andre deler av samme bølge. Det er nærmest umulig å holde disse to begrepene fra hverandre i alle situasjoner, derfor opplever vi iblant en litt ulogisk bruk av ordene.

Uansett navngiving, diffraksjon og interferens er noen av de mest bølge-spesifikke fenomenene vi kjenner til. Thomas Young’s dobbeltpalt er en av de mest omtalte eksperimentene i fysikken den dag i dag, og interferens er den viktigste grunnen til at man ikke kunne overse lysets bølgenatur for hundre år siden da Einstein med flere fant holdepunkter for at lyset også iblant synes å oppføre seg som partikler.

Interferens og diffraksjon er kanskje mest kjent som fenomener knyttet til lys, men vi finner de samme særtrekkene stort sett for alle typer bølger. Vi kan vise effekten på vannbølger, på lydbølger og til en viss utstrekning også ved bølger på en streng. Stående bølger kan med velvilje forstås som et interferensfenomen.

Basis for all interferens og diffraksjon er *superposisjonsprinsippet*:

Responsen på to eller flere samtidige stimuli vil ved en gitt tid og sted en være lik summen av responsen systemet ville hatt på hver av stimuliene enkeltvis.

Superposisjon innebærer med andre ord additivitet, matematisk uttrykt:

$$F(x_1 + x_2 + \dots + x_n) = F(x_1) + F(x_2) + \dots + F(x_n)$$

Dette innebærer at F er en lineær avbildning. Med andre ord: F må være en lineær funksjon!

I fysikken kjenner vi til at mange fenomener oppfører seg tilnærmet lineært. De mest slitte eksemplene er antakelig Ohm’s lov for resistans og Hooke’s lov for stekking av en fjær. Så lenge “utslagene” er små, gjelder (tilnærmet) en lineær sammenheng. Men vi vet at denne lovmessigheten ikke gir en god beskrivelse for større “utslag”. Da må “høyere ordens ledd” trekkes inn (uttrykket kan forstås med referanse til en Taylor-utvikling). Vi nevner dette for å minne om at superposisjonsprinsippet IKKE gjelder i enhver sammenheng. I dette kapitlet begrenser vi oss likevel nesten utelukkende til lineære systemer hvor superposisjon gjelder.

I dette kapitlet presenterer vi en matematisk beskrivelse av tre basis-situasjoner:

- Interferens fra en dobbeltpalt,
- Interferens fra et gitter (mange parallelle spalter), og
- Diffraksjon fra en enkeltpalt.

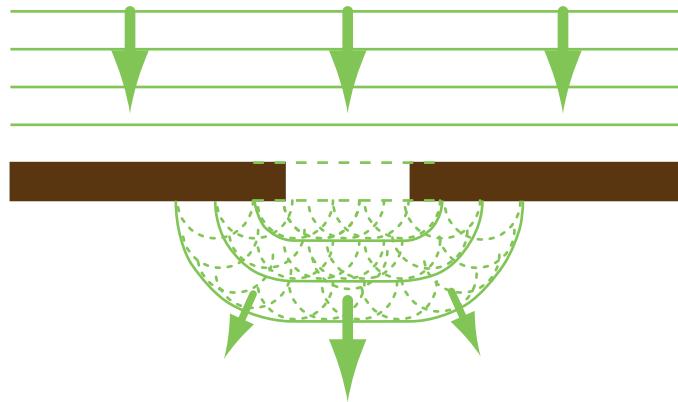
Ut fra disse eksemplene skal vi beskrive noen av de fenomenene vi ofte observerer i praksis.

13.2 Huygens prinsipp

Vår beskrivelse av interferens og diffraksjon er basert på Huygens prinsipp som sier at:

Ethvert punkt i en bølge kan sees på som en kilde til en ny bølge, såkalt elementærbølge, som brer seg ut i alle retninger. Når vi skal følge en bølgebevegelse, kan vi ta utgangspunkt i f.eks. en bølgefront og konstruere alle tenkelige elementærbølgene. Går vi én bølgelengde langs disse elementærbølgene, vil deres omhyllingskurve beskrive neste bølgefront.

Fresnel justerte litt på tankegangen ved å si at dersom vi skal finne bølgeamplituden et sted i rommet (også et godt stykke vekk fra en opprinnelig bølgefront), kan vi summere alle tenkelige bølger forutsatt at vi tar hensyn til både amplitude og fase (og hvorvidt noe kommer i veien for bølgen eller ikke).



Figur 13.1: I Huygens prinsipp tenker vi oss at ethvert punkt på en bølgefront er kilde til elementærbølgjer.

Huygen levde fra 1629 til 1695 og Fresnel fra 1788 til 1827, og vi kan undres over om et så gammelt tankegods kan være aktuelt i dag etter at vi har fått på bordet Maxwells ligninger, relativitetsteori og kvantefysikk. Merkelig nok er Huygens-Fresnels prinsipp fortsatt anvendelig og det er et bærende prinsipp i kvanteelektrodynamikk (QED) som er den mest nøyaktige teorien som finnes overhodet i verden i dag. Riktignok bruker vi litt andre ord på hva vi gjør i QED enn det Huygen og Fresnel gjorde, men matematisk sett er hovedidéen temmelig ekvivalent. I kvanteelektrodynamikken sies det at vi må følge alle mulige veier som en bølge kan gå fra en kilde til det stedet bølgen (eller sannsynlighetstettheten) skal evalueres. Brukes det en partikkelbeskrivelse, ligger faseinformasjonen likevel i bunnen gjennom selve kvantefeltet. Med andre ord, Huygens-Fresnels prinsipp er slitesterkt.

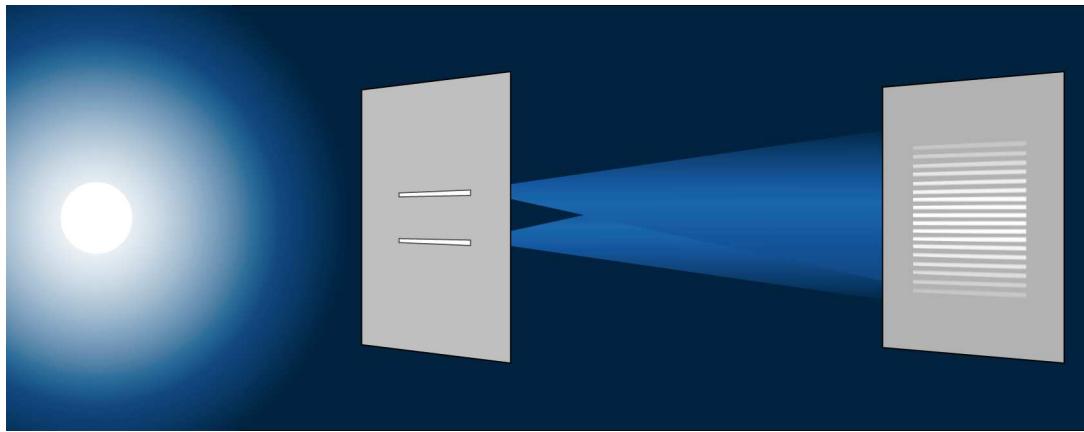
I hele kapitlet antar vi at lyset er “tilstrekkelig koherent”. Vi kommer tilbake til koherens i et senere kapittel, så vi nøyer oss i denne omgang med å si at lyset vi starter ut med (f.eks. i en spalt) kan beskrives som nærmest en matematisk perfekt sinusbølge uten noe endringer i amplitude eller frekvens etter som tiden går. Vi antar med andre ord fullstendig forutsigbarhet i fasen på Huygen-Fresnels-elementærbølgene når vi kjenner fasen i utgangspunktet.

13.3 Interferens fra en dobbeltpspalt

I 1801, da Thomas Young gjennomførte sitt berømte dobbeltpspalt-eksperiment, var det Newtons partikkellmodell for lys som rådde grunnen. Partikkellmodellen passet fint med at lysstråler syntes å gå i rette baner og ble reflektert fra speil slik de gjør. Og Newtons røde, grønne og blå partikler (for å si det litt enkelt) var et utmerket utgangspunkt for å forklare additiv fargeblanding.

Dersom Newtons lyspartikler går gjennom to smale, parallelle spalter, skulle vi forvente at vi ville se to striper på en skjerm plassert bak en dobbeltpspalt. Men hva var det Young observerte? Han så *flere* parallele striper! Dette var det nærmest umulig å forklare ut fra Newtons partikkellmodell. Young, og siden Fresnel og andre, kunne imidlertid nokså lett forklare dette fenomenet, og vi skal straks se på matematikken.

De to spaltene antas å være smale (ofte 1-1000 ganger bølgelengden), men “uendelig” lange slik at vi kan betrakte hele problemet som to-dimensjonalt.



Figur 13.2: *Eksperimentelt oppsett ved Youngs dobbeltpspaltforsøk. Spaltestørrelser og stripermønstre er kraftig forstørret sammenlignet med avstanden mellom lyskilde, spalter og skjerm.*

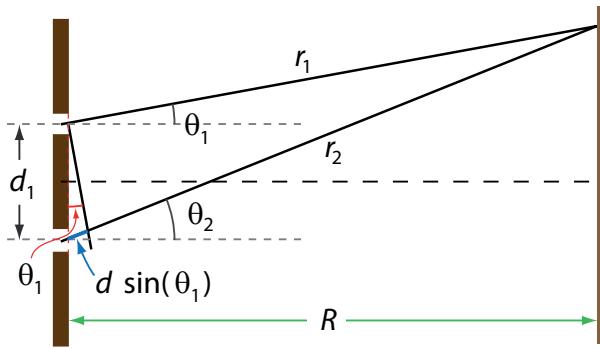
Vi antar at lys kommer inn med bølgefront parallellt med spaltene slik at lyset starter med identisk fase overalt i “utgangsplanet” i begge spaltene. Vi tenker oss videre at hver av spaltene sender ut elementærbølger, og av grunner nettopp nevnt, vil disse bølgene ha en bølgefront som har form som en del av en sylinderoverflate med spalten som sylinderakse. I et snitt vinkelrett på spalten får vi da en rent todimensjonal beskrivelse (se figur 13.3).

Vi har med lys å gjøre, det vil si en elektromagnetisk bølge. Bølgen er transversal og beskrives av et elektrisk og magnetisk felt som har hver sin retning i rommet. Vi antar at vi betrakter interferensfenomenet så langt unna spaltene at vi kan se bort fra *forskjell* i retningen i rommet for elektrisk felt som stammer fra spalt 1 sammenlignet medfeltet som stammer fra spalt 2. Vi nøyer oss derfor med å addere elektrisk felt som skalarer ut fra korrekt intensitet og fase.

Vi ønsker å finne elektrisk felt et sted på skjermen i en retning θ relativt til normalvektoren midt mellom spaltene (se figur 13.2). Bidragene fra de to spaltene er da:

$$E_1(\theta_1) = E_{1,0}(r_1, \theta_1) \cos(kr_1 - \omega t - \phi)$$

$$E_2(\theta_2) = E_{2,0}(r_2, \theta_2) \cos(kr_2 - \omega t - \phi)$$



Figur 13.3: Skjematisk lysgang fra dobbeltspaltene til et gitt punkt på skjermen bak. I virkeligheten er avstanden R fra spalter til skjermen svært mye større enn avstanden d_1 mellom spaltene. Se tekst for detaljer.

hvor ϕ er en vilkårlig fasevinkel når rom og tid er gitt. Siden skjermen med spaltene og skjermen der vi fanger opp bildet er svært langt fra hverandre sammenlignet med avstanden mellom spaltene, vil vinklene θ_1 og θ_2 være svært nær identiske, og vi erstatter dem begge med θ :

$$\theta_1 \approx \theta_2 = \theta$$

Av samme grunn kan vi anta at amplitudene er identiske, dvs:

$$E_{1,0}(r_1, \theta_1) = E_{2,0}(r_2, \theta_2) = E_0(r, \theta)$$

Den totale amplituden i retning θ er da (ifølge superposisjonsprinsippet):

$$E_{tot}(\theta) = E_0(r, \theta)[\cos(kr_1 - \omega t - \phi) + \cos(kr_2 - \omega t - \phi)]$$

Vi bruker så en generell relasjon for cosinus:

$$\cos(a) + \cos(b) = 2 \cos\left(\frac{a+b}{2}\right) \cos\left(\frac{a-b}{2}\right)$$

og får:

$$E_{tot}(\theta) = 2E_0(r, \theta) \cdot \cos\left(k\frac{r_1 + r_2}{2} - \omega t - \phi\right) \cos\left(k\frac{r_1 - r_2}{2}\right)$$

Superposisjon skjer alltid "på amplitudenivå" (det vil si en reell fysisk størrelse, ikke en abstrakt størrelse så som energi og intensitet). Fysiske målinger er likevel ofte basert på intensitet. Når vi betrakter lys på en skjerm med øynene våre, er lysintensiteten vi fornemmer proporsjonal med intensiteten i bølgen.

Intensiteten for en plan elektromagnetisk bølge i fjernfeltsonen er gitt ut fra Poynting vektor, men skalarverdien er gitt ved:

$$I = cED = c\epsilon E^2$$

hvor c er lyshastigheten, E elektrisk felt, D elektrisk fluksstetthet (elektrisk forskyvning) og ϵ er elektrisk permittivitet. Da følger:

$$I(\theta, t) = c\epsilon E_{tot}^2(\theta, t) = 4c\epsilon E_0^2(r, \theta) \cdot \cos^2\left(k\frac{r_1 + r_2}{2} - \omega t - \phi\right) \cos^2\left(k\frac{r_1 - r_2}{2}\right)$$

Dette er en såkalt momentan intensitet som varierer med tiden innenfor en periode. Vi er mest interessert i tidsmidlet intensitet. Det første cosinus²-leddet varierer med tiden, og

tidsmidlet av \cos^2 er $1/2$. Følgelig:

$$I(\theta) = 2c\epsilon E_0^2(r, \theta) \cos^2 \left(k \frac{r_1 - r_2}{2} \right)$$

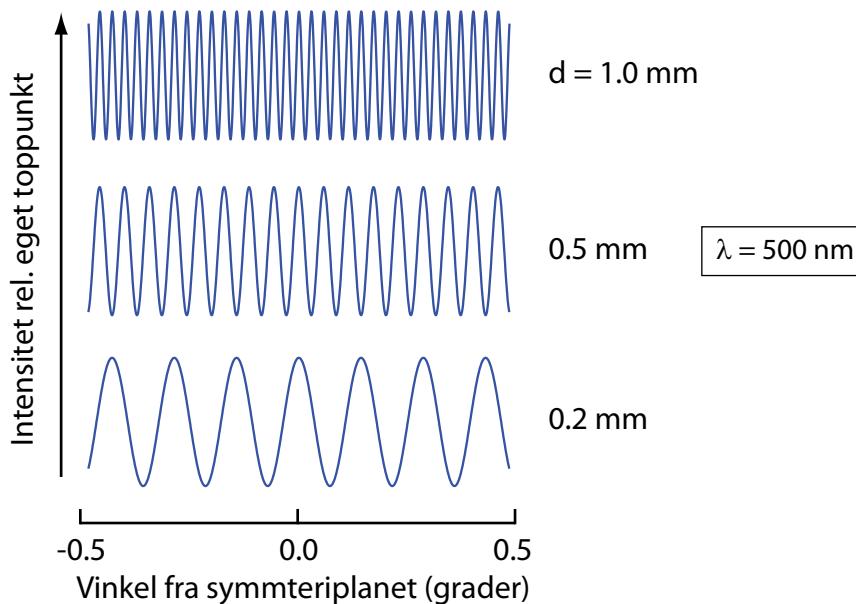
Vi definerer

$$r_1 - r_2 = \Delta r = d \sin \theta$$

hvor d er avstanden mellom spaltene. Videre trekker vi inn bølgelengden gjennom relasjonen $k = \frac{2\pi}{\lambda}$.

Da følger intensitetsfordelingen for lyset som har passert en dobbeltspalt:

$$\bar{I}(\theta) = 2c\epsilon E_0^2(r, \theta) \cos^2 \left(\frac{d \sin \theta}{\lambda} \pi \right) \quad (13.1)$$



Figur 13.4: Stripemønsteret på en skjerm etter dobbeltspaltene. Avstand mellom spaltene er angitt.

Når $\theta = 0$ får vi maksimal intensitet. Minima får vi når argumentet til cosinusfunksjonen er $\pi/2, 3\pi/2, 5\pi/2 \dots$:

$$\frac{d \sin \theta}{\lambda} \pi = (2n + 1) \frac{\pi}{2}$$

n er her et naturlig tall (inklusiv null). Det vil si minima får vi når:

$$\sin \theta = \frac{\lambda}{d} (n + \frac{1}{2})$$

Maksima får vi *omtrent* når:

$$\sin \theta = \frac{n\lambda}{d}$$

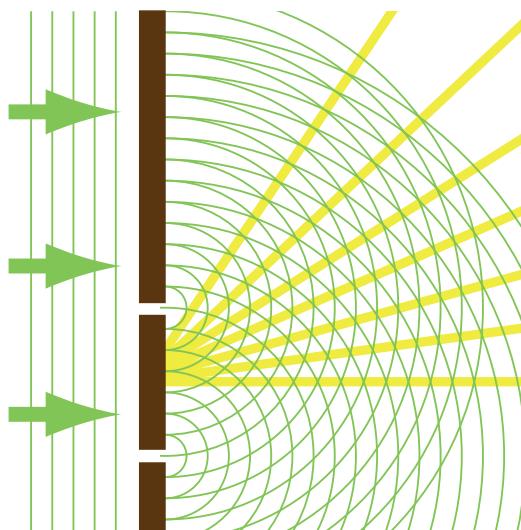
Ordet "omtrent" brukes siden det eksakte uttrykket for maksima også avhenger av hvordan $E_0^2(r, \theta)$ varierer med θ .

Vi kan merke oss at vanligvis, i alle fall for lys, er avstanden mellom spaltene stor i forhold til bølgelengden. Det vil si at vinkelen mellom to minima (eller mellom to maksima)

vanligvis er ganske liten. Det betyr at vi i prinsippet kan få et interferensmønster som består av svært mange parallele lyse stripene på skjermen med mørke partier mellom. Det blir altså ikke bare *to* stripene slik en partikkkelmodell for lys ville gitt.

Hvor mange stripene får vi egentlig? Vel, det avhenger av $E_0^2(r, \theta)$. Dersom vi bruker Huygens prinsipp og bare bruker én elementærbølge, skulle denne ha samme intensitet i alle retninger (der bølgen kan bre seg ut). Men spalten kan ikke være infinitesimal smal. Da ville praktisk talt ikke noe lys sluppet gjennom. Når spalten har en endelig bredde, skal vi egentlig la elementærbølgene starte ut i ethvert punkt i spalten. Disse elementærbølgene vil sette opp en totalbølge for spalt 1 og en totalbølge for spalt 2 som ikke vil ha samme elektrisk felt i alle retninger θ . Vi kommer til å behandle dette problemet nedenfor (diffraksjon fra én spalt).

Siden $E_0^2(r, \theta)$ bare vil være stor for et relativt smalt vinkelområde, får vi et begrenset antall linjer på skjermen når vi samler opp lyset fra dobbeltpalten. Vi skal se eksempler på noen forløp siden.



Figur 13.5: Retningen til interferenslinjene kan demonstreres ved å legge to sirkelmønstre oppå hverandre, men med sentrene et lite stykke fra hverandre (svarende til avstanden mellom spaltene).

I figur 13.5 viser vi til slutt en nokså vanlig måte å illustrere interferens ved en dobbeltspalt. Med sentrum i hvert av de to spaltene (og i et plan normalt på spaltene og midt mellom dem) er det tegnet inn bølgefronter, karakterisert ved at elektrisk felt f.eks. er maksimalt i en retning normalt på planet vi betrakter. Alle steder der bølgetopper fra en spalt treffer bølgetopp fra den andre spalten, vil vi få konstruktiv interferens og maksimalt elektrisk felt. Dette er steder der sirklene krysser hverandre.

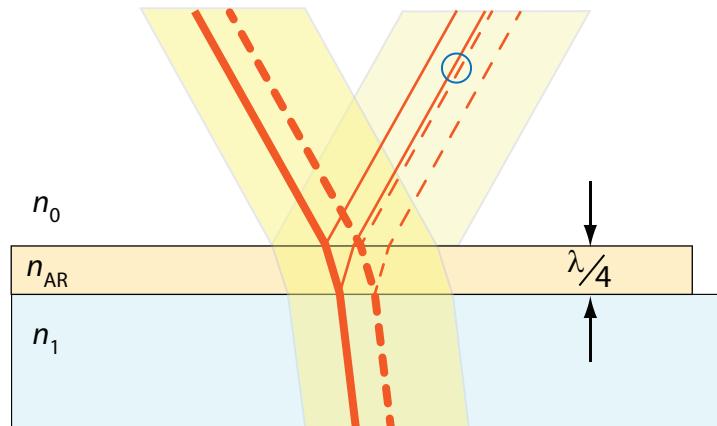
Steder der en bølgetopp fra en spalt treffer en bølgebunn fra den andre spalten (dvs midt mellom to sirkler fra denne spalten), vil vi få destruktiv interferens, og tilnærmet et neglisjerbart elektrisk felt.

Vi kan merke oss fra figuren at posisjoner med konstruktiv interferens ligger langs linjer som stråler ut omtrent midt mellom de to spaltene. Det er i disse retningene vi får stripene i interferensbildet fra en dobbeltpunkt. Midt mellom disse er det destruktiv interferens og lite eller ikke noe lys.

Det er instruktivt å få demonstrert hvordan vinklene til stripene endrer seg etter som vi endrer avstanden mellom sentrum i sirkelmønstrene.

13.3.1 Interferensfiltre, interferens fra en tynn film

Vi har tidligere sett at når vi sender lys inn mot en plan grenseflate mellom luft og glass, reflekteres om lag 5 % av lyset i overflaten (enda mer etter hvert som innfallsvinkelen øker). En slik refleksjon ødelegger kontrasten og bildekvaliteten generelt dersom linser f.eks. i en kikkert eller et foto-objektiv ikke har antirefleksbehandling. Men hvordan kan vi lage et antirefleks-belegg på en linse?

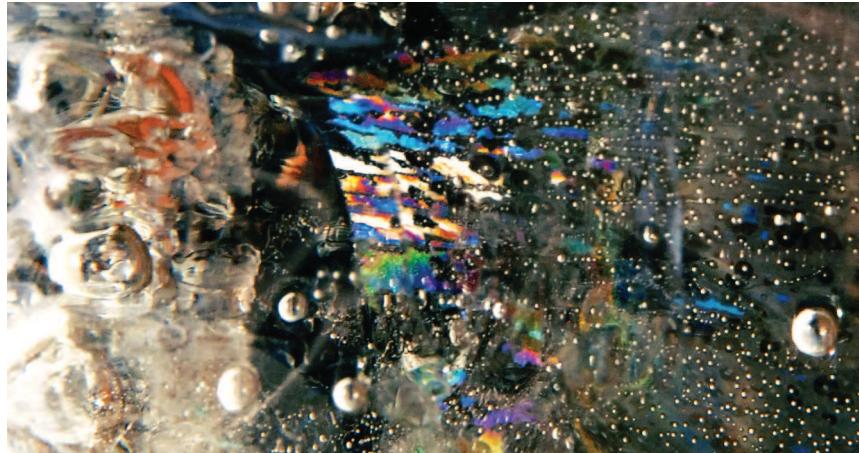


Figur 13.6: En antirefleks-behandling av en linse eller brille består i et tynt gjennomsiktig lag med brytningsindeks nær midt mellom luft og glass. Laget må være om lag en kvart bølgelengde tykt for de bølgelengdene der filteret har best effekt. En strålebunt som kommer litt på skrå inn mot overflaten er tegnet inn for å få fram summasjon mellom en del av bølgen som reflekteres på overflaten av antireflekslaget (stiplet) og en del av bølgen som reflekteres fra overflaten av selve glasset (heltrukket). Overlappet mellom disse er markert med en sirkel.

Figur 13.6 viser skjematisk hvordan vi kan gå fram. Vi legger et tynt lag med et eller annet gjennomsiktig stoff utenpå glasset, og velger et stoff som har en brytningsindeks omtrent midt mellom den til luft og glass. Vi får da reflektert omtrent like mye lys fra grenseflaten luft - belegg som fra belegg - glass. Dersom vi ser bort fra enda en refleksjon (i returstrålen) ser vi at lys som reflekteres fra øvre og nedre lag vil ha samme retning når de går tilbake til luften. De to "strålene" vil superponere. Dersom de to superponerer med motsatt fase, vil de langt på vei slokke hverandre ut. Det betyr at lyset som faktisk reflekteres (totalt sett) vil bli vesentlig mindre intenst enn dersom belegget ikke var på plass.

Ved å velge alle parametre med omhu, kan vi bestemme hvorvidt det skal bli destruktiv eller konstruktiv interferens. I det første tilfellet får vi som allerede vist et antirefleksende lag. I det andre tilfellet får vi økt refleksjon. Da benyttes ofte et belegg som består av flere lag oppå hverandre, og parametrenene velges slik at lys som reflekteres overalt kommer i fase med andre bidrag til refleks, og at det lys som transmitteres fra ulike lag alltid kommer i motfase med andre transmisjonsbidrag. På denne måten er det mulig å lage speil som kan ha mer enn 99.9 % refleksjon for en bestemt bølgelengde og for en bestemt retning for en lysstråle inn mot speilet, mens vi ved andre bølgelengder kan se tvers gjennom speilet! Det er ganske artig å oppleve slike speil!

I naturen og i hverdagen dannes tynne filmer spontant f.eks. i tynne sprekker eller tynne luftsjikt mellom f.eks. to glassplater. Legger vi for eksempel et "urglass" (svakt buet glass til et lommeur) oppå en plan glassflate, får vi ringer med henholdsvis konstruktiv og destruktiv interferens mellom lys som reflekteres i grenseflatene glass1-luft og luft-glass2.



Figur 13.7: Fargespill i en isklump med en sprekk.

Siden effekten er bølgelengdeavhengig, blir sirklene fargede, og de går under betegnelsen Newton-ringer.

I figur 13.7 er det vist et annet eksempel på samme sak. Det er en isbit hvor det er oppstått en tynn sprekk etter et slag mot biten, og fargespillet kommer tydelig fram.

13.4 Mange, parallelle spalter (optisk gitter)

Dersom vi har mange, parallelle spalter med samme innbyrdes avstand d , og samler opp lyset på en skjerm langt unna spaltene (sammenlignet med d), får vi en situasjon som kan analyseres på omtrent samme måte som for dobbeltpalten. Forskjellen ligger i at vi må summere bidrag fra alle N spaltene.

Resultantfeltet blir da:

$$\begin{aligned} E_{tot}(\theta) &= E_1 + E_2 + \dots + E_N \\ &= E_0(r, \theta)(\cos(kr_1 - \omega t - \psi) + \cos(kr_2 - \omega t - \psi) + \dots + \cos(kr_N - \omega t - \psi)) \end{aligned}$$

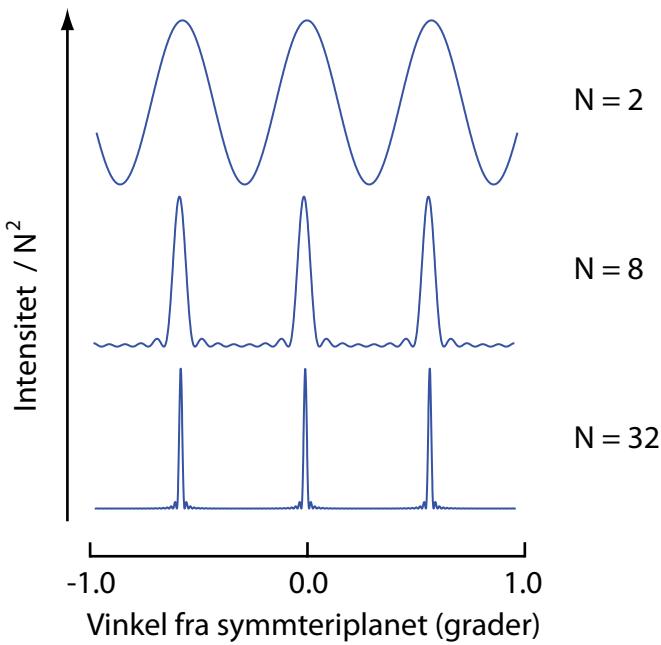
For å forenkle regningen videre innser vi at absolutt fase ψ i forhold til valgt posisjon og tid er uinteressant. Når vi siden bare skal se på tidsmidlet intensitet, er det bare faseforskjeller som skyldes ulik ganglengde for de ulike elementærbølgene som spiller inn. For en gitt vinkel θ vil gangforskjellen mellom to nærliggende elementærbølger være gitt ved $d \sin \theta$. Denne gangforskjellen representerer en faseforskjell ϕ , og vi har allerede ovenfor vist grunnlaget for at denne faseforskjellen er gitt ved $\phi = 2\pi d \sin \theta / \lambda$.

Merk at tar vi utgangspunkt i én spalt, vil faseforskjellen til den neste være ϕ , til den deretter 2ϕ , den neste 3ϕ osv. Da kan vi skrive resultantfeltet på denne forenkleda måten:

$$\begin{aligned} E_{tot}(\theta) &= E_0(r, \theta)(\cos(\omega t) + \cos(\omega t + \phi) + \cos(\omega t + 2\phi) + \dots + \cos(\omega t + (N-1)\phi)) \\ E_{tot}(\theta) &= E_0(r, \theta) \sum_{n=0}^{N-1} \cos(\omega t + n\phi) \end{aligned}$$

Vi bruker så Eulers formel $e^{i\theta} = \cos \theta + i \sin \theta$ og får:

$$\sum_{n=0}^{N-1} \cos(\omega t + n\phi) = \operatorname{Re} \sum_{n=0}^{N-1} e^{i(\omega t + n\phi)} = \operatorname{Re}(e^{i\omega t} \sum_{n=0}^{N-1} e^{in\phi})$$



Figur 13.8: Intensitetsfordeling vs vinkel for 2, 8 og 32 spalter.

Fra matematikken vet vi at summen av en endelig geometrisk rekke med koeffisient k er gitt ved:

$$1 + k + k^2 + \dots + k^{N-1} = \sum_{n=0}^{N-1} k^n = \frac{k^N - 1}{k - 1}$$

Vi anvender denne relasjonen for ledetet $\sum_{n=0}^{N-1} e^{in\phi}$ (k svarer til $e^{i\phi}$) og får:

$$\begin{aligned} \sum_{n=0}^{N-1} \cos(\omega t + n\phi) &= \operatorname{Re}(e^{i\omega t} \sum_{n=0}^{N-1} e^{in\phi}) = \operatorname{Re}(e^{i\omega t} \frac{e^{iN\phi} - 1}{e^{i\phi} - 1}) \\ &= \operatorname{Re}(e^{i\omega t} \cdot \frac{e^{iN\phi/2}}{e^{i\phi/2}} \cdot \frac{e^{iN\phi/2} - e^{-iN\phi/2}}{e^{i\phi/2} - e^{-i\phi/2}}) \\ &= \operatorname{Re}(e^{i\omega t} \cdot e^{iN\phi/2-i\phi/2} \cdot \frac{2i \sin \frac{N\phi}{2}}{2i \sin \frac{\phi}{2}}) \\ &= \operatorname{Re}(e^{i(\omega t + N\phi/2 - \phi/2)} \cdot \frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}}) \\ &= \cos(\omega t + N\phi/2 - \phi/2) \cdot \frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \end{aligned}$$

Kombinerer vi dette med tidligere uttrykk, blir elektrisk felt i retning θ :

$$E_{tot}(\theta) = E_0(r, \theta) \cos(\omega t + \frac{N\phi}{2} - \frac{\phi}{2}) \cdot \frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}}$$

På samme måte som for dobbeltspalten er vi interessert i intensiteten i interferensmønsteret vi kan observere. Igjen har vi:

$$I(\theta, t) = c\epsilon E_{tot}^2(\theta, t)$$

Når tidsmidlet beregnes, er $\overline{\cos[2(\omega t + \frac{N\phi}{2} - \frac{\phi}{2})]} = \frac{1}{2}$ som før. Følgelig:

$$I(\theta) = \frac{1}{2} c \epsilon E_0^2(r, \theta) \left[\frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \right]^2$$

$$I(\theta) = I_0(r, \theta) \left[\frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \right]^2$$

Vi kan vise (ved bruk av L'Hôpital's regel) at når ϕ går mot null, vil uttrykket inni firkantparantesen gå mot N . Det vil si at intensiteten i stripene som finnes ved $\phi = 0$ blir N^2 ganger intensiteten vi hadde hatt bare fra én spalt.

Andre maksima finner vi for $\sin \frac{\phi}{2} = 0$ (forutsatt at vi ser bort fra vinkelavhengigheten til $E_0^2(r, \theta)$). Siden vi har definert ϕ ved $\phi = 2\pi d \sin \theta / \lambda$, følger det at maksima vil forekomme når:

$$\sin(\pi d \sin \theta / \lambda) = 0$$

Det vil si:

$$m\pi = \pi d \sin \theta / \lambda$$

hvor $m = \dots, -2, -1, 0, +1, +2, \dots$

$$\sin \theta = \frac{m\lambda}{d}$$

Vi ser at posisjonene til intensitetsmaksimaene er uavhengig av antall spalter (N).

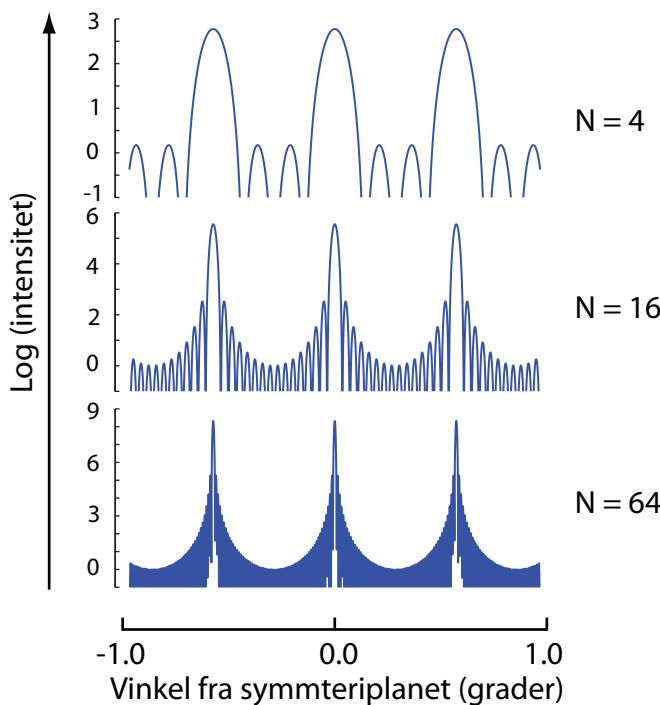
Figur 13.8 viser hvordan intensitetsfordelingen er for litt forskjellig antall spalter. Vi ser at det mest karakteristiske trekket er at toppene blir mer og mer markante når antall spalter øker.

Det kan vises at halvverdibredden for toppene er gitt ved:

$$\Delta\theta_{1/2} = \frac{1}{N \sqrt{(\frac{d}{\lambda})^2 - m^2}} \quad (13.2)$$

hvor m som før angir linjens "orden". Vi ser at sentrallinen $m = 0$ har minst linjebredde, og at linjebredden øker når vi betrakter linjer lenger og lenger vekk fra sentrum. (Spørsmål: Kan $(d/\lambda)^2 - m^2$ bli negativt?)

I figur 13.9 har vi tegnet de samme kurvene som i figur 13.8, men nå med logaritmisk y-akse. Hensikten er å vise detaljer i småtoppene mellom hovedtoppene. Vi ser at den nærmeste småtoppen til en hovedtopp er om lag tre log-enheter (nærmest en faktor 1000) mindre enn hovedtoppen. Det er ikke dramatiske avvik fra denne regelen selv om vi endrer antall spalter betydelig. Imidlertid ser vi at bredden på hver hovedtopp avtar med antall spalter, - også dersom vi tar med et par småtopper på hver side av hovedtoppen. Videre viser det logaritmiske plottet at intensiteten til hovedlinjene i forhold til minimumspunktet omrent midt mellom, øker dramatisk med antall spalter. Dette har sammenheng med at toppene får en intensitet lik N^2 ganger intensiteten fra hver enkelt spalt.



Figur 13.9: Intensitetsfordeling vs vinkel for 4, 16 og 64 spalter, men nå tegnet i logaritmisk skala langs y-aksen for å kunne studere detaljer nær nulllinjen mellom hovedtoppene.

13.5 Diffraksjon fra én spalt

Anta at vi nå har en enkel spalt som blyses fra én side med planpolariserte bølger med bølgefront parallellt med spaltens "flate". Vi kan *modellere* spalten som et optisk gitter hvor spaltene ligger så tett og har en så stor bredde at de går helt over i hverandre. Dersom spalten har en bredde a , kan vi altså tenke oss at spalten består av N smale parallele spalter med innbyrdes avstand $d = a/N$.

Det finnes to ulike beregningsmetoder for lysintensiteten på en skjerm etter spalten. Den enkleste metoden er basert på en tilnærming der skjermen antas å være svært langt unna spalten, sammenlignet både med spaltens bredde og bølgelengden. Dette tilfellet kalles Fraunhofer-diffraksjon, og er karakterisert ved at styrken på elektrisk felt fra hver av de tenkte delspaltene er tilnærmet identisk på skjermens plass, og at vinkelen fra en delspalt til en gitt posisjon på skjermen er tilnærmet lik vinkelen fra en annen delspalt til samme posisjon.

Dersom avstanden mellom spalten og skjermen ikke er svært stor relativt til spaltens bredde og/eller bølgelengden, må vi bruke mer nøyaktige uttrykk for både styrke på feltbidrag og vinkler. Dette tilfellet kalles Fresnel-diffraksjon, og er vanskeligere å håndtere enn Fraunhoferdiffraksjon. Med numeriske metoder er det likevel overkommelig, noe vi kommer tilbake til senere i kapitlet.

La oss nå gå tilbake til den enkle Fraunhofer-diffraksjonen hvor vi altså betrakter en spalt som sammensatt av N smale parallele spalter som ligger kant i kant. Vi kan nå bruke samme uttrykk som for optisk gitter, forutsatt at vi erstatter d med a/N . I uttrykket for faseforskjellen ϕ får vi nå følgende sammenheng:

$$\phi = 2\pi \frac{d \sin \theta}{\lambda} = 2\pi \frac{a \sin \theta}{N\lambda} = \frac{\beta}{N}$$

hvor

$$\beta = 2\pi \frac{a \sin \theta}{\lambda}$$

Siden lysintensiteten fra hver enkelt tenkt spalt vil være $I_0(r, \theta)/N$, blir den totale intensitetsfordelingen:

$$I(\theta) = \frac{I_0(r, \theta)}{N} \left[\frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \right]^2 = \frac{I_0(r, \theta)}{N} \left[\frac{\sin \frac{\beta}{2}}{\sin \frac{\beta}{2N}} \right]^2$$

Når N velges meget stor, vil β/N være så liten at $\sin \frac{\beta}{2N} \approx \frac{\beta}{2N}$. Da kan vi skrive intensitetsfordelingen slik:

$$I(\theta) = \frac{I_0(r, \theta)}{N} \left[\frac{\sin \frac{\beta}{2}}{\frac{\beta}{2N}} \right]^2$$

$$I(\theta) = NI_0(r, \theta) \left[\frac{\sin \frac{\beta}{2}}{\frac{\beta}{2}} \right]^2$$

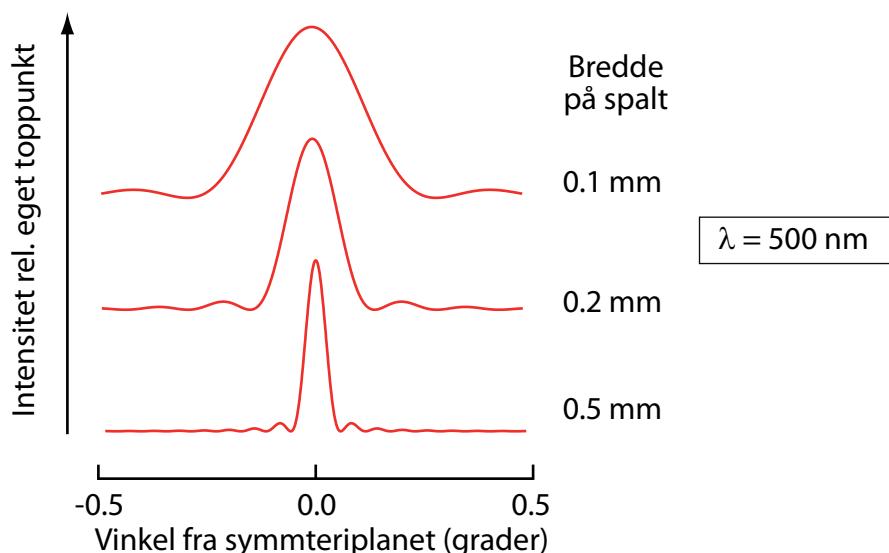
Det kan her synes som om maksimal intensitet går til uendelig, noe den selvfølgelig ikke gjør. En nøyere analyse viser at det som redder oss er at intensitetsfordelingen $I_0(r, \theta)$ er mye videre enn fordelingen $I(\theta)$. En bred spalt fører til en smalere lysbunt langt unna spalten enn en smal spalt, noe vi kan upresist angi som $NI_{0,max,smal} \approx I_{max,bred}$.

Vi er først og fremst interessert i *formen* til intensitetsfordelingen og dropper detaljer mhp sammenligninger av absolutt intensitet. Da får vi:

$$I(\theta) = I_{max} \left[\frac{\sin \frac{\beta}{2}}{\frac{\beta}{2}} \right]^2 \quad (13.3)$$

hvor

$$\beta = \frac{2\pi a}{\lambda} \sin \theta$$



Figur 13.10: *Intensitetsfordeling for stripene etter en enkelt spalt.*

Dette uttrykket kan i første omgang se temmelig likt ut som intensitetsfordelingen fra et optisk gitter. Hva ligger forskjellen i?

Vinkel mellom sentraltoppen og første minimum for et optisk gitter er gitt ved:

$$\phi = 2\pi \frac{d \sin \theta}{\lambda} = \frac{2\pi}{N}$$

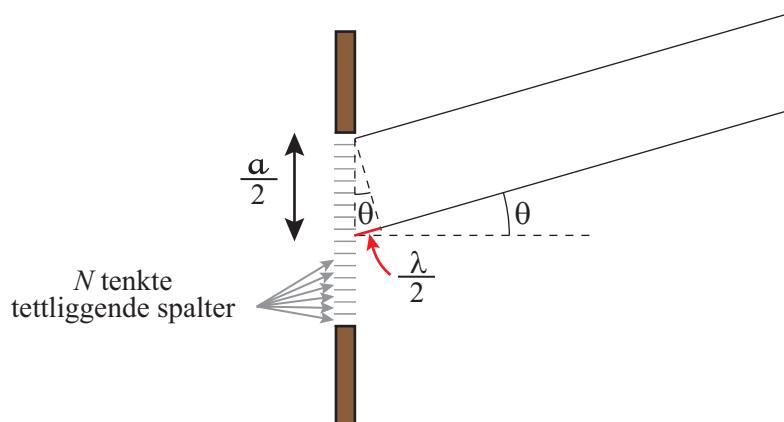
$$\sin \theta = \frac{\lambda}{Nd}$$

For enkeltpalten tenkte vi oss imidlertid at spalten var delt opp i N tettliggende spalter, og avstanden mellom disse tenkte spaltene er da a/N . Setter vi dette uttrykket inn i det forgående, finner vi et uttrykk for vinkelen som svarer til første minimum i intensiteten. Vi får samme svar dersom vi startet med ligning (13.3) og vet at $x = \pi$ svarer til første minimum i $\sin(x)/x$ -funksjonen. Følgelig:

Første minimum i intensiteten fra en enkeltpalt finnes ved vinkelen θ der

$$\sin \theta = \frac{\lambda}{a}$$

Dette uttrykket er uavhengig av N , noe vi forventet.



Figur 13.11: Geometriske forhold som viser hvilken retning intensiteten til diffraksjonen fra en enkeltpalt vil være null. For ethvert valg av tenkte spaltepar med avstand mellom hverandre lik $a/2$ vil gangforskjellen i lysveien være lik en halv bølgelengde (hvilket gir destruktiv interferens).

Dette resultatet kan vi finne også ved en helt annen betraktnign. Figur 13.11 viser hvordan vi kan tenke oss at to og to tenkte spalter virker sammen for å få destruktiv interferens for alt lyset gjennom spalten. Vi skjønner også ut fra figuren at minimum for diffraksjonen fra én spalt alltid må finnes ved større vinkler enn diffraksjonen fra to eller flere separate spalter (siden avstanden d mellom spalter nødvendigvis må være større eller lik spaltebredden i et optisk gitter). Med andre ord: Vinkelavstanden til første minimum for et optisk gitter kan lett bli mye mindre enn for vinkelavstanden til første minimum i diffraksjonsmønsteret.

Vi kan beregne halvverdibredden for intensitetsfordelingen fra enkeltpalten ved å benytte ligning (13.2) for et optisk gitter, men igjen erstatte spalteavstanden d med vår tenkte spalteavstand a/N . Da følger:

$$\Delta\theta_{1/2} = \frac{1}{N\sqrt{(\frac{a}{N\lambda})^2 - m^2}}$$

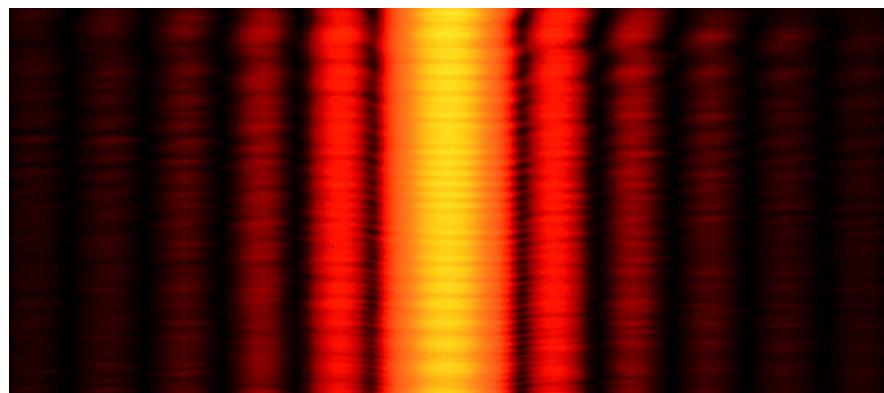
$$= \frac{1}{\sqrt{\left(\frac{a}{\lambda}\right)^2 - (Nm)^2}}$$

Halvverdibredden for sentraltoppen i diffraksjonen fra en enkeltpalt er da ($m = 0$):

$$\Delta\theta_{1/2} = \frac{\lambda}{a}$$

Vi ser at uttrykket (selvfølgelig) er uavhengig av vår tenkte N .

Intensitetsfordelingen for enkeltpalten kan typisk se ut omtrent som vist i figur 13.10 og 13.12. Det er en markant klokkeformet topp med svake stripene på siden. Det kan lett vises at vi ikke får flere markante topper enn sentraltoppen (siden nevneren aldri blir null unntatt for sentraltoppen).

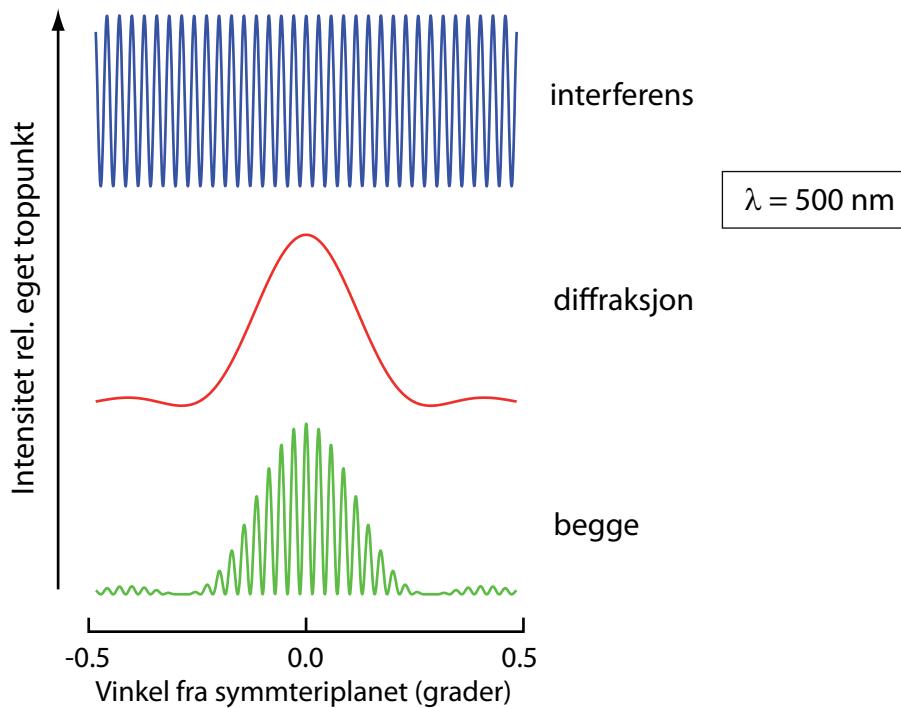


Figur 13.12: Eksempel på observert intensitetsfordeling for stripemønsteret fra en enkeltspalt. Sentrallinjen er overeksponert for at sidelinjene skal komme godt fram.

13.6 Kombinert effekt

I utviklingen av uttrykket for intensitetsfordelingen fra en enkeltpalt, gjorde vi ikke noe særlig nummer av at styrken på det elektriske feltet vil variere med vinkelen θ . I uttrykkene for dobbeltpalt og optisk gitter la vi mer vekt på dette. Grunnen er at det faktisk er den underliggende diffraksjonen fra hver enkelt spalt som danner bakteppet for $E_0^2(r, \theta)$! Vi får ikke det mest tydelige stripemønsteret fra en dobbeltpalt eller fra et optisk gitter til å strekke seg lenger ut enn den sentrale toppen i diffraksjonsbildet fra hver enkeltpalt.

I praksis vil vi derfor alltid få en kombinert effekt av diffraksjon fra en enkeltpalt og interferens fra to eller flere samtidige spalter. Figur 13.13 viser den kombinerte effekten av diffraksjon fra hver enkelt av de to parallelle spaltene, og interferens som skyldes at vi har to spalter. Eksemplet er valgt slik at det skal svare til et optimalt dobbeltpalteksperiment der det er et betydelig antall godt synlige linjer innenfor den sentrale diffraksjonstoppen.



Figur 13.13: Beregnet intensitetsfordeling for stripemønsteret fra en dobbeltpunktspalt når hver av spaltene er 200 bølgelengder brede og avstanden mellom spaltemidtene er 2000 bølgelengder.

13.7 Diffraksjon i en videre ramme

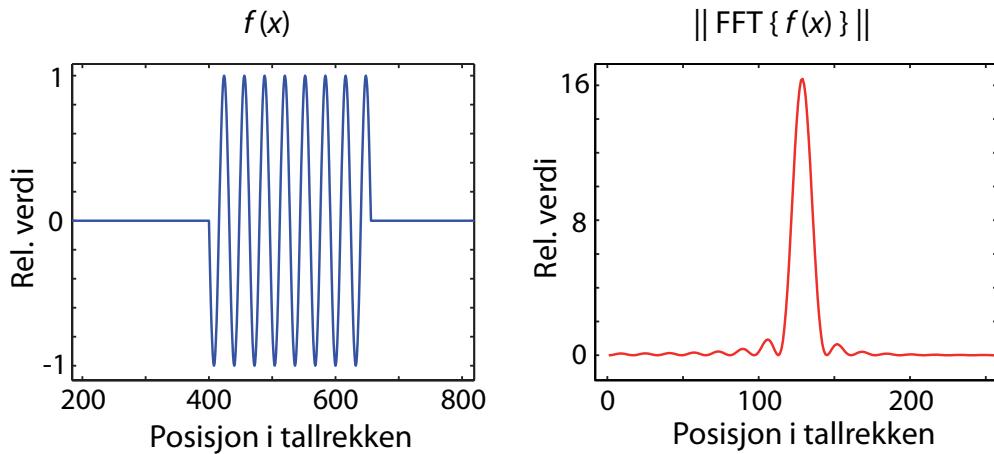
Vi har nettopp utledet intensitetsfordelingen av lys på en skjerm etter at det har passert en smal spalt. Intensitetsfordelingen like etter spalten kan betraktes som en romlig firkantpuls. Intensiteten som fanges opp på en skjerm et godt stykke unna viste imidlertid en intens klokkeformet senterlinje med smålinjer på hver side (se figur 13.10). De to nærmeste sidelinjene har intensiteten 4.72 % og 1.65 % av senterlinjens intensitet. Er det noe magisk med denne forandringen fra en firkantent til en klokkeformet intensitetsfordeling? På en måte er det faktisk det.

I figur 13.14 er det vist kvadratetet til den fouriertransformerte til en sinuskurve multiplisert med en firkantfunksjon. Den fouriertransformerte kurven har eksakt samme form som intensitetsfordelingen vi beregnet for diffraksjon fra en enkeltspalt. Dette er et eksempel på en del av optikken kalt “Fourier-optikk”.

Dersom vi multipliserer en sinusfunksjon med en gaussisk kurve i stedet for en firkantkurve, blir kvadratet av den fouriertransformerte en rent gaussisk kurve. Starter vi eksperimentelt med en gaussisk intensitetsfordeling i en stråle, vil strålen både kunne gjøres smalere eller bredere, ved hjelp av linser og diffraksjon, og likevel beholde sin gaussiske intensitetsfordelingsform. Med andre ord vil diffraksjon ikke føre til noen topper utenfor midtlinjen når strålen har en gaussisk intensitetsfordeling.

Det kan vises mer generelt at intensitetsfordelingen for diffraksjon fra en spalt henger nøye sammen med intensitetsfordelingen til lysstrålen vi starter ut med. Med andre ord kan intensitetsfordeling anses som en form for “randbetingelser” når en bølge brer seg utover / treffer materialer som setter begrensinger for bølgebevegelsen.

I moderne optikk benyttes ofte laserstråler med gaussisk intensitetsfordeling på tvers



Figur 13.14: Multipliseres et sinussignal med en firkantpuls, får vi et signal som vist til venstre i denne figuren (bare det interessante området er tatt med). Her er det brukt 4096 punkter i beskrivelsen, sinussignalet har 32 punkter per periode, og firkantpulsen er valgt slik at vi får åtte hele perioder innenfor firkanten. Dersom dette signalet fouriertransformeres, og vi beregner lengden på hvert av de komplekse tallene som fremkommer, og attpåtil beregner kvadratet av hvert tall, får vi kurven vist til høyre i figuren (bare det interessante området er tatt med). Kurven har eksakt samme form som kurven vi beregnet for diffraksjon fra en enkeltspalt.

av strålen. Da vil stråleformen beholdes selv etter at strålen stadig er gjenstand for diffraksjon.

Det er utviklet en nydelig formalisme basert på matriser (kalt ABCD-metoden) som kan benyttes for å beregne hvordan diffraksjon endrer størrelsen til en laserstråle (forutsatt at intensitetsprofilen er gaussisk). I formalismen inngår først og fremst to størrelser som betyr alt for hvordan en slik stråle utvikler seg. Den ene er diameter for strålen (diameter mellom punkter der intensiteten har falt til $1/e^2$ av peakverdien). Den andre parameteren er krumningsradius for bølgefronten som funksjon av posisjon langs strålen. Formalismen baserer seg på ”små vinkler”. Dette til orientering.

[♠ ⇒ Test deg selv:

Opplysningene som er gitt i figurteksten til figur 13.14 henger sammen med selve figuren. Har du lyst å teste hvor mye du husker fra fouriertransformasjon, kan du forsøke å svare på følgende spørsmål:

1. Kan du forklare hvorfor toppen i høyre del av figuren havner der den er?
2. Er det noen sammenheng mellom posisjonen hvor firkantpulsen opptrådte i venstre del av figuren, og posisjon/intensitet i høyre del av figuren? Begrunn som vanlig svaret!
3. Dersom firkantpulsen i venstre del bare var halvparten så bred som i vårt tilfelle, hvordan ville du da forventet at høyre figur ville se ut?

⇐ ♠]

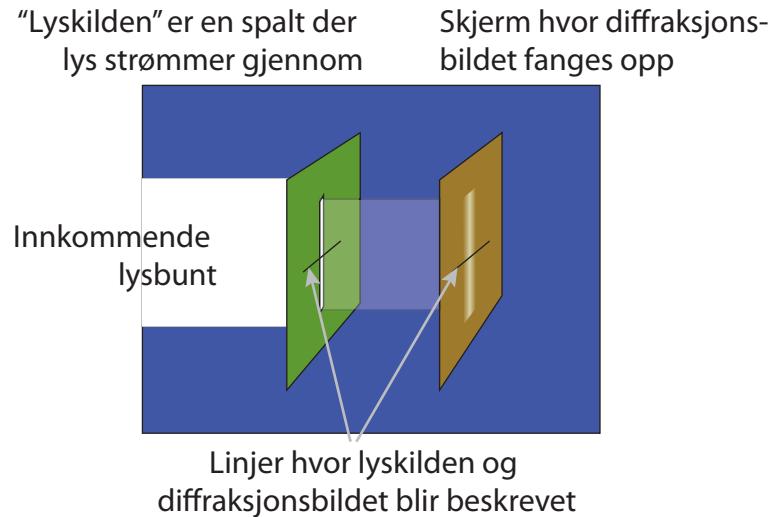
13.8 Numerisk beregning av diffraksjon

Utledningene vi har gjennomført hittil er basert på analytisk matematikk. Det har gitt oss matematiske uttrykk for intensitetsfordelinger ved ulike diffraksjonsfenomener. Disse

uttrykkene er gull verdt. Imidlertid baserer uttrykkene seg på tilnæringer som bare representerer grensetilfeller av en langt mer kompleks virkelighet.

Vi skal nå se hvordan numeriske metoder kan hjelpe oss til å beregne diffraksjon for et langt større variasjonsområde for de parametrene som inngår.

For enkelhets skyld tar vi utgangspunkt i beregning av diffraksjon fra en spalt. Lys sendes normalt inn på en plan flate hvor det er en rektangulær åpning, en spalt, der lengden er mye større enn bredden. Det fysiske systemet har en høy grad av sylindersymmetri, og vi nøyser oss derfor å betrakte elektrisk felt og intensiteter langs en éndimensjonal linje på tvers av spalten (se figur 13.15).



Figur 13.15: Skisse som viser hvor vi beskriver lyskilden og diffraksjonsbildet ved beregning av diffraksjon fra en spalt.

13.8.1 Den grunnleggende modellen

Modelleringen ved vår numeriske beregning er den samme som vi benyttet ved utledning av de analytiske løsningene, bare at vi slipper å gjøre så drastiske tilnærninger som der. Figur 13.16 viser hvordan vi går fram.

Vi tar utgangspunkt i elektromagnetiske bølger som stammer fra N kildepunkter langs en linje på tvers av spalten. Punktene har posisjoner x_n som varierer fra $-a/2$ til $a/2$ siden spaltens bredde er a (se figur 13.16). Amplituden til det elektriske feltet er A_n , slik at den elektromagnetiske bølgen i punktet x_n er

$$\vec{E}_n = A_n e^{i(kz - \omega t + \theta_n)} \vec{u}_n$$

der symbolene har sin vanlige mening, bortsett fra u_n som bare er en enhetsvektor som angir det elektriske feltets retning (vinkelrett på bølgens bevegelsesretning på det angitte stedet). θ_n er et ledd som gir relativ fase fra ett punkt til et annet på tvers av spalten. Dersom bølgefronten på den innkommende lysbunten er parallel med spaltens plan, er alle θ_n identiske, og parameteren kan i så fall sløyfes.

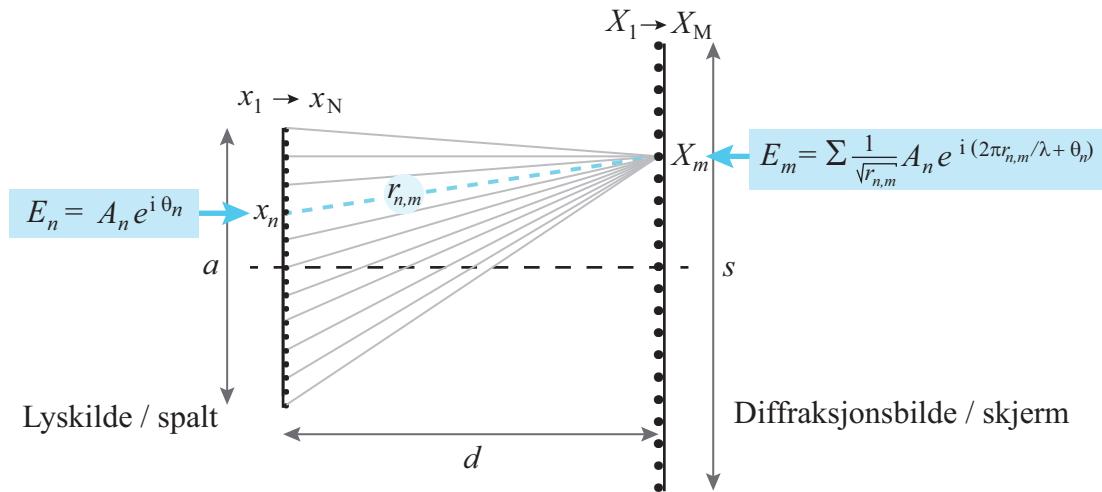
I vår modellering av diffraksjon vil vi ta utgangspunkt i elektrisk felt ved samtidighet i hele spalten. Da vil ledetet $e^{-i\omega t}$ være et konstant faseledd som faller bort når intensiteter skal beregnes til slutt. Vi dropper derfor å ta med dette ledetet. På tilsvarende vis fjerner

vi allerede i utgangspunktet det tilsvarende ledet ved beregningene av feltet på skjermen hvor diffraksjonsbildet fanges opp.

Dersom vi velger $z = 0$ i spalten, ender vi opp med en forenklet skrivemåte for elektrisk felt i de ulike punktene på tvers av spalten:

$$\vec{E}_n = A_n e^{i\theta_n} \vec{u}_n \quad (13.4)$$

La oss så betrakte diffraksjonsbildet. Det blir fanget opp på en skjerm parallel med spalten, i en avstand d fra spalten. Numerisk beregner vi diffraksjonsbildet i M punkter symmetrisk plassert i forhold til midtpunktet på spalten. Beregningene spenner over en bredde s slik at posisjonen til de valgte punktene X_m går fra $-s/2$ til $s/2$. Vi må selv velge en passe verdi for s for å fange opp de interessante delene av diffraksjonsmønsteret (men heller ikke så mye mer).



Figur 13.16: Skisse som indikerer hvordan Huygen-Fresnels prinsipp brukes ved beregning av diffraksjon fra en spalt.

Det elektriske feltet i et vilkårlig punkt X_m vil være summen av bidrag fra elektromagnetiske bøller som kommer fra alle punktene x_n i spalten. Siden avstanden $r_{n,m}$ mellom de aktuelle punktene endrer seg etter som vi gjennomløper alle x_n , vil bidragene ha forskjellig fase. Dessuten gjør avstandsforskjellene sitt til at amplituden av det elektriske feltet vil bli redusert. Totalt sett får vi da følgende uttrykk for summasjon av alle bidrag til det elektriske feltet i punkt X_m :

$$\vec{E}_m = \sum \frac{A_n}{\sqrt{r_{n,m}}} e^{i(2\pi r_{n,m}/\lambda + \theta_n)} \vec{u}_{n,m}$$

Uttrykket er problematisk, for det er ingen enkel måte å finne retningen $\vec{u}_{n,m}$ på hvert enkelt elektrisk feltbidrag. Vi blir derfor mer eller mindre tvunget til å behandle elektrisk felt som skalare størrelser i en slik formalisme. Som allerede nevt tidligere i kapitlet, er dette ikke et stort problem når vi betrakter diffraksjonsbildet langt fra spalten. Svært nær spalten vil imidlertid den skalare tilnærmingen være en klar feilkilde i vår beregninger.

Det grunnleggende uttrykket for numerisk beregning av diffraksjon fra en spalt, er da:

$$E_m = \sum \frac{A_n}{\sqrt{r_{n,m}}} e^{i(2\pi r_{n,m}/\lambda + \theta_n)} \quad (13.5)$$

hvor

$$r_{n,m} = \sqrt{d^2 + (X_m - x_n)^2} \quad (13.6)$$

Intensiteten i det aktuelle punktet er proporsjonal med kvadratet av det elektriske feltet.

Merk her at vi har brukt kvadratroten av avstanden ved beregning av redusert elektrisk feltstyrke. Det har sammenheng med at vi har sylinderSymmetri. Sender vi ut lys langs en linje, vil intensiteten gjennom enhver sylinderflate med sentrum i linjen bli den samme. Sylinderflaten har areal $2\pi r L$ dersom sylinderen har lengde L . Siden intensitet er proporsjonal med elektrisk feltstyrke kvadrert, må da elektrisk felt i seg selv avta med kvadratroten av avstanden. Hadde vi hatt sfærisk geometri, ville intensiteten fordelt seg på en kuleflater med areal $4\pi r^2$, og det elektriske feltet ville avtatt som $1/r$.

13.8.2 Ulike løsninger

Beregninger basert på uttrykkene (13.5) og (13.6) kan i enkelte sammenhenger bli krevende siden det inngår beregninger av sinuser, cosinuser og kvadrater og kvadratrøtter i hvert eneste ledd. Dessuten må det $N \times M$ beregninger til. For moderne datamaskiner er dette godt overkommelig for rett fram beregninger av diffraksjon. Likevel, dersom diffraksjonsberegnene inngår i mer omfattende beregninger av bildedannelser basert på fourier optikk med mere, er uttrykkene ovenfor faktisk litt for regnemaskinkrevende selv i dag.

Historisk sett har det derfor vært utformet ulike forenklinger i forhold til uttrykkene ovenfor for å få regnetiden ned. I mange aktuelle situasjoner hvor vi studerer diffraksjonsbilder av lys, er $a \ll d$ og $s \ll d$ i figur 13.16. Vi kan da bruke en Taylorutvikling i uttrykket for $r_{n,m}$ i stedet for ligning (13.6). Resultatet er (kan for moro skyld forsøke å utlede uttrykket selv):

$$r_{n,m} = \sqrt{d^2 + (X_m - x_n)^2} \approx d \left(1 + \frac{1}{2} \frac{(X_m - x_n)^2}{d^2} - \frac{1}{8} \frac{(X_m - x_n)^4}{d^4} \right) \quad (13.7)$$

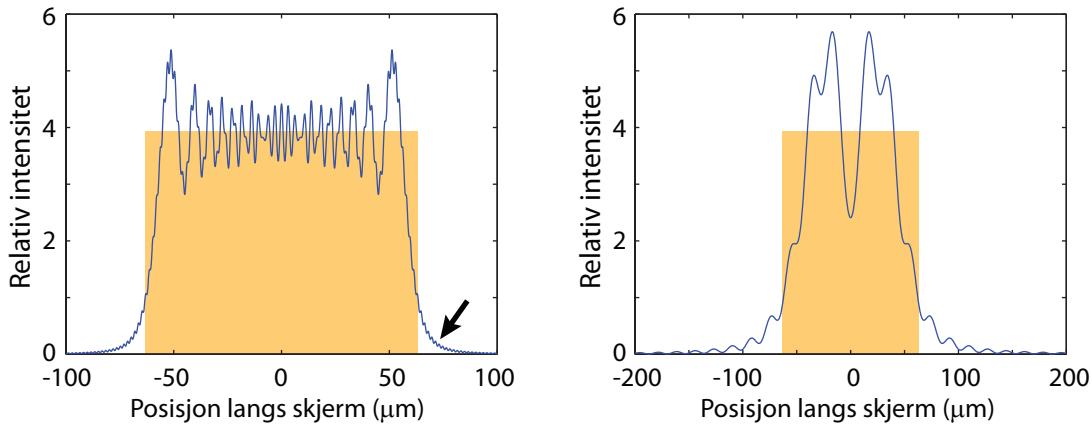
I ligning (13.5) er det viktigste leddet der $r_{n,m}$ inngår leddet $e^{i2\pi r_{n,m}/\lambda}$. Setter vi inn det tilnærmede uttrykket for $r_{n,m}$, får vi:

$$e^{i2\pi r_{n,m}/\lambda} \approx e^{i2\pi d/\lambda} \cdot e^{i\pi \frac{(X_m - x_n)^2}{d}/\lambda} \cdot e^{-i\pi \frac{1}{4} \frac{(X_m - x_n)^4}{d^3}/\lambda} \quad (13.8)$$

I ulike situasjoner vil noen av disse leddene praktisk talt være konstanter, og nettopp dette er utgangspunkt for noen historiske klassifiseringer av diffraksjon.

Vi skal nå forsøke å gi en oversikt over ulike varianter av beregningsnøyaktighet:

- 1. Mindre enn noen få bølgelengder unna kantene på spalten.** Her må vi bruke Maxwells ligninger, og ta med polarisering og strømmer i materialet som lager spalten. "Evanescent waves" inngår i løsningen.



Figur 13.17: Diffraksjon fra en spalt beregnet ut fra Huygen-Fresnels prinsipp. Figurens venstre del svarer til at vi er temmelig tett på spalten. Høyre del svarer til litt lengre vekk fra spalten, men likevel ikke så langt som Fraunhofer diffraksjon som ble behandlet i med analytisk matematikk tidligere i kapitlet. Bredden på spalten er markert med gult rektangel.

2. **For $d^3 \leq 2\pi \frac{a^4}{\lambda}$** . Dette er et problematisk område hvor Maxwells ligninger må brukes for de minste d , mens uttrykkene (13.5) og (13.6) begynner å fungerer rimelig bra for de største d som tilfredsstiller grensen som er gitt.
3. **For $d^3 \gg 2\pi \frac{a^4}{\lambda}$ har vi Huygen-Fresnel diffraksjon.** Uttrykkene (13.5) og (13.6) fungerer. Selv om vi setter $1/\sqrt{r_{n,m}}$ lik $1/\sqrt{d}$ og vi sløyfer siste ledd i rekkeutviklingen i ligning (13.7), blir resultatet tilfredsstillende.
4. **For $d \gg \pi \frac{a^2}{\lambda}$ har vi Fraunhofer diffraksjon.** Uttrykkene (13.5) og (13.6) fungerer. Selv om vi bruker de samme tilnærmingene som for Huygen-Fresnel diffraksjon og atpå tilsetter $(X_m - x_n)^2 \approx X_m^2 + 2X_m x_n$ i midtre ledd i rekkeutviklingen i ligning (13.7), blir resultatene tilfredsstillende.

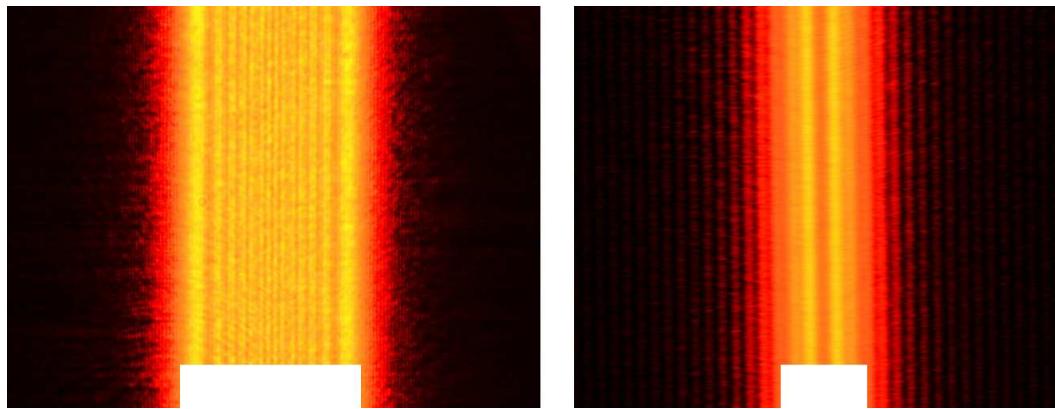
Figur 13.17 viser numeriske beregninger basert på uttrykkene (13.5) og (13.6) direkte, for et tilfelle hvor vi er relativt nær spalten (Huygen-Fresnel sonen) og for et tilfelle i overgangen mellom Huygen-Fresnel og Fraunhofer-sonene.

Legg merke til at når vi er nær spalten (Huygen-Fresnel sonen), vil diffraksjonsbildet på skjermen ha omtrent samme størrelse som spalten. Noe av intensiteten i kanten av spalten siver likevel ut i skyggepartiet (markert med pil i figuren), slik at det er en helt gradvis intensitetsfordeling mellom skygge og full lysintensitet. Vi får karakteristiske stripemønstre i bildet av spalten. Det er større “romlig bølgelengde” på disse stripene nær kanten på spalten enn mot midten. Det er kun svake stripemønstre i skyggepartiet på hver side av bildet av spalten.

Figur 13.18 viser et fotografi av to diffraksjonsbilder som har lignende trekk som den numeriske beregningen.

Helt tilsvarende har vi gjengitt beregninger og et eksempel på diffraksjonsbilde i grenseland mellom Huygen-Fresnel og Fraunhofer-sonene i høyre del av figurene 13.17 og 13.18. Vi har her litt bølgefornemmelse både i bildet av spalten og i lyset i skyggesonene.

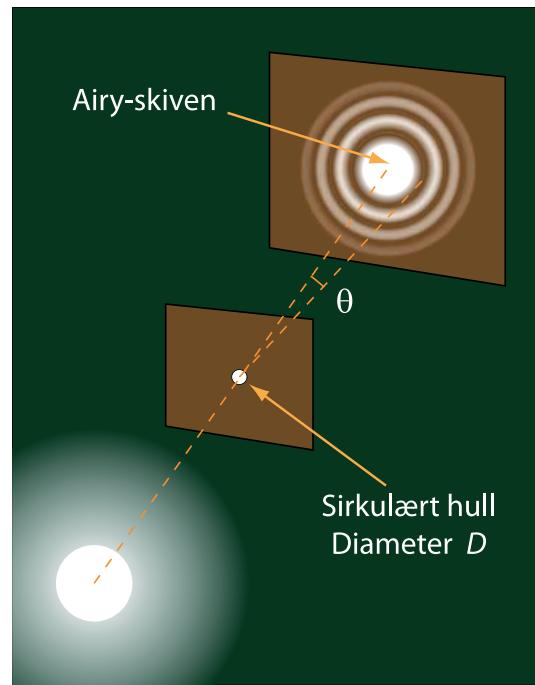
Diffraksjonsbildet for ren Fraunhofer-sone er akkurat den samme som vi har utledet analytisk, og eksempler på resultater er allerede gitt i figur 13.10 og et fotografi i figur 13.12. I det tilfellet har vi bare bølgefornemmelse i sonen utenfor den sentrale toppen.



Figur 13.18: *Fotografi av diffraksjonsbildet av en spalt ved omtrent de avstandene som beregningene i figur 13.17 tilsvarer. Størrelsen på spalten er markert nederst.*

13.9 Diffraksjon fra et rundt hull

Når en plan bølge sendes inn mot et sirkulært hull, får vi også diffraksjon (se figur 13.19 og 13.20), men det er vanskeligere å sette opp en matematisk analyse av det problemet enn for spalter. Resultatet er at bildet som kan samles opp på en skjerm viser en markant sentral klokkeformet topp, med svake ringer rundt. Den sentrale toppen synes å danne en sirkulær skive som går under navnet “Airy skiven”.



Figur 13.19: *Eksperimentelt oppsett for å observere diffraksjon fra et rundt hull.*

Matematisk sett er intensiteten i en vinkelavstand θ vekk fra senterlinjen gitt ved:

$$I(\theta) = I_{max} \left[1 - J_0^2 \left(\frac{1}{2} k D \sin \theta \right) - J_1^2 \left(\frac{1}{2} k D \sin \theta \right) \right]$$

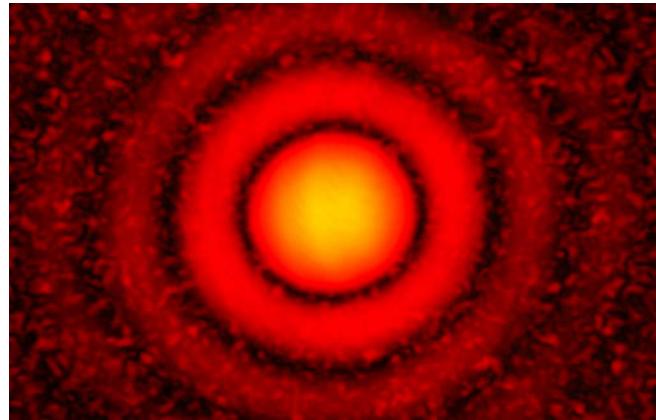
hvor J_0 og J_1 er de to første Besselfunksjonene, D er diameteren til hullet og k bølgetallet. (Verdiene av Besselfunksjoner kan lett beregnes numerisk.)

Vinkelen til første minimum er gitt ved:

$$\sin \theta = \frac{1.22\lambda}{D}$$

hvor D er diameteren på hullet. Siden vinkelen vanligvis er meget liten, kan vi ofte bruke tilnærmingen:

$$\theta = \frac{1.22\lambda}{D}$$



Figur 13.20: *Airy-skiven slik den ser ut med en del overeksponering i det midtre partiet for å få fram sirklene rundt. Overeksponering er vanskelig å unngå siden maksimum intensitet i første ring er bare 1.75 % av maksimum intensitet i den sentrale toppen. Det er kommet med noen "specles", antakelig pga spredt laserlys i rommet.*

Dette uttrykket og hele fenomenet har vidtgående konsekvenser, og vi skal nevne noen.

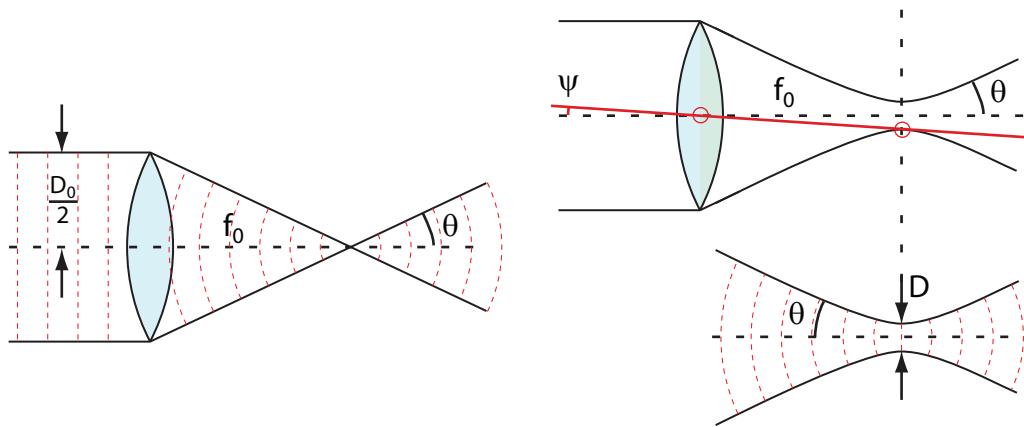
13.9.1 Bildet av stjerner i et teleskop

Lys fra en stjerne kommer inn mot et teleskop. Lyset kan betraktes som en plan bølge når det når objektivet. Objektivet fokuserer lyset med linser og/eller speil. I geometrisk optikk får vi inntrykk av at vi kan samle alle lysstrålene fra et fjernet objekt i ett punkt, brennpunktet, slik det er indikert i venstre del av figur 13.21. I det minste skulle det være mulig dersom vinkeldiameteren til objektet er svært liten, slik det er når vi betrakter stjerner på stjernehimmelen. Det er feil!

Lysbunten fra en stjerne vil følge en form lignende den som er vist i høyre del av figur 13.21. Lysbunten får en minste diameter D som er vesentlig større enn det vi skulle forvente ut fra vinkeldiameteren til objektet (stjernen). Årsaken er diffraksjon.

Vi har tidligere påpekt at Maxwellts ligninger kan kjøres like godt forlengs som baklengs. Vi kan da tenke oss å *starte* med intensitetsfordelingen på strålens smaleste punkt. Anta at strålen der har en diameter D og konstant intensitet over hele denne diameteren. Det kan vises at lysstrålen har en plan bølgefront der diameteren er minst i figur 13.21, og det var det som også var utgangspunkt for vår behandling av diffraksjon tidligere i dette kapitlet.

Vi har sett at diffraksjon i et slikt tilfelle vil føre til at lysbunten sprer seg utover med



Figur 13.21: I geometrisk optikk tenker vi oss at paralleltt lys inn til en linse samles i brennpunktet. Fra diffraksjonslæren vet vi at en lysstråle aldri kan bli punktformig, men at lysbunten går via en minste diameter i "midjen" til strålen (der vi finner Airy-skiven) og utvider seg deretter (på begge sider) slik vi beskriver i diffraksjon. Et korrekt bilde er kombinasjonen av disse to (nederst). Størrelsen på midjen (Airy-skiven) medfører at det finnes en minste vinkel ψ hvor det er mulig å skille to objekter. De røde stippled linjene antyder bølgefrontens posisjon (flater med samme fase i det elektriske feltet).

en vinkel θ (relativt til senterlinjen i strålen)

$$\theta \approx \frac{1.22\lambda}{D}$$

Denne voksende, kjegleformede lysbunten som skyldes diffraksjon fra lysstrålens smaleste punkt når vi tenker oss lyset kjørt baklengs, må samsvare med den avtakende kjegleformede lysbunten som skyldes den konvekse linsen når vi tenker oss lyset kjørt forlengs.

[♣ ⇒ Argumentasjonen er i virkeligheten litt mer komplisert enn så, fordi lysstrålen ikke har en konstant intensitet over hele tverrsnittet der den har minst diameter. Tvert om er intensitetsfordelingen temmelig lik den som er vist i figur 13.20. Det morsomme er likevel at denne intensitetsfordelingen nettopp fører til en konstant intensitetsfordeling når strålen baklengs når teleskopobjektivet. Det er symmetrien i forlengs og baklengs fouriertransformasjon som sørger for dette. ← ♣]

Vinkelen θ gitt ut fra geometrisk optikk settes lik θ basert på diffraksjon fra et hull:

$$\frac{D_o/2}{f_o} = \tan \theta \approx \sin \theta \approx \theta = \frac{1.22\lambda}{D}$$

hvor D_o er diameteren på objektivet og f_o er brennvidden på objektivet.

Radien i en Airy-skive i fokalplanet vil da bli:

$$\frac{D}{2} = \frac{1.22 \cdot \lambda f_o}{D_o}$$

Airykiven til en stjerne vi ha denne utstrekningen, selv om vinkelutstrekningen på himmelen er forsvinnende liten. Alle stjernene vil i prinsippet lage like store lysende skiver i fokalplanet, men intensiteten vil avspeile stjernenes lysstyrke.

To stjerner som ligger nær hverandre på himmelen vil danne delvis overlappende skiver i fokalplanet. Er overlappet meget stort, vil vi ikke kunne se at det er to skiver. Vi vil vurdere dem som én. Er overlappet lite, vil vi skjønne at det er to skiver; med andre ord at det er to stjerner.

Lord Rayleigh formulerete dette omtrent slik:

Når to objekter (eller detaljer i objekter) betraktes i et teleskop, vil grensen for å kunne skille de to objektene fra hverandre være at sentralmaksimum i Airy-skiven fra det ene objektet faller sammen med første diffraksjonsminimum fra det andre objektet.

Denne beskrivelsen er kjent som Rayleighs oppløsningskriterium. Vinkel ψ i figur 13.21 svarer til at vi såvidt kan se at det er to Airy-skiver. Med andre ord: Med et objektiv med diameter D_o og brennvidde f_o kan vi såvidt skjelne to stjerner (eller andre nærmiljøformige objekter) fra hverandre dersom vinkelavstanden mellom stjernene minst er

$$\psi \approx \frac{D/2}{f_o} = \frac{1.22\lambda}{D_o} \quad (13.9)$$

Eksempler:

Som vi nettopp har sett er vi ikke i stand med en kikkert til å skille detaljer som har en vinkelavstand mindre enn $\frac{1.22\lambda}{D_o}$, uansett hvor mye vi forstørrer bildet. For en prismekikkert med objektiv på ca 5 cm diameter, vil minste vinkelavstand vi kan løse opp med 500 nm lys være

$$\frac{1.22 \cdot 500 \cdot 10^{-9}}{0.05}$$

som svarer til 0.00069 grader. For Mount Palomar teleskopet, med et speil på 5 m diameter, er beste oppløsning 1/100 av denne vinkelen. Mount Palomar-teleskopet kan løse opp detaljer som er ca 50 m fra hverandre på månen, mens en prismekikkert kun vil kunne løse opp detaljer som ligger 5 km fra hverandre.

Diameteren på pupillen i øyet vårt er om lag 5 - 10 mm i mørke. Det betyr at vi uten hjelpebidrag bare kan skille detaljer på månen som er minst 25 - 50 km fra hverandre.

I en prismekikkert er forstørrelsen nesten bestandig så liten at vi ikke får sett Airy-skivene. I et teleskop hvor vi kan endre okularer slik at forstørrelsen kan bli ganske stor, er det vanlig å se Airy-skivene. En stjerne ser ikke ut som et punkt når den betraktes med stor forstørrelse gjennom et teleskop. Stjernen ser ut akkurat som diffraksjonsbildet fra en liten sirkulær åpning i en skjerm, med en sentral skive (Airy-skiven) omgitt av svake ringer. Ringene er ofte så lyssvake at det er vanskelig å få øye på dem.

Mange kikkerter og teleskop har dårlig optikk slik at f.eks. sfæriske feil, kromatisk feil eller andre uperfektheter gjør at vi ikke får fram noe fin Airy-skive dersom vi forstørrer opp bildet av en stjerne. Vi får i stedet en mer eller mindre uregelmessig lysende flate som

dekker et enda større vinkelområde enn Airy-skiven ville ha gjort. For slike teleskoper klarer vi ikke å løse opp så fine detaljer som Rayleigh-kriteriet tilskier.

Kikkerter som er så perfekte at det er Airy-skiven som setter begrensingen på oppløsningen, sies å ha *diffraksjonsbegrenset optikk*. Dette er et kvalitetsstempel!

13.9.2 Divergens i en lysstråle

Ved Alomar-observatoriet på Andøya er det installert en ozon-lidar der en laserstråle sendes 8-90 km opp i atmosfæren for å observere sammensetning og bevegelser av molekyler der oppe. Lysstrålen bør være så smal som mulig langt der oppe, og vi kan lure på hvordan dette kan oppnås.

Første valg ville kanskje være å anvende en smal laserstråle direkte fra en laser. Strålen er da typisk 1 - 2 mm i diameter. Hvor bred ville denne strålen bli f.eks. i en høyde av 30 km?

Vi bruker relasjonen for diffraksjon fra et sirkulært hull og finner divergensvinkelen θ :

$$\sin \theta = \frac{1.22 \cdot \lambda}{D}$$

Dersom vi anvender lys med bølgelengde 500 nm og har strålediameteren 2.0 mm i starten, får man:

$$\sin \theta = \frac{1.22 \cdot 500\text{e}^{-9}}{0.002} = 3.05\text{e}^{-4}$$

Vinkelen er liten, og dersom diameteren på strålen ved 30 km høyde kalles $D_{30\text{ km}}$ følger da:

$$\frac{D_{30\text{ km}}/2}{30\text{ km}} = \tan \theta \approx \sin \theta = 3.05\text{e}^{-4}$$

$$D_{30\text{ km}} = 18.3\text{ m}$$

Med andre ord, laserstrålen som var 2 mm i diameter ved bakken har vokst til 18 m diameter i 30 km høyde!

Andre valg vil være å utvide laserstrålen slik at den starter ut mye bredere enn de 2 mm. Anta at vi utvider strålen slik at den faktisk er $D = 50\text{ cm}$ i diameter ved bakken. Anta at bølgefronten er plan ved bakken slik at strålen i starten er parallell (såkalt "midje") og etter hvert divergerer.

Hvor stor blir da diameteren ved $R = 30\text{ km}$ høyde?

Vi må da være litt omhyggelig når vi angir divergensvinkelen, og får:

$$\frac{D_{30\text{ km}}/2 - D/2}{R} \approx \tan \theta \approx \sin \theta = \frac{1.22 \cdot \lambda}{D}$$

Løser vi denne ligningen mhp $D_{30\text{ km}}$ får vi 57.3 cm. Med andre ord, strålen som startet ut som 50 cm bred, er bare blitt 57.3 cm bred i 30 km høyde!!! Dette er vesentlig bedre enn om vi startet med en 2 mm tynn stråle!

Vi kan imidlertid gjøre det *enda* bedre! Vi kan velge å plassere laseren (lyskilden) ikke nøyaktig i brennpunktet for det 50 cm speilet som vi brukte i stad (som et ledd på å gjøre strålen bred). Plasserer vi laseren litt utenfor brennpunktet, vil strålen faktisk konvergere før den når "midjen" (som svarer til Airy-skiven) og deretter divergerer igjen. Se figur 13.21. Hvor liten kan vi få midjen (Airydisken) i 30 km høyde?

Vi kan da regne baklengs og anse ”midjen” i 30 km høyde være kilde til en divergerende stråle (på begge sider av midjen, siden vi har symmetri her). I så fall vil strålen akkurat ha divergert til D lik 50 cm på speilets plass (tenker oss altså at strålen går baklengs). Regnestykket vil da se slik ut:

$$\frac{D/2 - D_{30 \text{ km}}/2}{R} \approx \tan \theta \approx \sin \theta = \frac{1.22 \cdot \lambda}{D}$$

$$D_{30 \text{ km}} = 42.7 \text{ cm}$$

Med andre ord, vi kan til og med få en mindre beam enn det vi startet ut med.

Konklusjon: Starter vi opp med en 2 mm diameter laserstråle ved bakken, blir den 18 m i diameter ved 30 km høyde. Starter vi derimot med en stråle på 50 cm diameter og fokuserer til riktig høyde, er strålen ”bare” 43 cm i diameter i samme høyde. Energitettheten i tverrsnittet er da over 400 ganger så stor som i det første tilfellet.

13.9.3 Andre eksempler

1. Som nevnt ovenfor representerer Airy-skiven en intensitetsfordeling som er ”klokkeformet”. Formen skyldes at lyset vi startet ut med hadde jevn belysning i hele tverrsnittet av det sirkulære hullet i en skjerm (firkant-funksjon i intensitet). Hva skjer dersom intensitetsfordelingen over hullet ikke er uniform?

I moderne optikk bruker vi ofte, som allerede nevnt, laserstråler som har en såkalt ”Gaussisk intensitetsprofil”. Intensiteten avtar da som en Gaussfunksjon fra aksen og utover. Det kan vises at sender du en slik stråle gjennom speil og linser, vil den Gausiske formen beholdes, selv om halvverdibredden kan variere f.eks. alt etter hvor langt vi er fra linser. Den Gausiske formen er på en måte en ”egenfunksjon” når vi beregner strålens/bølgens vandring i rommet vha Maxwells ligninger. Vi får ikke noe diffraksjonsringer rundt sentralstrålen.

2. Det finnes i dag ulike måter å komme unna diffraksjonsbegrensingen på. Det kan vi oppnå ved å bruke materialer som har en finstruktur som er mindre enn bølgelengden. Sender vi f.eks. lys gjennom et hull i et materiale, og lar diameteren til hullet være mindre enn bølgelengden, kan vi observere finere detaljer enn bølgelengden i materialer som holdes nær hullet. Derimot, opererer vi med de elektromagnetiske bølgene i ”fjernsonen” der randbetingelser spiller liten rolle, vil diffraksjon bestandig sette begrensning i oppløsningen. Vi rekker ikke å gå inn på disse finurlighetene, men det er en stadig voksende gren av moderne fysikk.
3. I øyet vårt er igjen oppløsningen begrenset av diffraksjon. Pupillens åpning er typisk 6 mm eller mindre under daglige gjøremål. Det setter en begrensning på hvor liten vinkelavstand to detaljer i synsbildet vårt vi kan holde adskilt. Det *kunne* vært en annen begrensning i øyets oppløsingsevne dersom hver enkelt synscelle var stor i forhold til diffraksjonens Airy-skive, men også øyet er et eksempel på at evolusjonen har elsket fram løsninger som er optimale, fysisk sett. Synscellene har omrent perfekt størrelse i forhold til diffraksjonen som forekommer.
4. I et kamera er det ikke nødvendigvis like god tilpasning. Dersom vi velger en bildebrikke som gir mange pixler per bilde, betyr det ikke nødvendigvis at vi kan *utnytte* denne oppløsningen. Dersom Airy-skiven for det valgte objektivet er større enn størrelsen på et pixel i CMOS-brikken, er det et tegn på at konstruksjonen ikke er optimal. Mange mobiltelefonkameraer reklamerer med stor oppløsning for

det innebygde kameraet, men kan slett ikke utnytte oppløsningen til CMOS-brikken på grunn av at optikken ikke er god nok. Det er morsomt å teste dette på egen mobiltelefon!

- Bredden i sentraltoppen i diffraksjonsbildet fra en enkelt spalt er som vi har sett gitt ved:

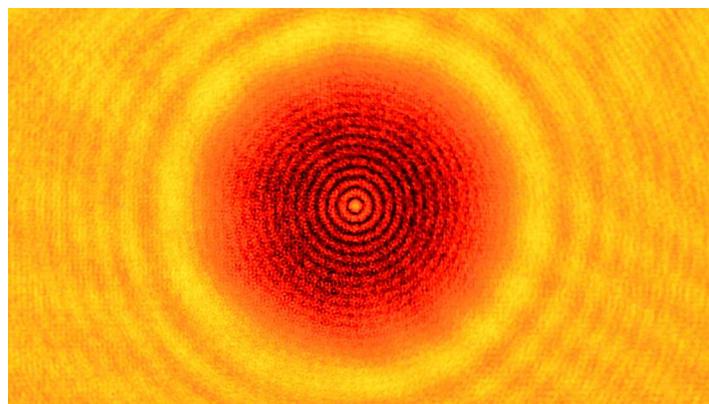
$$\Delta\theta_{1/2} = \frac{\lambda}{a} \quad (13.10)$$

I kvantefysikken har dette resultatet iblant tatt til inntekt for Heisenbergs uskarphetsrelasjon. Det er imidlertid mange svakheter ved en slik betraktningsmåte! Uttrykket i ligning (13.10) kan vi riktignok få til å stemme, men alle de flotte detaljene vi kan vise til i dette kapitlet, kan ikke Heisenbergs uskarphetsrelasjon gi oss!. Vi har oppnådd våre diffraksjonsrelasjoner ved å betrakte lyset som bølger og brukt Huygens prinsipp.

Også i andre deler av denne boka har vi fått relasjoner som minner om Heisenbergs uskarphetsrelasjon. I alle disse situasjonene er det bølgeegenskaper som ligger bak.

Det er derfor ikke så rart at mange i dag oppfatter Heisenbergs uskarphetsrelasjon som en naturlig konsekvens av bølgenaturen til lys og materie, og at den bare sekundært har noe med måleusikkerhet å gjøre.

- Diffraksjon har spilt en viktig rolle i vår oppfatning av lys. På begynnelsen av 1800-tallet viste Poisson at dersom lys hadde bølgenatur og oppførte seg etter Huygens prinsipp, skulle vi forvente å se en lys flekk i skyggebildet av en kule (eller sirkulær skive). Arago gjennomførte eksperimentet og fant at det faktisk var en lys flekk i midten (se figur 13.22). Fenomenet går nå under navnet Arago's flekk (eller Poisson-Arago's flekk)



Figur 13.22: Fotografi av Aragos flekk i skyggebildet fra en kule. Kula ble holdt på plass ved at den var limt opp med en liten (!) dråpe lim på et tynt stykke mikroskopi-dekkglass. I tillegg til Aragos flekk ser vi en rekke detaljer som skyldes diffraksjon, både i skyggepartiet og det belyste partiet. Merk at det er ingen klar grense mellom skygge og lys.

13.9.4 Diffraksjon ved to og tre dimensjoner

I vår behandling av interferens og diffraksjon har vi hittil bare sett på summering av bølger fra elementærbølge-kilder som ligger langs en rett linje. Det er en normal situasjon for interferens og diffraksjon for lys.

For andre typer bølger kan vi finne spredesentra som danner to- eller tre-dimensjonale mønstre. Mest kjent er kanskje røntgendiffraksjon. Når røntgenstråler sendes inn mot en krystall av et eller annet stoff, vil enkeltatomer spre røntgenstrålene slik at elementærbølgene kommer fra hvert enkelt atom i det området røntgenstrålen går gjennom.

Atomene i en krystall ligger i et regelmessig mønster. Plukker vi ut atomer som ligger på en linje / i et plan, vil elementærbølgene fra disse atomene gi interferens-linjer eller interferenspunkter som kan beregnes med tilsvarende ligninger som de vi har vært gjennom i dette kapitlet.

Både i fysikk og kjemi gir såkalt "røntgendiffraksjon" informasjon som kan brukes for å bestemme strukturen i de krystallene som undersøkes. Det er denne type undersøkelser som ligger bak omtrent alt vi har av detaljert informasjon hvordan atomene ligger i forhold til hverandre i ulike stoffer.

Figur 13.23 illustrerer at punkter som ligger regelmessig i forhold til hverandre, danner linjer som kan gi interferens/diffraksjon i mange ulike retninger.



Figur 13.23: *Fotografi fra en militærgravplass i San Diego, hvor gravsteinene er plassert meget regelmessig. I enkelte retninger ser vi mange gravsteiner på linje. Dersom disse gravsteinene sendte ut elementærbølger, ville vi få interferens-mønstre liknende de vi har diskutert for optisk gitter i dette kapitlet. Vi ville få en rekke interferenslinje-sett, men midtpunktet for hvert sett ville svare til de ulike retningene til linjene vi ser i bildet. Avstanden mellom linjene i hvert enkelt sett ville avhenge av avstanden mellom kildepunktene langs den retningen vi betrakter.*

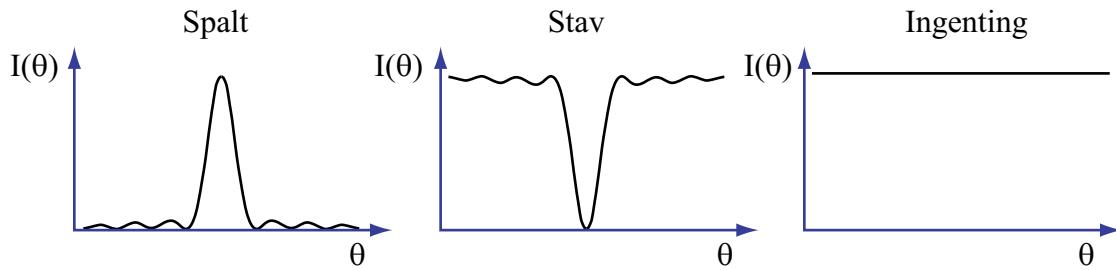
13.10 Babinet's prinsipp

Superposisjonsprinsippet kan brukes på en litt spesiell måte der vi utnytter symmetrier.

Vi har utledet hvordan intensitetsfordelingen blir når vi sender lys gjennom en smal spalt. Hvordan ville interferensbildet sett ut ved den komplementære strukturen, som er en stav med nøyaktig samme størrelse som spalten? Det sier Babinet's prinsipp noe om:

Sender vi en plan bølge mot en lang spalt, får vi et diffraksjonsbilde med en sentral kraftig lysende stripe med lyse stripene på hver side. Sender vi en plan bølge mot en lang stav med samme bredde og lengde som spalten, får vi en mørk stripe omgitt av andre mørke stripene. Summen av de elektromagnetiske bølgene fra de to tilfellene, må være identisk med de elektromagnetiske bølgene vi hadde hatt dersom verken skjerm med spalt eller stav var til stede.

Figur 13.24 viser prinsippet (figuren er en forenkling av virkeligheten siden vi bare angir intensiteter).

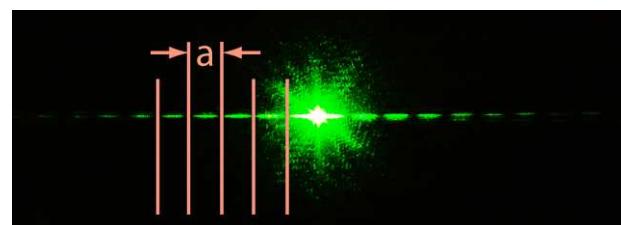


Figur 13.24: Intensitetsfordelingen fra en spalt og en stav er komplementære.

Dersom vi sender en relativ smal laserstråle inn mot en spalt og dernest mot en tråd med samme tykkelse som spalten, kan vi igjen bruke Babinets prinsipp for å finne ut (omtrent) hvordan de to diffraksjonsbildene vil forholde seg til hverandre. Forholdene er likevel temmelig forskjellig fra den meget brede lysstråle/planbølge-situasjonen vi viser i figur 13.24. Utenfor den smale laserstrålen er nemlig intensiteten praktisk talt lik null når ikke spalten eller tråden finnes i lysveien. Men med spalt eller tråd inne, får vi stripene ut i det området det ellers ikke ville vært noe lys.

Dette kan forstås ved at superposisjonsprinsippet bare kan anvendes på amplitudenivå, ikke på intensitetsnivå. Riktignok gikk det rimelig greit å jobbe på intensitetsnivå for plane bølger og spalt og stav (i alle fall overfladisk sett), men når vi bruker en smal laserstråle forstår vi at utslagene for spalten og tråden *utenfor* laserstrålen må ligge på nøyaktig samme sted. På amplitudenivå må de to oppsettene gi eksakt samme elektrisk felt i ethvert valgt punkt, men feltet må ha motsatt fortegn for spalsituasjonen sammenlignet med stavsjtuasjonen.

Det betyr at vi kan bruke teorien for diffraksjon fra en spalt også ved analyse av diffraksjonsbildet fra en tråd. Figur 13.25 viser diffraksjonsbildet fra et enkelt hår plassert i strålen fra en laserpenn. Med meget enkle midler kan vi ved å måle avstanden mellom minimumspunktene mellom lysflekkene bestemme tykkelsen på håret, forutsatt at vi kjenner bølgelengden til laseren. En oppgave sist i kapitlet gir et konkret eksempel på hvordan en konkret måling kan falle ut.



Figur 13.25: Diffraksjon fra et menneskehår i en smal laserstråle.

13.11 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for Huygen/Fresnels prinsipp.
- Utlede betingelsen for konstruktiv interferens fra en dobbeltspalt (når spalten antas å være meget smal).
- Beskrive interferensmønsteret fra en dobbeltspalt, og angi hvorfor forsøket til Thomas Young fikk stor historisk betydning.
- Gi hovedideen for en vanlig antirefleksbehandling av optikk.
- Angi kvalitativt hvordan interferensbildet endrer seg når flere enn to parallelle, identiske spalter benyttes.
- Med litt hjelp kunne utlede et matematisk uttrykk for intensitetsfordelingen for flere enn to parallelle, identiske spalter.
- Forklare kvalitativt intensitetsfordelingen i et diffraksjonsbilde fra en smal enkeltspalt når vi betrakter bildet langt fra spalten.
- Med litt hjelp kunne utlede et matematisk uttrykk for intensitetsfordelingen fra en smal enkeltspalt ved Fraunhofer-betingelser.
- Beregne ved hjelp av numeriske metoder interferensbilder også for Fresnel-diffraksjon.
- Angi hvordan diffraksjonsbildet ser ut for lys som går gjennom et sirkulært hull.
- Gjøre rede for hvordan diffraksjon setter begrensinger for hvor nær to stjerner kan være på himmelen før vi ikke lenger klarer å skille dem når vi betrakter dem gjennom et teleskop.
- Beregne maksimal oppnåelig vinkelopløsning for linser i mange ulike sammenhenger (øyet, kameralinser, teleskop m.m.).
- Kjenne til Babinets prinsipp.
- Kjenne til såkalt Aragos flekk (også kalt Poissons flekk), og hvorfor dette fenomenet fikk en historisk betydning.

13.12 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Er det mulig å gjennomføre et Youngs dobbeltpalteksperiment med lyd? Drøft mulig eksperimentelt oppsett og hvorvidt det er forskjell på longitudinell og transversell bølge i denne sammenheng.
2. Vi bruker superposisjonsprinsippet “på amplitudenivå” i stedet for “på intensitetsnivå”. Forklar hvorfor.
3. Vi har et teleskop og ønsker å sjekke om et objekt vi observerer er en dobbeltstjerne eller ikke. Vi trenger med andre ord *litt* større oppløsning, og vi antar at teleskopet har såkalt “diffraksjonsbegrenset” optikk. Hva mener vi med dette uttrykket? Kan vi øke oppløsingsevnen ved å “blende ned” slik at vi bare bruker en sentral del av objektivet? Eller kan vi øke oppløsingsevnen ved å sette inn et filter som slipper gjennom lys enten i det blå området eller i det røde området?
4. I et diffraksjonsekspertiment med en enkeltpalt og lys med bølgelengde λ er det ikke noe intensitetsminimum. Hva kan vi da si om bredden på spalten?
5. En vanlig regnbue får vi når dråpene er over en viss størrelse. For svært små dråper blir regnbuen nesten hvit. Hvor små tror du dråpene må være for at det skal skje?
6. I en stereohøyttaler brukes gjerne en basshøyttaler med relativt stor diameter, men en diskanhøyttaler som er bare noen få cm i diameter. Forsök å gi en forklaring på dette valget ut fra det du vet om diffraksjon.
7. Hvorfor er et optisk gitter (med mange spalter) bedre enn en dobbeltpalt dersom den skal brukes i et spektrometer hvor vi skal kunne måle bølgelengder?
8. Diffraksjon fra en enkeltpalt har betydning også for interferensbildet fra et optisk gitter. Forklar sammenhengen.
9. Forsök å beskrive essensen i figur 13.21 i kompendiet. Legg spesiell vekt på hva som er likhet og ulikheter mellom venstre og høyre del av figuren.

Regneoppgaver

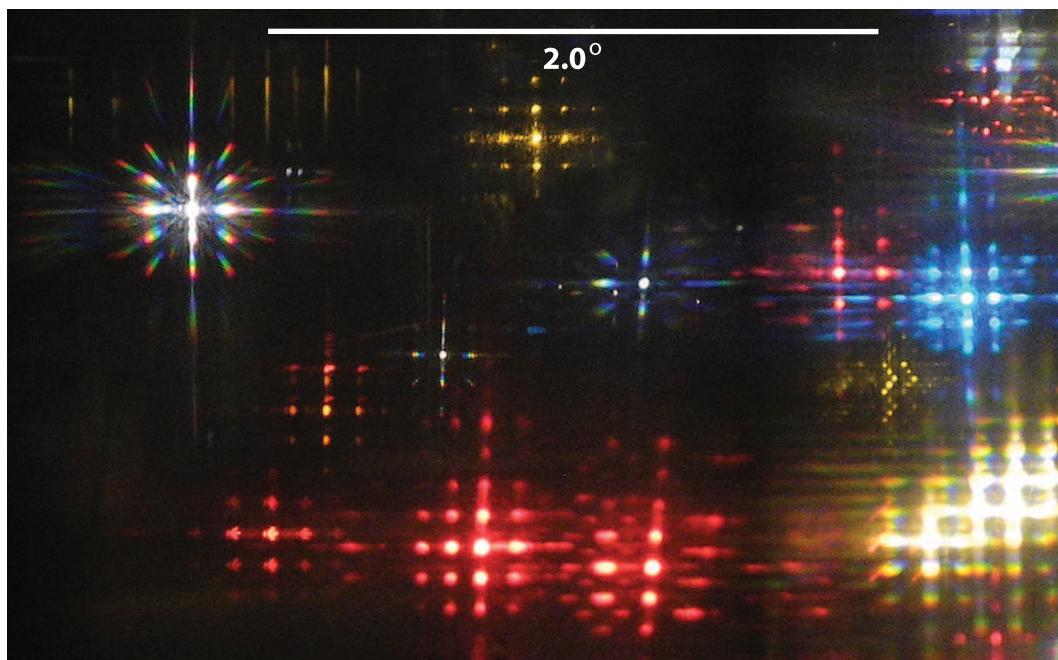
10. To koherente kilder for radiobølger er plassert 5.00 m fra hverandre, og bølgene har en bølgelengde på 3.00 m. Finn punkter på en linje som går gjennom de to kildene hvor vi har konstruktiv og destruktiv interferens (dersom slike punkter finnes).
11. Vi har to lyskilder som sender ut monokromatisk lys. De to kildene ligger i samme synsretning fra oss, den ene $2.04 \mu\text{m}$ direkte bak den andre (men begge synlig). Bølgelengden på lyset kan varieres i intervallet 400 - 700 nm. Ved en gitt bølgelengde synes vi lyset fra de to kildene blir ekstra sterkt. Ved hvilken bølgelengde skjer dette?
12. To spalter med innbyrdes avstand 0.450 mm plasseres 7.5 m fra en skjerm og blyses med koherent lys med bølgelengde 500 nm. Hvor stor avstand er det mellom andre og tredje mørke linje i interferensstripene på skjermen?

13. Et antirefleksbelegg på en linse har brytningsindeksen $n=1.42$ (og glassets er 1.52). Hva er minste tykkelse belegget kan ha for at rødt lys med bølgelengde 650 nm skal ha minimal refleksjon?
14. I et Young dobbeltspaltforsøk plasseres et stykke glass med brytningsindeks n og tykkelse L foran én av spaltene. Beskriv kvalitativt hva som skjer med interferensmønsteret, og utled dernest et matematisk uttrykk som viser intensiteten til interferensmønsteret ut fra vinkel (definert som vanlig).
15. Vi bruker en 10 cm diameter bikonveks linse med brennvidde 50 cm for å fokusere lyset fra sola slik at linsen fungerer som "brennglass". Lyset samler seg ikke i ett punkt, men i en skive med diameter d . Det er to bidrag til størrelsen på skiven, nemlig at sola bli avbildet av linsen og at linsen fører til diffraksjon. Bestem de to bidragene for å se hvilket som er viktigst i dette tilfellet.
16. Et digitalt speilreflekskamera har en CMOS-brikke som er 15.8×23.6 mm stor og har 2592×3872 pixler. Et 35 mm brennvidde objektiv brukes med blender 3.3 til 22. Hvor stor er største og minste Airy-skiven fra objektivet? Angi svaret både i absolutt mål og relativt til pixelstrørrelsen.
17. Vi betrakter diffraksjonsbildet fra et menneskehår holdt i strålen til en grønn laserpenn med bølgelengde 532 nm. Det er 16.2 cm mellom to minimumspunkter med 11 lyse områder mellom når laserpennen (håret) er 185 cm fra skjermen hvor målingene ble foretatt. Hvor stor diameter har håret? Er verdien du kommer fram til rimelig ut fra tilgjengelig info om diametre til menneskehår?
18. Et optisk gitter har sitt tredjeordens lyse bånd ved vinkelen 78.4 grader for lys med bølgelengde 681 nm. Bestem hvor mange linjer gitteret har per centimeter. Bestem også vinklene for første og andre ordens bånd. Finnes et fjerde ordens bånd?
19. Vi lyser med en vanlig He-Ne laser med bølgelengde 632.8 nm vinkelrett inn på en CD. "Rillene" i en CD ligger $1.60 \mu\text{m}$ fra hverandre. For hvilke vinkler kommer refleksjonene fra CDen?
20. Hubbel Space Teleskopet har en apertur (åpning) på 2.4 m og brukes for synlig lys (400 - 700 nm). Arecibo radioteleskopet på Puerto Rico er 305 m i diameter (bygget i en dal) og brukes for radiobølger med bølgelengde 75 cm.
 - a) Hva er minste kraterstørrelse på Månen som kan skilles fra et nabokrater med de to teleskopene? (Avstanden til Månen er om lag ti omkretser rundt Jorden, nærmere bestemt 3.84×10^8 m.)
 - b) Anta at vi ønsker å gjøre Hubbel om til en spionsatellitt som går i en ny bane rundt Jordas. Dersom vi skulle kunne lese av nummerskilt til biler med teleskopet, hvilken høyde måtte den nye banen til Hubbel da være?
21. Betrakt Månen med bare øynene. Forsøk å merke deg en minste struktur du kan skjelne. Finn et bilde av Månen og gjenfinn strukturen der. Bestem avstanden over strukturen og sammenlign denne med det du skulle forvente ut fra Rayleighs opplosningskriterium.
22. Lag et dataprogram hvor du kan beregne diffraksjon fra en enkeltspalt også i en avstand som er liten i forhold til spaltbredden (såkalt Fresnel-diffraksjon). Vis at intensitetsfordelingen like etter spalten følger en buet linje slik vi ser nær "midjen" i lysbunten i nedre høyre del av figur 13.21.

23. Ta et bilde av et fjernt lyspunkt med mobiltelefonen din. Analysér bildet for å se om du kan påvise Airy-skiven. Det krever at du kan blåse opp bildet du tok slik at du får sett enkeltpixler i bildet. Forsøk å beregne hvor stor Airyskiven skulle forventes å bli.
24. Ta utgangspunkt i figur 13.23. Anta at avstanden mellom gravstøttene sides lengst er a og at de ligger en avstand b bakenfor hverandre. Bestem vinkelen mellom hver rad av støtter som ligger bak hverandre slik vi ser i fotografiet. Bestem også avstanden mellom nærliggende støtter langs de linjene vi ser (avstanden som vil svare til spalteavstanden i et optisk gitter).
25. Fra et hotellvindu ble det observert interferenslignende mønster når små lyspunkter ute ble betraktet gjennom lette gardiner (som vi delvis kan se tvers gjennom, se figur 13.26). Eksempler på lysfenomenet vi observerte natterstider gjennom gardinene er vist i figur 13.27. Bildet endres ikke selv om vi er nær eller lenger fra gardinet.
- Angi hvilke detaljer i bildet av lysfenomenene som tyder på at det faktisk er diffraksjon/interferens som er ansvarlig for det vi ser.
 - Gjennomfør beregninger som kan støtte opp om en slik konklusjon. (Det er nok anslagsvis 20 % usikkerhet i de målene som er angitt i figurene.)



Figur 13.26: Bildet fra et lett gardin som det var mulig å se gjennom. Detaljer viser hvordan fibrene i gardinet lå i forhold til hverandre. Staven i midtre del er opprinnelig 2.0 mm lang.



Figur 13.27: Bilde av fjerne lyspunkter observert gjennom gardinet i forrige figur. Staven angir en vinkel på 2.0 grader.

Kapittel 14

Koherens, dipolstråling og laser

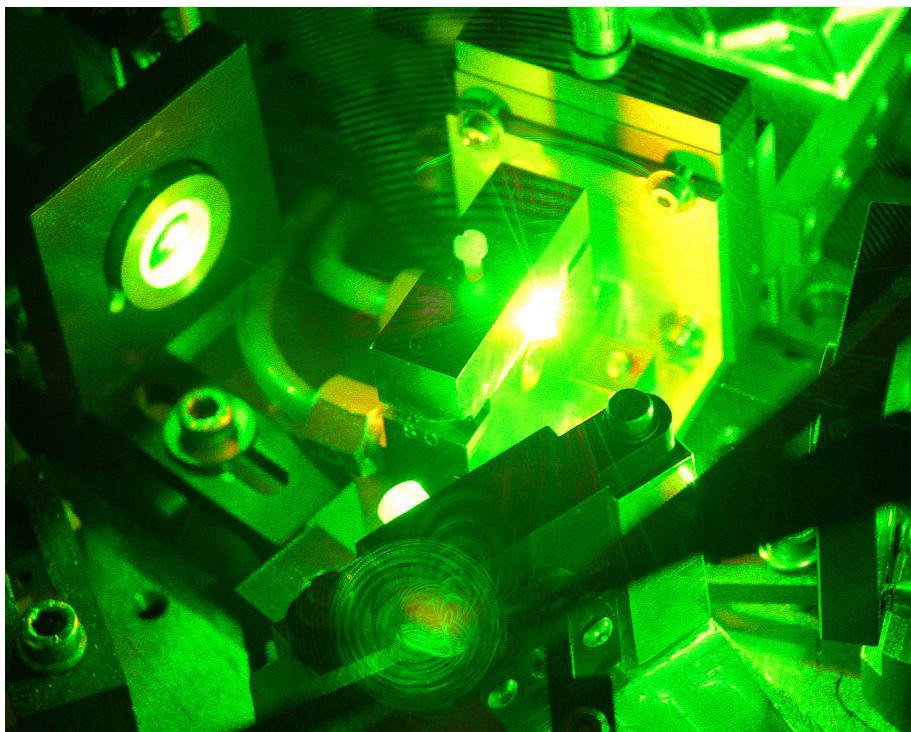


Foto fra innsiden av en laser bygget ved Niels Bohr instituttet ved København universitet. Lysintensiteten er stor, og strålen kommer så nær en "klassisk elektromagnetisk bølge" som det lar seg gjøre.

Kapitlet tar for seg tre ulike temaer, men nesten utelukkende på tegner og forteller-nivå. Likevel er kapitlet slett ikke av de enkleste. Koherens er et litt uvant begrep for mange, og krever litt innsats for at forståelsen skal sitte. Autokorrelasjonsfunksjonen står her sentralt. "Stråling" fra en elektrisk ladning i bevegelse er heller ikke triviell. Det kan lønne seg å tenke tilbake på kapittel 8 om hvilke forandringer i elektrisk og magnetisk felt i tid og rom som må til for at en bølge skal transportere med seg energi. Sist, men ikke minst, sier vi litt om lasere. Dette er et kjempeområde innen fysikk, og vi berører det bare så vidt. Populasjonsinvertering og stimulert emisjon er her sentrale begreper!

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 09052013.

14.1 Koherens, en kvalitativt tilnærming

“Koherens” er et ord, uttrykk eller begrep som få nordmenn har et forhold til. Ordet er antakelig brukt mer i det engelske språk, og i en dictionary kan vi finne forklaringer på “coherence” så som: “The quality of being logically integrated, consistent, and intelligible”, eller: “The mutual relationship between sets of electromagnetic or sound waves in which their amplitudes are exactly equivalent and rise and fall together.”

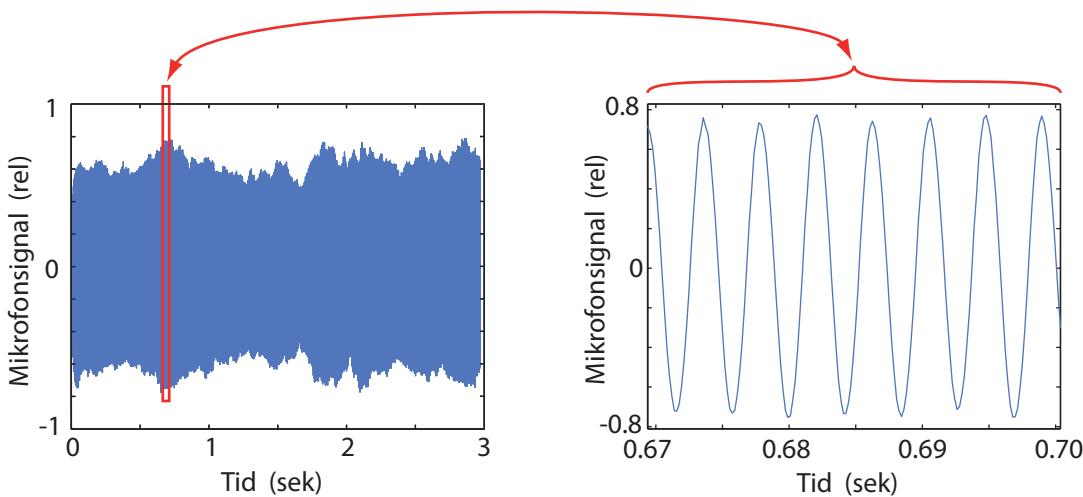
Når vi har med koherens å gjøre, er det mulig å peke på en eller annen egenskap, f.eks. amplitude og/eller fase for bølgers vedkommende, som kan gjenfinnes i alle delsystemene som er koherente med hverandre.

Koherens er imidlertid nøyne knyttet til statistikk, og statistikk kommer inn i bildet når vi har med variabilitet å gjøre.

“Matematikk er fysikkens språk”, sier vi ofte. Matematikkens funksjoner er imidlertid å anse som Platonske idealiseringer, uten variabilitet, mens virkeligheten, den fysiske verden, ofte er mer kompleks enn et enkelt matematisk bilde tilsier. Et godt eksempel på dette finner vi nettopp når vi betrakter svingninger og bølger. Vi har hittil beskrevet bølger på en idealisert måte, nemlig som:

$$f(x, t) = A \cos(kx - \omega t + \phi) \quad (14.1)$$

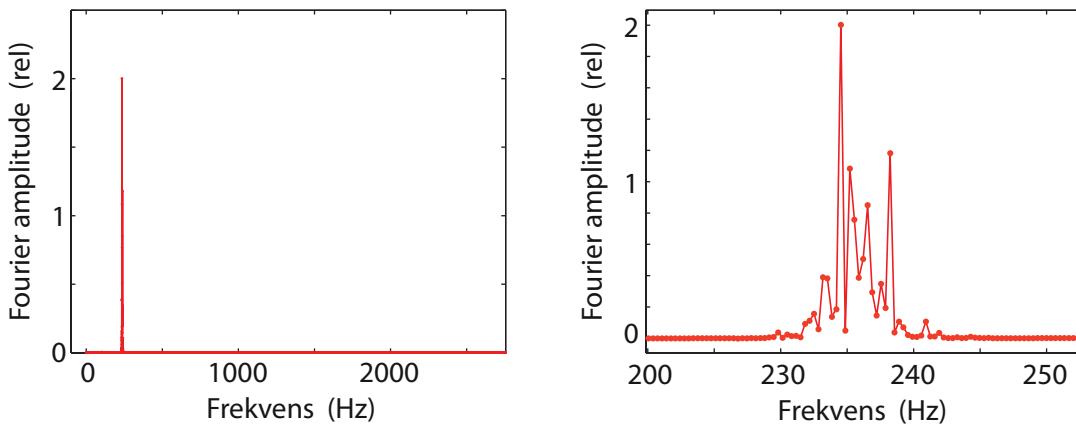
Her er det underforstått at amplituden A er tilnærmet konstant, likeså (vinkel)frekvensen ω . Vi kan finne bølger der denne beskrivelsen er en god tilnærming, men andre ganger er forskjellen mellom fysisk realitet og matematisk idealisme ganske iøynefallende. Ta for eksempel en menneskelig stemme som synger en “iiiiiiiii”, og lyden registreres gjennom en mikrofon og digitaliseres. Et eksempel på hvordan signalet kan se ut er vist i figur 14.1.



Figur 14.1: Tidsbildet av mikrofonopptak av en mannsstemme som synger en iiii. Til venstre vises hele tidsstrekken som ble samplet, til høyre vises et utdrag på 30 ms.

Samplingsfrekvensen var 5512 Hz og 16384 punkter ble samplet, svarende til totalt nesten 3 sek.

Når vi følger svingningen som mikrofonen oppfanger på sin plass, ser vi at amplituden ikke er konstant, men varierer en god del. Når vi tar et tidsutsnitt over bare noen få perioder, ser vi at innenfor et slikt lite intervall kan vi ha en tilnærmet konstant amplitude.



Figur 14.2: Frekvensanalyse av tidssignalet vist i forrige figur. Til venstre er hele frekvensspekteret (under halve samplingsfrekvensen) vist, mens til høyre er det vist detaljer rundt hovedlinjen i frekvensspekteret.

Vi ser forøvrig at tidsforløpet også innenfor en kort periode ikke er en 100 % ren sinus, men har et lite innslag av høyere harmoniske.

Når vi betrakter signalet over en lengre periode, innser vi at den matematiske beskrivelsen er nokså forskjellig fra den fysiske. Dette skyldes at såvel amplituder A_i og (vinkel)frekvenser ω_i i praksis ikke er konstanter, men varierer noe med tiden. Denne variasjonen gir opphav til en rekke fenomener, blant annet “koherens”.

Vi får en indikasjon på variasjonen som forekommer i amplitud og vinkelfrekvens ved å studere frekvensspekteret nøyne (se figur 14.2). Vi har tidligere vist at dersom vi har en helt ren sinus eller cosinusfunksjon, og samplingstiden er et eksakt heltallig antall perioder av signalet, vil frekvensspekteret i en FFT bare ha *ett* punkt forskjellig fra null. Det vil si at frekvensspekteret er helt skarpt innen den oppløsningen vi har (begrenset bare av lengden av tidsstrenge vi analyserer).

Vi har også sett at dersom den totale samplingstiden ikke er *eksakt* lik et heltall ganger med periodetiden, vil vi få *litt* bidrag i frekvensbildet som smøres ut over flere punkter i frekvensspekteret. Dette er imidlertid en effekt som bare gir bidrag til noen ganske få punkter ved siden av grunnfrekvensen.

I figur 14.2 ser vi at et 30-talls punkter i frekvensspekteret er klart forskjellig fra null. En slik relativt bred topp, for et *vedvarende* signal, viser tydelig at frekvensen varierer underveis mens signalet formes.

Hvordan vil denne lille variasjonen i frekvens påvirke fysiske fenomener i praksis? At en sangstemme varierer litt med tiden, er på mange måter positivt, fordi det skaper mer “liv” i lyden enn om vi hadde hatt et “rent” lydbilde som angitt i ligning (14.1). Summerer vi imidlertid to reelle bølger, vil det fremkomme interessante detaljer knyttet til koherens.

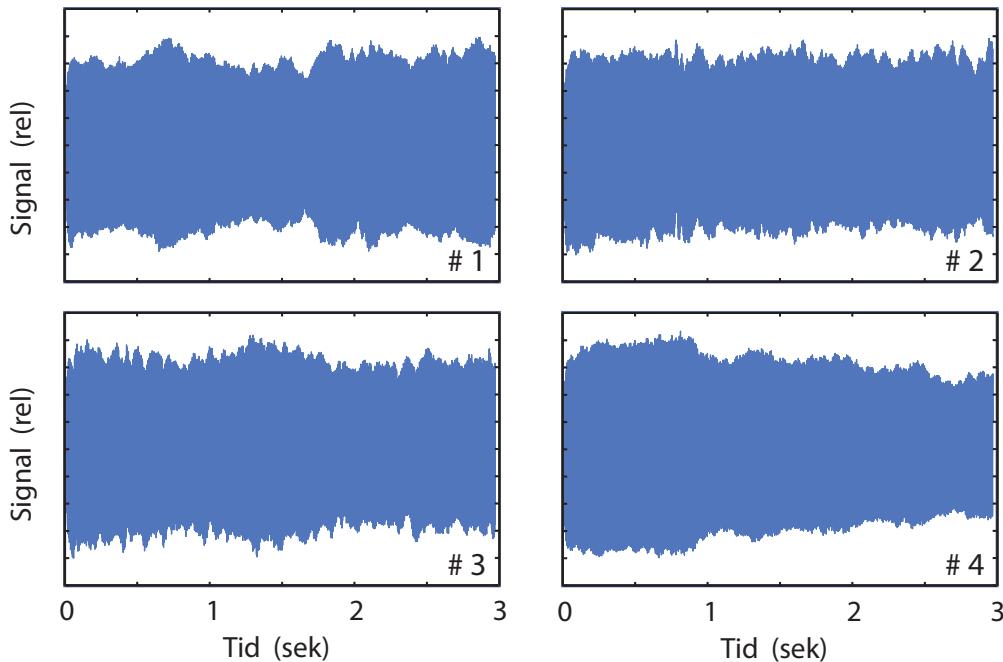
Siden alle fenomener knyttet til interferens og diffraksjon er basert på addisjon av flere signaler, betyr det at *all interferens og diffraksjon påvirkes i høy grad av koherens*. I forrige kapittel, da vi beskrev interferens fra en dobbeltspalt, satte vi opp følgende uttrykk for elektrisk felt fra hver av de to spaltene:

$$E_1(\theta_1) = E_{1,0}(r_1, \theta_1) \cos(kr_1 - \omega t - \phi)$$

$$E_2(\theta_2) = E_{2,0}(r_2, \theta_2) \cos(kr_2 - \omega t - \phi)$$

Vi sa da at ϕ er en vilkårlig fasevinkel, og vi antok at det var samme ϕ på begge spaltene

siden vi hadde en plan bølgefront inn fra baksiden inn mot spaltene. Det er imidlertid ingen selvfølge at det er samme fase ved begge spaltene. Koherens handler nettopp om dette. Det er da på tide å undersøke litt nærmere hva vi mener med “koherens” og i hvilke andre sammenhenger fenomenet dukker opp.



Figur 14.3: Tidssignalet for fire ekvivalente opptak der det synges “iiiiiiiii” med samme frekvens (midtfrekvensen $< 1 \text{ Hz}$ forskjell).

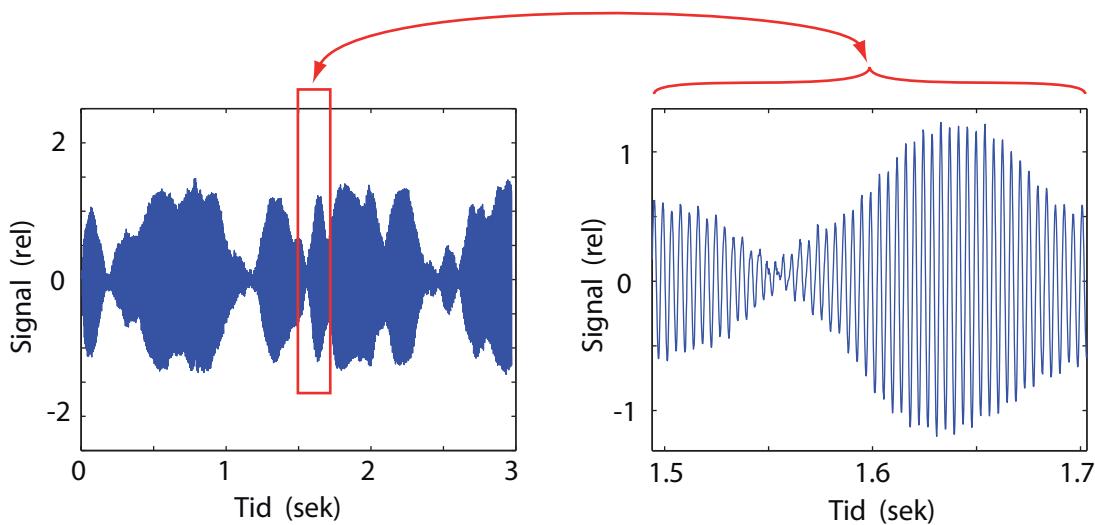
14.2 Sum av to reelle bølger

Lys er elektromagnetiske bølger med så høy frekvens at vi ikke er i stand til å følge endringene i elektrisk felt i detalj (bortsett fra ved helt spesielle, indirekte metoder). Det er derfor lettere å ta utgangspunkt i et annet bølgefenomen med lavere frekvens for å vise hva koherens går ut på. Vi velger lydbølger.

Vi har tatt opp signalet fra en mikrofon for fire sangere som etter tur sang “iiiiiiiiii” med samme tonehøyde (i virkeligheten var det en og samme sanger som sang inn samme tone flere ganger etter hverandre). Signalene var omrent like sterke, men for å forenkle den videre diskusjonen, har vi normalisert alle signalene til samme midlere intensitet i løpet av samplingstiden. Resultatet er angitt i figur 14.3. Senterfrekvensene var identiske innen 1 Hz.

Summerer vi signalene fra to sangere, ser vi noe interessant (se figur 14.4). Det dukker nå opp kraftige fluktasjoner i amplituden i sum-signalet, fra nesten null til om lag to ganger amplitudene til signalene vi startet ut med. Årsaken til variasjonen er nokså enkel å forstå. Lydbølgene fra hver av sangerne er ikke matematisk idealiserte bølger med en helt eksakt lik og konstant frekvens hele tiden. I enkelte tidsperioder vil signalene fra de to sangerne på mikrofonens plass være i fase med hverandre, og amplituden går opp til summen av bidragene fra hver bølge. I andre tidsperioder vil imidlertid signalene fra de to sangerne være i motfase med hverandre, slik at signalet nærmest forsvinner.

Klarer sangerne å holde frekvensen svært nær konstant i tid, og samme frekvens, vil



Figur 14.4: Tidssignalet for summen av to stemmer (to opptak). Legg merke til endringene i styrken med tiden.

det ta lang tid før det blir et skifte fra konstruktiv til destruktiv interferens mellom de to. Dersom sangerne på den annen side ikke er dyktige til å holde samme frekvens, vil det ta kort tid mellom konstruktiv og destruktiv interferens.

Dette er signaler fra en menneskelig stemme, og ikke en maskin. Det medfører at variasjonen i frekvens vil være uregelmessig og signalet vil ikke for noen del av tidsstrenge være eksakt lik den fra en annen bit av tidsstrenge. Det betyr at vi ikke kan finne noe entydig tid mellom full konstruktiv til full destruktiv interferens. Vi må operere med statistiske størrelser.

Den gjennomsnittlige tiden det tar fra fra konstruktiv til destruktiv interferens når vi adderer to bølger med samme (gjennomsnittlige) frekvens, kaller vi *koherenstiden* for signalet. Dette er bare en omtrentlig definisjon, vi kommer tilbake til en mer presis definisjon siden.

Den lengden bølgen har beveget seg i løpet av koherenstiden, kaller vi *koherenslengden*.

Koherens er altså knyttet opp til bølger, og forteller noe om kilden til bølgen, om mekanismene som lå bak da bølgen ble til. Studier av koherenstider/lengder er derfor nyttige når vi ønsker å lære mer om hvordan naturen oppfører seg.

14.3 Sum av flere bølger

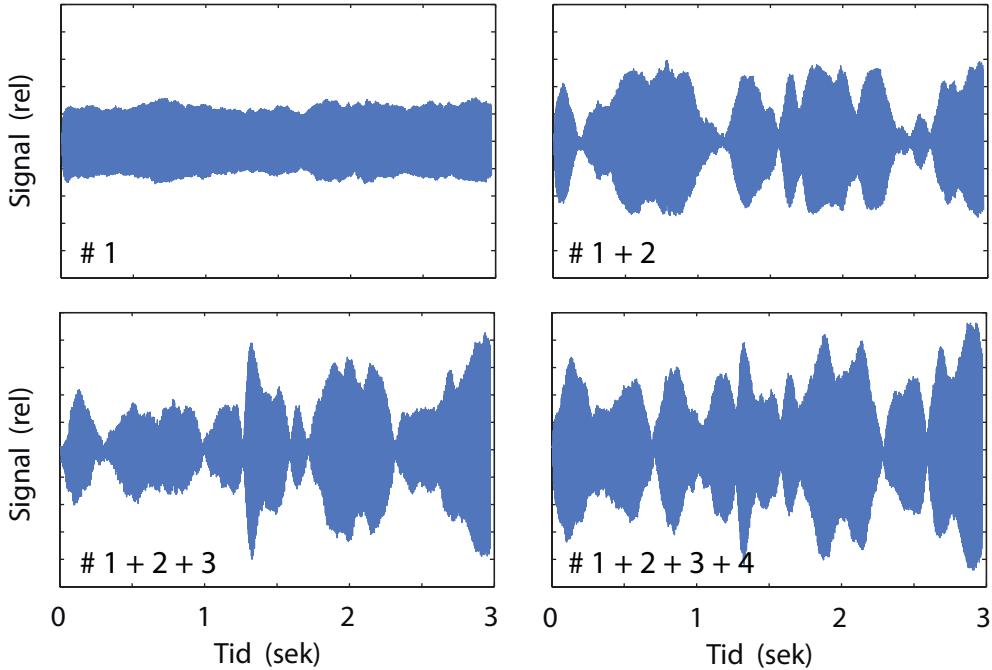
Dersom du har sunget i kor har du sikkert lagt merke til at lydstyrken når flere synger sammen ikke øker så mye som vi kanskje skulle tro. En solist med god stemme høres slett ikke så verst selv i vekselsang med et kor på kanskje 30 personer. Hvorfor blir ikke den ene stemmen totalt overdøvet av de 30?

Vi kan illustrere hva som skjer ved å ta utgangspunkt i mikrofonignalene (y_1, y_2, y_3 og y_4) fra fire sangere (se figur 14.3). Signalene fra hver av sangerne er digitalisert hver for seg, og ulike sum-signaler kan lages i ettertid ved å summere signaler fra enkeltsangere. For eksempel er signalet "1+2" i figur 14.5 laget slik: $y = y_1 + y_2$ i en leddvis addisjon. Flere resultater av summering er vist i figur 14.5.

Vi regner ut midlere intensitet av signalene før og etter summasjon av signaler og forsøker å finne en lovmessighet. Intensitet er her et relativt mål, og er beregnet ut fra ligningen:

$$I = \sum_{n=1}^N y^2(n) \quad (14.2)$$

hvor N er antall samplinger i hvert av signalene y_1 til y_4 , og y er sum av en eller flere enkeltsignaler som antydet ovenfor.



Figur 14.5: Tidssignalet for én, to, tre og fire samtidige sangere som synger samme tone.

Dersom midlere intensitet til hvert av startsignalene normaliseres til “1.0”, blir intensiteten av summene (figur 14.5) som følger:

Signal	Rel.midl.intensitet
1	1.00
1 + 2	2.04
1 + 2 + 3	2.26
1 + 2 + 3 + 4	3.65

Det er ingen perfekt linearitet, for det er betydelige variasjoner i det statistiske materialet vi starter ut med. Likevel ser vi at intensiteten i grove trekk synes å øke omtrent lineært med antall bidrag. Vi har gjentatt forsøket med andre dataopptak og fluktasjone ne er betydelige, men hovedtendensen er klar. Ti sangere vil altså gi en lydstyrke omtrent ti ganger den hver enkelt sanger kan lage alene (dersom de var nokså like). Dette virker naturlig.

14.4 Koherente bølger

Hvordan ville resultatet blitt dersom meget nær perfekte sinusformede bølger ble addert til hverandre? Bølger med en slik egenskap kalles *koherente*.

Det viser seg at *summen av m identiske sinussignal i fase, får en amplitud som er m ganger så stor som hvert av bidragene. Intensiteten vil da bli m² ganger intensiteten til hvert enkelt signal.*

Dersom sangerne våre hadde sunget eksakt likt og hele tiden i fase, ville fire sangere gitt en intensitet lik 16 ganger intensiteten til hver enkelt. Dette er ganske forskjellig fra 4 ganger intensiteten vi i praksis fant med våre reelle signaler. Dette illustrerer en viktig side ved koherente bølger.

Imidlertid, “*There Ain’t No Such Thing As A Free Lunch*”. Vi får ingenting gratis. Fire hypotetiske sangere som synger koherent lyd i fase vil gi fire ganger så mye lydenergi som fire som ikke synger koherent. Hvordan kan energiregnskapet tilfredsstilles?

Energiregnskapet går bra dersom vi tar i betrakning også romlige forhold. Fire hypotetiske sangere som synger koherent vil *ikke* gi fire ganger lydenergien overalt i rommet, bare på de stedene der signalene fra alle fire er i fase med hverandre. Andre steder i rommet, der signalene er i motfase, vil lydenergien være bortimot null. Slik vil det ikke være med de reelle sangerne. Det finnes ingen steder i rommet hvor det er permanent konstruktiv eller destruktiv interferens for disse sangerne. Vi vil høre lyden fra de reelle sangerne overalt.

Integreres lydenergien overalt i rommet, vil den bli omtrent den samme uansett om sangerne synger koherent eller ikke-koherent.

Disse romlige betraktingene har visse paralleller med intensitetsfordelingen for stripene fra flere spalter. Når det blir flere og flere spalter, blir stripene mer og mer intense selv om totalt lysstrøm ut fra spaltene er uforandret. Dette skyldes at stripene blir smalere og smalere.

14.5 Koherenstidsbestemmelse

Et perfekt koherent signal er et hvor vi kan forutsi fasen til en svingning framover i tid med stor sikkerehet så lenge vi ønsker. En matematisk funksjon:

$$f(t) = A \sin(\omega t + \phi)$$

er et eksempel på et perfekt koherent signal. En sangstemme vil alltid ha en grense for hvor lang tid vi kan forutsi fasen i signalet. Det samme gjelder for et hvilket som helst fysisk signal, også et elektrisk felt i lyset fra en laser. For noen signaler kan denne tiden være meget lang, ja mange tusen år (i kosmologien), mens andre ganger kan den være nede i under en femtosekund (10^{-15} sek). Det er denne tiden vi kan kalle “*koherenstiden*”.

For signaler der vi kan følge svingningene i detalj (ser både topper og bunner etter som bølgen passerer), kan vi gi en relativt presis måte å bestemme koherenstiden. Først beregnes en første ordens korrelasjonsfunksjon:

$$F(\tau) = 1/T \int_0^T f(t)f(t + \tau)dt \quad (14.3)$$

Er svingningen beskrevet ved diskrete tidspunkt, ser uttrykket ut som følger:

$$F(\tau_j) = 1/N \sum_{i=1}^N f(i)f(i + j) \quad (14.4)$$

hvor $\tau_j = j/f_s$ der j er et heltall og f_s er samplingsfrekvensen.

Vi multipliserer med andre ord bølgen med en tidsforskjøvet bit av samme bølge, punkt for punkt, og summerer over alle punktene og normaliserer.

Siden korrelasjonsfunksjonen vi har definert innebærer bare ett signal, hvor vi sammenholder en del av signalet med en annen del av det samme signalet, kalles den for "autokorrelasjonsfunksjonen".

Antar først perfekt sinussignal

Dersom f er en perfekt sinus, vil en τ_j som svarer til en forskyvning på $2\pi n$ (n er et heltall) i sinussignalet, medføre at $f(i)f(i + j)$ i ligning (14.4) i praksis er en \sin^2 -funksjon. Autokorrelasjonsfunksjonen F vil da rett og slett bli middelverdien av \sin^2 , hvilket er $1/2$.

På lignende måte kan vi vise at når τ_j svarer til $(2n + 1)\pi$ i det perfekte sinussignalet f , vil beregningen innebære middelverdien av $-\sin^2$, og svaret ville bli $-1/2$. For en forskyvning på en kvart bølgelengde i forhold til de tilfellene vi allerede har nevt, vil F bli middelverdien av et sin cos-uttrykk, som er lik null.

Autokorrelasjonsfunksjonen for en perfekt sinus vil derfor bli en regelmessig funksjon som varierer fra $+1/2$ til $-1/2$ ganger amplituden på signalet (med noe normalisering i tillegg). Autokorrelasjonsfunksjonen får rett og slett en sinusform.

Et reelt fysisk signal

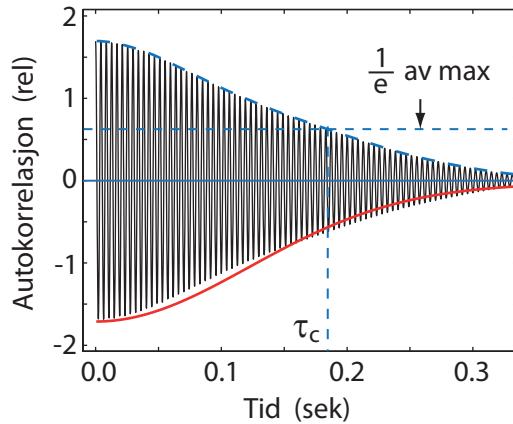
For et reelt signal vil det imidlertid bli annerledes. Når vi i beregningen av autokorrelasjonen forskyver et signal med en avstand *større enn en karakteristisk tid kalt koherenstiden*, vil det bare være en svak eller ingen korrelasjon mellom fasene i $f(t)$ og i $f(t + \tau)$ i ligning (14.3). Da vil vi for deler av integrasjonsintervallet få \sin^2 -sammenhenger mens vi i andre områder får $-\sin^2$ -sammenhenger, og hele integralet blir betydelig mindre enn dersom $\tau = 0$. Det er ikke vanskelig ut fra variasjonen i figur 14.3 å forstå at vi vil finne litt forskjellig F hver gang en beregning gjennomføres i praksis.

Tar vi gjennomsnittet av mange tidsforløp, vil autokorrelasjonsfunksjonen ofte tilnærmet kunne skrives:

$$F(\tau) = F_0 \cos(\bar{\omega}\tau) e^{-(\frac{\tau}{\tau_c})^2} \quad (14.5)$$

hvor $\bar{\omega}$ svarer til middelverdien av vinkelfrekvensene i det opprinnelige signalet. τ_c kalles korrelasjonstiden og svarer (med et visst forbehold) til det vi også kaller "koherenstiden" for vår bølge/swingning. Vi kan med noen forbehold si at koherenstiden defineres ved ligning (14.5). Når $\tau = \tau_c$ vil omhyllingskurven til $F(\tau)$ i figur 14.6 være lik F_0/e .

Vi har beregnet korrelasjonsfunksjonen til vårt signal, og resultatet er vist i figur 14.6.



Figur 14.6: Korrelasjonsfunksjonen for vår opprinnelige tidsfunksjon, eller for en av de sammensatte signalene i figur 14.3. Resultatet varierer en del med hvilket startpunkt som velges, og med lengden på tidsstrenge vi midler over. Ved en såpass kort tidssteng som vi har her, konvergerer aldri omhyllingskurven til korrelasjonsfunksjonen monoton mot null. Likevel kan vi lese av en omtrentlig korrelasjonstid, også kalt koherenstiden. [Korrelasjonstiden er her beregnet ved å lage den stiplete blå "omhyllingskurven" øverst, og bestemme tiden da denne kurven krysser nivået $1/e$ av max. Den røde kurven nederst viser omhyllingskurven av den teoretiiske relasjonen i ligning (14.5).]

Korrelasjonstiden for denne "iiiiii"-sangen er altså ca 0.18 sekund. Det vil si at vi med rimelig god grad av sikkerhet kan angi fasen til signalet bortimot 0.2 sekund framover i tid dersom vi kjenner fasen nå.

Lydhastigheten i luft er om lag 340 m/s. Det betyr at vi kan forutsi korrelasjon i fasen i det aktuelle lydsignalet innenfor en strekning på ca $\Delta L = 340 \cdot 0.18$ meter, det vil si ca 60 m. Denne strekningen kaller vi *koherenslengden* for vår bølge (mer presist: *Den temporære koherenslengden*, også kalt *tidskoherenslengden*).

Det betyr at det vil være mulig å kombinere ulike lydsignaler fra vår sang og finne interferens og diffraksjon for alle mulig gangforskjeller på opp til ca 60 meter. Forsøker vi å vise diffraksjonsfenomener basert på en bølge som har gått over 60 meter lenger enn en annen bølge (begge bølgene startet ut fra samme mannsstemme som brukt i figurene våre), så ville diffraksjon og interferens bli betydelig dårligere enn om forskjellen i gangavstand var vesentlig mindre. Selvfølgelig er det ingen skarp grense ved 60 meter, men tallet forteller at gangforskjeller på opp til 20-40 m ikke er noe problem, mens gangforskjeller over f.eks. 120 m vil gi store problemer dersom vi ville påvise interferens el.l.

[♣ ⇒ En kommentar:

Uttrykkene som er gitt er ikke helt korrekte. Det er en del detaljer som kommer inn i en grundigere

behandling. Hovedideene er likevel de samme som det vi har presentert her. Den heltrukne røde omhyllingskurven i nedre kant av figur 14.6 angir formen til omtrent beste Gaussiske forløp av type ligning (14.5). Vi ser at det ikke er noe god overensstemmelse mellom teori og praksis. En vesentlig grunn til dette er at resultatet varierer betydelig fra signal til signal. En annen grunn er at den gaussiske formen på omhyllingskurven for autokorrelasjonen forutsetter en del om mekanismene for frekvensendringene vs. tid, og disse forutsetningene er ikke tilfredsstilt for den menneskelige stemme. ← ♠]

Det finnes mer omtrentlige måter å bestemme koherenslengder på enn det vi har skissert her. Da vi tidlig i boka arbeidet med fouriertransformasjoner, så vi at det var en sammenheng mellom hvor lenge et tidssignal varte (Δt) og bredden på frekvensspekteret (Δf). Sammenhengen var:

$$\Delta t \Delta f \approx 1$$

(Dette er mer eller mindre identisk med Heisenbergs uskarphetsrelasjon.)

Når vi snakker om koherens, kan vi på sett og vis si at et ikke-koherent signal kan anses som en etterfølgende rekke av tilnærmet koherente bølger som bare varer en tid omtrent lik koherenstiden. Det er derfor ikke så overraskende at det er en tilnærmet, men enkel relasjon mellom bredde i frekvensspekteret (Δf) og koherenstiden (τ_c). Den er:

$$\tau_c \Delta f \approx 1$$

Det finnes også en lignende sammenheng, kalt *Wiener-Kinchine teoremet* som sier at den fouriertransformerte av første ordens autokorrelasjonsfunksjon til en funksjon er lik frekvensspekteret til funksjonen. Vi går ikke inn i detaljer om denne siste relasjonen.

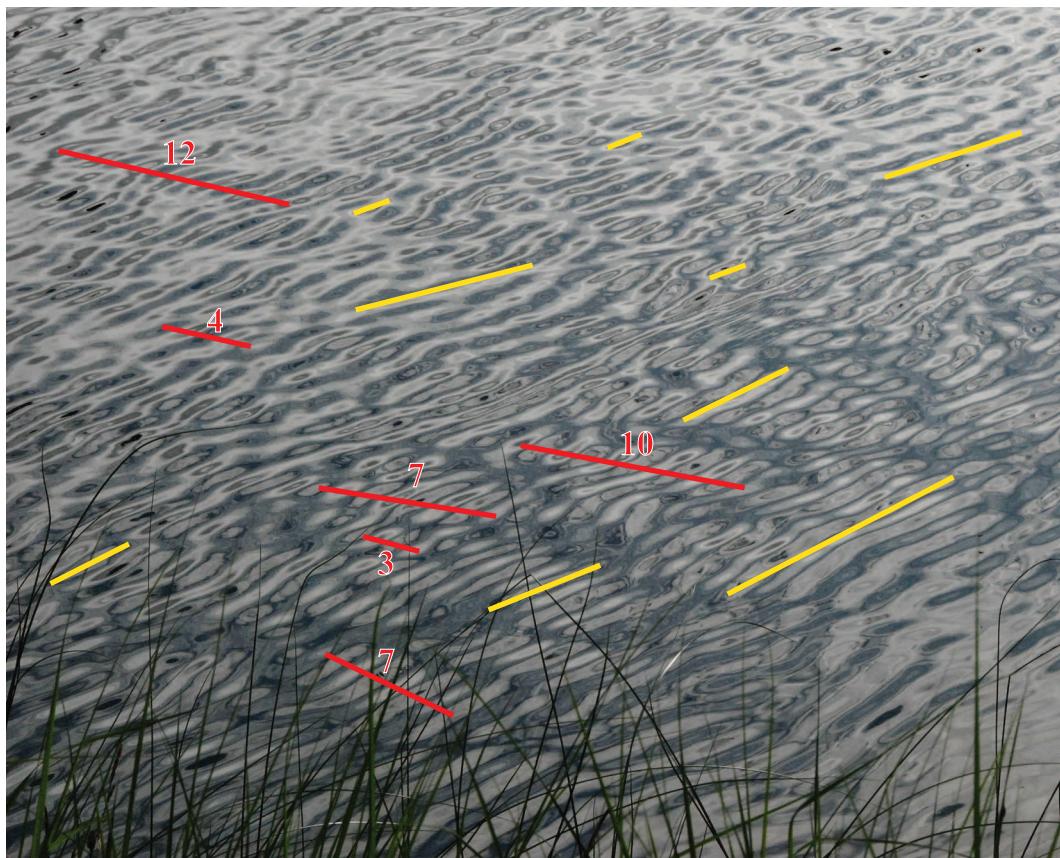
14.6 Anskueliggjøring av koherens

Det er ingen enkel sak å få en god forståelse av koherens. Vi velger derfor å ta med et fotografi fra overflatebølger på vann for å illustrere koherens på en annen måte enn hittil.

Innenfor små flekker på overflaten har vi en temmelig “ren” bølge (se figur 14.7). Innen disse flekkene er det mulig å forutsi relativt greit innbyrdes faseforhold i den retningen bølgen brer seg (røde streker). Innenfor flekkene er fasen og amplituden på bølgen omtrent konstant i en retning normalt på den retningen bølgen beveger seg (gule streker). Flekkene har svært ulik størrelse. I bølgeretningen varierer den mellom et par til tolv bølgelengder. Det betyr at den temporære (tidsmessige) koherenslengden er i størrelsesordne 5-7 bølgelengder, men at det slett ikke er en nøyktig bestemt størrelse. I retningen normalt på bølgevandringsretningen er de gule stripene gjennomsnittlig omtrent like lange som gjennomsnittet for de røde stripene. Det betyr at den spatielle (romlige) koherenslengden er omtrent like lang som den temporære i dette tilfellet. Perspektiviske forhold gjør at det er vanskelig å angi hvor brede flekkene er der bølgen er nokså ensartet i hele sin bredde.

Dersom vi betrakter bølger i tre dimensjoner, vil “flekkene” hvor det var temmelig veldefinerte bølger, bli erstattet med små volumer der det er temmelig veldefinerte (og tilnærmet plane) bølger.

Flekkene eller volumene med nokså veldefinerte bølger vil imidlertid endre seg i tid. Det gjør at kompleksiteten blir ytterligere forværret. Det er ikke vanskelig å forstå at det er litt av en statistikk-utfordring å beskrive denne dynamiske situasjonen som vi svært ofte finner i praksis når bølger brer seg i rommet.



Figur 14.7: Overflatebølger på vann ved ett tidspunkt. I små flekker på overflaten har vi en temmelig “ren” bølge. Se teksten for videre omtale.

En liten detalj i figur 14.7 kan det være verdt å minne om. I ethvert punkt har vannoverflaten i et bestemt øyeblikk en nokså veldefinert høyde. Eller sagt på en annen måte: Vannoverflaten har ikke flere verdier samtidig! Vi har så lett for å tenke oss at det “finnes flere bølger samtidig”, men på ett og samme sted har det lokale lufttrykket bare én verdi i et gitt øyeblikk for lydbølger i luft, og på ett og samme sted har det elektriske feltet bare én verdi og bare én retning i et øyeblikk for summen av samtlige elektromagnetiske bølger på dette stedet.

Dette er et faktum det kan være vel verd å grunne litt over!

14.7 Måling av koherenslengde for lys

Synlig lys har en så høy frekvens at vi ikke klarer å registrere sinussvingningen av elektrisk felt etter som bølgen passerer. Vi kan da ikke bruke matematikken angitt ovenfor direkte.

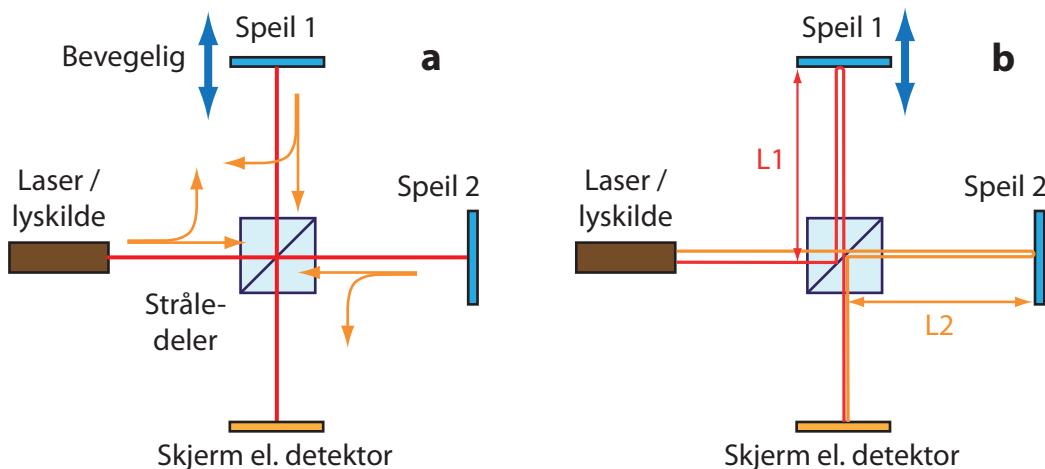
Vi kan imidlertid gjennomføre en *analog* beregning av en størrelse som er nær beslektet med autokorrelasjonsfunksjonen. Det gjøres ved at en lysbunt splittes i to ved en såkalt “beam splitter”. De to delstrålene føres så sammen igjen, men først etter at den ene går en lengre vei enn den andre. Når strålene føres sammen, adderes elektrisk felt fra de to lysbuntene og magnetisk felt fra de to lysbuntene, og vi betrakter intensiteten av summen.

Vi betrakter rett og slett:

$$\begin{aligned}
 G(\tau) &= 1/T \int_0^T (f(t) + f(t + \tau))^2 dt \\
 &= 1/T \int_0^T (f^2(t) + 2f(t)f(t + \tau) + f^2(t + \tau)) dt \\
 &= 1.0 + 2/T \int_0^T f(t)f(t + \tau) dt
 \end{aligned} \tag{14.6}$$

Her er $f(t)$ å betrakte som f.eks. elektrisk felt i strålen etter at den er delt i to og at amplituden er normalisert til 1.0.

Det betyr at vi rett og slett kan endre på gangveien til den ene dellysstrålen sammenlignet med den andre før de kombineres, og at vi da får ut autokorrelasjonsfunksjonen akkurat som i figur 14.6, bortsett fra at hele kurven er forskjøvet slik at minimum ligger på null (intensiteten kan ikke være negativ). Figur 14.8 viser prinsippet i et såkalt Michelson interferometer som ofte brukes i slike målinger. Gangforskjellen for de to delstrålene er $\Delta L = 2L_1 - 2L_2$.



Figur 14.8: I et Michelson interferometer sendes en lysstråle inn til en stråledeler. Halvparten av strålen går til et fast speil og blir reflektert herfra, mens den andre halvparten går til et flyttbart speil. Halvparten av lyset som reflekteres fra speilene sendes til en skjerm eller detektor, hvor elektrisk felt fra de to bidragene adderes. I høyre del av figuren er lysveien for de to delstrålene markert skjematiske.

Lys fra termiske lyskilder, f.eks. glødelamper, kan ha en koherenslengde på bare noen få bølgelengder (det vil si bare noen få mikrometer). Lys med liten koherenslengde kalles "ikke-koherent". Lys fra en god laser kan ha en koherenslengde på opp til flere hundre meter. En laser som koster noen få tusen kroner har typisk en koherenslengde på noen centimetre (dvs i størrelsesorden 100 000 bølgelengder). Lys med lang koherenslengde kalles "koherent". Det er ingen skarp grense mellom ikke-koherent og koherent lys.

[♣ ⇒ Albert Abraham Michelson (1852-1931)]

var en eminent eksperimentalfysiker. Han er kanskje mest kjent i forbindelse med Michelson-Morley-eksperimentet i 1887 som viste at lys ikke brer seg gjennom en ether. Michelson målte lyshastigheten med stor presisjon. Videre utviklet han stjerne-interferometre og klarte derved å måle diameteren på fjerne

stjerner, og å måle avstanden mellom stjernepar (“binary stars” på engelsk). Han fikk i 1907 Nobelpriisen i fysikk, og var den første amerikaner som fikk denne prisen. $\Leftarrow \spadesuit$

Hva kan koherensmålinger fortelle oss?

$\spadesuit \Rightarrow$ [Så langt har vi omtalt en herrestemme som sang en “iiiiiii”, og koherenstiden var om lag 0.18 sek. Det er ikke godt å si hva som er hovedgrunnen til at denne personen ikke klarer å holde faserenhet ut over denne tiden. Vi kan tenke oss at en veludannet sanger kanskje vil ha lenger koherenstid. Imidlertid er det vanskelig å se for seg at koherenstiden kan være lengre enn tiden mellom hver gang sangeren trekker pusten. Hver gang hun trekker inn pust vil nemlig bevegelsen til stemmebåndet bli ganske forstyrret, og faseinformasjon vil forsvinne.

På samme måte vil det være et brudd i fasen hver gang vi klimprer på en gitar. Det blir da markante sprang i fase. Slike markante sprang vil gi en noe annen korrelasjonsfunksjon enn mer gradvise endringer i fase. Det betyr at vi ved å studere korrelasjonsfunksjoner av ulike typer, kan si litt om hvilke mekanismer som ligger bak faseendringene. Med andre ord, å studere statistikk knyttet til bølgene, vil kunne gi oss informasjon om det systemet vi studerer, ofte informasjon som vanskelig kan oppnås på annet vis.

I optikken snakker vi om “fotonstatistikk”, og denne er allerede blitt en viktig del av kvanteoptikk og studier av kvantefysiske systemer. La oss ta et eksempel: Når lys sendes ut fra en gass, vil molekylene kunne bevare faserelasjoner i lyset som sendes ut så lenge molekylet beveger seg fritt. En gang i blant kolliderer imidlertid et molekyl med et annet, og en kollisjon fører til at det blir et sprang i fasen på det lyset som sendes ut. Dette påvirker koherenstiden til lyset, og derved gir det oss muligheter til å studere kollisjoner på en måte som ellers vanskelig kunne oppnås.

Det er vanskelig å forestille seg hvordan koherens skal forstås dersom vi anser lys som nærmest punktformige partikler. Derimot blir koherens-studier spesielt interessante dersom vi tenker oss at vi kan modellere fotoner som “bølgepakker” med endelig utstrekning i rommet. Har korrelasjonslengden i så fall noe sammenheng med koherenslengden? Flere modeller kan være av interesse i denne sammenheng, og kanskje noen går an å teste ut?

Tenker vi oss en lysbunt satt sammen av mange endelig store fotoner, skjønner vi at det holder ikke bare å vurdere hvor langt fotonet strekker seg ut i stråleretningen (longitudinal retning). Det er også interessant å vurdere hvor langt på tvers av stråleretningen hvert enkelt foton strekker seg ut.

Går vi tilbake til lydsignalet vi først og fremst har konsentrert oss om i dette kompendiet, har bølgen en svært stor utstrekning i “tvers-retningen”. For talefrekvenser blant menn er bølgelengden gjerne i størrelsesorden 1 m (1 m akkurat ved 340 Hz, eller ca 1.4 m for grunnfrekvensen i sangen som lå bak figur 14.1 og 14.2). Strupehodet og munnen er små i utstrekning sammenliknet med disse bølgelengdene, og lyden vil gå ut fra munnen med nær samme intensitet i alle retninger, i alle fall i alle retninger framfor ansiktet.

Kan lys også ha en lignende oppførsel i visse sammenhenger? Eller er alltid lys mye mer “konsentrert” i planet vinkelrett på fotonets bevegelsesretning?

Når et molekyl sender ut lys, er utstrekningen på kilden (dvs selve molekylet eller en del av dette) i størrelsesorden 10-20 Å stort. Bølgelengden for det synlige området er i størrelsesorden 5000 Å. Kilden er enda mindre relativt til bølgelengden enn en munnen til en mansstemme og lyden han lager. Igjen skulle det forventes at fotonet hadde en betydelig utstrekning på tvers av bevegelsesretningen, mens vi i andre sammenhenger tenker oss fotonet som nærmest punktformig.

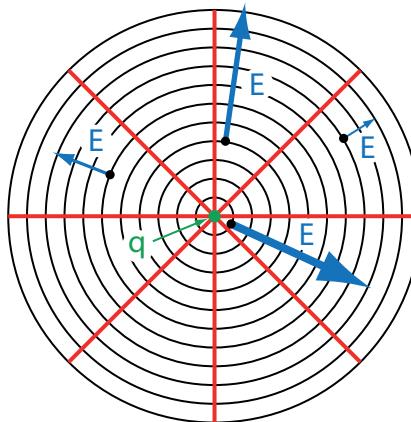
Det er ingen konsensus blant nåtidens fysikere mhp hvordan vi skal oppfatte lys. Det gjør det spesielt spennende å jobbe i dette forskningsfeltet, for en god idé eller en tilfeldighet kan om vi er heldig føre til meget interessante resultater.] $\Leftarrow \spadesuit$

14.8 Stråling fra en elektrisk ladning

Koherens er knyttet opp mot mekanismene for hvordan til bølgene genereres. Vi har tidligere såvidt diskutert mekanismer for å få bølger på en streng, lydbølger og overflatebølger på vann. For elektromagnetiske bølger har vi hittil bare konstatert at f.eks. plane bølger er løsninger av Maxwells ligninger i fjernfeltsområdet i vakuum (i alle fall uten frie ladninger). Men hva er vanligvis selve *kilden* eller mekanismen bak generering av elektromagnetiske bølger? Vi skal bare såvidt berøre dette enorme feltet innen fysikk. Først skal vi se hvordan ladning i bevegelse kan gi bølger, og dernest skal vi se litt på noen av hovedtrekkene bak laseren.

Vi kan beregne hvordan vi kan lage elektromagnetiske radiobølger ved å sende en vekselstrøm inn mot en antenn. I dette tilfellet har vi med frie ladninger og frie strømmer å gjøre, og Maxwells ligninger gir oss en ikke-homogen annen ordens partiell differensialligning for det elektriskefeltet \vec{E} og en tilsvarende ligning for magnetfeltet \vec{H} . Beregninger av denne typen kan gjøres med Finite Element Metoder såvidt nevnt tidligere. Vi går ikke inn i detaljer her.

Vi velger her en “tegner og forteller”-fremstilling i stedet for en rigid matematisk behandling, men håper det vil være tilstrekkelig for å få fram et hovedtrekk.



Figur 14.9: En ladning i ro har sfæriske ekvipotensialflater rundt seg (sorte sirkler), og radielle elektriske feltlinjer (røde streker). Det elektriske feltet er tegnet inn i fire punkter i planet. Feltet er sterkt nær ladningen og avtar med økende avstand..

Figur 14.9 viser skjematiske at en ladning q i ro har elektriske feltlinjer som peker radielt utover (dersom q er positiv). Ekvipotensialflatene er kuleskall med sentrum i ladningen.

Beveger ladningen seg med konstant hastighet, vil ekvipotensialflatene ifølge relativitetsteorien bli “sammenklemt”, det vil si svakt diskosformet med minste akse i retningen bevegelsen skjer. I systemet hvor ladningen er i ro, er det bare et elektrisk felt. I et system hvor ladningen er i bevegelse, vil det være både et elektrisk og magnetisk felt. Når vi derimot snakker om å generere bølger, må vi trekke inn såkalte *retarderte potensialer*. En ordentlig behandling av dette temaet følger i senere kurs. Vi ser bare på noen overfladiske trekk her.

Vi bygger på en antakelse at *endringer* i elektrisk og magnetisk felt forflytter seg i rommet med lyshastigheten. Vi ser ikke en supernova når den skjer, men først etter at lyset har kommet seg den lange veien fra novaen til oss. Det vi ser i dag er supernovaen slik den var for akkurat tiden d/c siden, der d er avstanden mellom oss og supernovaen og c er lyshastigheten.

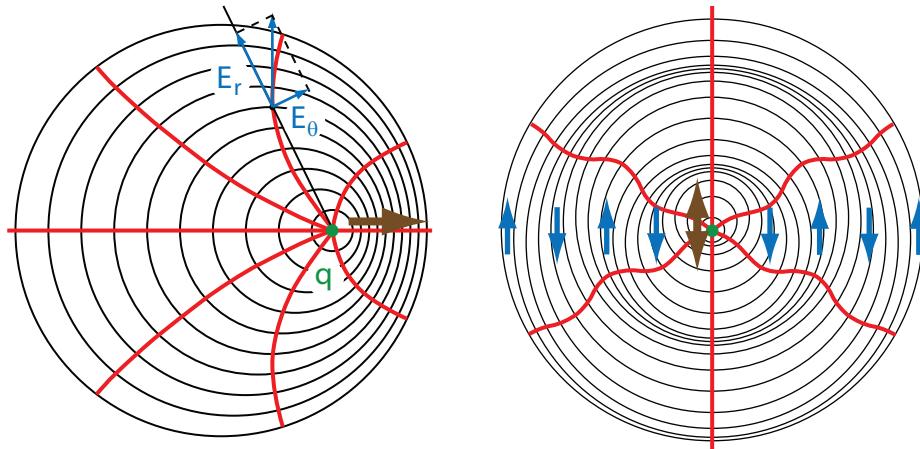
Slik er det også når vi flytter en ladning i rommet. Feltet et sted i rommet har da en fordeling som svarer til ladningens plassering på tidspunktet

$$t' = t - d/c$$

der t er tiden "nå" og d er avstanden fra ladningen til punktet vi betrakter feltet i, på tidspunktet t' .

Dersom vi da tegner opp ekvipotensialflater fra en ladning med konstant hastighet, vil flatene ha relative posisjoner som antydet i venstre del av figur 14.10. Effekten er sterkt overdrevet idet den aktuelle ladningen faktisk ville ha en hastighet over halve lyshastigheten slik figuren nå er tegnet.

Det elektriske feltet er normalt gitt som gradienten til det elektriske potensialet, og den hovedregelen beholder vi også når vi bruker retarderte potensialer. Da får vi elektrisk feltlinjer som er krummet, som vist i figuren.



Figur 14.10: Venstre del: En ladning i en konstant rettlinjet bevegelse har sfæriske ekvipotensialflater rundt seg, men bare i den forstand at ekvipotensialflatene i en viss avstand har sentrum i ladningens posisjon på tidspunktet d/c tidligere, der d er avstanden fra ekvipotensialflatene og ladningen på det tidligere tidspunktet. Forskyvningen av ekvipotensialflatene i forhold til hverandre fører til at de elektriske feltlinjene ikke lenger blir rent radielt rettet, men får en tangentiel komponent i tillegg. Feltlinjene i samme retning (og motsatt retning) som ladningen beveger seg, får ingen tangentiel komponent. Høyre del: En oscillerende ladning vil gi ekvipotensialflater som ligger litt forskjøvet i forhold til hverandre, som antydet. Det medfører at elektriske feltlinjer i tangentiel retning endrer seg som en bølge.

Anta at vi står i ro og en ladd partikkel kommer farende forbi oss i konstant hastighet. Da vil vi oppleve et elektrisk (og magnetisk) felt på vårt sted som har en tidsutvikling der det elektriskefeltet først har samme retning som ladningen beveger seg i (for positiv ladning), via et mye sterkere felt vinkelrett på denne retningen idet ladningen passerer, og ender som et svakt felt i motsatt retning av bevegelsen. Dette er en "puls" av elektrisk (og magnetisk) felt, og ikke en bølge i vanlig forstand.

En observatør som eventuelt fulgte med i ladningens konstante hastighet, vil beskrive det elektriske feltet som statisk, på en liknende måte som i vår figur 14.9. En slik situasjon kvalifiserer ikke for en utstråling av energi. I vårt eget referansesystem, hvor ladningen er i bevegelse, vil elektrisk felt bygges opp i ett sted av rommet, mens det skjer en helt ekvivalent nedbygging av felt et annet sted i rommet. Riktignok vil området som har

høyest feltenergi forflytte seg, på samme måte som ladningen, men denne forflytningen er av lokal karakter, og representerer ikke energi som fjerner seg fra området rundt ladningen.

For å få en bølge som brer seg ut over nærområdet til ladningen, må vi etterstrebe en situasjon lignende den vi hadde for en elektromagnetisk bølge i kapittel 8. Elektrisk (og magnetisk) felt må oscillere og ha en retning vinkelrett på bølgens bevegelsesretning. For å få til dette ved vår ladning i bevegelse, må vi ha en ladning som utsettes for en *akselerasjon*. Ladningen kan f.eks. oscillere fram og tilbake i rommet, gjerne i en harmonisk bevegelse. Det elektriskefeltet et stykke fra vil da oscillere som skissert i høyre del av figur 14.10. Denne tidsendringen i elektrisk felt vil ha både en radiell og en tangentiell komponent relativt til radiusvektor fra ladningen til punktet vi betrakter.

Komponenten i radiell retning (når vi er et stykke unna ladningen sammenlignet med utslaget i ladningens oscillasjon) vil (nesten) ikke endre seg med tiden. Denne komponenten vil derfor (nesten) ikke gi opphav til noe bølge som brer seg utover.

Komponenten vinkelrett på den radielle retningen vil derimot svinge (nesten) som en sinus med tiden. Denne komponenten vil kunne gi opphav til en elektromagnetisk bølge som brer seg ut i rommet.

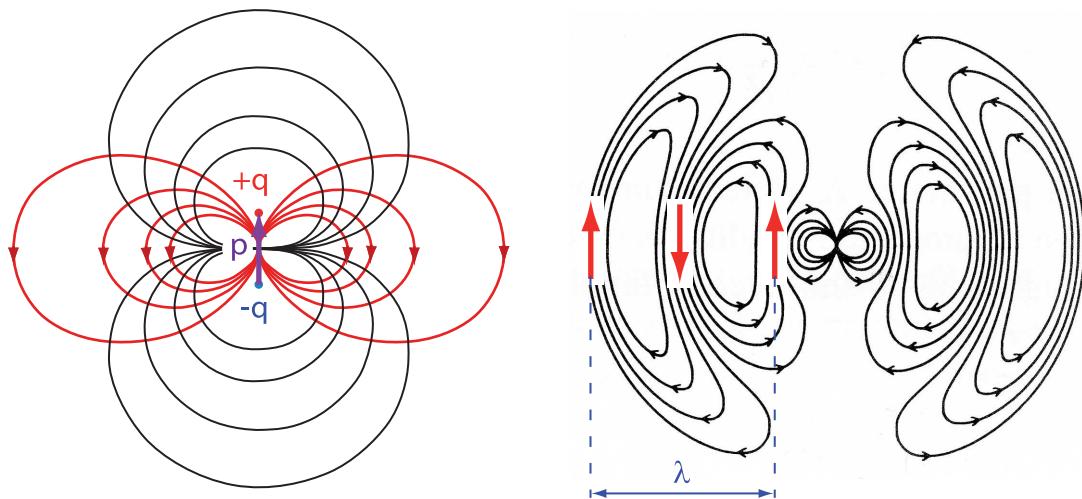
Krumningen på det elektriskefeltet øker med hastigheten til ladningen mens den oscillatorer. Den tidsderiverte av denne igjen bestemmer hvor stor $\frac{\partial E}{\partial t}$ blir. Disse to faktorene tilsammen medfører at utstrålt energi blir proporsjonal med frekvensen til svingningen i annen potens. Vi sier derfor ofte at utstrålingen er proporsjonal med *akselerasjonen* til ladningen.

♠ ⇒ [Noen side-bemerkninger:

Det er nærliggende å tenke seg at elektrisk felt fra en ladning "stråler" utover hele tiden. Det ligger liksom i kortene siden vi opererer med retarderte potensialer der vi tenker oss at feltet på et sted skyldes ladningen der den var for en stund siden. Et konstant utstrålende elektrisk felt ville imidlertid lett kunne stride mot energibevaring m.m. Heldigvis er det ikke nødvendig å tenke i de baner. Det er *endringer* i elektrisk og magnetisk felt som forplanter seg med lyshastigheten. Før en bestemt endring har bredd seg ut og nådd fram til et gitt sted, er det feltfordelingen som har opphav fra forhold *før* endringen som gjelder. Det elektriskefeltet fra en ladning i ro er i likevekt med seg selv. Det er en løsning av Maxwells ligninger, og det skjer ingen endringer i felt og ingen transport av energi. Først så snart bevegelser og spesielt akselerasjon forekommer, blir ting annerledes.] ← ♠

14.8.1 Dipolstråling

En alternativ måte å generere elektromagnetiske bølger på er å bruke en elektrisk (eller magnetisk) dipol som varierer i tid. Dette er en meget effektiv måte å lage bølger på. Det kan vi skjønne ved å betrakte elektrisk feltfordeling fra en permanent elektrisk dipol (se venstre del av figur 14.11). Det elektriskefeltet er rettet vinkelrett på den radielle retningen i planet normalt på dipolens retning.



Figur 14.11: Venstre del: En statisk elektrisk dipol består av to identiske ladninger, men med motsatt fortegn, plassert litt fra hverandre. Det er tegnet inn såvel ekvipotensialflater (sorte, med buk oppover og nedover) som elektriske feltlinjer (røde, med buk utover mot sidene). Dipolens fysiske utstrekning er kraftig overdrevet i forhold til feltlinjemønsteret. Høyre del: En oscillerende elektrisk dipol vil skape et elektrisk felt i rommet rundt som indikert. Feltmønsteret beveger seg utover med lyshastigheten. (Høyre del er en videreføring av en figur i P.Lorrain, D.R.Corson, F.Lorrain: *Electromagnetic fields and waves*, 3rd ed.)

Dersom vi endrer polariteten til dipolen på en harmonisk måte, får vi et elektrisk felt i dette ekvatorialplanet som vil variere akkurat slik vi ønsker det for å generere en elektromagnetisk bølge som kan bre seg ut i rommet (elektrisk felt vinkelrett på bevegelsesretningen). I retningen dipolen peker (og motsatt retning) er det elektriskefeltet et stykke fra dipolen omrent lik null, og det har ingen komponent på tvers av den radielle retningen. I disse to retningene blir det praktisk talt ikke noe utsendt noen bølger.

I høyre del av figur 14.11 er det vist et diagram over elektrisk feltfordeling nær en dipolantenne ved et gitt tidspunkt. Det elektriskefeltet er kraftigst der feltlinjene ligger tettest. Hele mønsteret beveger seg utover med lyshastigheten, og nye looper danner seg nær antennen, to ganger for hver periode (retning på feltet skifter retning i de to systemene av looper som dannes hver periode i dipolvariasjonen). En animasjon av forløpet (og mye annen informasjon) er tilgjengelig på Wikipedia under oppslagsordene “dipole radiation”.

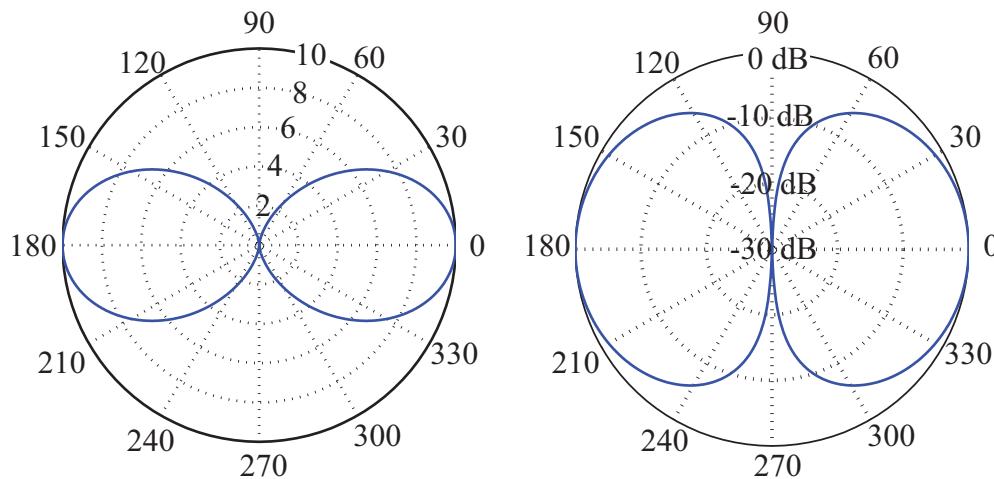
♠ ⇒ [En kommentar:

Høyre del av figurene 14.10 og 14.11 har en viss relasjon til hverandre, men er likevel forskjellige. Forsök å påpeke forskjeller og likheter.] ← ♠

Det er vanlig å tegne *retningsdiagram* for antenner. Et retningsdiagram angir relativ tidsmidlet intensitet på de utsendte bølgene for ulike retninger i rommet, i et såkalt “polardiagram”. Figur 14.12 viser retningsdiagrammet i et vertikalplan som går gjennom en enkel vertikal dipolantenne. Et retningsdiagram kan gis med en lineær skalering i radiell retning (venstre del av figuren), men mest vanlig brukes det en logaritmisk skala i radiell retning (høyre del av figuren).

♠ ⇒ [En kommentar:

Anta at vi lager et polardiagram med logaritmisk skalering i radiell retning. I så fall må vi velge hvilken intensitet som skal svare til sentrum i diagrammet og skalere verdiene slik at denne intensiteten skaleres



Figur 14.12: Retningsdiagram for en enkel vertikal dipolantenne. Diagrammet gjelder bare i en avstand fra antennen som er stor i forhold til antennens lengde. Det er brukt lineær skala for intensiteten i radiell retning i venstre del av figuren og en logaritmisk skalering til høyre. Intensitetene er alle relative til den maksimale verdien. Diagrammet leses slik (venstre del): Intensiteten ved 0 grader er satt til "10". Da er intensiteten ved 30 grader ca 7.3 og ved 60 grader ca 2.8.)

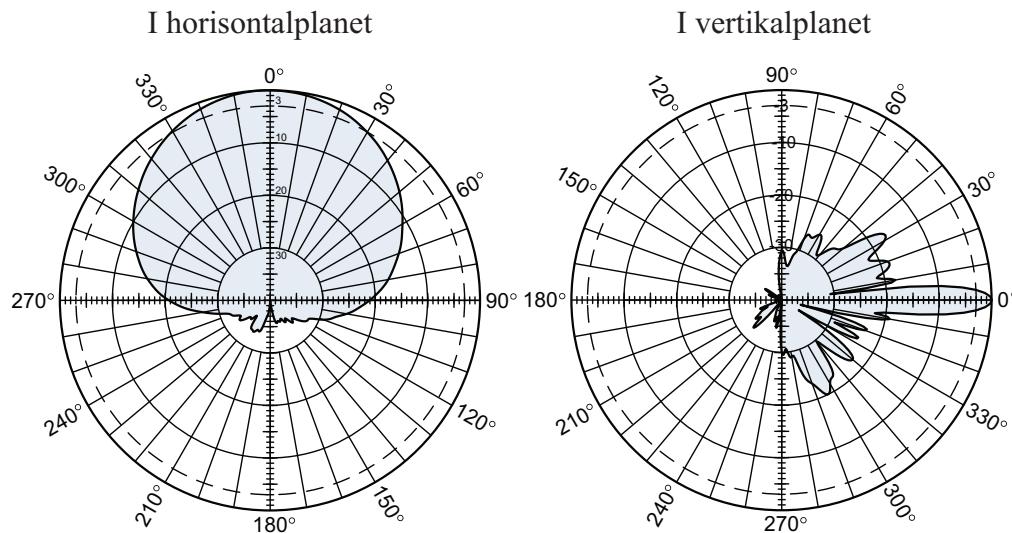
til 1.0 ($\log_{10}(1.0) = 0.0$). I figur 14.12 har vi valgt at en intensitet lik 1/1000 av maksimal verdi skal svare til null utslag. Intensiteter mindre enn dette ville bli negative i en logaritmisk skala og ville dukke opp på motsatt side av diagrammet. For å unngå misforståelser fjerner vi negative verdier før plotting.

Matlabprogrammet som ble brukt for å lage figur 14.12 var som følger:]⇐ ♠

```
\footnotesize{
function antennediagram3
N = 1024;
theta = linspace(-pi/2.0,3.0*pi/2.0,N);      % Vinkler
costheta = cos(theta);
intensitet = costheta.*costheta;
intensitet = log10(intensitet*1000.0);
for i = 1:N
    if(intensitet(i)<0)
        intensitet(i)=0;
    end;
end;
polar(theta,intensitet);}}
```

En dipolantenne kan på sett og vis sees på som en enkeltpalt med svært liten bredde. Utstrålingen blir identisk i alle retninger vinkelrett på dipolens retning. Setter vi imidlertid inn to dipoler ved siden av hverandre, og mater begge antennene med identisk signal, vil strålingsdiagrammet se ut omtrent som for en dobbeltpalt. Ved å sette mange identiske antenner etter hverandre på en rekke, får vi et strålingsdiagram omtrent som for en enkeltpalt. Ved å sette inn reflektorer og direktorer, kan vi påvirke strålingsdiagrammet ytterligere, og et eksempel er gitt i figur 14.13. Det er et antennediagram for en antennen som brukes mye i basestasjoner for mobiltelefoni her i landet. Merk at diagrammet er radielt med desibelmål i radiell retning.

Det er langt på vei samme type tenkning som ligger bak utstrålingsdiagrammene for



Figur 14.13: Retningsdiagrammer for en vanlig basestasjonsantenne for GSM 900 mobiltelefon i landet (Kathrein 80010621). Det ene diagrammet gir vinkelfordeling i vertikal retning, det andre for horisontal retning.

antennen som for lysintensitetsfordeling ved en og flere spalter osv. Det bærende element i beregningene er interferens mellom tilstrekkelig koherente bølger. Da kan vi bruke tenkningen om forskjeller i veilegning for å addere ulike bidrag med korrekt innbyrdes fase.

14.9 Lasere

Lasere er i dag en av de viktigste lyskildene i naturvitenskap, og lasere brukes også i teknologi så som CD og DVD-spillere og i laserskrivere. Lasere brukes ved kutting av metaller og andre materialer, og i medisinen f.eks. ved omforming av hornhinna på øynene våre og andre operasjoner. Ja, til og med min tannlege har gått over til å bruke lasere ved "boring" i tennene. Noen bilfabrikanter bruker nå lasere som frontlykter. Bruksområdet for lasere er imponerende, og øker stadig!

Ordet laser er en forkortelse for Light Amplification by Stimulated Emission of Radiation. Theodore Maiman ved Hughes Research Laboratories klarte å lage verdens første laser (se figur 14.14). Dette skjedde 16. mai 1960. Laseren feiret derfor sitt 50-årsjubileum i 2010. Det er imidlertid svært mange fysikere som har vært involvert i utviklingen og utnyttelsen av laseren, og det var Charles H Townes som i 1964 fikk Nobelprisen i fysikk "for the development of laser principles". Også andre Nobelpriser i fysikk er temmelig nært knyttet opp mot laseren på en eller annen måte. Det er derfor naturlig at vi kommer litt inn på prinsippene bak en laser. Det er imidlertid bare hovedprinsipper som blir berørt, og også her brukes en "tegner og forteller"-tilnærming.

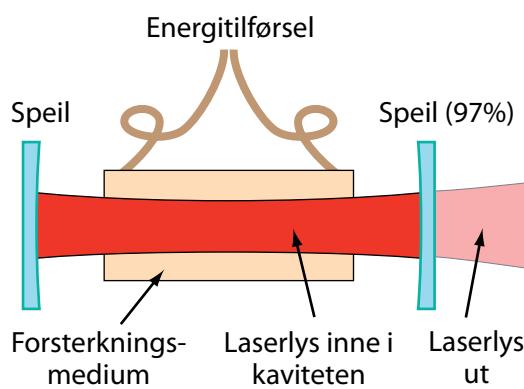
En laser er basert på såkalt *stimulert emisjon*. Einstein hadde allerede i 1917 vist at vi kan ha stimulert emisjon fra f.eks. atomer. Med det mener vi at vi ikke behøver å eksitere et atom og vente til det finner det for godt å sende ut lys idet atomet går tilbake til grunntilstanden. Vi kan ved å tilføre litt lys til atomet faktisk trigge/stimulere atomet til å gå tilbake til grunntilstanden. Laseren er basert på at et medium lar seg stimulere til emisjon og at lyset som emisjonen fører til i sin tid er med på å stimulere utsendelse av enda mere lys fra mediet. På denne måten får vi en positiv tilbakekobling i prosessen som gjør at den nærmest går av seg selv. Energien som går til spille på ulikt vis må imidlertid



Figur 14.14: Foto av den første laseren. Et blitz-rør omkranser en rubinstav. Rubinstaven er belagt med et nær 100 % reflekterende speil i ene enden og ca 95 % reflekterende speil i andre enden. Laseren ga fra seg pulset koherent lys. Fotografiet ble frigitt fra Hughes Research Laboratories i forbindelse med 50-årsjubileet i 2010.

kompenseres for, så det trengs en ytre energikilde for at prosessen skal vedvare.

Figur 14.15 viser hovedingrediensene i en laser. Den inneholder et forsterkningsmedium som kan tilføres energi fra en ytre energikilde. Mediet ligger i en optisk kavitet ("hulrom" eller "boks") begrenset av to speil. Lys som dannes i mediet vil i utgangspunktet spre seg ut i alle retninger, men lys som treffer speilet har en tendens til å bli reflektert tilbake gjennom mediet, treffer speilet på andre siden, og går så fram og tilbake og setter opp en stående bølge av elektriske og magnetiske felt i kavitetten. Lyset som reflekteres fram og tilbake kan virke som stimuli for å få mediet til å gi fra seg enda mere lys.



Figur 14.15: Hovedbestanddelene i en vanlig laser (prinsipielt).

I vanlig lys kommer lyset fra mange atomer eller molekyler som opptrer temmelig uavhengig av hverandre. Da får vi en tilstand som svarer til våre reelle sangere. Lyset fra lyskilden øker omtrent proporsjonalt med antall atomer/molekyler som sender ut lys, og lyset kommer i "alle" retninger.

I en laser derimot, befinner atomene/molekylene som sender ut lys seg i kavitetten mellom de to speilene. Det elektromagnetiskefeltet vil øke gradvis og mer og mer lys blir

dannet på grunn av emisjonen. Når vi starter laseren er det spontan emisjon i atomene som dominerer, men etter hvert blir det mer og mer stimulert emisjon. En viss tid er det fortsatt mindre lys som dannes enn det som forsvinner i form av ulike typer tap. Før eller senere overgår stimulert lys tap av lys, og da løper systemet på en måte løpsk. Effekten av lyset i kavitetten stiger med flere størrelsesordener i en bestemt retning, og linjebredden til emisjonen blir flere størrelsesordener mindre enn den var fra først av. Den magiske grensen kalles på engelsk “lasing threshold”, grensen for at laseren skal lase.

Grunnen til at hele prosessen skal gå som beskrevet er at det kraftige elektromagnetiske feltet i kavitetten sørger for at alt lyset som kommer fra de ulike atomene har temmelig lik fase som de stående bølgene mellom speilene. Da vil amplitudene av elektrisk og magnetisk felt adderes direkte, og intensiteten på lyset blir kvadratet av antall atomer/molekyler som sender ut lys.

Siden lysbølgene i en laser befinner seg i en “kavitet” (hulrom med speil i begge ender, se figur 14.16), vil lyset som nevnt danne stående bølger. Da vil frekvensen bli meget presis, på en lignende måte som lyden fra en gitarstreng som er festet i begge ender er temmelig presis. Dersom avstanden mellom speilene er 30 cm for en HeNe-laser med bølgelengde ca 633 nm, det vil være om lag 473940 bølgelengder mellom speilene. Selve linjebredden på den energiovergangen vi bruker i neon-gassen (kolliderer med andre molekyler) er såpass bred at det da i blant kan forefinnes flere samtidige bølgelengder i kavitetten. Med 473940 bølgelengder mellom speilene vil bølgelengden være 632.9915 nm, men med en bølgelengde mer eller mindre enn dette, vil bølgelengden være 632.9902 og 632.9929 nm henholdsvis. Vi snakker om “moder” for laserlyset. Ved mekanisk oppvarming av laserkavitetten, vil avstanden mellom speilene endre seg litt. I så fall vil bølgelengdene også endre seg, og vi får såkalt “mode hopping”.

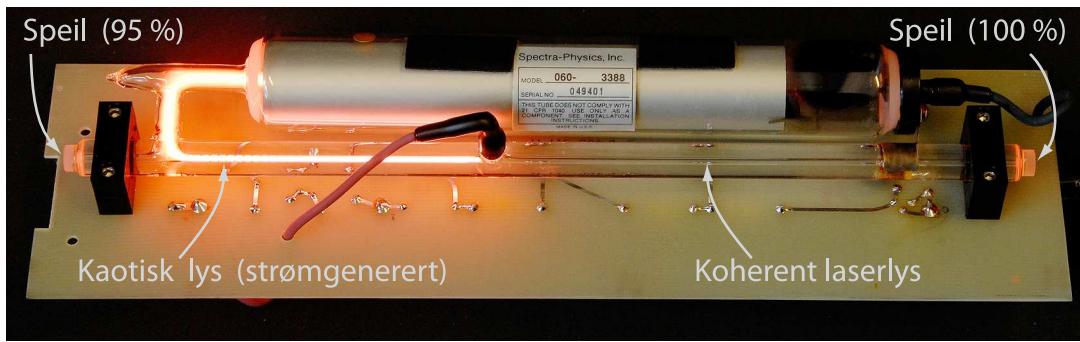
I enkelte sammenhenger lages det lasere som spiller aktivt på de ulike modene laseren kan operere i. Vi kan da oppnå mange bølgelengder som ligger med tilnærmet samme bølgelengdeavstand fra hverandre, og fenomenet kalles “frekvens-kam”. Theodor Hänsch fikk i 2005 Nobelprisen fordi han hadde laget en “frequency comb synthesizer” som gjorde det mulig for første gang å måle oscillasjonene i lys med ekstrem presisjon. Metoden danner basis for våre mest moderne atomur.

Selv om frekvensen til laseren skulle endre seg litt etter som tiden går, blant annet fordi temperaturfluktasjoner gjør at lengden på kavitetten endrer seg litt med tiden, vil ikke dette ødelegge for hovedprosessen. Alt lyset som kommer fra ulike atomer/molekyler inne i laseren vil fortsatt være i fase fordi feltet i kavitetten tvinger nytt lys som sendes ut til å være i fase med hovedfeltet til enhver tid. Av den grunn vil aldri et av atomene bidra med lys som er i motfase til lys fra et annet atom. Lyset blir koherent. Dette er en idealisert beskrivelse, men virkeligheten ligger ikke så langt fra det ideelle i gode lasere.

Laserlyset slipper ut av kavitetten ved at et av speilene i enden av kavitetten ikke reflekterer 100 %, men bare kanskje 95 - 98 %. Det lyset som da slipper ut er ganske annerledes enn lys fra f.eks. en glødelampe.

Dersom vi sammenligner fasen til den elektromagnetiske bølgen rundt omkring i et plant tverrsnitt normalt på strålen, vil fasen overalt i planet være nesten identisk. Vi omtaler dette som romlig (spatsiell) koherens. Høy grad av romlig koherens vil si at bølgefronten er meget plan. Dette er ulikt lys som stammer fra mange atomer som ikke har noe faserelasjon med hverandre, f.eks. lys fra en glødelampe. Slikt lys har liten romlig koherens, hvilket betyr at bølgefronten bukler mye på seg.

Den stimulerte emisjonen i kavitetten er temmelig stabil, og frekvensen er så veldefinert



Figur 14.16: Et fotografi som viser innmaten til en vanlig laboratorielaser av HeNe-typen. En fortynnet blanding av helium og neon befinner seg i en kavitet mellom to speil. En del av kaviteten er felles med et mer eller mindre standard lysrør a la det vi finner i farget reklamebelysning natterstid (“neon-reklame”). Elektrisk strøm gjennom dette lysrøret gir kaotisk, ikke-koherent lys som sendes ut i alle retninger. Energi tappes fra de atomene som er eksitert av den elektriske strømmen og kollisjonene den medfører. Denne energien brukes for å bygge opp en kraftig, koherent lysstråle mellom de to speilene. Vi ser ikke denne strålen fra siden, fordi lyset i kavitetten er nær perfekt rettet langs aksen mellom speilene, og da slipper ikke stort ut til sidene.

hele tiden, at vi kan forutsi fasen i laserlys-strålen mange bølgelengder framover. Denne form for koherens kalles temporær eller “tidskoherens”.

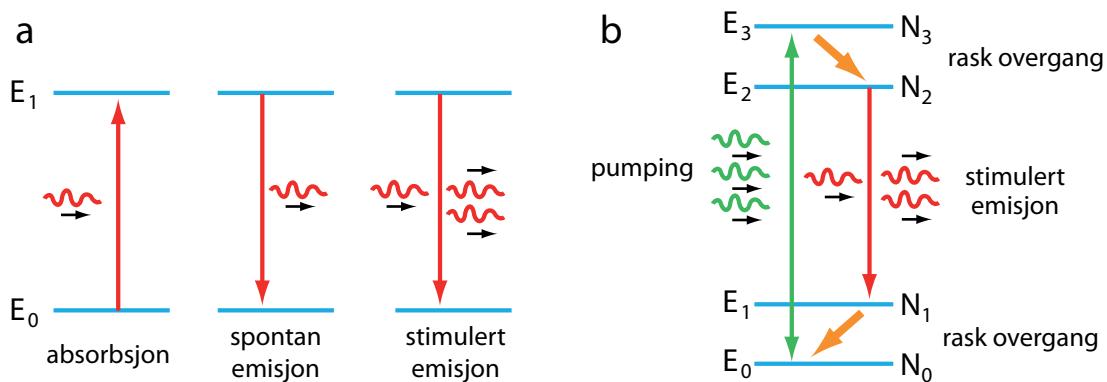
Siden bølgefronten er svært godt definert på tvers av strålen samtidig som vi kan forutsi fasen til laserlyset lange avstander langs selve strålen, er en laser en ekstremt mye bedre lyskilde ved interferens og diffraksjons-eksperimenter sammenlignet med såkalt termisk lys (“ikke-koherent” lys). Det betyr også at en laserstråle vil holde seg samlet som en meget vel avgrenset stråle der diffraksjon holdes på et minimum. Lyset i en laserstråle er noe av det nærmeste vi kan komme til en matematisk idealisert bølgebeskrivelse i praksis. Laserlys kalles derfor iblant for “klassisk lys”, men en slik betegnelse forvirrer mer enn den er til nytte.

14.9.1 Populasjonsinvertering

Når vi skal forklare lasere, kommer vi ikke helt utenom en detalj kalt populasjonsinvertering. Vi skal ikke gå i detalj, siden dette temaet ikke er så viktig i vårt sammenheng. Her er likevel en kort gjennomgang.

Et atom kan befinner seg i flere ulike energitilstander. Vi tegner ofte energitilstandene skjematisk som i venstre del av figur 14.17. Grunntilstanden markeres gjerne med E_0 og første eksitere tilstand E_1 . Et atom kan eksiteres fra grunntilstanden til en eksitert tilstand bl.a. ved å plassere det i et elektromagnetisk felt med frekvensen $\nu = (E_1 - E_0)/\hbar$ der \hbar er Plancks konstant. Et atom i eksitert tilstand kan falle tilbake til grunntilstanden helt av seg selv (kalles spontan emisjon). Vi kan også stimulere overgangen med et elektromagnetisk felt med samme frekvens som angitt for absorpsjon.

Ved absorbsjon stjenes lys fra en lysstråle for å eksitere atomet, mens ved stimulert emisjon frigjøres lys fra atomet. Det er samme sannsynlighet for den ene overgangen som den andre per atom, forutsatt at det er i den aktuelle utgangstilstanden. Skal vi få frigjort mer lys enn vi putter inn (slik det kreves i en laser), må det være flere atomer i den eksiterte tilstanden enn i tilstanden atomet faller tilbake til ved emisjon. Populasjonen av energinivåene følger vanligvis Boltzmannstatistikk. Da er det flere atomer i en lav energitilstand enn i en høyere. Det vil si at det vanligvis er umulig å danne en laser ut fra atomer i termisk likevekt.



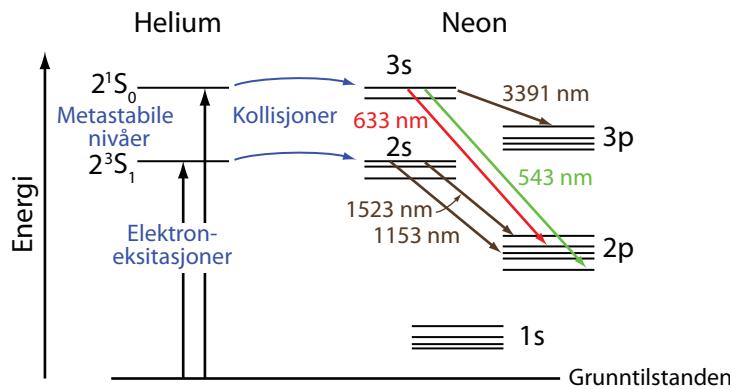
Figur 14.17: Venstre del: To energitilstander i et atom, og skjematiske overganger mellom disse (meget forenklet). Høyre del: Populasjonsinvertering kan oppnås ved pumping mellom andre energinivåer enn hvor laserlyset skapes. Se tekstu for detaljer.

I høyre del av figur 14.17 er det vist en måte å få til høyere populasjon i en energitilstand enn i en lavere. Prinsippet brukes i neodymium YAG lasere og er basert på fire energinivåer. Atomene eksiteres ved hjelp av kraftig lys fra en eller annen lyskilde, fra grunntilstanden til fjerde energinivå (E_3). Atomet går da spontant raskt over til energitilstanden E_2 , men holder seg her. Det er også en rask spontan overgang fra E_1 ned til grunntilstanden. Overgangen fra E_2 til E_1 er imidlertid ikke raskt, og etter en del pumping, blir det flere atomer i E_2 enn i E_1 . Vi har fått en populasjonsinversjon!

Sender vi nå et (svakt) lys med frekvensen $\nu = (E_2 - E_1)/\hbar$, vil vi få mer utsendt lys enn absorbert lys fra atomene, og det ligger an til at vi kan danne en laser. Intensiteten på laseren vil likevel være begrenset av hvor raskt vi kan pumpe atomer fra grunntilstanden til E_4 når en laserovergang har funnet sted.

En av de mest vanlige laserne i laboratoriesammenheng er helium-neon laseren. Her er mer kompliserte energinivåer involvert. En skisse (til orientering) er vist i figur 14.18. I dette tilfellet eksiteres heliumatomer ved at det sendes en elektrisk strøm gjennom gassblanding av helium og neon. Elektroner med betydelig hastighet står for eksitasjonen. Helium har to eksiterte nivåer som er "metastabile" slik at helium kan være i disse tilstandene ganske lenge før de faller ned til lavere energier. Dersom et slikt eksitert heliumatom kolliderer med et neonatom, vil den eksiterte energien kunne overføres fra helium til neon. Neonatomet kan så deksiteres videre bl.a. gjennom en overgang som gir lys ved 632.8 nm. Det er dette røde lyset vi kjenner igjen fra en HeNe-laser.

I dag er det mange ulike måter å lage en laser på. Folk flest har vel opptil flere lasere i sitt hjem, i og med at CD og DVD-spillere bruker lasere. I tillegg har mange også en laserpeker. I alle disse eksemplene benyttes halvleder-laserdioder. Lyset fra slike laserdioder er kontinuerlig i tid. Det finnes også lasere som bare gir fra seg til dels meget korte lyspulser. Pulslengden kan være helt ned i det såkalte femtosekund-området. Bølgelengden er ikke



Figur 14.18: Energitilstander som er involvert i en vanlig HeNe-laser.

vel definert for slike korte laserpulser!

Relevans for oss?

♣ ⇒ Da vi gikk gjennom generering av elektromagnetiske bølger ved hjelp av en oscillerende ladning eller oscillerende dipol, baserte vi oss på Maxwells ligninger. Prosessen ble beskrevet som kontinuerlige funksjoner, og vi fikk en elektromagnetisk bølge som varte ved så lenge oscilleringen foregikk.

Når vi forklarte laseren brukte vi energinivåer og hopp fra en energitilstand til en annen. Et slikt bilde bygger på kvantefysikk, men egentlig bare en kvantefysikk som er basert på energiegentilstander der vi bruker perturbasjonsteori for å se på sannsynligheter for overganger. Hvordan skal slike energidiagrammer som angitt i figur 14.17 oppfattes? Når en overgang først foregår, foregår den da øyeblikkelig, eller tar overgangen en del tid? Dette spørsmålet er det vanskelig å få et godt svar på!

Vi tegner gjerne inn "fotoner" som små bølgepakker i slike diagrammer, hvilket innebærer at vi antyder at det kommer ut en liten bølge når et foton frigjøres fra et atom. Men hvor lang er så denne bølgens? Kan vi ha bølger som kommer ut som har en fasehukommelse (koherenslengde) som svarer til flere hundre tusen bølgelengder, men som selv har bare nesten ingen utstrekning?

Og hvordan kan det ha seg at elektroner ved lavere frekvenser gir en kontinuerlig, vedvarende bølge i Maxwells formalisme, mens elektromagnetiske bølger ved lys-frekvensene plutselig ikke kan beskrives som vedvarende bølger (men som foton-partikler)?

Det er en kobling mellom kvantemekanikk og det klassiske bildet av oscillerende elektriske dipoler: Det er nettopp dipolmomenter i atomer, beregnet kvantemekanisk etter en operator som har en klar klassisk analogi, at vi kan regne ut sannsynligheter for at et atom skal sende ut lys.

Til tross for at fysikere i dag ofte omtaler lys som "fotoner" som har en litt uklar partikkelnatur, kan de aller fleste fenomener hvor lys er involvert forklares ut fra bølgemodellen for lys. Det er svært få fenomener hvor vi må bruke en partikkelmanmodell. Men hva menes med bølger og hva menes med partikler når alt kommer til alt? Det kan hende bølge-partikkkel-dualismen og de tilsvarende paradoksene som følger av en slik oppfatning, kan forsvinne dersom vi forsøker å bli litt mer presise i vår beskrivelse.

Det er nå vel 100 år siden forrige gang fysikere skiftet fra en grunntanke til en annen grunntanke om lys. Kanskje det er på tide at det skjer et nytt skifte? ← ♣]

14.10 Litteratur

Det er store mengder stoff om koherens, elektromagnetisk dipolstråling og lasere på web. Wikipedia har gode artikler om "Coherence" (innen fysikk), "Laser" og "Lasing threshold".

14.11 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for hva som skiller en reell bølge fra en idealisert enkel matematisk beskrivelse av en bølge.
- Gjøre rede for hva som menes med koherenslengde og koherenstid.
- Kjenne til begrepet autokorrelasjonsfunksjon.
- Forklare kvalitativt at linjebredde i et frekvensspekter har en sammenheng med koherenslengder.
- Forklare hvordan vi kan måle koherenslengder med et Michelson interferometer.
- Forklare kvalitativt at en ladning i oscillerende bevegelse fører til at det sendes ut elektromagneiske bølger.
- Angi en kvalitativ sammenheng mellom en sammensatt radiofrekvens-antenne og diffraksjon av lys fra to eller flere spalter i en skjerm.
- Forklare hvorfor en laser i utgangspunktet får en (temporær og longitudinal) koherenslengde som er langt større enn termisk lys fra f.eks. en glødelampe.
- Forklare kvalitativt hvorfor populasjonsinversjon er viktig for å få en laser til å fungere.

14.12 Oppgaver

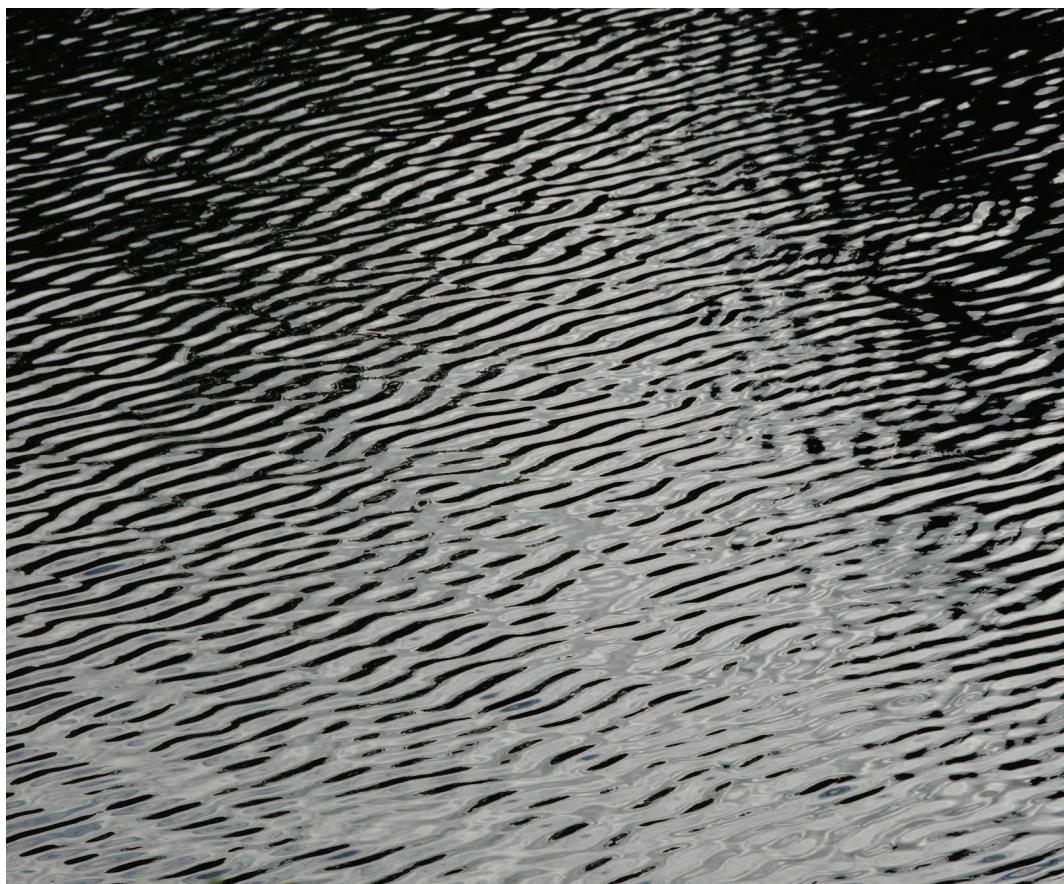
Forståelses- / diskusjonsspørsmål

1. Forsøk å forklare hva som ligger bak at signalet i figur 14.4 varierer så mye i amplitude selv om signalene vi startet ut med hadde mye jevnere amplitude.
2. Vil en sanger som synger en “død” tone forventes å ha en større eller mindre koherenstid for stemmen enn en sanger som har betydelig “vibrato” i sangen sin?
3. Angi fordeler og ulemper ved at folk som synger i et kor har relativt korte koherenstider for stemmen (lyden).
4. Ut fra figur 14.5 skal du kunne estimere koherenstiden for sangerne som var involvert. Forklar hvorfor det er mulig, og forsøk å angi en verdi for koherenstiden ut fra disse figurene.
5. Noen tror at når vi snakker om koherente og ikke-koherente bølger, er det snakk om to vel avgrensede typer bølger. I virkeligheten er det en kontinuerlig overgang fra “ikke-koherent” til “koherent”. Forklar.
6. Det er fordeler og ulemper knyttet til både koherent lys og ikke-koherent lys. I hvilke sammenhenger ville du foretrekke det ene og når ville du foretrekke det andre? Begrunn som vanlig svarene.
7. Forsøk å beskrive med egne ord hva vi mener med koherenslengde og koherenstid for en bølge. Hva er forskjellen mellom romlig koherens (spatiell koherens) og tidmessig koherens (temporær koherens)?
8. Hva er hovedtanken bak “retardert potensial”?
9. Hva er viktigst for å skape en elektromagnetisk bølge fra en oscillerende ladning eller oscillerende dipol, enten den radielle komponenten av det elektriskefeltet eller den tangentiente komponenten? Forklar!
10. For at en laser skal virke må det stilles krav til lys som stammer fra stimulert emisjon. Hvilke krav?
11. Forklar hvordan populasjonsinversjon kan oppnås i et fire-eneginivå-skjema.

Regneoppgaver

12. a) Bestem hvor mange bølgelengder koherenslengden er for den lyden som er beskrevet i figurene 12.1 - 12.5.
 b) Anta at koherenslengden er like mange hele bølgelengder også for toner i hele intervallet 100 - 8000 Hz. Hvor lang ville i så fall koherenslengden være (i antall meter) for lyd med frekvensene 100, 1000 og 8000 Hz?
 c) Anta at vi sender samme tonesignalet til to identiske stereohøyttalere plassert 4.0 meter fra hverandre (begge vender i samme retning, vinkelrett på forbindelseslinjen mellom dem). Forsøk å lage et slags kart over områder foran høyttalerne hvor vi kan forvente å observere interferanseffekter og hvor vi ikke skal forvente slikt. Velg tre-fire ulike frekvenser for kartleggingen.

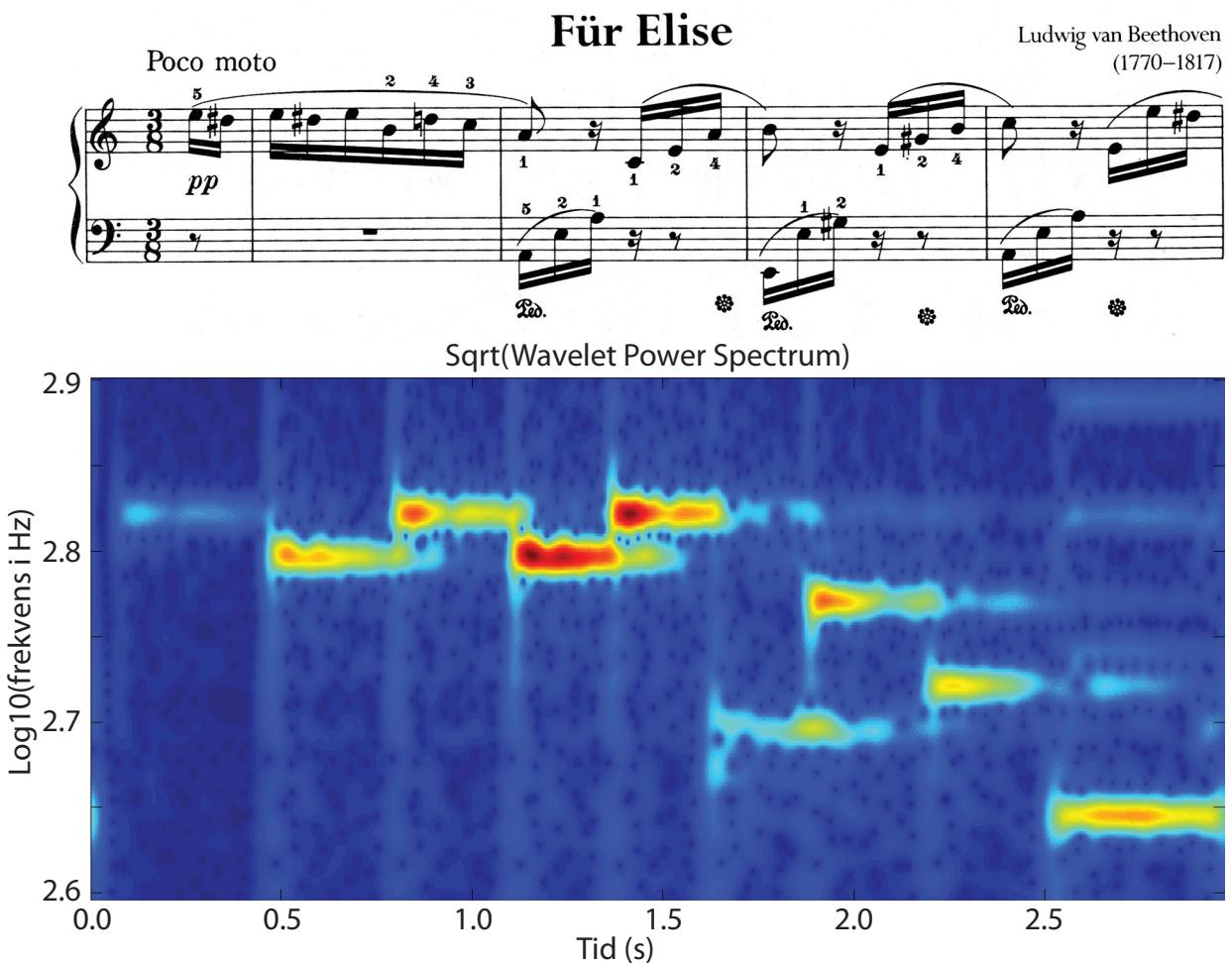
13. I figur 14.13 er det gitt et antennediagram for en mye brukt basestasjonsantenne. I et konkret tilfelle står en slik antenne på en mast 22 m over bakken. Bestem intensiteten på bakkenivå i en avstand 30 m fra masten (målt langs bakken) i forhold til intensiteten 500 m fra antennen i den retningen i horisontalplanet hvor intensiteten er størst. Gjennomfør samme beregning dersom det i stedet var brukt en enkel dipolantenne. Er det gunstig at basestasjonsantennen har den intensitetsprofilen den har, eller ville det vært en fordel om det ble brukt en enkel dipolantenne i stedet?
14. Kan du foreslå en måte å lage et interferometer for lyd som i funksjon tilsvarer et Michelson interferometer for lys? (Skal kunne brukes for å måle koherenslengder for lyd fra f.eks. ulike musikkinstrumenter.)
15. Bestem koherenstiden til egen stemme. Nærmere bestemt går oppgaven ut på følgende:
 - a) Lag et dataprogram hvor du kan digitalisere lyd, beregne autokorrelasjonsfunksjonen og plottet en utvalgt del. Bruk plottet til å anslå omtrentlig koherenstid for signalet. Dersom du har laget et program for digitalisering av lyd ved arbeidet med kapittel 4, kan du vinne mye tid på å utnytte dette også her.
 - b) Fortell spesielt hvordan du valgte å utnytte datastrengen du fikk ved digitaliseringen i analysen. Nærmere bestem: Hvordan valgte du å la indeksene i og j i ligning (14.4) løpe i forhold til den totale datastrengen?
 - c) Bestem omtrentlig koherenstiden til egen stemme når du synger "iiiii" med så jevn stemme du klarer. Gjennomfør dette for 2 - 3 ulike tonehøyder. Synes koherenstiden å endre seg mye med tonehøyden?
 - d) Fortell spesielt litt om hvilke problemer du muligens har hatt i bestemmelsen av koherenstid ut fra plottet av autokorrelasjonsfunksjonen. Du forstår nå kanskje bedre kommentarene i figurteksten til figur 14.6? Kan du antyde hvordan en grundigere statistikk ville kunne gi bedre resultater?
 - e) Digitaliser en annen lyd og bestem koherenstiden også for denne. (Forslag til lyd: Egen stemme, samme tonehøyde som du har brukt i punkt c, men at du nå synger "ooooooooo" i stedet for "iiiii". Alternativt: Lyd fra et piano, gitar eller et annet musikkinstrument.) Finner du noe interessante forskjeller eller likheter sammenlignet med det du fant i punkt c?
16. Bestem koherenslengden for en laser ved å bruke et Michelson interferometer.
17. Søk på internett for å danne deg et bilde av status for lasere i røntgenområdet. Hvor langt er utviklingen kommet? Hvilke anvendelser har en røntgenlaser?
18. Forsøk i figur 14.19 å markere områder hvor bølgene er relativt veldefinerte. Hvor store er disse områdene omtrentlig? Og hvor stor del av hele vannoverflaten har du tatt hensyn til i markeringen av disse områdene? Angi lengder i "omtentlige bølgelengder" som mål. [For å få et korrekt bilde av koherenslengder må egentlig hele overflaten tas med i den statistiske behandlingen.]



Figur 14.19: *Overflatebølger på vann ved ett tidspunkt.*

Kapittel 15

Wavelettransformasjon



Ønsker du deg en samtidig frekvens- og tidsanalyse, er wavelettransformasjon en spennende metode. Moderne wavelettransformasjon har ofte store fordeler framfor klassisk fouriertransformasjon. Atpå til er det utfordrende analogier mellom waveletanalyse og Heisenbergs uskarphetsrelasjon. Håper ditt møte med wavelets blir en spennende reise inn i nytt terreng!

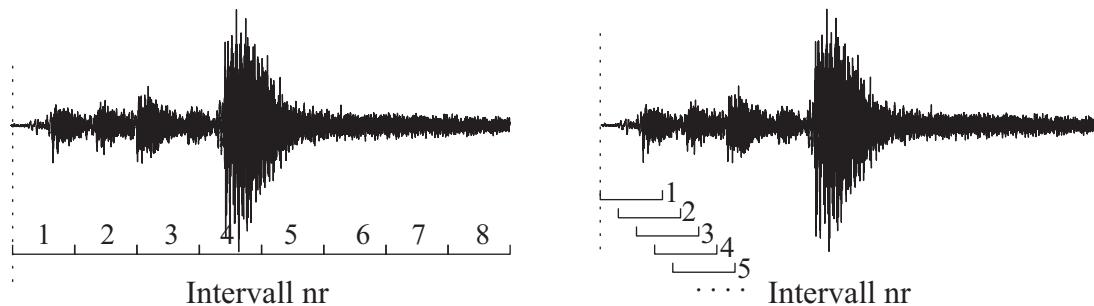
¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 16042013.

15.1 Hva slags info kan wavelettransformasjon gi oss?

Anta at vi synger en “a” og holder tonen hele tiden mens vi digitaliserer lyden. Fourieromvendes signalet vil vi se et frekvensspekter med grunntone og overharmoniske temmelig likt det vi finner fra analysen av en ren tone fra et musikkinstrument. For “stasjonære” signaler (endrer ikke nevneverdig karakter underveis) er fouriertransformasjon en fantastisk metode for å hente ut verdifull informasjon.

Anta derimot at vi digitaliserer skarpe trommelyder med nesten helt stille partier innimellom og foretar fouriertransformasjon. Da vil vi se at fouriertransformasjon egentlig er en dårlig metode ved analyse av signaler som endrer karakter med tiden. Når vi tar ut absoluttverdien fra en fouriertransformasjon, forsvinner enhver tidsinformasjon, mens frekvensoppløsningen er god. Men hva betyr egentlig frekvens i et signal som endrer karakter så mye underveis? Det spørsmålet bør vi tenke litt over, slik at vi ikke bruker en slagkraftig metode på et datamateriale der metoden egentlig ikke egner seg!

Vi kan bøte på dette problemet noe ved å stykke opp signalet slik at vi har et noenlunde ensartet signal innenfor hver bit vi analyserer. I så fall anvender vi en metode som kalles stykkevis fouriertransformasjon (se figur 15.1). Den formen for waveletanalyse vi skal diskutere i dette kapitlet ligner litt på stykkevis fouriertransformasjon, men har en rekke fordeler framfor denne. Vi skal i dette kapitlet gi en innføring i waveletanalyse og vise hvordan metoden kan optimaliseres.



Figur 15.1: For å få tidsinformasjon ved analyse av en lang tidssteng med data, kan vi stykke opp det totale tidsintervallet og foreta fouriertransformasjon for intervall etter intervall. Intervallene kan velges slik at de ikke overlapper hverandre (venstre del) eller slik at de overlapper hverandre (høyre del). Se teksten for vurdering av disse teknikkene.

Benyttes stykkevis FT ved at den totale datamengden deles opp i f.eks. m like store biter, skal det noe til at vi får ensartede signaler innenfor hver av bitene. Dersom vi i stedet velger å stykke opp i m biter med varierende lengde slik at det blir et noenlunde ensartet signal innen hver bit, vil det bli store forskjeller i frekvensspekteret fra bit til bit: Analyseresultatet vil bli kritisk avhengig av hvordan vi stykker opp den totale datamengden.

En bedre metode kan kanskje være å velge ett “vindu” (f.eks. 256 datapunkter), og gjøre analysen først for de første 256 datapunktene i den lange datastrengen vi har, deretter flytte vinduet litt (f.eks. 16 punkter), og gjøre anlysen på punkt 17 til og med punkt $17+255=272$, for så å flytte vinduet enda et hakk for neste analyse.

Ved en slik “glidende filter”-metode unngår vi hopp i resultatene som skyldes tilfeligheter i hvordan intervallene velges. Ulempen er at vi må gjennomføre til dels mange tilsynelatende unødvendige beregninger. Vi får med andre ord ganske mye overflødige data (engelsk: “redundancy”) i resultatene.

Den største ulempen med en glidende stykkevis fouriertransformasjon er likevel at den relative frekvensoppløsningen blir svært forskjellig for høye og lave frekvenser. Vi husker fra et tidligere kapittel at ved en diskret fouriertransformasjon av en tidsstrep med varighet T blir frekvensoppløsningen $1/T$. Anta at vi f.eks. velger $T = 0.5$ s. Da blir frekvensoppløsningen 2 Hz. *Relativ* frekvensoppløsning kan defineres som frekvensoppløsning dividert på signalets frekvens. Ved 50 Hz vil da den relative frekvensoppløsningen bli $2/50 = 4\%$, mens ved et 5 kHz signal vil den relative frekvensoppløsningen bli $2/5000 = 0.04\%$. Vi ser altså at frekvensangivelsen vil bli langt mer nøyaktig for høye frekvenser enn for lave.

“Kontinuerlig waveletanalyse” ligner på en glidende, stykkevis fouriertransformasjon, men gir samme relative frekvensoppløsning for alle frekvensene. Hemmeligheten er å bruke ulik lengde på tidsstrengeen alt etter hvilken frekvens vi analyserer.

Resultatet kan fremstilles som et tredimensjonalt plot. Langs x-aksen har vi tid, angitt f.eks. ved midtpunktet i analysevinduet vi bruker. Langs y-aksen har vi frekvensskalaen, og som den tredje komponenten har vi f.eks. intensiteten for bølgen (fourierkomponenten for denne frekvensen kvadrert). Denne informasjonen kan gis i en eller annen form for 3D-graf, f.eks. som “stack-plot”, kvotekurver, surface mesh, eller fargekoding for å nevne noen anskuelsesformer.

Det kan nevnes at det også finnes en diskret wavelettransformasjon hvor vi foretar så få transformasjoner som overhodet mulig uten tap av informasjon. En slik transformasjon er mye mer effektiv enn den kontinuerlige, og brukes i teknologiske sammenhenger der det er viktig at ting går fort (for eksempel i MP3-spillere). Ulempen med en diskret wavelettransformasjon er at resultatet er langt vanskeligere å forstå enn et vanlig fourierspekter. Det er hovedgrunnen til at vi ikke går inn på den metoden her.

Waveletanalyse er et omfattende fagfelt innen matematikk/informatikk, og det gis egne kurs om emnet ved mange universiteter. Vi kommer ikke til å gå i detalj om den strengt matematiske eller datatekniske siden av wavelets. Hensikten med å ta med wavelets i denne boka, er å gjøre studenten oppmerksom på at fouriertransformasjon *ikke* egner seg for ikke-stasjonære signaler, og samtidig peke på en analysemetode som ofte er langt å foretrekke i slike sammenhenger. Dessuten kan arbeid med wavelets bidra til en dypere forståelse av tidsavgrensede fenomener og de tilsvarende frekvenser. Blant annet er det nære analogier mellom Heisenbergs uskarphetsrelasjon og waveletanalyse.

En del av dere vil nok bruke waveletanalyse i masteroppgaven eller i et evt. PhD-prosjekt (og senere jobber). Av den grunn legger vi vekt på å vise hvor waveletanalyse er nyttig og når metoden ikke har mye å gi. Wavelets brukes bl.a. for å analysere solflek-kaktivitet (og endringer i syklusen med tiden), El Niño sørlige oscillasjoner i Stillehavet, isbre-sykler, ruhet, kornstørrelseanalyser, analyse av f.eks. kreftceller vs normale celler og mye mer.

Teknologisk er det en omfattende bruk av wavelets bl.a. i jpeg-komprimering av bildefiler, og i mp3-komprimering av lyd.

15.2 Kort historikk

Den franske matematikeren Joseph Fourier (1768-1830) "oppdaget" fouriertransformasjon for vel 200 år siden. (Fourier arbeidet forøvrig med varmestrømning, og var visstnok den første som oppdaget drivhuseffekten.)

Fouriertransformasjon benyttes i stor grad i analytisk matematikk. I tillegg fikk transformasjonen en enorm utbredelse i dataverdenen etter at J.W.Cooley og J.W.Tukey i 1965 oppdaget den såkalte "Fast Fourier Transform" (FFT) som gjør det mulig å foreta en fouriertransformasjon svært mye raskere enn tidligere. Ved FFT benyttes symmetriene i sinus og cosinusfunksjonene for å redusere antall multiplikasjoner ved utregningen, men for å få dette til, må antall datapunkter være en heltalls potens av 2, dvs $N = 2^n$.

Det hevdes at Cooley-Tukeys Fast Fourier Transform egentlig ble oppdaget av Carl Friedrich Gauss ca 1805, men glemt og delvis gjennoppfunnet flere ganger før 1965. Sukssessen til Cooley og Tukeys gjenoppdaging skyldes nok at datamaskinen gjorde sitt inntog omtrent på denne tiden.

Waveletanalyse er av langt yngre dato. Riktignok ble wavelets introdusert allerede ca 1909, men metoden ble først tatt i bruk fra ca 1980 av. Det er langt større spillerom for spesielle varianter av waveletanalyse enn ved fouriertransformasjon. Det er både en fordel og ulempe. Vi kan langt på vei skreddersy en waveletanalyse slik at den passer optimalt for de dataene vi vil analysere. Uelmpen er at den store variasjonsmuligheten medfører at vi må bruke hodet litt mer ved waveletanalyse enn ved fouriertransformasjon, både når transformasjonen skal gjennomføres, og når vi tolker resultatene. Men resultatene blir ofte desto mer interessante!

15.3 Kort om matematikken bak

15.3.1 Oppfrisking av fouriertransformasjon

Vi har gått gjennom fouriertransformasjon i et tidligere kapittel, men la oss repetere de matematiske uttrykkene også her.

La $x(t)$ være en integrerbar funksjon av tid. Da kan vi beregne en ny funksjon $X(\omega)$, hvor ω er vinkelfrekvens, på følgende måte:

$$X(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt \quad (15.1)$$

Det morsomme med denne funksjonen er at vi kan ta en tilsvarende "omvendt" transformasjon:

$$x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(\omega)e^{i\omega t} d\omega \quad (15.2)$$

og ende opp med eksakt den opprinnelige funksjonen igjen. Merk fortegnskiftet i den komplekse eksponensialfunksjonen.

Det finnes tre ulike, vanlige måter å fordele faktoren foran integraltegnet på, vi har valgt den varianten som gir mest symmetriske uttrykk.

Fra likningene (15.1) og (15.2) ser vi at når $x(t)$ er reell, vil $X(\omega)$ være kompleks. Dette er nødvendig for at $X(\omega)$ både skal kunne angi hvor kraftig svingning vi har ved ulike frekvenser, og i tillegg angi innbyrdes fase til de ulike frekvenskomponentene. (En symmetri i X medfører at x etter den omvendte transformasjonen blir reell, slik den var opprinnelig.)

Det bør forøvrig nevnes at x og X generelt sett ikke behøver å være funksjoner av tid og frekvens. Det finnes mange ulike typer funksjoner og sammenhenger hvor fouriertransformasjon anvendes. I vårt sammenheng begrenser vi oss imidlertid (nesten utelukkende) til tid og frekvens.

Uttrykkene ovenfor anvendes ved analytiske beregninger. Når vi bruker datamaskin, kjenner vi ikke fullstendig til hvordan $x(t)$ varierer i tid. Vi kjenner bare verdien av x i diskrete tidspunkt t_n . I disse tidspunktene har x verdiene x_n hvor n er en indeks som varierer fra 1 til N , dersom x er beskrevet i N tidspunkt. Vi antar at det er valgt ekvidistante tidspunkt slik at det er en fast tid mellom to nærliggende tidspunkt. Total tid x er beskrevet over er da $T = N * \delta t$ hvor δt er tiden mellom to tidspunkt i beskrivelsen (detaljer diskutert i et tidligere kapittel).

Når fouriertransformasjon gjennomføres på diskrete data, brukes en diskret transformasjon. Denne kan angis slik:

$$X_k = \frac{1}{N} \sum_{n=1}^N x_n e^{-i2\pi f_k t_n} \quad (15.3)$$

hvor $k = 0, 1, 2, \dots, N - 1$. Videre er $f_k = 0, f_s/N, 2f_s/N, \dots, f_s(N - 1)/N$ der f_s er samplingsfrekvensen. Endelig er $t_n = 0, T/N, 2T/N, \dots, T(N - 1)/N$ der $N/T = f_s$.

Det er kanskje ikke så lett å se ut av uttrykket, men det vi egentlig gjør for å bestemme den fouriertransformerte for en frekvens f_k , er å multiplisere (ledd for ledd) den digitaliserte funksjonen x_n med en cosinusfunksjon med frekvensen f_k og summere alle leddene som da framkommer. (For imaginærddelen av den fouriertransformerte multipliserer vi med en sinusfunksjon med frekvensen f_k .)

Den tilsvarende ”omvendte” transformasjonen er da:

$$x_n = \sum_{k=1}^N X_k e^{i2\pi f_k t_n} \quad (15.4)$$

for $n = 1, 2, 3, \dots, N$.

15.3.2 Formalisme ved wavelettransformasjon

Wavlettransformasjon kan angis tilsvetende på nokså analog måte som en fouriertransformasjon:

La $x(t)$ være en integrerbar funksjon av tid. Da kan vi beregne en ny funksjon $\gamma_K(\omega_a, t)$ som gir informasjon om frekvenser og tid samtidig.

ω_a kan betegnes som “analyse-vinkelfrekvens”. K er en “skarphets”-parameter (også kalt “bølgetallet”, se siden) knyttet til hvorvidt vi ønsker å ha høy presisjon i tidsangivelser (K liten) eller høy presisjon i frekvensangivelser (K stor).

Enkeltverdier for den nye wavelettransformerte funksjonen kan finnes på følgende måte:

$$\gamma_K(\omega_a, t) = \int_{-\infty}^{\infty} x(t + \tau) \Psi_{\omega_a, K}^*(\tau) d\tau \quad (15.5)$$

Her er $\Psi_{\omega_a, K}(\tau)$ selve waveleten. Asteriksen sier at det er den kompleks konjugerte av uttrykket for waveleten vi må bruke.

Det spesielle med waveletanalyse er at vi kan velge mellom nærmest uendelig mange forskjellige wavelets alt etter hva vi ønsker å få fram i analysen. I vår sammenheng kommer vi bare til å bruke såkalt Morlet wavelets. Matematisk kan den relle delen av en Morlet wavelet uttrykkes som en *cosinusfunksjon* (pluss et litt konstant korreksjonsledd) konvolert med en gaussisk funksjon (“gaussisk omhyllingskurve”). Den imaginære delen er en *sinusfunksjon* konvolert med samme gaussisk funksjon som i stad.

En Morlet-wavelet kan beskrives som:

$$\Psi_{\omega_a, K}(\tau) = C \{ \exp(-i\omega_a \tau) - \exp(-K^2) \} \cdot \exp(-\omega_a^2 \tau^2 / (2K)^2) \quad (15.6)$$

hvor C er en ”normeringskonstant”. Når vi beskrver $\Psi_{\omega_a, K}(\tau)$ numerisk, kan vi med fordel bruke følgende uttrykk for C :

$$C = \frac{0.798}{f_s K} \omega_a \quad (15.7)$$

hvor f_s er samplingsfrekvensen.

[♣ ⇒ Noen kommentarer:

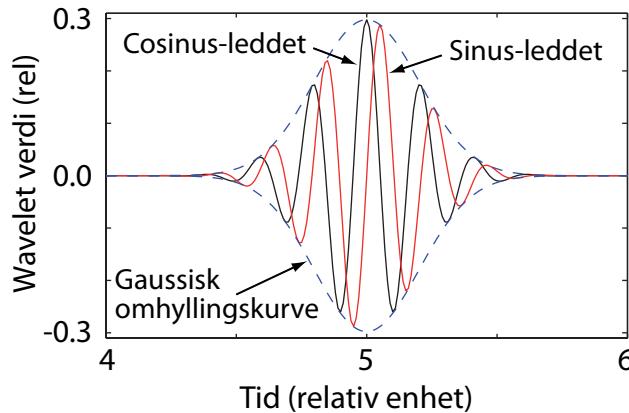
Det har ikke satt seg en ensartet beskrivelse av wavelets ennå. Forskjellige kilder angir formalismen på ulike måter, der blant annet uttrykk så som ”scaling-parameter”, ”Mother-” og ”daughter-wavelets” er sentrale begreper. Vi har valgt å bruke en fremstilling som ligger nær opp til en artikkel av Najmi og Sadowsky (se litteraturlisten) fordi denne ligger temmelig nær opp til øvrig formalisme i denne boka. ”Konstanten” C har jeg imidlertid valgt ut fra prøving og feiling ut fra ønsket om at en wavelettransformasjon av et rent sinussignal skal gi amplituden til sinussignalet uansett valg av parametrene ω_a , f_s og K (avvik fra det perfekte er som oftest mindre enn ca 1 %). Det aktuelle uttrykket for en Morlet wavelet kommer vi ikke til å bruke i praksis, bortsett fra et illustrativt eksempel. Ved effektiv wavelettransformasjon kommer vi til å ta utgangspunkt i den fouriertransformerte av waveleten, og beskriver den direkte. Detaljer gis i teksten som følger. ⇐ ♣]

I waveletanalysen sjekker vi om signalet vi studerer inneholder ulike frevenser ved ulike tider. Vinkelfrekvensen vi analyserer for ved en spesifikk wavelet er ω_a . Parameteren τ angir tiden der en spesifikk wavelet har et maksimum, og svarer til senter for det lille tidsintervallet vi undersøker.

Parameteren K er en reell konstant og kan kalles ”bredden” på waveleten. Noen kaller den ”bølgetallet”, fordi den angir omtrentlig antall bølger som det er plass til under den gaussiske omhyllingskurven for waveleten (omhyllingskurven er gitt i siste ledd i ligning (15.6)). Det anbefales at K er 6.0 eller større.

På grunn av det midtre leddet i ligning (15.6) ser vi at waveleten Ψ er kompleks.

Figur 15.2 viser et eksempel på en Morlet wavelet. Vi ser at den bærer navnet med rette: “wavelet” kan nemlig oversettes med “liten bølge”.



Figur 15.2: *Eksempel på en Morlet wavelet for $K = 6$. Både den reelle delen (cosinus-ledd) og den imaginære delen (sinus-ledd) er gitt.*

Vær sikker på at du gjennomskuer hvordan waveleten dannes, nemlig som en gaussisk konvoluttering av en kompleks harmonisk funksjon sentrert rundt tiden τ .

Merk at uttrykket i ligning (15.6) er en generell beskrivelse. Når denne skal implementeres i et datamaskinprogram, og analysere et konkret signal, må vi kjenne til samplingsfrekvensen som ble brukt. Denne kommer inn i normeringskonstanten C . Dersom det konkrete signalet er beskrevet i N ekvidistante punkter i tid, er den totale tiden samplingen har foregått lik $T = N/f_s$.

Vi velger da å beskrive enhver Morlet-wavelet vi bruker i analysen ved hjelp av en array med samme samplingsfrekvens og samme lengde på arrayen som det konkrete signalet vi skal analysere.

Dersom vi sammenligner ligning (15.1) med ligning (15.5), ser vi at uttrykkene ligner mye på hverandre. Vi integrerer opp produktet av en funksjon x og en bølge. Begge er derved knyttet til et “indreprodukt” innen matematikken, men som sagt, vi skal bare touche matematikken med en harelabb her.

Det er imidlertid flere forskjeller enn vi først skulle tro. En vesentlig forskjell ligger i at wavelettransformasjonen fører til en tredimensjonal beskrivelse (verdien av γ som funksjon av både ω_a og t), mens en beskrivelse basert på fouriertransformasjon bare er todimensjonal (verdien av X som funksjon av frekvens).

Også for en wavelettransformasjon er det mulig å foreta en ”omvendt” transformasjon. Det er essensielt når wavelets brukes i jpeg bildekompresjon og mp3 musikkfil-komprimering. Vi tar imidlertid ikke med detaljer angående denne formalismen i vår sammenheng. (Interesserte henvises til siste referanse i litteraturlisten i slutten av kapittelet.)

15.3.3 “Diskret kontinuerlig” wavelettransformasjon

Først litt om bruk av ordene “diskret” og “kontinuerlig”. Et digitalisert signal vil vi kalle diskret, fordi vi bare har et endelig antall måleresultater (ekvidistante i tid). Vi vil likevel betegne den spesielle wavelettransformasjonen som beskrives i dette kapitlet, som “kontinuerlig”, i betydning at det “glidende filteret” bare forskyves ett punkt fram i det digitaliserte signalet hver gang en ny beregning gjennomføres. Et alternativ ville være å forskyve waveleten med f.eks. halve waveletbredden.

Wavelettransformasjon brukes nesten utelukkende på diskrete signaler, siden beregningsene er så omfattende at de nesten er umulige å gjennomføre analytisk (unntatt i svært enkle modell-beskrivelser).

For digitaliserte signaler (diskrete signaler) kan selve Morlet waveleten skrives som:

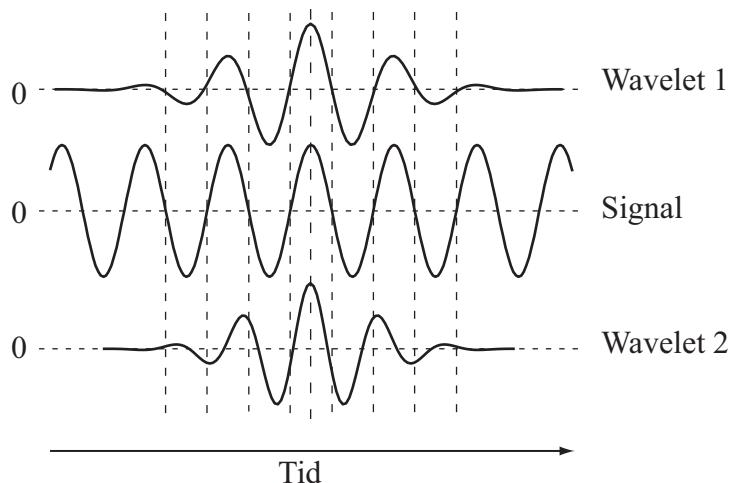
$$\Psi_{\omega_a, K, t_k}(t_n) = C \{ \exp(-i\omega_a(t_n - t_k)) - \exp(-K^2) \} \cdot \exp(-\omega_a^2(t_n - t_k)^2 / (2K)^2) \quad (15.8)$$

Her er det antatt at signalet vi skal analysere er beskrevet i ekvidistante punkter ved hjelp av tallrekken x_n for $n = 1 \dots N$. Tiden t_k angir *midtpunktet til waveleten* (!).

Selv wavelettransformasjonen vil da kunne beskrives slik:

$$\gamma_K(\omega_a, t_k) = \sum_{n=1}^N x_n \Psi_{\omega_a, K, t_k}^*(t_n) \quad (15.9)$$

La oss bildeliggjøre prosessen litt for at vi skal bedre forstå hva den går ut på. I figur 15.3 viser vi et utsnitt av en tidsstreng sammen med to ulike valg av wavelets. Wavelettransformasjon består i å multiplisere signalet med waveleten, tidspunkt etter tidspunkt, og danne en sum av produktene.



Figur 15.3: Et sinusformet signal (i midten) sammen med en wavelet med samme periodetid (øverst) og en wavelet med noe kortere periodetid (nederst).

For wavelet 1 ser vi at signalet skifter fortegn omtrent på samme sted som waveleten skifter fortegn. Det vil si at produktet i ethvert punkt blir positivt, og summen av produkter blir derfor ganske stor (svarer til at $\int \cos^2(\omega t) dt$ er positiv). For wavelet 2 skifter signalet fortegn på andre steder enn for waveleten. Noen av produktene blir derved positive og noen negative. Summen av produkter blir betydelig lavere enn for det første tilfellet (svarer til at $\int \cos(\omega_1 t) \cos(\omega_2 t) dt$ ofte er nær null når $\omega_1 \neq \omega_2$).

Vi har derved forsøkt å anskueliggjøre at wavelettransformasjonen av en regulær sinusbølge vil ha et maksimum når “bølgelengden” til waveleten svarer til “bølgelengden” til signalet i det tidsintervallet hvor vi foretar analysen.

For å analysere signalet x_n for andre bølgelengder, trenger vi å endre waveleten, og det gjøres blant annet ved hjelp av parameteren ω_a .

15.4 Praktisk gjennomføring

15.4.1 Eksempel på råmetoden for wavelettransformasjon

Vi skal nå vise i praksis et eksempel på råmetoden for wavelettransformasjon (gitt i ligning (15.9) og (15.8)). Vi tar også med Matlab-koden som er brukt for å generere de nødvendige figurene i tilfelle noen ønsker å forfölge eksemplet i litt andre retninger. Ellers er koden på de neste tre sidene nokså uinteressante.

Signalet vi genererer skifter mellom 100 og 200 Hz med faste intervaller (se figur 15.4). De ytterste bitene av signalet settes lik null. Merk at når vi genererer et signal med variabel frekvens i løpet av den tiden signalet eksisterer, må vi sørge for å unngå brudd i signalet akkurat i det tidspunktet frekvensen endrer seg. Vi får en god beskrivelse dersom vi tar utgangspunkt i *fasen* til signalet til enhver tid, og oppgraderer fasen for hvert nytt trinn i tidsbeskrivelsen. Dette er implementert i programkoden, og kontinuiteten i signalet er demonstrert i detaljutsnittet i figur 15.4.

```
function firkantSign5

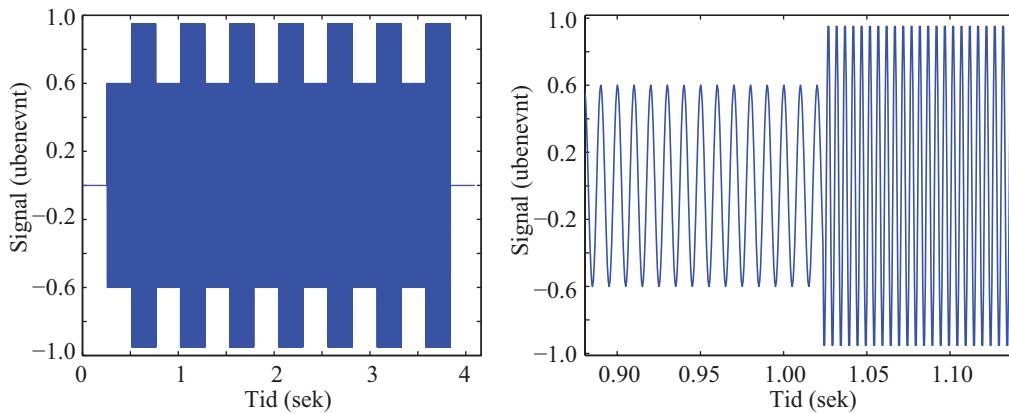
% Genererer et signal som veksler med faste intervaller mellom to
% frekvenser, hver av disse har konstant amplitude. Nullstiller første og
% siste del av signalet. Oppintegrerer fasen ved genereringen for å unngå
% diskontinuiteter når signalet skifter fra en frekvens til den neste.
N = 4096*2;
f_sampl=2000.0;
T = N/f_sampl;
t = linspace(0,T,N);
delta_t = 1.0/f_sampl;
fase = 0.0;
omega1 = 200.0*2*pi;
omega2 = 100.0*2*pi;
k = 1;
for ii = 1:16
    if (mod(ii,2)>0) delta_fase = omega1*delta_t;
        A = 1.0;
    else delta_fase = omega2*delta_t;
        A = 0.6;
    end;
    for j = 1:512
        fase = fase+delta_fase;
        signal(k) = A*sin(fase);
        k = k+1;
    end;
end;
```

```

signal(1:512) = 0.0;
signal(N-512:N) = 0.0;
% % FOR UTTESTING (genererer et enkelt harmonisk signal):
% omeg = 2.0*pi*100;
% signal = sin(omeg*t);

% Plotter signalet, beregner så den fouriertransformerte og plotter denne
plot(t,signal,'-b');
xlabel('Tid (sek)');
ylabel('Signal (rel enheter)');
figure;
FTsignal = fft(signal);
f = linspace(0,f_sampl*(N-1)/N, N);
plot(f,abs(FTsignal),'-b');
xlabel('Frekvens (Hz)');
ylabel('Fourierkoeffisient (rel enhet)');
title('Den fouriertransformerte av signalet');
xlim([-0.0,300]);
% ylim([-0.2,2.2]);

```



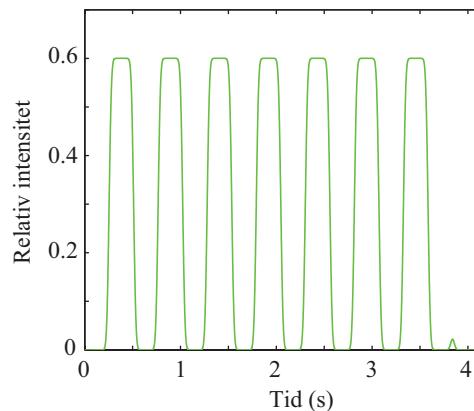
Figur 15.4: Det genererte tidssignalet vi har brukt i eksemplet vårt. Venstre del viser hele signalet, mens en detalj av dette er vist i høyre del. Amplituden på 100 Hz signalet er 0.6 mens amplituden på 200 Hz signalet er 1.0.

Vi implementerer så wavelettransformasjonen direkte ut fra ligningene (15.9) og (15.8). Analysefrekvensen for waveleten vi har valgt er $\omega_a = 2\pi 100$ (som svarer til 100 Hz i selve signalet). For hvert nytt punkt i wavelettransformasjonen som beregnes, forsøker vi toppunktet til waveleten fra å ligge helt i ytre høyre kant til helt i ytre venstre kant. Resultatet er vist i figur 15.5. Vi ser at vi får en verdi på ca. 0.6 for de tidspunktene at det opprinnelige signalet hadde en frekvens lik analysefrekvensen.

En detalj er imidlertid verd å merke seg: Kurven i figur 15.5 har avrundete hjørner. Dette skyldes at waveleten har en endelig utstrekning i tid, og derfor vil ”oppdage” en 100 Hz sekvens allerede før waveletens toppunkt er innenfor 100 Hz-området. Likeså vil waveleten ”oppdage” områder hvor det ikke er 100 Hz selv når toppunktet til waveleten såvidt ligger innenfor 100 Hz-området.

Vi kommer tilbake til denne effekten i stor detalj siden.

Merk altså at en wavelettransformasjon slik vi har gjennomført den i dette eksemplet bare gir wavelet-transformasjonen for én analysefrekvens. Dersom vi ønsker å få med også andre frekvenser, må dette gjøres ved å la prosedyren vi brukte her inngå i en løkke slik at alle ønskelige analysefrekvenser blir tatt med.



Figur 15.5: Den wavelettransformerte av tidssignalet for en analysefrekvens på 100 Hz. Parameteren K var 12, hvilket betyr at waveleten var grovt sett om lag $12 \times (1/100) s = 0.12 s$ lang. Bredden på waveleten fører til avrunding av skarpe hjørner i diagrammet.

Det er imidlertid masse multiplikasjoner og summasjoner som skal gjøres i en slik råversjon av wavelettransformasjon. Det er smartere måter å gå fram på, og det skal vi se nærmere på nå, men vi gir resten av programkoden for råmetoden først:

```
% Fortsettelse av forrige kodesnutt:
% Går nå over til å lage en Morlet wavelet for én valgt analysefrekvens
% for å gjennomføre en analyse ved rett-fram-metoden (tidkrevende)

f_analyse = 100.0; % Frekvensen vi vil analysere for
omega_a = 2.0*pi*f_analyse; % Omega_analyse
K = 12; % Parameter som angir ca antall perioder
% innen omhyllingskurven
C = 0.7980*omega_a/(f_sampl*K); % Normeringskonstant (ulike valgmuligheter)
tx = linspace(-T,T,2*N+1); % Lager én wavelet vi kan plukke ut deler av
harmSign = (cos(omega_a.*tx) + i.*sin(omega_a.*tx)); % Array for
% lagring av komplekst harmonisk signal
harmSign = C.*(harmSign - exp(-K*K)); % Trekker fra et lite korreksjonsledd
arg = tx.*omega_a/K; % Array for gaussisk omhyllingskurve
arg2 = -0.5.*arg.*arg;
morlet = exp(arg2).*harmSign; % Kombinerer for å lage Morlet wavelet

figure; % Plotter Morlet-waveleten om ønskelig
utsnitt = 240; % Plotter bare 2*utsnitt punkter av totalen
plot(tx(N-utsnitt:N+utsnitt),real(morlet(N-utsnitt:N+utsnitt)),'-b');
hold on;
plot(tx(N-utsnitt:N+utsnitt),imag(morlet(N-utsnitt:N+utsnitt)),'r');
xlim([-0.13,0.13]); % Kan begrense plottet dersom man ønsker det, men
ylim([-0.048,0.048]); % pass på at du ikke mister oversikten!
xlabel('Tid (s)');
ylabel('Relativ intensitet');
title('Morlet wavelet');
```

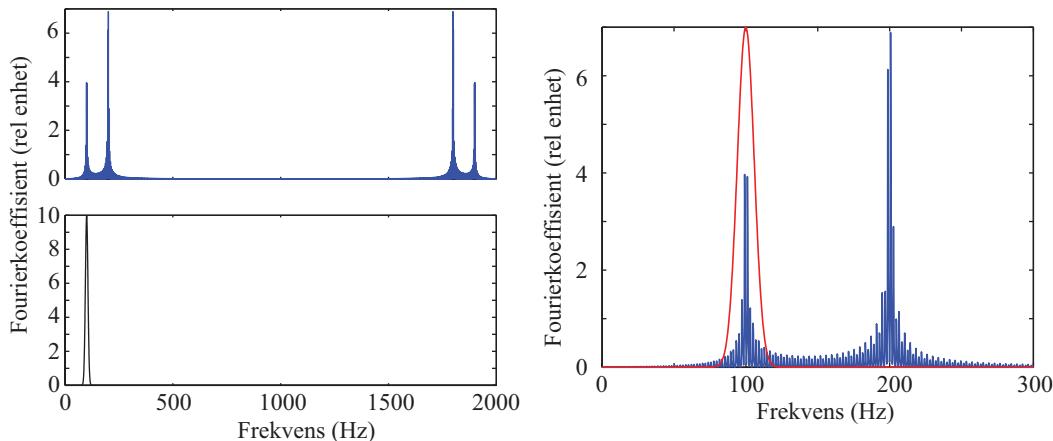
```
% Gjennomfører en waveletanalyse for den ene valgte analysefrekvensen (men
% waveleten morletT har sitt sentrum til "alle tider")
morletT = zeros(1,N);
for ii = 0:N-1
    morletT = morlet(N+1-ii:2*N-ii);
    produkt = morletT.*signal;
    integral = 0.0;
    for j = 1:N
        integral = integral + produkt(j);
    end;
    WT(ii+1) = integral;
end;

figure;           % Plotter wavelettransformasjonen for denne ene frekvensen
plot(t,abs(WT),'-g');
xlabel('Tid (s)');
ylabel('Relativ intensitet');
title('Wavelettransformen for én frekvens');
```

15.4.2 En mye mer effektiv algoritme

Da vi behandlet fouriertransformasjon i et tidligere kapittel, var vi såvidt inne på filtrering av signaler ved å maskere ut deler av et fourierspekter til et signal før vi transfonerer tilbake til en tidsbeskrivelse igjen.

I figur 15.6 viser vi øverst til venstre frekvensspekteret til signalet i figur 15.4. Nedenfor er det vist et frekvensspekter som har en gaussisk form og som er sentrert ved den ene av de to toppene i frekvensspekteret. Den gaussiske funksjonen viser ingen “folding” slik den fouriertransformerte til signalet gjør.



Figur 15.6: Den fouriertransformerte av tidssignalet vårt, samme med en gaussformet funksjon som kan brukes ved en frekvensfiltrering. Se teksten for detaljer.

Til høyre er det vist et utsnitt i frekvensskalaen, og den klokkeformede kurven er tegnet inn i samme diagram som den fouriertransformerte av signalet.

Poenget er da som følger:

Dersom vi multipliserer den fourieromvendte av signalet med den klokkeformede kurven, har vi bare tatt vare på frekvensbidrag i et relativt smalt frekvensområde sentrert rundt 100 Hz. Den delen som overlever inneholder informasjon om 100 Hz-delen av det

oppriinnelige signalet.

Dersom vi multipliserer den komplekse fourieromvendte av signalet (altså UTEN at absoluttverdien beregnes), vil vi beholde faseinformasjon i signalet. Det betyr at dersom vi foretar en omvendt fouriertransformasjon, vil 100-Hz-signalet komme ut med en tidsangivelse som oppriinnelig.

Det er på sett og vis denne type informasjon vi ønsker å få ut ved waveletanalyse! Men hvordan skal vi lage den gaussformede funksjonen? Her er vi heldige:

Den fouriertransformerte til en Morlet wavelet (ligning (15.8)) kan gis på følgende måte:

$$\hat{\Psi}_{\omega_a, K}(\omega) = 2\{\exp(-[K(\omega - \omega_a)/\omega_a]^2) - \exp(-K^2)\exp(-[K\omega/\omega_a]^2)\} \quad (15.10)$$

Vi ser at dette nettopp er en klokkeformet (gaussisk) funksjon (bortsett fra et ganske ubetydelig korreksjonsledd for de fleste valg av K). Toppunktet i klokkefunksjonen finnes ved analysefrekvensen.

Dette er et rent reelt uttrykk og er egentlig bare absoluttverdien av den fouriertransformerte av waveleten. Absoluttverdien av den fouriertransformerte av waveleten er den samme uansett hvilken tid toppunktet til Morlet waveleten er plassert.

Merk at den fouriertransformerte av en Morlet-wavelet bare har én topp i frekvensspekteret (ingen speiling!).

Det kan nå vises at wavelettransformasjon gjennomført med rå-metoden beskrevet ovenfor kan erstattes med følgende prosedyre:

- Beregn den fouriertransformerte av tidssignalet vi skal analysere.
- Beregn direkte den fouriertransformerte til en Morlet wavelet med den analysefrekvensen den vi er interessert i.
- Multipliser disse med hverandre, punkt for punkt.
- Foreta en invers fouriertransformasjon.
- Absoluttverdien av denne vil da gi informasjon om hvilke tidspunkt det oppriinnelige signalet inneholdt frekvenser lik analysefrekvensen.

Den fouriertransformerte X_k av tidssignalet beregnes ut fra ligning (15.3). Denne transformasjonen behøver vi bare gjøre én gang.

Den fouriertransformerte til en Morlet wavelet med analysefrekvensen f_a kan skrives slik:

$$\hat{\Psi}_{f_a, K}(f_k) = 2\{\exp(-[K(f_k - f_a)/f_a]^2) - \exp(-K^2)\exp(-[Kf_k/f_a]^2)\} \quad (15.11)$$

Merk at det må være overensstemmelse mellom frekvensene f_k i den fouriertransformerte av signalet og den fouriertransformerte av waveleten. Ved å bruke den angitte prosedyren kan vi bygge opp en hel horisontal linje (alle tidspunkt) i én jafs. Ved å endre Morlet waveleten slik at den svarer til neste analysefrekvens, får vi bygget opp stadig nye horisontale linjer i waveletdiagrammet inntil vi har dekket så mange analysefrekvenser vi ønsker. Siden Fast Fourier Transform er så effektiv som den er, blir metoden vi nettopp har skissert mye raskere å gjennomføre enn rå-metoden vi omtalte ovenfor.

Programkode for å vise hvordan dette kan implementeres er gitt i det følgende. Koden bygger på de to forgående programsnuttene er kjørt forut for denne.

```
% Foretar til slutt en effektiv wavelettransformasjon basert på FFT/IFFT
% Beregner igjen FFT av tidsstrenge (holder å gjøre det én gang!)
FTsignal = fft(signal);
f = linspace(0,f_sampl*(N-1)/N, N);

% % Plotter frekvensspekteret dersom det ønskes
% figure; % Plotter FFT (absoluttverdier only)
% nmax = floor(N/2); % Plotter iblant bare litt av FFTen
% plot(f(1:nmax),abs(FTsignal(1:nmax)));
% xlabel('Frekvens (Hz)');
% ylabel('Relativ intensitet');
% title('Vanlig frekvenspektrum');

% Genererer (den fouriertransformerte av en wavelet) direkte, og plotter
% denne (de neste 11 linjene er bare tatt med for å kunne plotte wavelets)
fktr = (K/f_analyse)*(K/f_analyse);
FTwl = exp(-fktr*(f-f_analyse).*(f-f_analyse));
FTwl = FTwl - exp(-K*K)*exp(-fktr*(f.*f)); % Lite korreksjonsledd
FTwl = 2.0*FTwl; % Faktor
figure;
plot(f,FTwl,'-k');
xlim([-0.0,300]); % Kan begrense plottet dersom man ønsker det, men
% ylim([-0.2,2.2]); % pass på at du ikke mister oversikten!
xlabel('Frekvens (Hz)');
ylabel('Relativ intensitet');
title('Den fouriertransformerte av waveleten');

% NÅ setter vi i gang med den virkelige effektive waveletanalysen!
M = 100; % Antall frekvenser som skal inngå i analysen
fstart = 70.0;
fslutt = 300.0;
ftrinn = (fslutt/fstart)^(1/M);
f_analyse = fstart;
fbrukt = zeros(1,M);

% Loop over alle frekvenser som inngår i analysen
for jj = 1:M
    fktr = (K/f_analyse)*(K/f_analyse);
    FTwl = exp(-fktr*(f-f_analyse).*(f-f_analyse));
    FTwl = FTwl - exp(-K*K)*exp(-fktr*(f.*f)); % Lite korreksjonsledd
    FTwl = 2.0*FTwl; % Faktor
    % Beregner så en hel linje i scalogrammet i én jafs!
    WLdiagram(jj,:) = abs(ifft(FTwl.*FTsignal));
    %scalogram(jj,:) = sqrt(abs(ifft(FTwl.*FTsignal)));
    % Bruker denne siste varianten for å få svake partier bedre synlig
    fbrukt(jj) = f_analyse; % Lagrer frekvensene som faktisk er brukt
    f_analyse = f_analyse*ftrinn; % Beregner neste frekvens
end;

% Klargjør for for å kunne vise COI (cone of interest) i plottet
maxverdi = max(WLdiagram);
mxv = max(maxverdi);
% mxv % Max verdi i WLdiagram. Brukes for å velge ok farge for markering
for jj = 1:M
    m = floor(K*f_sampl/(pi*fbrukt(jj)));
    WLdiagram(jj,m) = mxv/2;
    WLdiagram(jj,N-m) = mxv/2;
end;
```

```
% Plotter waveletdiagrammet
figure;
imagesc(t,log10(fbrukt),WLdiagram,'YData',[1 size(WLdiagram,1)]);
set(gca,'YDir','normal');
xlabel('Tid (sek)');
ylabel('Log10(frekvens i Hz)');
title('Wavelet Power Spektrum');
%title('Sqrt(Wavelet Power Spektrum)'); % Når sqrt av scalogram blir brukt
colorbar('location','southoutside');

input('Lukk alt');
close all;
```

15.5 Viktige detaljer

15.5.1 Faseinformasjon og skalering av utslaget

I vanlig fouriertransformasjon foretar vi i prinsippet to transformasjoner samtidig, nemlig en av typen

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \cos(\omega t) dt \quad (15.12)$$

og en av typen

$$X(\omega) = -i \int_{-\infty}^{\infty} x(t) \sin(\omega t) dt \quad (15.13)$$

Grunnen er at vi har både sinus og cosinusledd er at vi må kunne fange opp f.eks. et sinussignal i x uansett hvilken fase det har.

I vanlig fouriertransformasjon har vi *ett* startpunkt for analysen. Det betyr at det er lett å finne hvilken relativ fase de ulike frekvenskomponentene har.

I kontinuerlig waveletanalyse har vi ulike startpunkt og lengder på analysevinduet underveis i beregningene. Det gjør det langt vanskeligere å holde orden på faser. Dette er nok en av grunnene til at vi nesten utelukkende bare tar utgangspunkt i absoluttverdien av wavelettransformasjonen i en eller annen variant når waveletanalysen skal angis. (Dersom vi imidlertid skulle foreta en invers wavelettransformasjon etterpå, måtte vi selvfølgelig tatt være på faseinformasjonen.)

Det er flere måter vi kan angi styrken i et waveletdiagram. Ofte brukes *kvadratet* av absoluttverdiene, hvilket gir *energien* i signalet.

Erfaringsmessig liker jeg ikke å bruke kvadratet av absoluttverdien, fordi forskjellen mellom de sterke og svake partiene da ofte blir så stor at vi mister informasjon om de svake partiene. Da er det ofte bedre å heller bruke absoluttverdien direkte ("amplitudenivå").

Jeg foretrekker imidlertid ofte å bruke *kvadratroten* av absoluttverdien. Da får jeg fram de svake partiene enda bedre enn om absoluttverdien ble plottet.

Vi står fritt i å velge hvordan resultatet fra wavelettransformasjonen blir plottet, men må ta følgen av vårt valg når vi skal hente kvantitative verdier ut av diagrammene.

15.5.2 Frekvensoppløsning vs tidsoppløsning

Vi skjønte ut fra figur 15.3 at når bølgelengden i signalet x er eksakt lik bølgelengden inne i waveleten, vil wavelettransformasjonen gi maksimal verdi. Endrer vi litt på bølgelengden i waveleten ved å endre på analysefrekvensen, vil transformasjonen gi en lavere verdi, men likevel ikke null verdi. En waveletanalyse vil med andre ord ikke bare gi utslag for en frekvens som svarer til signalets, men også for nærliggende frekvenser.

Det er viktig å vite hvor langt ut denne “smitteeffekten” går, og det er tema for dette underkapitlet.

La oss da ta utgangspunkt i at en wavelettransformasjon innebærer en “digital filtrering” av et signal, slik vi anskueliggjorde i figur 15.6. Hvor skarp filteringen er bestemmes av bredden på den gaussiske klokkefunksjonen som brukes i filtreringen. Vi trenger da å finne en sammenheng mellom bredden i frekvensbildet og bredden av waveleten i tidsbildet.

I figur 15.7 viser til venstre tre ulike valg av waveleter (beregnet ut fra ligning (15.8)), og til høyre er den fouriertransformerte av waveleten (beregnet ut fra ligning (15.10)).

Vi vet fra tidligere at frekvensspekteret fra fouriertransformasjon av et sinussignal konvolutert med en gaussisk omhyllingskurve, selv har en gaussisk omhyllingskurve. Det får vi på ny bekreftet gjennom figur 15.7.

Bredden for tidsutstrekningen til waveleten kan bestemmes ved å ta utgangspunkt i omhyllingskurven (ut fra ligning (15.8)). Dersom vi definerer bredden som tidsforskjellen mellom toppunktet og et punkt hvor amplituden på omhyllingskurven har sunket til $1/e$ av maksimalverdien, er (halv)bredden:

$$\Delta t_{1/e} = 2K/\omega_a$$

Den tilsvarende bredden i den fouriertransformerte av waveleten er meget nær (ut fra ligning (15.10))

$$\Delta f_{1/e} = f_a/K = \omega_a/(2\pi K) \quad (15.14)$$

Det interessante er at

$$\Delta t_{1/e} \Delta f_{1/e} = (2K/\omega_a) \cdot (\omega_a/(2\pi K)) = 1/\pi$$

Dersom vi beregner “standardavviket” for tid og frekvens ut fra mer statistiske mål, slik:

$$\sigma_t^2 = \frac{\int t^2 \Psi^2(t) dt}{\int \Psi^2(t) dt}$$

og

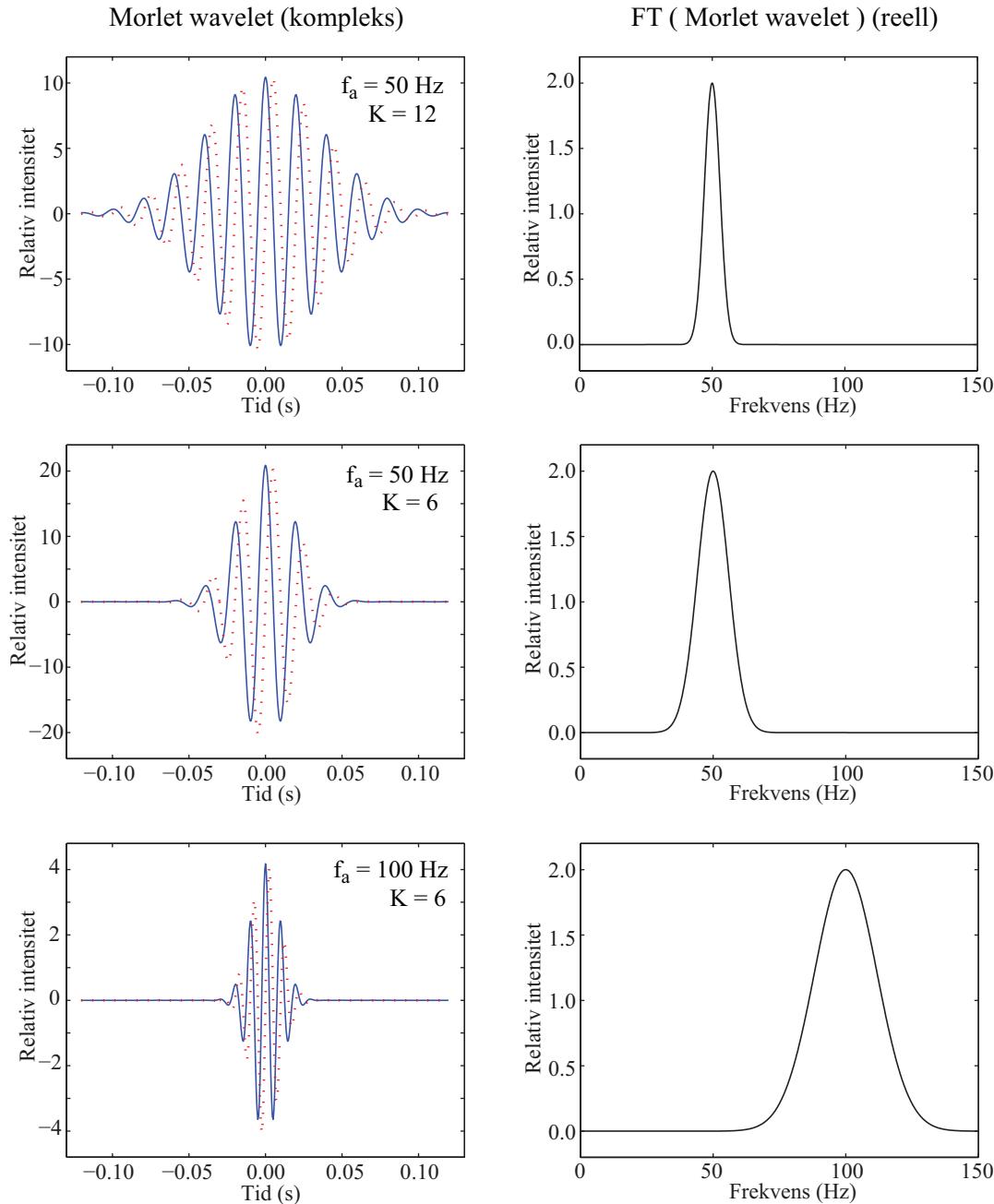
$$\sigma_f^2 = \frac{\int f^2 \hat{\Psi}^2(f) df}{\int \hat{\Psi}^2(f) df}$$

kan det vises at

$$\sigma_t^2 \sigma_f^2 = \frac{1}{2\pi} \quad (15.15)$$

Denne relasjonen er analog til Heisenbergs uskarphetsfunksjon. Eksempler i tråd med denne relasjonen er vist i figur 15.7.

Relasjonen er svært viktig for waveletanalyse. Dersom vi lar en wavelet strekke seg over en lang tid, vil bredden i frekvensdomenet være liten og visa versa. Med andre ord: *Vi kan ikke både få en nøyaktig tidsangivelse av detaljer i et signal samtidig som vi får en nøyaktig frekvensangivelse.*



Figur 15.7: Tre ulike wavelets som indikerer hvordan parametrene analysefrekvensen ω_a og “bølgetallet” K virker inn på waveleten. En wavelet har en avgrenset utstrekning i tid (venstre del). Vi kan angi en bredde på omhyllingskurven f.eks. ut fra at verdien har sunket til $1/e$ av toppverdien.. Fourieromvendes denne waveleten, får vi frekvensresponsene vist til høyre. Legg merke til både posisjon i frekvensspekteret og bredden på de gaussformede kurvene. Bredden på frekvensresponsen kan angis på lignende måte som i tidsbildet. Det viktige er at breddene i tidsdomenet multiplisert med breden i frekvensdomenet er en konstant, hvilket innebærer at dersom den ene økes, vil den andre avta og visa versa.

En interessant følge av ligning (15.14) er at

$$\Delta f_{1/e}/f_a = 1/K$$

Med andre ord, i en waveletanalyse holder vi oftest K konstant i hele analysen. Da er den relative usikkerheten i frekvensangivelsene konstant i hele diagrammet.

For å utnytte dette, er det normalt å velge en logaritmisk frekvensakse, i betydning at frekvensene vi velger å ta med i analysen forholder seg til hverandre som

$$(f_a)_{k+1} = (f_a)_k \cdot f_{\text{faktor}}$$

Vi har valgt logaritmisk akse for de valgte analysefrekvensene i alle eksemplene i dette kapitlet, men det er selvfølgelig mulig å velge analysefrekvensene også ut fra en lineær skala, i alle fall dersom forskjellen mellom minste og største analysefrekvens er liten (f.eks. en faktor to eller mindre).

Sammenligning wavelets vs stykkevis FT

Skulle vi bruke stykkevis FT, ville et fast tidsintervall medføre at vi har svært få (eller ingen hele) periodetider innenfor intervallet for lave frekvenser, men ganske mange periodetider for høye frekvenser. Det betyr at vi ville få en elendig frekvensoppløsning for de laveste frekvensene (målt som relativ frekvens), men en langt bedre frekvensoppløsning for de høyere frekvensene. Det betyr at vi ville ende opp med en analyse som ikke ville være optimal.

Prosedyren som brukes i waveletanalyse gir en optimal tidsoppløsning for *alle* frekvenser. Men vi *kan* likevel velge å vektlegge tidsoppløsning *noe* på bekostning av frekvensoppløsning og omvendt alt etter hva vi ønsker å studere. Det gjør at metoden blir et meget slagkraftig hjelpemiddel i mange sammenhenger.

Kunne vi valgt en stykkevis FT der vi faktisk brukte skaleringsprinsippet med å stykke opp tidsstrengen i mindre biter når vi analyserte høye frekvenser enn ved lave? Det ville kreve en enorm mengde fouriertransformasjoner, men ville da gi omtrent samme resultat som en waveletanalyse. Resultatet ville likevel ikke bli like godt. Vi ville nemlig med skalerte FT-intervaller i prinsippet hatt en ren waveletanalyse, men med en wavlet som har en annen form enn Morlet. Det ville være en firkantpuls som da var omhyllingskurven. Frekvensresponsen ville da ha en form

$$\left(\frac{\sin(x)}{x} \right)^2$$

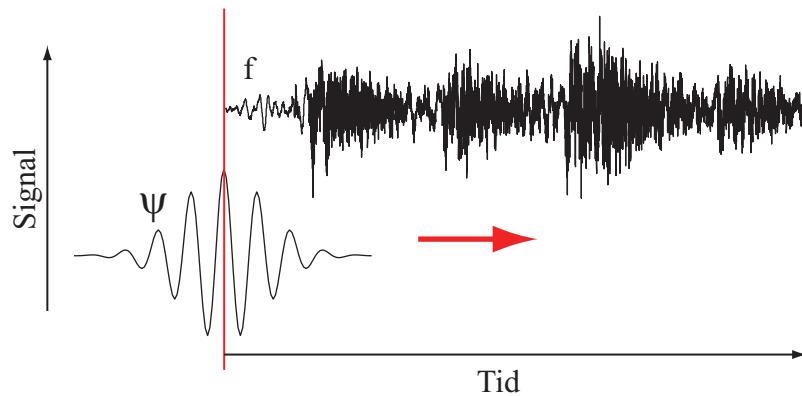
i stedet for en gaussisk omhyllingskurve. Resultatet ville bli en hale utenfor den sentrale toppen som gjør at vi får en dårligere frekvensbestemmelse enn ved Morlet wavelets.

I denne sammenheng kan vi trekke analogier til diffraksjon. Dersom vi sender lys inn mot en smal spalt og har *samme* intensiteten over hele spalten, vil diffraksjonsintensiteten nettopp være gitt som $(\frac{\sin(\theta)}{\theta})^2$. Dersom vi derimot sendte en lysstråle med *gaussisk* intensitetsprofil gjennom spalten, ville vi fått en gaussisk intensitetsprofil på skjermen uten ekstra stripers ved siden av den sentrale toppen.

15.5.3 Randproblem

Når vi foretar en wavelettransformasjon, multipliserer vi i prinsippet et signal med en wavelet og summerer alle produktene. Vi flytter så waveleten og gjør det samme på ny. Dette gjentas om igjen og om igjen fra den situasjonen at waveletens midtpunkt ligger helt i den ene enden av signalet til waveletens midtpunkt ligger i andre enden av signalet.

Vi endrer så analysefrekvensen for waveleten og gjør det samme på ny.

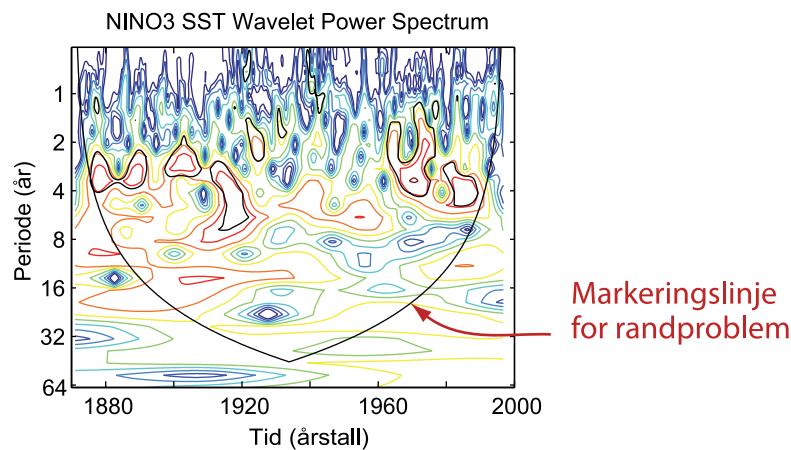


Figur 15.8: Det er ikke mulig å få et korrekt waveletresultat for tider og frekvenser der ikke hele waveleten kommer innenfor dataområdet under beregningene.

Her oppstår det imidlertid et problem. Så lenge wavleten ikke er fullstendig innenfor dataområdet, vil vi måtte forvente et annet resultat enn om hele waveleten ble brukt i beregningene. Dette er anskueliggjort i figur 15.8. For den posisjonen waveleten har i forhold til dataene i denne figuren, vil bare om lag halvparten av waveleten benyttes i praksis. Det betyr at summen av produktene må forventes å bli langt lavere (i størrelsesorden halvparten) av hva summen ville blitt dersom vi hadde fullt overlapp.

Av den grunn kan vi ikke stole noe særlig på datapunktene i en endelig wavelettransformasjon dersom waveleten som er brukt ikke ligger fullstendig innenfor datastrengen vi analyserer. Det er derfor vanlig å markere ytterområdet mhp tid hvor vi har et randproblem.

I figur 15.9 er det vist et eksempel på et waveletdiagram etter analyse av temperaturoscillasjoner i det sydlige Stillehav. Det er brukt kvoter sammen med farger for å markere “energien” i ulike former for svingninger (periodisitet) etter som de har utviklet seg de siste vel hundre år.



Figur 15.9: Eksempel på et waveletdiagram for temperaturoscillasjoner i det sydlige stillehav. Figuren er produsert med data og programvare tilgjengelig fra <http://paos.colorado.edu/research/wavelets/> tilgjengelig april 2013.

I dette diagrammet er det tegnet inn en krum V-formet linje som starter langt nede og midt i waveletdiagremmet med symmetriske buede linjer som går opp og ut mot kanten. Disse linjene markerer området hvor det aller meste av waveletene er fullstendig innenfor datastrengen: Alt over denne buedelen er gyldige data. Alt under streken er data vi må

ta med en klype salt.

I programeksemplene gitt i dette kapitlet har vi valgt å legge inn en markering som svarer til at bare den ytre delen av waveleten som har verdi mindre enn $1/e$ av maksimum på omhyllingskurven ligger utenfor diagrammet. Vi har bare med et så lite frekvensområde at vi i egne eksempler ikke får fram hele den buete V-en, men bare et smalt horisontalt bånd av den totale V-en. Alle deler av waveletdiagrammet som ligger mellom disse markeringene har kun ubetydelige feil på grunn av randproblematikken. Vi gir detaljer nedenfor om hvordan markeringene settes opp.

15.6 Optimalisering

Waveletanalyse er mer krevende enn vanlig fouriertransformasjon. Vi må velge hva slags wavelet vi vil bruke. Selv om vi holder oss til Morlet wavelets, må vi bestemme oss for hvilket “bølgetall” vi vil bruke.

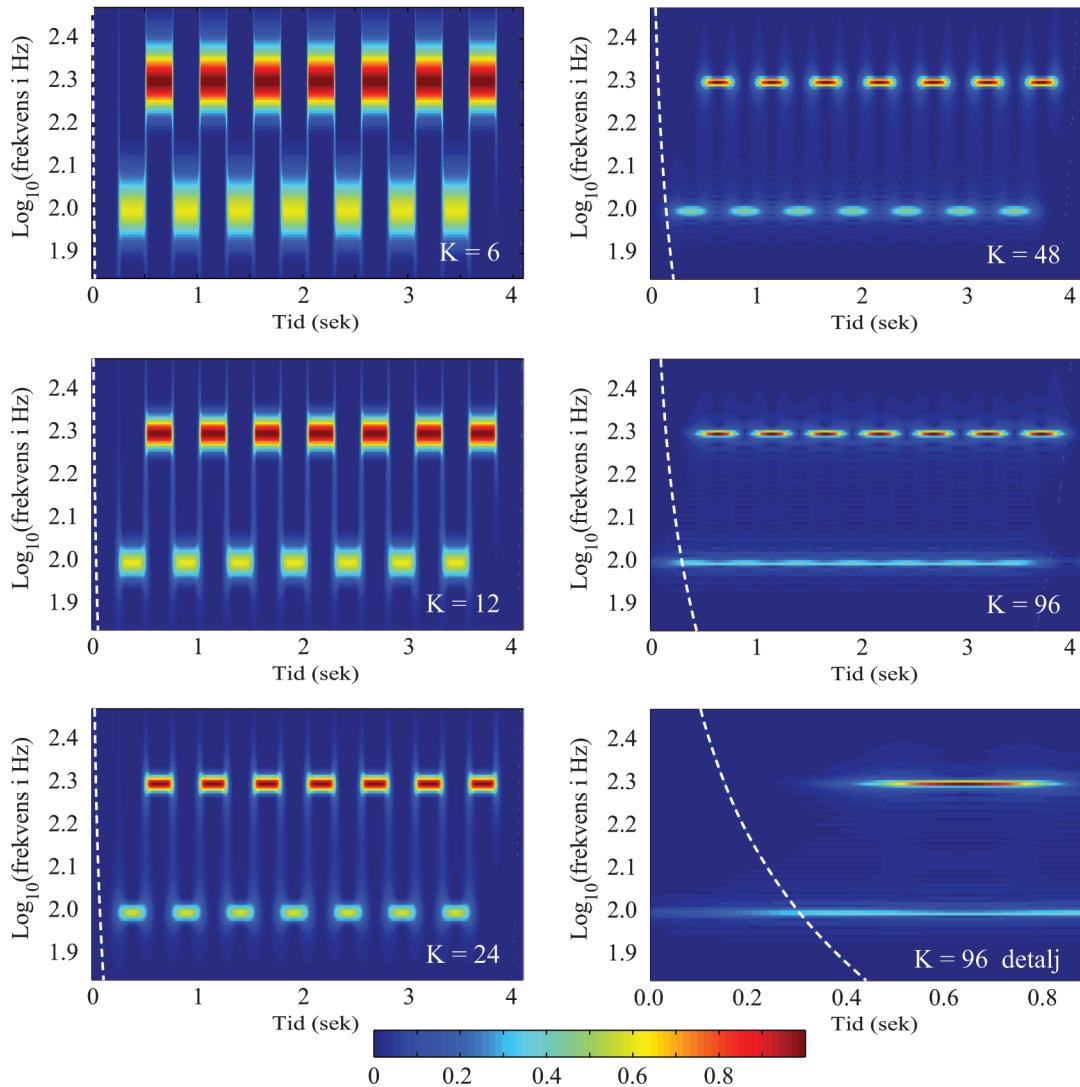
Vi har tidligere sett at ved å øke bølgetallet K , vil waveleten ha en betydelig verdi over et større tidsrom enn ved lavt bølgetall (ved samme analysefrekvens). Videre har vi sett at når “bredden” på waveleten i tidsdomenet er stor (det vil si stor K -verdi), vil “bredden” på den fouriertransformerte av waveleten bli liten. Produktet av bredden på waveleten i tidsdomenet og bredden av waveleten i frekvensdomenet er jo konstant.

Følgen er at det finnes ingen Ole Brumm-løsninger: “Ja takk, begge deler” innen waveletanalyse. Ønsker vi å få en nøyaktig angivelse av tidsforløp, vil små K -verdier være å foretrekke. Ønsker vi å få så nøyaktige frekvensangivelser som mulig, bør K -veriden være høy. Vi ønsker i prinsippet så god tidsoppløsning og frekvensoppløsning som mulig, men må alltid velge et kompromiss.

Det optimale resultatet oppnås ofte dersom vi tar utgangspunkt i signalet selv. Signalet har ofte innebygget en uskarphet i tid og/eller en uskarphet i frekvens. Vi kan aldri få en bedre oppløsning i tid ved waveletanalyse enn den oppløsningen signalet selv har. Tilsvarende for analyse av frekvens.

I figur 15.10 er det vist waveletdiagrammer av signalet vi drøftet ovenfor som vekslet mellom 100 og 200 Hz. Fem ulike bølgetall K er brukt. Vi ser at for lave K -verdier er tidsoppløsningen meget presis, men frekvensbestemmelsen er elendig. For høye K -verdier er det motsatt: Frekvensoppløsningen er god, men tidsoppløsningen er elendig.

I dette tilfellet er det egentlig ikke mye mer å hente i frekvensoppløsning når vi går fra $K = 48$ til $K = 96$. Det betyr at frekvensoppløsningen i signalet selv (siden varigheten av hver periode med “konstant frekvens”) svarer omtrent til oppløsningen vi oppnår med en K -verdi litt større enn 48. (Strengt tatt er det forskjellig frekvensoppløsning på 100 Hz- signalene og 200 Hz signalene siden det er ulikt antall perioder innen hver av disse periodene med konstant frekvens.)



Figur 15.10: Waveletdiagrammer for tidssignalet i figur 15.4 for seks ulike “bølgetall” K . Se teksten for detaljer.

I figur 15.10 har vi gjort markeringen av venstre side av randproblemsområdet ekstra tydelig. Vi ser da klart at randproblemet øker med økende K -verdi. Det kan være interessant å merke seg at randproblemmarkeringen endrer seg med analysefrekvensen. Videre er det nyttig å vite at avstanden fra sidekanten til randproblemmarkeringen også indikerer utsmøring i klare tidsangivelser i waveletdiagrammet. All tidsinformasjon i analysen blir smurt ut om lag med en tidsforskjell som nettopp svarer til avstanden fra randen til randproblemmarkeringen.

Hva er da “beste valg” av alle analysene gitt i figur 15.10? Vel, det kommer an på hva vi ønsker å få fram av opplysninger. Diagrammet for $K = 6$ demonstrerer at endringen fra 100 til 200 Hz (og omvendt) foregår meget skarpt i tid. Diagrammet for $K = 96$ viser at frekvensen er så ensartet som den kan være innenfor hver av tidsintervallene. Skulle vi velge en slags generell optimalisering, kunne kanskje $K = 48$ eller deromkring være et godt valg.

Optimalisering i frekvensoppløsning (programmeringsteknisk)

En annen form for optimalisering ligger i valg av frekvensområde for analysen. I en digital fast fourier analyse får vi automatisk “alle” frekvenser mellom null og samplingsfrekvensen (men bare halvparten er nyttig pga folding). For en kontinuerlig waveletanalyse velger vi oftest å innskrenke frekvensområdet til det området der frekvensinnholdet er av interesse.

I figur 15.10 valgte vi bare å ta med frekvenser mellom 70 og 300 Hz i analysen. Grunnen er at vi visste at signalet bare inneholdt frekvenser nær opp til 100 og 200 Hz. Det kan ofte være en fordel å starte med en vanlig fouriertransformasjon for å sikre oss at vi velger et frekvensområde som egner seg.

Det er imidlertid viktig også å tenke på hvor mange mellomliggende frekvenser vi skal ta med i analysen. I denne sammenheng må vi gå tilbake til “bredden” av wavleten i frekvensdomenet. Denne bredden var som vi tidligere har sett:

$$\Delta f = f_a/K$$

Denne “bredden” var bestemt ved at den gaussformede frekvenskurven var kommet ned til $1/e$ av maksimum verdi. Vi ønsker ikke å gå så store steg i frekvens fra en analysefrekvens til den neste, men kanskje bare en brøkdel av dette.

Praktisk testing viser at et optimalt valg av forskjellen mellom en analysefrekvens og den neste er da om lag

$$f_{a,neste} - f_{a,naa} = f_{a,naa}/8K \quad (15.16)$$

Dersom vi skal spenne over et frekvensintervall $[f_{start}, f_{slutt}]$ kan det da enkelt vises at vi bør bruke M analysefrekvenser i en logaritmisk orden, hvor

$$f_{slutt} = \left(1 + \frac{1}{8K}\right)^{M-1} \cdot f_{start}$$

Antall analysefrekvenser er da:

$$M = 1 + \log(f_{slutt}/f_{start}) / \log\left(1 + \frac{1}{8K}\right) \quad (15.17)$$

Optimalisering i tidsoppløsning (programmeringsteknisk)

Et kontinuerlig waveletdiagram kan iblant bestå av svært mange punkter. Dersom vi f.eks. tar utgangspunkt i lyd som er digitalisert med en samplingsfrekvens på 44.1 kHz, og vi studerer lyd med frekvens i området 100 - 10 000 Hz, kan vi i praksis plukke ut bare hvert fjerde punkt i tidsdimensjonen uten problemer. Når vi atpå til vet at wavleten har en bredde på om lag K ganger periodetiden for analysefrekvensen, innser vi at vi kan fjerne enda flere punkter i tidsdimensjonen uten at det vil oppdages i et waveletdiagram.

Det kan iblant være av interesse å optimalisere et waveletdiagram mhp tidsangivelsen. Ikke minst er dette viktig for å få plottefiler som er små små at de lett lar seg innlemme i rapporter og liknende.

I praksis viser det seg at vi kan nøye oss med å gjengi hvert P -te punkt i tidsdimensjonen i et waveletdiagram uten at nevneverdig informasjon går tapt når P er gitt ved:

$$P = \text{HeltallsverdienAv} \left(\frac{K}{24} \frac{f_s}{f_{a,max}} \right) \quad (15.18)$$

15.7 Et realistisk eksempel

Figur 15.11 viser et eksempel på en optimalisert waveletanalyse. Signalet er en lydfil i CD-kvalitet som gir lyden av en gjøk som synger sitt “ko ko”. Signalet er gitt i tre varianter, nemlig som et rent tidssignal, som frekvensspekter etter en vanlig FT, og endelig analysert ved å bruke wavelettransformasjon.

I den totale arbeidsgangen er første trinn å velge ut et passe utsnitt fra lydfilen. Det gjøres ved at vi velger startpunkt og totalt antall punkter som skal hentes ut fra den tilgjengelige datafilen. Dernest foretas en fourieranalyse. Fra fourierspekteret ser vi at lyden stort sett bare inneholder frekvenser mellom 450 og 750 Hz. Av den grunn er waveletanalysen begrenset til dette frekvensintervallet.

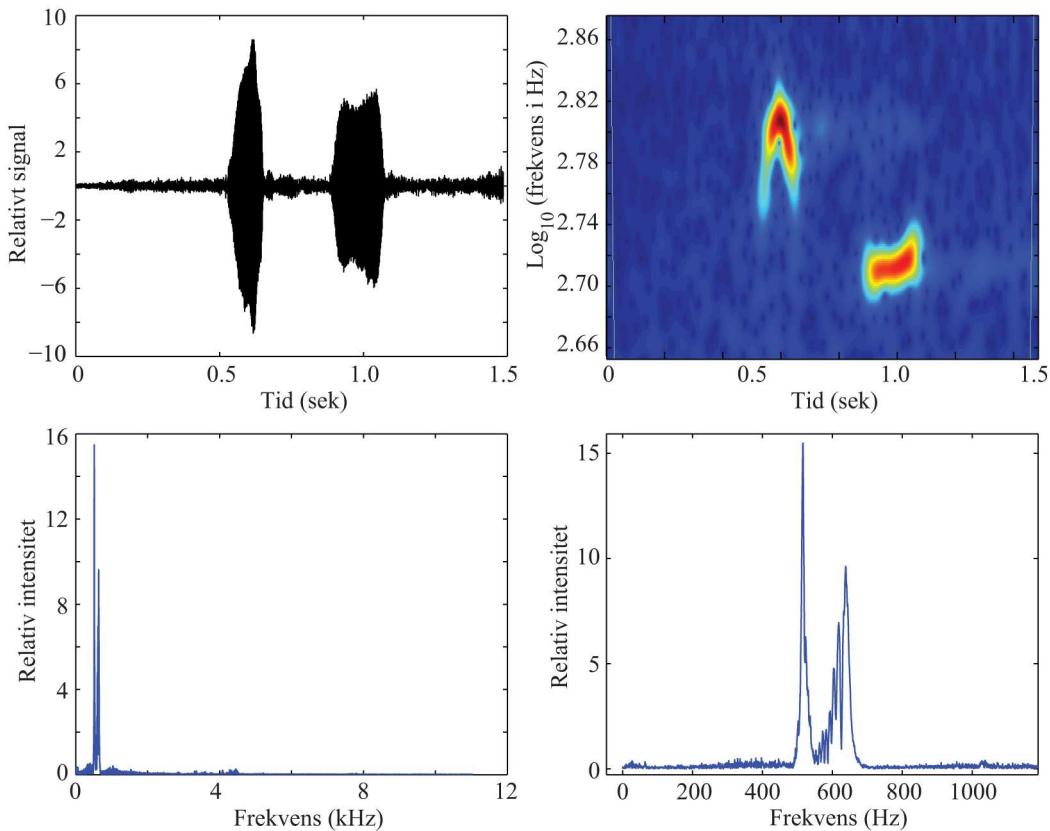
Til slutt må vi forsøke med ulike K -verdier og velge et “beste” kompromiss mellom god tidsangivelse og frekvensangivelse samtidig. Vi må da bestemme oss for hvorvidt vi vil prioritere tidsoppløsning (ved å ha en liten K), men da på bekostning av en nokså bred frekvensrespons, eller godta en litt dårligere tidsoppløsning (ved å velge en større K) for å få en noe bedre frekvensoppløsning. Hva som er optimalt kommer an på selve signalet vi analyserer og vil avhenge av hva den enkelte vektlegger mest. I vårt eksempel ble $K = 40.0$ benyttet.

Legg merke til de nydelige detaljene som kommer fram i waveletanalysen. Har du f.eks. vært klar over at lyden i det første “ko”-et faktisk endrer seg betydelig den korte stunden lyden varer? Detaljer i waveletanalyse av fuglesang gjør det mulig for ornitologer å gjenkjenne fugler individuelt. Detaljene er finere enn hva en menneskehørsel klarer å oppfatte.

Det bør være innlysende at wavelettransformasjon gir langt mer interessante data enn en vanlig fouriertransformasjon for en slik type lyd.

Vi har tatt med et Matlab program som kan brukes for å analysere biter av lydfiler. Programmet kan kopieres direkte fra pdf-filen slik at du slipper å skrive den inn på nytt.

[♣ ⇒ Noen kommentarer:



Figur 15.11: Gjøkens “ko ko” analysert i tidsbildet, i frekvensbildet (med et utsnitt) og i kombinasjonen: Frekvens og tid i form av et waveletdiagram.

Programmet er selvsnekret og bærer preg av at når jeg selv gjør programmering, er det for å få gjort en jobb. De av dere som er mer finesmackere i programmering får heller lage deres eget program som er mer i tråd med moderne trender og krav. Selv startet jeg med hullkort og hadde spesialtillatelse til å bruke hele 150 kB av RAM av totalt 250 kB (!) i Norges største datamaskin på den tiden (Regnemaskinen Blindern/Kjeller) til x antall millioner kroner. Jeg er derfor fornøyd når et program fungerer, selv om koden ikke er så elegant som den kunne vært! ↪ ♠]

```
function WaveletTransform7

% Program som foretar en waveletanalyse av et lydsignal som foreligger som
% en wav-fil. Laget for FYS2130 våren 2013 av Arnt Inge Vistnes.
% Denne versjonen er fra 18042013.

% Navn på noen lydfiler (her angitt sammen med egnede parametre for analyse)
c = 'Svartrost2.wav'; % Nstart 17000, 64 k, 1500-8000 Hz, K12-96
% c = 'bokfink.wav'; % Nstart 34 000, 128 k, K 48, 2000-8000 Hz (minst)
% c = 'gjok.wav'; % Nstart 12000, 450-750 Hz + noe støy ved > 4000 Hz

% PARAMETRE VI MÅ SETTE: (Parametre for waveletanalysen settes nedenfor!)
N = 1024*64; % Lengde på data-utsnitt (lydfil) (helst 2^n) Her: "64k"
nstart = 17000; % Startpunkt for utsnitt fra lydfil

% Leser passe utsnitt av lydfil, spiller den av og plotter tidsbildet
nslutt = nstart+N-1;
[y, fs] = audioread(c, [nstart nslutt]); % Les array y(N,2) fra fil
% 'fs' er vanligvis 44100 (samplingsfrekvens ved CD kvalitet)
h = zeros(N,1); % Plukker ut bare én kanal fra stereosignalet lest
```

```

h = y(:,1);
sound(h,fs);           % Spiller av utsnittet som er brukt
T = N/fs;              % Total tid lydutsnittet tar (i sek)
t = linspace(0,T*(N-1)/N,N);
plot(t,h,'-k');
title('Opprinnelig tids-funksjon');
xlabel('Tid (sek)');
ylabel('Signal (rel enhet)');

%*****



% Foretar så en effektiv wavelettransformasjon basert på FFT/IFFT. På dette
% punktet må totalt antall punkter N og samplingsfrekvens fs være kjent.

% Beregner først FFT av tidsstrekken h (gjøres bare én gang!)
FTsignal = fft(h);

% Plotter frekvensspekteret (absoluttverdier only)
f = linspace(0,fs*(N-1)/N, N);
nmax = floor(N/2);          % Plotter bare opp til halve samplingsfrekvens.
figure;
plot(f(1:nmax),abs(FTsignal(1:nmax)));
xlabel('Frekvens (Hz)');
ylabel('Relativ intensitet');
title('Vanlig frekvensspektrum');

% Dernest selve waveletanalysen

% INPUT PARAMETRE vi selv må velge følger her:
K = 12;                  % Morlet-wavelet-bredde (kan være 6 - 400)
fmin = 1500.0;            % Minimum frekvens i waveletanalysen (i Hz)
fmax = 8000.0;            % Maximum frekvens

% Beregner # analysefrekvenser, skriver til skjerm, klargjør frekvensene
M = floor(log(fmax/fmin) / log(1+(1/(8*K)))) + 1;
AntallFrekvenserIAnalyse = M
ftrinn = (fmax/fmin)^(1/(M-1));
f_analyse = fmin;

% Allokerer plass til waveletdiagrammet og array for lagring av frekvenser
WLdiagram = zeros(M,N);
fbrukt = zeros(1,M);

% Løkke over alle frekvenser som inngår i analysen
for jj = 1:M
    faktor = (K/f_analyse)*(K/f_analyse);
    FTwl = exp(-faktor*(f-f_analyse).*(f-f_analyse));
    FTwl = FTwl - exp(-K*K)*exp(-faktor*(f.*f));    % Lite korreksjonsledd
    FTwl = 2.0*FTwl;                                  % Faktor (ulike valg!)
    % Beregner så en hel linje i waveletdiagrammet i én jaf!
    %WLdiagram(jj,:) = abs(ifft(FTwl.*transpose(FTsignal)));% Ett alternativ
    WLdiagram(jj,:) = sqrt(abs(ifft(FTwl.*transpose(FTsignal)))); % Ett annet
    % Bruker den siste varianten for å få svake partier bedre synlig
    fbrukt(jj) = f_analyse;                % Lagrer frekvensene som faktisk er brukt
    f_analyse = f_analyse*ftrinn;        % Beregner neste frekvens
end;

% Reduserer filstørrelse ved å fjerne mye av overflødig informasjon i tid.
% Dette gjøres kun for at filstørrelsen på plottene skal bli håndterbar.
P = floor((K*fs)/(24 * fmax)); % Tallet 24 kan endres ved behov

```

```

TarBareMedHvertXITid = P
NP = floor(N/P);
AntallPktITid = NP

for jj = 1:M
    for ii = 1:NP
        WLdiagram2(jj,ii) = WLdiagram(jj,ii*P);
        tP(ii) = t(ii*P);
    end;
end;

% Foreta en markering i plottet for å vise områder med randproblemer
maxverdi = max(WLdiagram2);
mxv = max(maxverdi);
for jj = 1:M
    m = floor(K*fs/(P*pi*fbrukt(jj)));
    WLdiagram2(jj,m) = mxv/2;
    WLdiagram2(jj,NP-m) = mxv/2;
end;

% Plotter waveletdiagrammet
figure;
imagesc(tP,log10(fbrukt),WLdiagram2,'YData',[1 size(WLdiagram2,1)]);
set(gca,'YDir','normal');
xlabel('Tid (sek)');
ylabel('Log10(frekvens i Hz)');
%title('Wavelet Power Spektrum'); % Velg denne når det er aktuelt,
title('Sqrt(Wavelet Power Spektrum)'); % men denne når sqrt blir brukt
colorbar('location','southoutside');

input('Lukk alt');
close all;

```

15.8 To ytterligere eksempler

Vi tar med to ytterligere eksempler på waveletanalyse. Det første er liknende den vi hadde for gjøkens ko ko. Vi har valgt kvitringen til en bokfink, som dominerer fuglesangen i april. Bokfinkens sang blir karakterisert på flere ulike måter. Selv liker jeg best karakteristikken “tit tit tit tit.... el-ske-de-viv”. Det morsomme med waveletanalysen er at lydbildet er langt mer komplisert enn det vi mennesker oppfatter. Det er en meget rask variasjon i frekvens innen hver “tit” som vi ikke oppfatter. K -verdi brukt i analysen var 48.0.

Det andre eksemplet er en waveletanalyse av en trompetlyd. Vi har valgt et utsnitt i tid der trompeten holder samme tone, og intensiteten på lyden opplever vi som temmelig konstant. Tidsbildet av lyden viser mer variasjon i intensitet enn det vi oppfatter. Frekvensbildet (frekvensspekteret) er imidlertid helt slik vi hadde forventet det. Det består av en rekke skarpe linjer som viser grunntonen og de harmoniske. Frekvensaksen er lineær, og derved er avstanden mellom to nærliggende harmoniske fast, nærmere bestemt lik frekvensen til grunntonen.

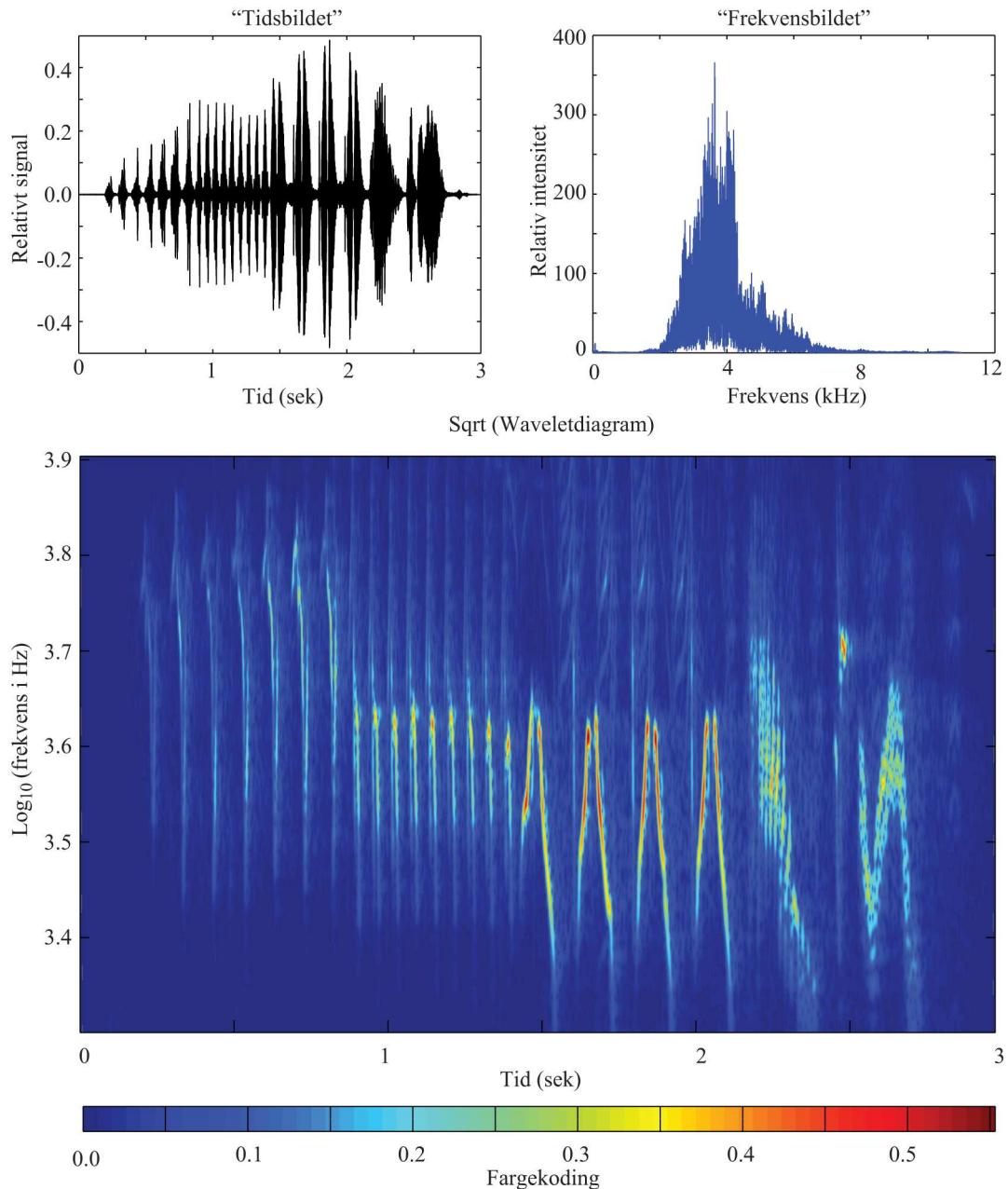
En waveletanalyse av et slikt signal klarer ikke å matche skarpheten i frekvensspekteret, så dersom vi først og fremst er interessert i frekvensen til grunntonen og de harmoniske *for en vedvarende tone*, er fourieranalyse metoden som bør velges.

Dersom vi er interessert i variasjoner i lyden over tid, egner imidlertid ikke fourierana-

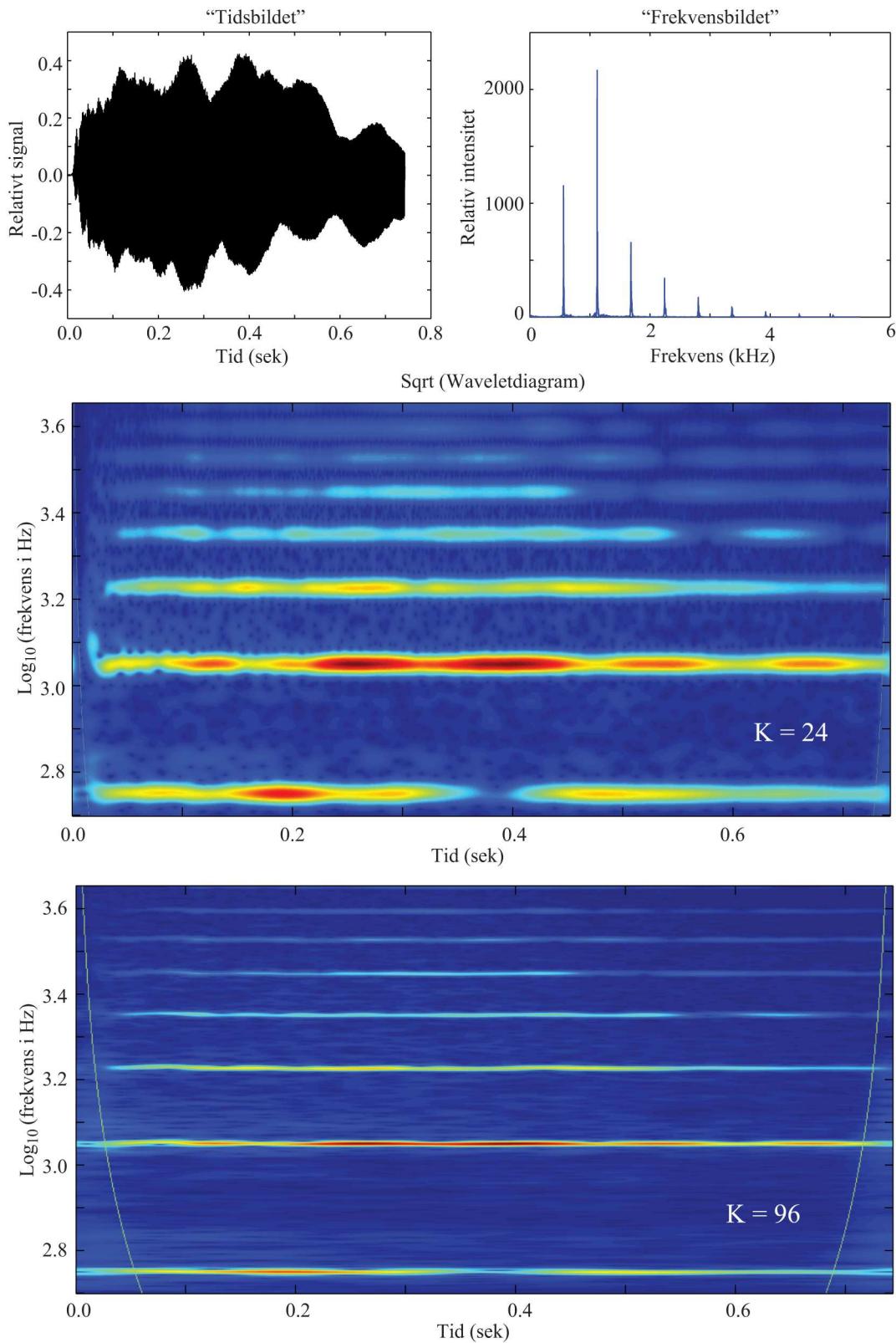
lysen seg. Da kommer waveletanalysen inn. Vi har tatt med to ulike varianter av analyse, basert på bølgetallene $K = 24$ og $K = 96$. I det første tilfellet er frekvensoppløsningen nokså dårlig, men tidsoppløsningen passer i forhold til signalet. I det siste tilfellet er frekvensoppløsningen god, men tidsoppløsningen dårlig.

Får vi noe ekstra ut av waveletanalysen framfor fourieranalysen? Ja, faktisk. Vi ser at styrken på grunntonen og de harmoniske varierer litt i tid. Vi ser også at det er en viss veksling mellom intensiteten til grunntonen og den første harmoniske: Når den ene er kraftig er den andre svak og visa versa. Dette gir liv til lydbildet, og viser et eksempel på at det er vanskelig å erstatte virkelig lyd med syntetisk lyd.

Legg forøvrig merke til at avstanden mellom de harmoniske ikke er konstant i et normalt waveletdiagram, siden vi der normalt har en logaritmisk frekvensakse.



Figur 15.12: Bokfinkens “tit tit tit tit tit.... el-ske-de-viv” analysert i tidsbildet, i frekvensbildet og i en wavelettransformasjon.



Figur 15.13: En ren trompetlyd analysert i tidsbildet, i frekvensbildet og i en wavelettransfomasjon. To ulike K -verdier er brukt i waveletanalysen.

15.9 Wavelet-ressurser på nett

1. A-H Najmi og J Sadowsky: "The continuous wavelet transform and variable resolution time-frequency analyses." Johns Hopkins APL technical digest, vol 18 (1997) 134-140. Tilgjengelig på <http://www.jhuapl.edu/techdigest/TD/td1801/najmi.pdf> den 15. april 2013.
2. <http://www.polyvalens.com/blog/>, "A really friendly guide to wavelets", med mere (C. Valens). Tilgjengelig 15. april 2013.
3. <http://tftb.nongnu.org/>, "Time-frequency toolbox". Tilgjengelig 15. april 2013.
4. <http://dsp.rice.edu/software/rice-wavelet-toolbox>, "Rice Wavelet Toolbox." Tilgjengelig 15. april 2013.
5. <http://www.cosy.sbg.ac.at/~uhl/wav.html>, Mengde wavelet-lenker. Tilgjengelig 15. april 2013.
6. Et 72 siders hefte av Liu Chaun-Lin: "A tutorial of the wavelet transform" (datert 23. februar 2010) er tilgjengelig på <http://disp.ee.ntu.edu.tw/tutorial/WaveletTutorial.pdf> tilgjengelig 15. april 2013. Heftet tar også for seg wavelets brukt i bildebehandling.

15.10 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for likheter og forskjeller mellom fouriertransformasjon og wavelettransfomasjon.
- Gjøre rede for hvilke signaler fouriertransformasjon er å foretrekke og hvilke signaler der wavelettransfomasjon foretrekkes. Begrunn hvorfor.
- Forklare hva vi kan lese ut av et gitt waveletdiagram.
- Forklare hvordan vi kan justere en wavelettransfomasjon for å fremheve detaljer i tid, eller detaljer i frekvens.
- Forklare kvalitativt analogier mellom wavelettransfomasjon og Heisenbergs uskarphetsrelasjon.
- Bruke et waveletanalyseprogram og optimalisere analysen.

15.11 Oppgaver

Forståelses- / diskusjonsspørsmål

1. Hva er viktigste forskjell mellom fouriertransformasjon og waveletanalyse?

2. I hvilke situasjoner gir fouriertransformasjon et temmelig ubruklig resultat?
3. Hvilke ulemper har wavelettransformasjon sammenlignet med fouriertransformasjon?
4. Når ble fouriertransformasjon tatt i bruk i stor stil (FFT), og når ble wavelettransformasjon tatt i bruk i betydelig grad?
5. Wavelettransformasjon har et “randproblem”. Hva menes med det? Hvor stor er randverdisonen?
6. Kan du skissere hvordan wavelettransformasjon kan tenkes brukt for å generere noter direkte fra et lydopptak? Hvilke problemer ser du for deg kan forekomme?

Regneoppgaver

7. a) Bereng en Morlet wavelet (i tidsdomenet) for analysefrekvensen 250 og 750 Hz når samplingsfrekvensen er 5000 Hz og K -parameteren er 16. Plot resultatet med korrekte angivelser av tid på x-aksen.
b) Beregn den fouriertransformerte av hver av de to waveletene. Bruk både en FFT direkte på Morlet-waveleten beskrevet i tidsbildet, og ved å beregne den fouriertransformerte direkte ved hjelp av ligning (15.10). Plot resultatene med korrekte angivelser av frekvens på x-aksen.
c) Kontroller at toppunktet kommer på det stedet du skulle forvente.
d) Gjenta punkt a-c også når K -parameteren er 50.
8. I denne oppgaven er det underliggende temaet analogien til Heisenbergs uskarphetsfunksjon.
a) Generer en numerisk tallrekke som representerer et signal

$$f(t_n) = c_1 \sin(2\pi f_1 t_n) + c_2 \cos(2\pi f_2 t_n)$$

hvor $n = 1, 2, \dots, N$, $N = 8192$, $f_1 = 1000$ Hz, $f_2 = 1600$ Hz, $c_1 = 1.0$, $c_2 = 1.7$, og samplingsfrekvensen er 10 kHz. Signalet skal være hele tiden vi betrakter signalet. Plot et passe utsnitt av signalet i “tidsbildet” (utslag som funksjon av tid) slik at detaljer kommer fram. Pass på å få korrekte tall samt tekst langs aksenene, gjerne også en overskrift.

- b) Beregn den fouriertransformerte av signalet. Plott et passe utsnitt av signalet i “frekvensbildet” (velg gjerne absoluttverdier av fourierkoeffisientene som funksjon av frekvens), med tall og tekst langs aksene som ovenfor.
- c) Beregn den wavelettransformerte av signalet (kan godt ta utgangspunkt i den siste versjonen av programmet gitt i dette kapitlet, eller du kan skrive programmet mer eller mindre fra scratch selv). Bruk Morlet wavelets, og la analysefrekvensen gå f.eks. fra 800 til 2000 Hz (logaritmisk fordelt som vanlig innen wavelettransformasjon). Plot etter tur resultatet for bølgetallet K lik 24 og 200. Kommenter resultatet.
- d) La så signalet være et harmonisk signal som før, men nå konvolvert med en gaussisk funksjon slik at vi får to “bølgepakker”:

$$f(t_n) = c_1 \sin(2\pi f_1 t_n) \exp(-[(t_n - t_1)/\sigma_1]^2) + c_2 \cos(2\pi f_2 t_n) \exp(-[(t_n - t_2)/\sigma_2]^2)$$

hvor $t_1 = 0.15$ s, $t_2 = 0.5$ s, $\sigma_1 = 0.01$ s og $\sigma_2 = 0.10$ s. Beregn den fouriertransformerte av signalet, og også den wavelettransformerte av signalet. Plot signalet i

tidsbildet, frekvensbildet (passe utsnitt) og den wavelettransformerte av signalet for $K = 24$ og 100 (test gjerne flere verdier!) og øvrige parametre som i deloppgave c). Kommenter resultatene!

e) I en lydfil med sangen fra en svarttrost som er tilgjengelig sammen med dette kapitlet, ber vi deg å bruke siste programsnutten i dette kapitlet (eller egen versjon) for å analysere en tidsstrek på vel 1.4 s. Parametre ved analysen: Filnavn: 'Svart-trost2.wav', Nstart = 17000, datastrengens lengde 64 k, frekvensområde 1500 - 8000 Hz, bølgetallet K lik 12 og 96 (og gjerne noen mellom disse verdiene også). Signalet består av fem ulike lydgrupper. Vi er først og fremst interessert i den fjerde av disse!

Plot signalet i tidsbildet, frekvensbildet og signalet analysert ved wavelettransformasjon for denne fjerde lydgruppen. Pass på å ta med også noen utsnitt av de opprinnelige plottene for å få fram detaljer. Dette gjelder i særdeleshet tidsbildet av den opprinnelige lyden! Forhåpentligvis vil du da gjenkjenne et signal vi har støtt på minst to ganger i tidligere kapitler. Du bør gjenkjenne hvordan vi kan lage et slikt signal matematisk.

Nøyne analyse av den fjerde biten av lydsignalet i (1) tidsbildet og (2) etter wavelettransformasjon gjør det mulig å se hvordan en nær analogi til Heisenbergs uskarphetsrelasjon spiller inn på en praktfull måte! For å få fullt utbytte bør du hente ut tidsforskjeller og frekvensforskjeller i diagrammene og sammenholde disse med waveletens utbredelse i tids- og frekvensbildet for de to valgte K -verdiene.

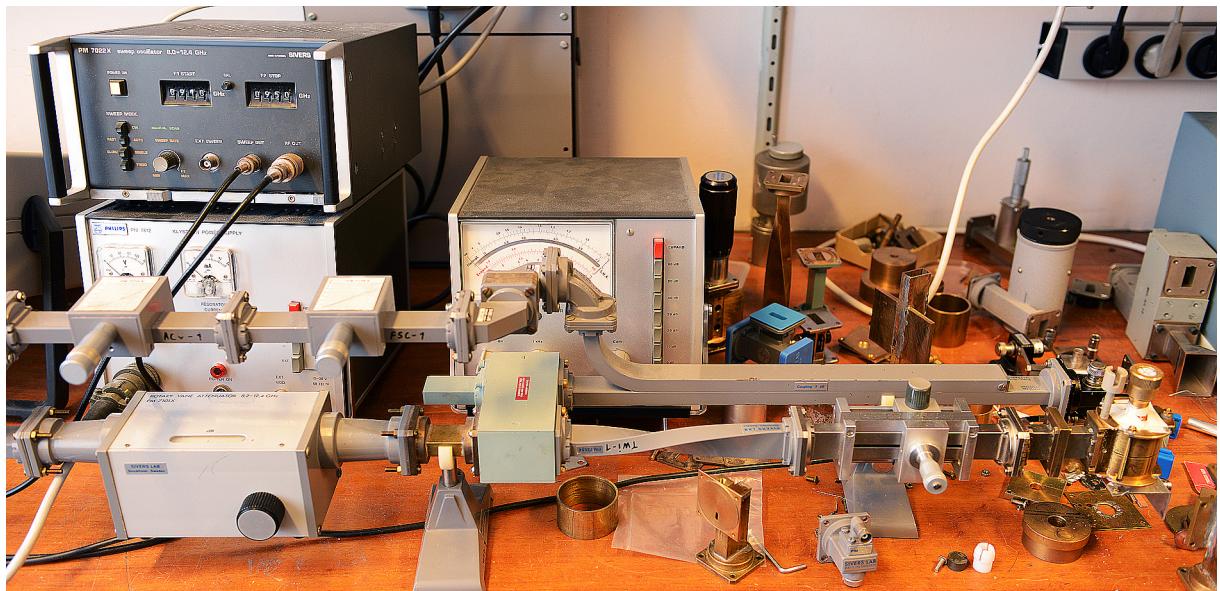
Dersom du er student og har tilbud om hjelp fra lærere, anbefaler vi sterkt at du diskuterer de aktuelle detaljene med læreren inntil at du gjennomskuer samspillet vi ønsker å få fram. Det er mye verdifull kunnskap å hente fra dette problemet, kunnskap som også kan være verdifull i mange andre deler av fysikken!

9. Foreta en waveletanalyse av lydfilen "bokfink" på kursets websider. Forsøk en K -faktor som er dobbelt så stor som i eksemplet i figur 15.12. Forsøk også en analyse med halvparten av K -verdien som ble brukt i den nevnte figuren. Beskriv forskjellene du ser.
10. Foreta waveletanalysen av bokfinklyden på ny for $K = 48$. Velg etter tur å bruke "power spectrum" (absoluttverdien etter omvendt fouriertransformasjon kvadrert), waveletanalyse "på amplitudenivå" (absoluttverdien etter omvendt fouriertransformasjon direkte) og waveletanalyse med kvadratrotten av waveletanalysen (kvadratrotten av absoluttverdien etter omvendt fouriertransformasjon). Gjør dine vurderinger av hvilken av disse metodene du liker best for akkurat dette signalet. Kan det hende at du for et annet signal ville foretrukket en av de andre anskuelsesformene (kvadrat, rett fram, eller kvadratrot-fremstillingene)?
11. Bruk kunnskapen fra kapittel 1 og 2 til å beregne tidsforløpet for en fjærpendel etter at den er satt i sving av en harmonisk kraft med frekvens lik resonansfrekvensen. Følg svingningene også en tid etter at den harmoniske kraften er fjernet. Foreta så en waveletanalyse av svingforløpet. Forsøk å optimalisere analysen mhp K -verdi. Finner du en tilsvarende sammenheng mellom Q -verdien for pendelsvingningen og K -verdien som gir optimalt waveletdiagram?
12. Velg selv en lydfil som du kan transformere til en .wav-fil, og velg et utsnitt som du kan ha lyst til å analysere. Optimaliser analysen, og fortell hvilken informasjon du får ut av diagrammet.
13. Finn data på nettet som viser en tidssekvens du synes kan være interessant å studere. Det kan være værdata, solflekker, strømforbruk eller hva du måtte finne på.

Analysér datasettet både ved tradisjonel fouriertransformasjon og med waveletanalyse. Hvilken metode synes du egner seg best for de dataene du valgte? (Bør ha data med en eller annen form for løs periodicitet med minst 20-30 perioder innenfor dataene du har tilgjengelig.)

Kapittel 16

Skinndybde og bølgeledere



En liten del av et mikrobølgeoppsett med bølgeledere for vel 9 GHz. Oppsettet brukes i elektron paramagnetisk resonans spektroskopি.

I dette kapitlet skal vi vise at elektromagnetiske bølger kan sendes gjennom hule rør omtrent som vi drikker brus fra et cola-glass! Våre rør kalles "bølgeledere". De fungerer fordi også er et annet fenomen kommer til nytte, nemlig det at høyfrekvente elektromagnetiske bølger ikke kan gå gjennom metallplater uten store tap (skinndybde). Det er pussig at Maxwells ligninger tillater at bølger kan følge bølgeledere med svært lite tap. Ved hjelp av enkle grep kan vi justere demping, faser, koblinger fra en bølgeleder til en annen og mere til, omtrent som om vi var rørleggere. Men Maxwells ligninger setter begrensinger for hvordan de elektriske og magnetiske feltene kan opptre.

Optiske fibre er et annet eksempel på bølgeledere. I et nydelig samspill mellom geometri, materialegenskaper og elektromagnetisme oppsummeres en rekke prinsipper fra tidligere kapitler i boka.

¹Copyright 2013 for tekst og figurer: Arnt Inge Vistnes. Versjon 15052013.

16.1 Husker du ...

Vi har tidligere i boka poengtert at løsningen av en bølgeligning i høy grad avhenger av randbetingelsene. I kapittel 8 kom vi tilbake til den samme påminnelsen for elektromagnetiske bølger. De velkjente plane elektromagnetiske bølgene finnes bare langt fra kilden og langt fra strukturer som kan forstyrre det elektriske og/eller det magnetiske feltet. Plane bølger er bare én løsning av Maxwells ligninger, og den løsningen er bare gyldig i medier uten frie ladninger, i fjernfeltsområdet.

Hva skjer dersom en elektromagnetisk bølge kommer inn mot en plan metallplate, eller et annet materiale med frie ladninger? Ladningene vil påvirkes av Lorentz-kraften og vil bevege seg. Bevegelsen vil sette opp et sekundærfelt som vil tendere å motvirke det opprinnelige feltet. Det er da naturlig å forvente at det elektromagnetiske feltet vil avta etter hvert som bølgjen trenger lengre og lengre inn i materialet. "Skinndybden" er en størrelse som forteller oss hvor langt inn i metallet bølgene trenger inn.

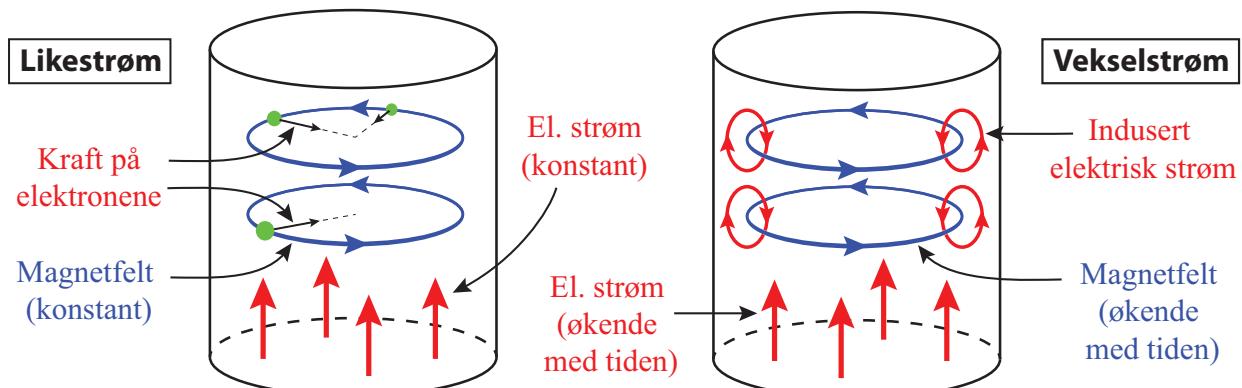
I andre tilfeller der geometrien er annerledes, kan det iblant finnes helt andre løsninger av bølgeligningen / Maxwells ligninger enn planbølger. Dette åpner opp for å transportere bølger uten særlig tap over store lengder, og bølgene sendes da gjennom såkalte "bølgeledere". Det er skinndybde og bølgeledere vi tar opp i dette kapitlet.

16.2 Skinndybde

Når elektromagnetiske bølger f.eks. sendes normalt inn mot en metallflate, vil bølgene bli ganske fort dempet etter som de trenger inn i metallet. Vi starter imidlertid med et enklere bilde for å få fram de underliggende mekanismene.

Når vi sender en elektrisk vekselstrøm gjennom en leder, vil ikke strømmen fordeles likt over hele tverrsnittet. Stømmen har en tendens til å bli størst i de ytre delene av lederen (i "skinnet"/"huden" av lederen). Tykkelsen på det laget der strømtettheten er størst, kaller vi skinndybden.

Når vi sender en vekselstrøm gjennom en sylinderisk metallede, er det relativt enkelt å forklare den viktigste virkningsmekanismen for skinneffekten.



Figur 16.1: Elektrisk og magnetisk felt inne i en sylinderisk metallede hvor det går en elektrisk strøm. Til venstre har vi en konstant likestrøm, til høyre en vekselstrøm. Retningene på de indukserte strømmene reflekterer forholdene mens strømmen er voksende med tiden.

Et øyeblikksbilde av strømmen og feltene denne genererer er vist i figur 16.1. Den elektriske strømmen vil generere sirkulaert orienterte magnetfelt normalt på og sentrert i lederens akse. Dersom det går likestrøm gjennom ledningen, vil elektronene bli påvirket av en kraft som trekker dem mot sentrum i lederen. Denne effekten gir en “Hall-effekt” hvor det blir en liten potensialforskjell mellom ytterste del av lederen og aksen i lederen. Potensialforskjellen fører raskt til et elektrisk felt som akkurat motvirker videre transport av elektroner mot sentrum av lederen. Bortsett fra denne “en gang for alle” effekten idet strømmen slås på, vil strømmen fordele seg jevnt over hele tverrsnittet ved likestrøm.

Ved vekselstrøm er det annerledes. Da vil økende strømstyrke føre til lokale strømsløyfer som vil forsøke å motvirke magnetfeltøkningen (“Lenz lov”). De lokale strømsløyfene fører til at strømøkningen i de sentrale deler av lederen blir motvirket, mens strømøkningen i ytre del av lederen blir forsterket. Dette er mekanismen bak skinneffekten for denne konfigurasjonen.

Skinneffekten fører til at vekselstrømmen ikke utnytter like godt hele tverrsnittet i lederen. Det betyr at resistansen i lederen for vekselstrøm er annerledes en resistansen for likestrøm.

Videre vil effektiviteten til de lokale strømsløyfene avhenge av frekvensen. Ved likestrøm er det ingen induserte strømsløyfer, men strømsløyfene vil bli mer og mer effektive ettersom frekvensen øker. Det fører til at det sjiktet der strømmen går, avtar med økende frekvens. Skinndybden er altså frekvensavhengig.

Vi skal om litt utlede et uttrykk for hvor stor skinndybden er, men kan allerede nå nevne at for aluminium, som ofte brukes i kraftledninger, er skinndybden 11-12 mm ved 50 Hz. Det betyr at for tykke kraftledninger med diameter på om lag 3 cm, vil det meste av strømmen gå i et ytre lag om lag 1 cm tykt, og i mindre grad i de sentrale deler av ledningen. Iblast lages slike kraftledninger hule fordi den midtre delen så allikevel ikke bidrar noe særlig til den totale ledningsevnen. Andre ganger legger man stålwire i midten av kabelen og aluminium rundt. Stålwiren gir økt styrke på ledningen, og den dårligere ledningsevnen spiller liten rolle siden strømtettheten i sentrum likevel er ganske beskjeden.

I stedet for én leder som er ekstra tykk ved overføring av store kraftmengder (stor strømstyrke), velger man iblant å legge to (“duplex”) eller tre (“triplex”) ledninger innenfor hver av de tre fasene i en kraftledning. De to eller tre ledningene holdes da i en konstant gjensidig avstand på 10-20 cm, for å “lure” skinndybdeeffekten.

16.2.1 Elektromagnetiske bølger inn mot en metallflate

For høye frekvenser er det iblant interessant å betrakte hva som vil skje dersom vi har en elektromagnetisk bølge som kommer inn mot f.eks. en metallbit/metalloverflate.

I kapittel 8 viste vi hvordan Maxwells ligninger under visse betingelser fører fram til følgende bølgeligning:

$$\frac{\partial^2 \vec{E}}{\partial t^2} = c^2 \frac{\partial^2 \vec{E}}{\partial z^2} \quad (16.1)$$

hvor

$$c = \frac{1}{\sqrt{\epsilon_r \epsilon_0 \mu_r \mu_0}} \equiv \frac{1}{\sqrt{\epsilon \mu}} \quad (16.2)$$

Symbolene regnes som kjent.

Når bølgen går vinkelrett inn i et medium hvor ledningsevnen $\sigma \neq 0$ (for eksempel et metall), blir strømtettheten også forskjellig fra null. Det kan vises at bølgeligningen under slike forhold får formen:

$$\frac{\partial^2 \vec{E}}{\partial z^2} = \mu\sigma \frac{\partial \vec{E}}{\partial t} + \mu\epsilon \frac{\partial^2 \vec{E}}{\partial t^2} \quad (16.3)$$

Vi kan gjette på en løsning hvor feltene avtar eksponentielt innover i metallet:

$$E = E_0 e^{i(kz - \omega t)} \quad (16.4)$$

Setter vi denne prøveløsningen inn i ligning (16.3), får vi:

$$k = \sqrt{\mu\omega} \sqrt{i\sigma + \epsilon\omega}$$

Vi merker oss at bølgetallet k nå er en kompleks størrelse, hvilket innebærer at E i ligning (16.4) får et eksponentielt avtakende ledd, slik vi forventet.

Dersom ledningsevnen er stor, eller mer presist: Dersom $\sigma \gg \epsilon\omega$, kan vi se bort fra det siste leddet i uttrykket for k . Da er:

$$k = \sqrt{i} \sqrt{\mu\sigma\omega}$$

Det kan vises at

$$\sqrt{i} = \frac{1}{\sqrt{2}}(1+i)$$

Da kan k skrives som følger:

$$k = \sqrt{\frac{\mu\sigma\omega}{2}}(1+i) \equiv \frac{1}{\delta}(1+i)$$

hvor δ kalles skinndybden. Setter vi dette inn i ligning (16.4), får vi:

$$E = E_0 e^{i(z/\delta - \omega t)} \cdot e^{-z/\delta}$$

Fysisk løsning er realverdien av uttrykket, som er:

$$E(z, t) = E_0 \cos\left(\frac{z}{\delta} - \omega t\right) \cdot e^{-z/\delta} \quad (16.5)$$

Spørsmålet er imidlertid om dette er en for enkel løsning. Vi forutsatte jo ovenfor $\sigma \gg \epsilon\omega$. Setter vi inn for de aktuelle størrelsene for kobber, får vi:

$$\frac{\sigma}{\epsilon\omega} = \frac{6.4 \cdot 10^{18}}{\omega} F^{-1} \Omega^{-1}$$

Det viser seg da at tilnærmingen vi gjorde holder for alle elektromagnetiske bølger fra omtrent røntgenområdet og lengre bølgelengder. Formelen er likevel bare gyldig for frekvenser som er godt unna betydelige atomære eller molekulære resonansfrekvenser, og godt unna normale kollisjonsfrekvenser for elektroner i deres vandring gjennom metallet vi betrakter. For ikke-metaller er det utledet en litt mer komplisert sammenheng mellom skinndybde og elektromagnetiske egenskaper til materialet, men vi går ikke inn på disse detaljene her.

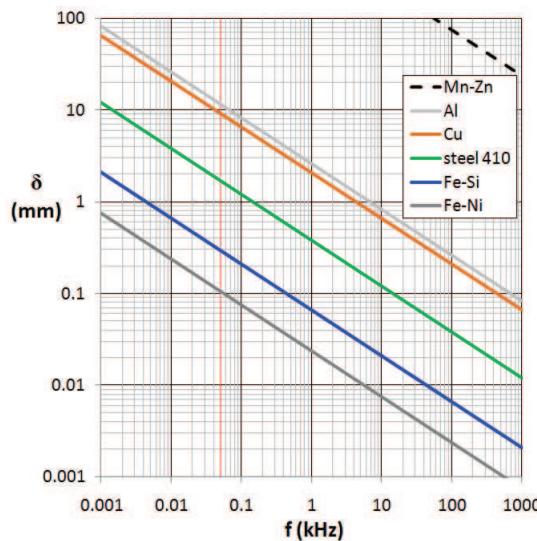
Ligning (16.5) synes å være ok for den geometrien vi valgte. Ligningen viser at den elektromagnetiske bølgen fortsetter innover i metallet, men at amplituden synker med en faktor $1/e$ for hver skinndybde-lengde vi går innover i metallet. Setter vi inn for materialegenskapene for kobber i uttrykket for skinndybden:

$$\delta = \sqrt{\frac{2}{\mu\sigma\omega}} \quad (16.6)$$

finner vi en skinndybde på

- 9 mm ved 50 Hz
- $66 \mu\text{m}$ ved 1 MHz
- 100 nm ved 30 GHz (radar)

Det betyr at bølgene ved radiofrekvenser og høyere blir "drept" ganske brutalt i den ytre delen av et metall. For lave frekvenser er dempingen langt dårligere.



Figur 16.2: Skinndybde som funksjon av frekvens for ulike metaller (idealisert). Et log-log-diagram er valgt for å dekke mange dekader. (Figuren er hentet fra Wikipedia med oppslagsord "Skin dept" i mai 2013.)

Figur 16.2 viser sammenhengen mellom skinndybden δ og frekvensen f for fem ulike metaller eller legeringer i et log-log diagram.

Av figuren ser vi at ved 1 MHz er skinndybden for aluminium $90 \mu\text{m}$, og for mobiltelefon-frekvenser på 0.9-1.8 GHz har skinndybden for aluminium sunket til ca $3 \mu\text{m}$! Det betyr at ved så høye frekvenser er det ikke mye å vinne mhp ac resistans om ledningene er mye tykkere enn dette. Stor overflate er viktigere enn totalt tverrsnitt. Ordet "skinndybde" synes i slike tilfeller å være et godt valg!

Skinndybde kommer inn også ved induksjonskomfyrer. Her brukes ofte en frekvens om lag 24 kHz. Brukes stålgryter hvor ledningsevnen ikke er spesielt høy, og den relative magentiske permeabiliteten er nær 1 (ikke-magnetisk materiale), blir skinndybden såpass liten at store deler av det elektromagnetiske feltet fra komfyren går rett gjennom bunnen på grytene. Først når vi har materialer som har en stor relativ magnetisk permeabilitet

(inneholder magnetiserbart jern), vil tilnærmet all energi i feltene fra komfyren bli avsatt som varme i bunnen på gryta.

I gryter og panner ment for induksjonskomfyrer, brukes gjerne magnetisk stål, f.eks. "karbonstål 1010" eller "rustfritt stål 432" som begge har en relativ magnetisk permeabilitet på ca 200. Ut fra ligning (16.6) ser vi at skinndybden da synker betraktelig sammenlignet med ikke-magnetisk materiale. Skinndybden ved 24 kHz blir bare 0.1-0.2 mm, og følgelig vil praktisk talt all energi fra komfyren bli avsatt som varme i bunnen av kjelen.

Kommentar

Utledningen av uttrykket for skinndybden må settes i perspektiv. Vi har vist at ligning (16.5) er en mulig løsning av Maxwells ligninger. Det er ikke derved sagt at løsningen i et konkret tilfelle faktisk *er* denne løsningen! Langt derifra! Vi lot som om løsningen kunne skrives som en plan bølge i betydning at løsningen ikke avhenger av x og y . For at det skal være aktuelt, må fysikken være slik at det ikke er noen randbetingelser som påvirker bølgen i x og y retning nær det stedet vi betrakter.

Hva menes da med "nær"? Vi har drøftet dette tidligere, og nærfeltet strekker seg gjerne typisk en beregnet bølgelengde utover.

Det betyr at dersom ligning (16.5) skal være en ok løsning for å beskrive hvordan elektrisk eller magnetisk felt nær en kraftledning dempes av en kobberplate, må kobberplaten være minst 6000 x 6000 km stor og uten skjørter!

Det andre vi har sett i tidligere kapitler er at løsningen også i høy grad avhenger av initialbetingelsene, i alle fall i nærfeltområdet. Tar vi igjen utgangspunkt i feltene fra en kraftledning, har vi sett at de er typiske nærfelter *ikke* følger lovemessigheten til plane elektromagnetiske bølger.

Med andre ord, løsningen (16.5) duger overhodet ikke for å beskrive feltene når vi setter opp en kobberplate i et forsøk på å dempe felter fra en kraftledning.

Derimot kan løsningen i visse tilfeller være ganske god når elektromagnetiske bølger møter en metallflate som er stor sammenlignet med bølgelengden, eller når metalltet kommer så tett inn mot kilden som i en induksjonskomfyr.

Min kommentar har først og fremst den hensikt å minne om at en generell løsning av en bølgeligning i utgangspunktet er lite verdt dersom vi ikke også trekker inn initialbetingelser og randbetingelser!

16.3 Bølgeledere

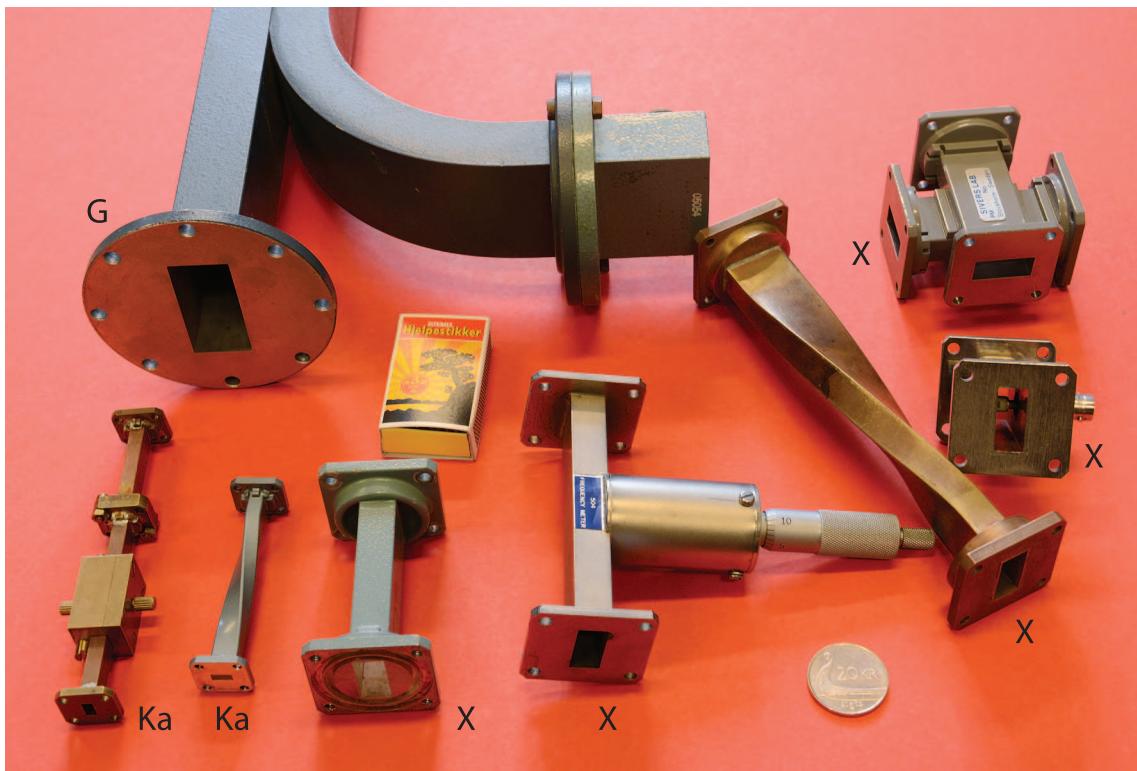
En bølgeleder er en mekanisk struktur som leder bølger fra et sted til et annet. I gamle båter var det gjerne et metallrør fra styrehuset/broen til maskinrommet. Pratet man inn i den ene enden av røret, kunne andre høre hva man sa selv om de var mange meter unna.

En enda mer kjent bølgeleder er legenes stetoskop. Lyd fra hjertet og lunger fanges opp i en liten trakt holdt mot hudens overflate, og lyden ledes til legens ører. Det er mer fysikk involvert i et stetoskop enn mange er klar over!

I vår sammenheng vil vi konsentrere oss om bølgeledere for elektromagnetiske bølger. I bunnen ligger da Maxwells ligninger og bølgeligningen vi utledet i kapittel 8, men nå

må differensialligningene løses med et sett randbetingelser helt forskjellig fra det vi hadde i fjernfeltet og som representerte plane elektromagnetiske bølger.

Bølgeledere for elektromagnetiske bølger er vanlige i mikrobølgeområdet, det vil si for frekvenser mellom 2 og 40 GHz (bølgelengder fra 15 cm og ned til 0.67 cm) [Området er egentlig enda større.] Mest vanlig er hule rektangulære metallrør, som vist på figur reffig:bl.



Figur 16.3: Fotografi av en del mikrobølgekomponenter der bølgeledere inngår. Utstyrekomponenter for tre ulike frekvensbånd er tatt med.)

Når Maxwell's ligninger skal løses for en slik geometri, er randbetingelsene som følger:

- Elektromagnetiske bølger går ikke gjennom metallet, men blir reflektert.
- Ethvert elektrisk felt som kommer inn mot en metalloverflate, må være (tilnærmet) vinkelrett på denne flaten.
- Ethvert magnetfelt som kommer inn mot en metalloverflate, må være (tilnærmet) parallel med flaten.

Det elektriske og magnetiske feltet kan selvfølgelig ha andre retninger inn mot metallet enn de vi nettopp listet opp. Randbetingelsene vi nevnte i stad er imidlertid valgt for å finne en løsning av Maxwell's ligninger som medfører så små strømmer som mulig i metallet. Det er nødvendig for at ikke bølgen skal miste for mye energi per lengde når den beveger seg gjennom bølgelederen.

Det finnes generelt en mengde ulike løsninger av Maxwell's ligninger for et rektangulært tverrsnitt i en bølgeleder. Elektrisk og magnetisk felt har til dels en svært forskjellig fordeling og retning i rommet sammenlignet med planbølge-løsningen i fjernfeltsonen vi diskuterte i kapittel 8.

For en gitt frekvens er det imidlertid bare et endelig antall mulige løsninger, og dersom bølgelengden er større enn to ganger den lengste dimensjonen i bølgelederens hulrom, er det faktisk ingen løsning. Når den lengste dimensjonen i hulrommet er mellom en halv og en hel bølgelengde, og den korteste dimensjonen er bare halvparten av den lengste, er det bare én mulig løsning av Maxwells ligning som svarer til en bølge. Det bølgemønsteret vi da får i bølgelederen er entydig bestemt. Vi sier at vi har *én-mode-overføring* (kvari engelsk-norsk!). Den laveste frekvensen som kan sendes gjennom en bølgeleder kaller vi "cutoff frekvensen".

Øker vi frekvensen på de elektromagnetiske bølgene slik at bølgelengden blir mindre enn halvparten av den lengste dimensjonen, er det minst to ulike løsninger av Maxwells ligninger. Da kan bølgen gå gjennom bølgelederen på (minst) to ulike måter. Vi får en fler-mode overføring.

I en rektangulær bølgeleder er gjerne den største dimensjonen omtrent dobbelt så stor som den minste dimensjonen. Det sikrer oss at polariteten til de elektromagnetiske bølgene bare kan være på én måte.

Noen typiske dimensjoner for bølgeledere (samme som i figur 16.3):

Bånd	Frekvens (GHz)	Bølgelengde (mm)	Dimensjon (mm)
G	3.95 - 5.85	51.3 - 75.9	22.15-47.55
X	8.2 - 12.5	24.0 - 36.6	10.16-22.86
Ka	26.5 - 40.0	7.5 - 11.3	3.55 - 7.11

16.3.1 Bølgemønsteret i en rektangulær bølgeleder

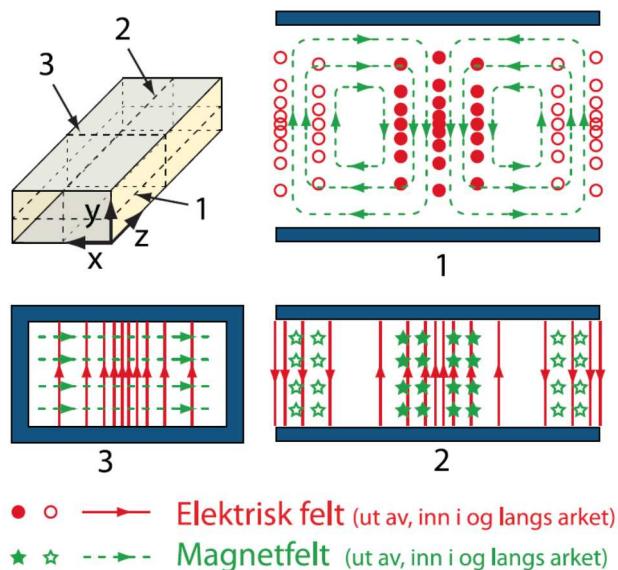
Figur 16.4 viser en prinsippskisse for feltfordelingen i en såkalt TE10 bølgeleder. TE står for "transverse electric". Det elektriske feltet er vinkelrett på den bredeste flaten i bølgelederen med rektangulært tverrsnitt. Feltfordelingen er ikke den samme som ved en plan elektromagnetisk bølge. Hva ligger forskjellen i?

Tenk deg en plan elektromagnetisk bølge slik vi behandlet den i kapittel 8. Dersom vi hadde en slik feltfordeling inne i den rektangulære bølgelederen, ville det elektriske feltet vært parallellt med to sidekanter. Et slikt felt ville medføre store strømmer av elektroner i metallveggen i bølgelederen, og derved et stort tap.

Feltfordelingen i en TE10 bølgeleder er slik at det elektriske feltet hele tiden er vinkelrett på den største innvendige flaten, men feltet avtar mot null når vi nærmer oss sideflatene. Derved blir det langt svakere elektriske strømmer i sideflatene enn det ville vært med en plan bølge. Den løsningen av Maxwells ligninger / bølgeligning som initialbetingelser og randbetingelser tvinger for en bølgeleder, kan likevel være minst like "vakker" som planbølgeløsningen

Iblast blir det sagt at bølgemønsteret i en bølgeleder svarer til at en plannbølge blir reflektert fram og tilbake mellom veggene i bølgelederen. Dette er en misvisende beskrivelse. Bølgene er løsninger av Maxwells ligninger under de gitte randbetingelsene, og er en særegen løsning. Når dimensjonen i bølgelederen blir stor i forhold til bølgelengden, blir det imidlertid mange ulike løsninger av Maxwells ligninger. I slike tilfeller gir det mening å sammenligne løsninger med reflekterte plane bølger gjennom bølgelederen.

Det elektriske feltet på tvers av bølgelederen har likevel sin forankring i elektriske



Figur 16.4: Feltfordelingen for en TE10 mode for det elektriske feltet inne i en rektangulær bølgeleder. For passe dimensjoner av bølgelederen i forhold til bølgelengden, overlever bare TE10 moden.

ladninger på overflaten inne i bølgelederen. Siden bølgen beveger seg langs bølgelederen, må disse ladningene også bevege seg. Derved blir det induserte strømmer i den innvendige overflaten av bølgelederen, slik det er antydet i figuren.

Den indre flaten av bølgeledere belegges gjerne med sølv eller gull for at ledningsevnen skal være så stor som mulig. Da blir tapet minimalt. Sølv- eller gullaget behøver bare å være noen få mikrometer tykt siden skinndybden ved disse frekvensene er så liten som den er.

Videre må det sørget for at det ikke er noen sprekker i strukturen som hindrer strømmene i overflaten. Merk deg hvordan strømmene går langs veggene. For å unngå å skjære av disse strømmene, kan vi bare lage lange hull bølgelederen dersom hullet lages på den brede siden. Ved å legge to bølgeledere inntil hverandre og lage et felles hull gjennom vegen (på bredsiden), kan en del av bølgen i den ene bølgelederen lekke over i den andre. Plasseres en halvlederdiode tvers over bølgelederen (og den ene enden ledes ut som en egen ledning), får vi en detektor som gir et signal som er proporsjonalt med intensiteten til bølgene som passerer (eksempel er gitt helt til høyre i figur 16.3).

Bølgelederne lages gjerne som rør med rektangulært tverrsnitt og flenser for å skru sammen ulike biter. Noen biter kan dreie feltet 90 grader, andre biter kan lage en 90 graders knekk på selve bølgelederen. Mikrobølgene følger rørsystemet opptil en god del meter fra generatoren (gjerne såkalt klystron) til antennen hvor mikrobølgene sendes ut.

Det finnes en hel rekke ulike "moder" som det elektriske feltet kan ha i rektangulære (og sirkulære) halvledere dersom bølgelengden er mindre enn den største dimensjonen i bølgelederen. Vi går ikke inn på andre moder enn TE10 i denne omgangen.

Det er en morsom utfordring å bruke Maxwells ligninger for å finne ut hvilken retning en TE10 bølge brer seg i når vi har en tegning av feltfordelingen i en bølgeleder (se oppgaver bak).

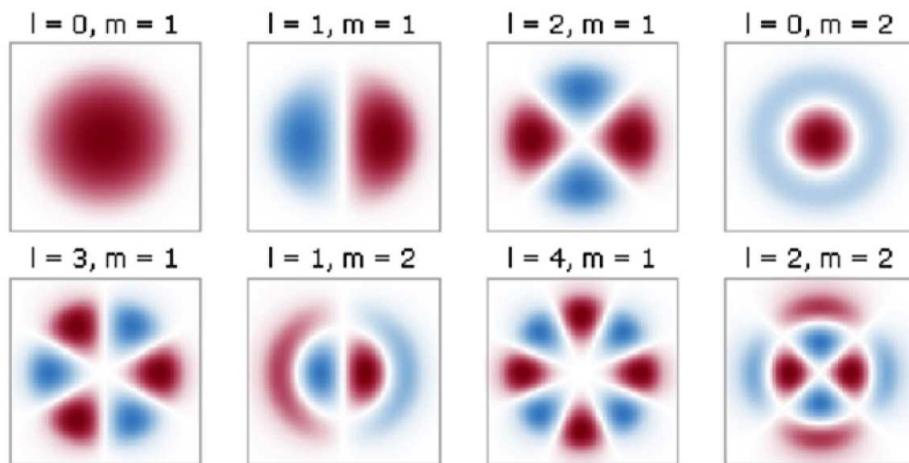
Elektromagnetiske bølger med frekvens i området 2-60 GHz har tradisjonelt blitt brukt

for radar, men nå er frekvensene også brukt for mobiltelefoni og dataoverføring. Spesielt for radarformål brukes det gjerne høye effekter på signalet som overføres fra en sender til selve radarantennen. Det er problematisk å sende slike signaler gjennom vanlige ledninger og coaxkabler, - bølgeleddere kan ofte tåle høyere effekter i overføringen.

16.4 Enkeltmode optisk fiber

Det er vanlig å høre at i en optisk fiber holder lyset seg inne i fiberen på grunn av totalrefleksjon (basert på Snels brytningslov). Vi har til dels gjort det samme tidligere i boka.

For optiske fibre med stort diameter i forhold til bølgelengden er det helt greit å bruke en slik forklaringsmodell. I det tilfellet tilfredsstiller grenseflaten mellom kjernen og kappen et godt stykke på vei forutsetningene vi gjorde da vi utledet refleksjonslovene basert på Maxwells ligninger.



Figur 16.5: Fordeling av elektrisk felt på tvers av en optisk fiber for åtte forskjellige “modes”. Kun den enkleste overlever i en “single mode fiber”. Rødt og blått forteller at retningen på det elektriske feltet er forskjellig i de to områdene. Modene klassifiseres ved hjelp av to tall som gir symmetriegenskapene til moden. Forsök å finne hva de to parametriene helt konkret forteller oss. (Figuren er hentet fra <http://www.rp-photonics.com/waveguides.html> i april 2010.)

Når diameteren til kjernen i den optiske fiberen krympes til ca seks ganger bølgelengden, blir det annerledes. Da kan vi ikke lenger betrakte lyset som plane bølger, for plane bølger vil ikke kunne overleve inne i en slik fiber.

Da er det andre løsninger av Maxwells ligninger som tvinger seg fram. I figur 16.5 er det vist tverrsnittet for flere mulige “moder” som lyset kan ha inne i en optisk fiber. En fullstendig beskrivelse av modene ville kreve en tredimensjonal skisse, men vi går ikke i detalj her.

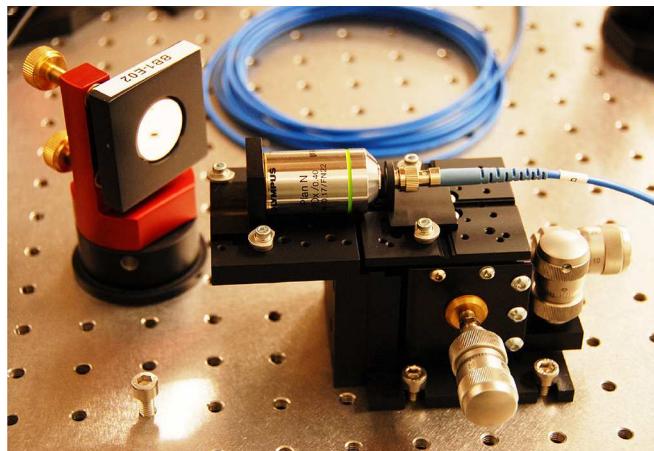
Poenget er at når diameteren til fiberen gjøres mindre og mindre, vil de høyere modene ikke kunne forplante seg langs fiberen. For en passe diameter vil bare den enkleste moden overleve. Skrumper vi inn diameteren enda mer, vil heller ikke den overleve. En optisk “single mode fiber” kan derfor brukes for å “renske opp” laserlys som ikke har en perfekt gaussisk intensitetsprofil.

Når lys i det infrarøde området sendes gjennom en single mode fiber som har de rette dimensjonene og har superrent glass i kjernen, er tapet utrolig lavt! Og siden lyset bare kan forflytte seg på en svært nøye definert måte (som ligger i “single mode”), kan pulser sendes mange, mange kilometre før pulsformen må renskes opp før signalet videresendes.

Det er slike optiske fibre som sikrer vårt imponerende internett. Med andre ord: En spesiell løsning av Maxwells ligninger hvor initialbetingelser og randbetingelser er alfa og omega, er det som holder internett gående! Ingen plane bølger der i gården!

En ulempe ved bruk av single mode fibre er at diameteren på kjernen er så liten at det er en utfordring å koble lys inn i fiberen. I vårt laboratorium bruker vi en single mode fiber for å renske opp laserlys med bølgelengden 405 nm. Den indre delen av fiberen (der lyset skal gå) er da bare $2.7\text{ }\mu\text{m}$ i diameter. Rundt denne kjernen strekker det seg en “cladding”-sone med lavere brytningsindeks ut til $125\text{ }\mu\text{m}$ diameter, og det legges en “coating”-sone utenpå der til en diameter $245\text{ }\mu\text{m}$. Starter vi med en laboratorielaser som normalt sender strålen med diameter minst 1 mm ut i fri luft, må strålen fokuseres kraftig. Det gjøres med et mikroskopobjektiv (se figur 16.6). Enden av fiberen må dernest plasseres akkurat i brennplanet for den fokuserte strålen, og fiberen må ha en retning som faller helt sammen med optisk akse for strålen. Det er en betydelig tålmodighetsprøve å få så mye av lyset inn i fiberen som mulig!

For telekommunikasjon er det utviklet spesialadaptere som gjør tilkoblingen langt enklere.



Figur 16.6: For å koble laserlys fra en laboratorielaser inn i en single mode optisk fiber, brukes mikroskopobjektiv og presisjonsskruer i tre dimensjoner.

16.5 Læringsmål

Etter å ha jobbet deg gjennom dette kapitlet bør du kunne:

- Gjøre rede for begrepet skinndybde når en vekselstrøm går gjennom en metalledning.
- Gjøre rede for begrepet skinndybde når elektromagnetiske bølger møter et metall.
- Kjenne til hvilke parametre som innvirker på størrelsen til skinndybden, og kjenne omtrent til skinndybder for noen få frekvenser og metaller.
- Gjøre rede for at en enkel analyse av skinndybde kan ha betydelige svakheter.
- Gi en enkel skisse av fordelingen av elektriske og magnetiske felt inne i en TE10 rektangulær bølgeleder.
- Forklare hvorfor Snels brytningslov ikke er relevant for å forklare hvordan en single mode optisk fiber fungerer.
- Antyde hvorfor single mode fibre er attraktive i forskning og teknologi.
- Gjøre rede for hvorfor det er problematisk å koble lys fra en åpen laboratorielaser inn i en single mode optisk fiber.

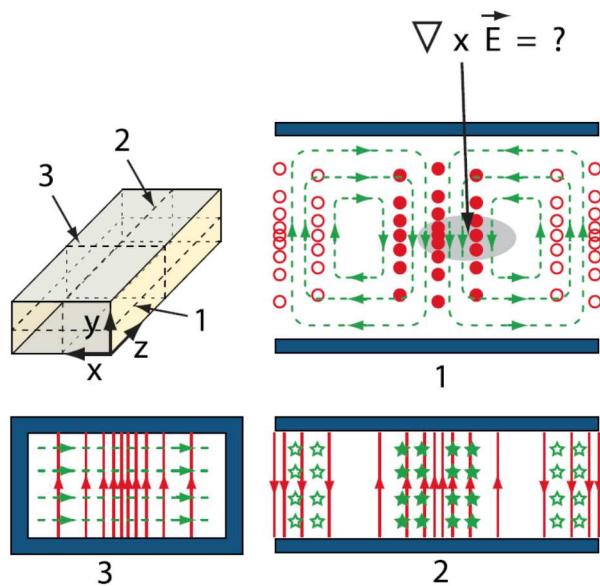
16.6 Oppgaver

Forståelses- / diskusjonsspørsmål

- Les læringsmålene for å se hva vi ønsker at du sitter igjen med etter å ha lest kapitlet.

Regneoppgaver

- Kan du ut fra feltfordelingen som er vist i figur 16.7 fortelle hvilken retning mikrobølgene brer seg i den rektangulære bølgelederen?



Figur 16.7: Betrakt feltfordelingen i det skraverte området og beregn den angitte størrelsen. Ved hjelp av Maxwells ligninger skal du da kunne slutte deg til tidsutviklingen framover.

- En enkelt-mode fiber beregnet på lys med bølgelengde mellom 450 og 600 nm, har en kjerne med diameter (mode field diameter) på om lag $3.5 \mu\text{m}$. Regn ut omtrent hvor mange prosent av lysintensiteten vi hadde mistet dersom vi satte en slik fiber inn i strålen fra en vanlig laboratorielaser uten å bruke et mikroskopobjektiv for å fokusere strålen inn på fiberen. Strålediameteren for mange laboratorielasere er om lag 1.5 mm.

Her kommer det i neste versjon av boka, blant annet en liste over konstanter som er nyttige ved løsning av oppgaver.

Arnt Inge Vistnes har lenge hatt interesse for svingninger og bølger. Han laget sitt eget teleskop mens han gikk på gymnaset. Han har vært interessert i fotografi, mørkeromsarbeid og linser fra barnsben av. Han har spilt ulike musikkinstrumenter og lagt merke til lovmessigheter i den sammenheng. Han har vært radioamatør og laget egne radioer og sendere, og lærte om antennediagrammer og utstrålt effekt. Han har forsket på biologisk effekt av elektromagnetiske felt/bølger i en rekke år, og innså i den tiden at mange fysikere har alvorlige misforståelser knyttet til disse fenomenene. Han har arbeidet på en forskningsgruppe ved Fysisk institutt ved Universitetet i Oslo som var opptatt av synsforskning og fargesyn, hvor man blant annet viste hvor viktig kanteffekter er. Dette er et eksampel på hvordan løsning av bølgeligningen i høy grad er avhengig av randbetingelser. Han har de siste årene forsket på lys, og har i den sammenheng interessert seg spesielt for lys med ulik grad av koherens, for diffraksjon og interferens, og for optikk generelt (og såkalte sammenfiltrede fotoner og filosofiske konsekvenser av disse).

Arnt Inge Vistnes har vært med på å innføre bruk av numeriske metoder ved løsning av fysikkoppgaver, og for å utforske fysiske fenomener også ved "numeriske eksperimenter". Numeriske metoder gir oss mulighet for å studere fenomener som i praksis er utilgjengelig med vanlig analytisk matematikk. På den måten får vi nå innsikt i viktige fenomener som "gårsdagens fysikere" aldri lærte seg.

Arnt Inge Vistnes synes det er meningsfylt å undervise, og setter aller mest pris på gode samtaler med studenter om fenomener hvor fysisk forståelse står i sentrum.