PROJEKT I MATEMATISK KOMMUNIKATION

VÅREN 2005

LUNDS TEKNISKA HÖGSKOLA
Lunds universitet

Matematikcentrum
Matematik

CENTRUM SCIENTIARUM MATHEMATICARUM

Förord

Detta är en sammanställning av de projektarbeten som gjorts inom kursen *Matematisk kommunikation*. Kursen (3p) är obligatorisk för förstaårsstudenter vid civilingenjörsprogrammet *Teknisk Matematik* vid Lunds Tekniska Högskola (LTH). De projektrapporter som återfinns här, samt tillhörande muntliga presentationer, har tillsammans med ytterligare två inlämningsuppgifter varit examinerande. Endast betygen *Godkänd* alternativt *Icke godkänd* har kunnat erhållas.

Målen med kursen är:

- Ge en introduktion till den matematiska idévärlden och genom enkla exempel påvisa kraften i stringent matematisk teori.
- Öka den matematiska allmänbildningen och ge en kort presentation av den matematiska kulturen och dess historia.
- Ge en inblick i aktuell och modern matematisk forskning.
- Öva förmågan att framlägga och presentera matematiska resonemang.

Många av de projekt som återfinns här innehåller stoff ifrån en högre nivå än den man som förstaårsstudent hunnit uppnå och samtliga projekt har handletts av forskare och lärare ifrån olika institutioner vid Lunds Universitet. Ett stort tack går till dessa handledare. Jag vill också speciellt tacka Pelle Pettersson och Johan Råde som båda har gjort en stor insats under kursen samt Jens Nilsson som har hjälpt till med sammanställningen av arbetena. Slutligen tackar jag alla studenter som medverkat i kursen för en god insats.

Lund i juni 2005,
Magnus Fontes.

Innehåll

Matematisk Kommunikation
Besluts och Spelteori

Ronja Månsson, Peter Jönsson, Simon Molin,
Huu-Phuoc Huynh, Sara Jönsson, Lina Brodén

18 maj 2005

Innehåll

Sammanfattning

Matematikens utveckling har lett till att mer avancerade spel kan analyseras. Det finns en mängd olika spel och beslutsteorier. Med hjälp av enkla exempel gås teorierna igenom mer ingående.

Inledning

Människan har i alla tider sysselsatt sig med någon form av spel. En kort tillbakablick på enkla spel från egyptiernas dynasti fram till dagens mer utvecklade spel ges. Spelen har blivit mer och mer avancerade med åren och i takt med matematikens framsteg genom bland annat teorierna i "Theory of Games and Economic Behavior" och Nash's jämviktsbegrepp har spelen analyserats djupare. Från enkla nollsummespel har utvecklingen gått vidare till att analysera mer avancerade spel. I texten tas olika exempel upp bland annat kommer det visas hur man sätter upp vinstmatriser för spel och hur lösningarna till dessa ser ut. Olika teorier för hur beslut ska fattas diskuteras med hjälp av några exempel.

1 Historia

1.1 Historisk sammanfattning fram till 1800

1.1.1 Förhistoria

Spår efter någon form av spel finns så långt tillbaka som man följt människan. Det man har hittat är flera fynd av benet astragalus vilket kan vara målat eller ristat i. Benet sitter i handen och är i form av en tärning, man kan säga att det är föregångaren till dagens tärningar. Astragalus har fyra platta sidor och två rundade som den inte kan vila på. Man vet inte riktigt vad det användes till från början. Teorierna är att det antingen har använts i någon form av chansspel, som valuta eller någon form av leksak. Under egyptiernas dynasti ca 3500 f.k. finns det dock säkra bevis på att astragalus har använts till någon form av brädspel. Man har bl.a. hittat brädspel, där man använde astragalus som tärning och flyttade en pjäs på ett bräde. När romarna började bygga upp sitt herravälde var spelandet spritt till alla samhällsklasser och en del av det vardagliga livet. Det finns romerska lagar som reglerar och i viss mån förbjuder spel. Vid denna perioden i historien d.v.s. århundradena före kristi födelse har spelen också blivit mer och mer avancerade och byggde säkert inte till hundra procent på chans längre. Man har hittat uppgifter på att det är olika svårt att få benet att landa på en viss sida. Man tror dock inte att man tog hänsyn till detta i spelens utförande. Detta gav i sin tur öppningar för de spelare som hade vetskap om att vissa sidor förekom oftare än andra. Detta kan då tolkas som tidiga spelteorier. Det bör också nämnas att även om astragalus förekom länge i historien utvecklades också olika former av tärningar, vissa liknande dem vi använder idag. Runt kristi födelse finns det redan skriven litteratur runt spelandet. Claudius (10 f.k. - 54 e.k.) skrev "How to win at Dice". Boken är tyvärr inte bevarad, men man är ganska säker på att boken i någon form behandlade ämnet spelteori.

1.1.2 1000-1600

Runt millenieskiftet började man komma så långt i sannolikhetsläran att man började fundera på hur många sätt tre tärningar kan falla när man kastar dem. Resonemanget om hur tre tärningar kan hamna skriver Richard de Fournival om i början av 1200-talet, men det är sannolikt att vetskapen om sådana här problem fanns redan innan. Att räkna på sannolikheter att välja ut t.ex. a saker ur någon grupp fanns inte under de Fournivals tid, man vet dock att redan grekerna försökte sig på sådana beräkningar. På 1200-talet hade man lite vetskap om binominalkoefficienterna men hade troligen stora problem med de teoretiska sanningarna bakom.

Facio Cardano (1444-1524) var advokat och undervisade bl.a. på universitetet i Milano. Han var också intresserad av sannolikheter och skrev boken ”liber de Ludo Aleae” (Boken om spel och chans). I denna bok gör han en hel del kommentarer om sannolikhet. Han skriver mycket om tärningsspel med två och tre tärningar och visar på olika sannolikheter i dessa spel. Boken innehöll en kort historisk tillbakablick på spelandet. Cardano var den första som beräknade teoretiska sannolikheter på ett korrekt sätt.

Under perioden 1550-1650 skrev många matematiker om spel och chans runt om i Europa. Att alla dessa skulle vara oberoende av varandra är inte

troligt. Bl.a. förekommer binominalkoefficienterna i många skrifter. Den enda matematikern under den här tiden som egentligen kom fram till något nytt var Galileo-Galilei (1564-1642). Någonstans mellan åren 1613-1623 skrev han "Sopra le Scoperte dei Dadi". I denna bok behandlar han sannolikheten att få olika siffersummor när man kastar tre tärningar. Han skriver bl.a. om att det är svårare att få summan 3 och 18 än att få t.ex. 6 och 7. Detta eftersom 3 och 18 bara kan fås på ett sätt, medan 6 och 7 kan fås på flera olika sätt. Han kommmer också fram till att när man kastar tre tärningar kan man få 216 olika utslag (6^3). I boken finns det också en tabell som visar på sannolikheten för att olika summor ska komma upp.

1.1.3 Fermat och Pascal

Nästa stora namn i historien är Fermat och Pascal. De största verk av dessa två personer inom spel och chans området kan man säga är breven som de skrev mellan sig. I dessa diskuterar de olika former av spel och bl.a. hur man ska fördela vinster om man skulle välja att avbryta spelet innan det är färdigt. Nedan ska vi ge ett exempel.

Pascal skriver i ett brev till Fermat 1656 om ett två personers spel. Varje person har satsat 32 pistoler och första spelaren till tre poäng vinner allt. Om hur själva spelet exakt går till nämns inte i brevet. Det är dock ett spel där båda spelarna har lika stor chans att vinna nästa poäng, ex. flippa krona. Brevet innehåller disskusioner om hur man ska fördela insatsen om man önskar avbryta spelet vid något tillfälle.

Pascal börjar med när spelare A har två poäng och spelare B en poäng. Om spelare A får nästa poäng vinner han alltså 64 pistoler, medan om B vinner har de båda två poäng och man kan säga att de har 32 pistoler var, eftersom de då har lika stor chans att vinna. Spelare A har alltså minst 32 pistoler oavsett om han får nästa pistol eller inte. Spelare A får då sina 32 pistoler och de återstående 32 borde de dela lika, d.v.s. spelare A får 48 pistoler och spelare B 16 pistoler. Brevet fortsätter med samma resonemang vid de andra ställningarna. Han kommer nu fram till mer generella lösningar på dessa typer av problem. Brevet innehåller också tabeller på olika fördelningar på samma typer av spel. Han skriver också om sannolikheter för problem som t.ex. hur stor är chansen att få en sexa på n försök.

1.1.4 1600-1800

Christian Huygens (1629-1695) är nästa stora namn i historien. Troligtvis var det så att Huygens inte hade vetskap om vad Pascal och Fermat hade kommit fram till i sin brevväxling. År 1657 kom "de Ratiociniis in Aleae Ludo" ut. Denna bok blev den ledande boken inom sannolikhetsläran i över ett halvt århundrande. I boken presenterar Huygens 14 teser. Dessa innehåller bl.a. sådant som Pascal och Fermat kommit fram till.

Den kände matematikern James Bernoulli intresserade sig tidigt för sannolikhetsproblem. Han läste Huygens bok "De Retiociniis in Aleae Ludo" och det är inte omöjligt att de också träffades. I början av 1680-talet började han även arbeta med vissa problem som publicerades i "Journal de Scavans". Ett känt exempel är:

A och B spelar tärning, den första som får en etta vinner. Spelets spelas så att A kastar först en gång följt av att B kastar en gång. Det fortsätter med att A kastar två kast och sedan B två kast o.s.v. tills någon får en etta. Hur stor är sannolikheten att A vinner?

Lösningen till detta problemet publicerade han inte förrän 1690 i "Acta Eruditorum". Hans mest kända verk på området är "Ars Conjectandi" som av olika anledningar inte publicerades förrän 1713, alltså åtta år efter James död.Boken innehåller flera exempel där han använder modern matematik på ett mycket kraftfullt sätt.

Det är ganska tydligt att han hade tänkt använda sin bok som material i sin undervisning eftersom den inte bara innehöll lösningar till problemen utan även kommentarer till möjliga fel man kunde göra. Någon som byggde vidare på James Bernoulli's arbete var Pierre Re'mond de Montmort. Han var väl insatt i James arbete och arbetade under vissa perioder ihop med en av James studenter nämligen hans bror Nicholas Bernoulli. År 1708 publicerades den första upplagan av "Essai Analyse". Den innehöll lösningar till James Bernoulli's problem men med lite moderna matematik. Montmort kollade också på olika former av kortspel och kom fram till mycket kraftfulla saker. Abraham de Moivre (1667-1754) var den man som skulle komma att utveckla Bernoullis och Montmorts arbeten mer. De Moivre arbetade i England och var också mycket god vän med Newton. Han publicerade den första upplagan av "Doctrine of Chances" 1718. Denna innehöll en del nytt och lite modernisering av tidigare arbeten. Det bör nämnas att de Moivre var en av de allra bästa matematikerna i världen under denna tiden. Många menar att den tredje upplagan 1756 av "Doctrine of Chances" är den första riktiga moderna boken om sannolikhet.

1.2 Spelteorins historia inom ekonomin

Inom ekonomivetenskapen är spelteorin en gren som handlar om samspelet mellan ekonomiska parter. Kortfattat kan spelteori beskrivas som en teori om under vilka förhållanden olika säljare på en marknad väljer att byta strategi för att sälja sina produkter och under vilka förhållanden det råder jämvikt mellan olika leverantörer. Begrepp som kartellbildning och orsakerna till att bryta sig ur en sådan kan också beskrivas med hjälp av spelteori.

Området kan sägas ha startat på riktigt i och med den stora spridningen av John von Neumanns och Oskar Morgensterns bok "Theory of Games and Economic Behavior" (1944). John von Neumann brukar kallas för spelteorins fader. I boken infördes bland annat teorin om förväntad nytta, maximin- och minimax-principerna.

"Theory of Games and Economic Behavior" är en klassisk bok som dagens spelteori är baserad på. Det som för mer en sextio år sedan började som ett anspråkslöst förslag i en artikel av John von Neumann och Oskar Morgenstern blev tillslut en bok på 616 sidor. I boken föreställde sig John von Neumann och Oskar Morgenstern ett genombrott av ekonomisk och samhällsvetenskaplig organisation baserad på en matematisk teori, spelteori. Boken kom att revolutionera nationalekonomin och spelteorin har sedan dess blivit vitt använd för att analysera mängder av verklighetsbaserade problem från vapenfrågor till optimalt politiskt val av presidentkandidater. Spelteorin är idag etablerad såväl inom den sociala vetenskapen som inom ett stort omfång av andra vetenskaper.

I och med boken så blev slutet av 40-talet och början av 50-talet en spännande period inom spelteorin. På Princeton gjorde John Nash grundjobbet bland annat för den generella non-kooperativa spelteorin. Lloyd Shapley definierade värdet för koalitions spel, tog initiativet till teorin om stokastiska spel och tillsammans med John Milnor utvecklade de första spelmodellerna med oändligt antal spelare. Harold Kuhn introducerade begreppen om beteende, strategi och perfekta spel. A. W. Tucker hittade på historien om "fångens dilemma".

Von Neumanns och Oskar Morgensterns bok accepterades först inte helt av ekonomerna. Det tog nästan ett kvarts sekel innan ekonomerna förstod att man hade stor nytta av spelteorin inom bland annat militära tillämpningar. I och med detta tog spelteorin en stor del inom ekonomisk teori.

1.2.1 John Nash

1994 fick John. F. Nash Jr tillsammans med John Harsanyi och Reinhard Selten nobelpriset i ekonomi för sina tillämpningar av spelteorin inom ekonomisk vetenskap (spelteoretiska jämviktsbegrepp). Nash klargjorde skillnaden mellan kooperativ och non-kooperativ spelteori och utvecklade för den senare det jämviktsbegrepp som har kommit att kallas Nash-jämvikt.

Nash-jämvikt avser en strategikombination som utmärks av att ingen spelare kan förbättra sitt utfall genom att byta strategi och kännetecknas av att ingen av de inblandade aktörerna skulle ha anledning att ändra sitt eget beslut om de i förväg känt till övriga aktörers beslut. Nash-jämvikt kan sägas vara en generalisering av de jämviktsbegrepp som tidigare införts.

Nash formulerade också en matematisk modell för förhandlingar och ett lösningsbegrepp, Nashs förhandlingslösning, som båda har varit mycket fruktbara både inom spelteori och inom ekonomisk analys. Nashs insats inom spelteorin har drag av genialitet. Vid tjugo års ålder hade han formulerat de begreppsliga verktyg som gjorde det möjligt att förverkliga spelteorins ursprungliga syfte att tjäna som underlag för den ekonomiska analysen. Ett tiotal år senare avbröts Nashs forskargärning av en lång period av psykisk ohälsa.

1.2.2 Efter Nash

Sedan John Nashs epokbildande arbeten från början av 1950-talet går det en skiljelinje mellan en kooperativ och en non-kooperativ gren inom spelteorin. Den kooperativa grenen, som hade en storhetsperiod under 1950- och 60-talen, försöker bestämma lösningar till spel genom att studera de utfall som olika grupper av spelare kan åstadkomma genom förhandlingar och koalitionsbildningar. Den non-kooperativa spelteorin intresserar sig i stället för den process som leder från spelarnas ömsesidiga förväntningar till deras strategival. Denna process resulterar ofta i något slags jämvikt, och en viktig fråga i sammanhanget är om denna jämvikt går att förena med ekonomisk effektivitet.

Den non-kooperativa ansatsen har varit klart dominerande sedan slutet av 1970-talet, delvis tack vare viktiga arbeten av John Harsanyi och Reinhard Selten. Selten har utvecklat Nashs jämviktsbegrepp för analys av dynamiska spel i vilka spelarna agerar vid flera olika tidpunkter. Han har visat att vissa Nash-jämvikter i sådana spel bygger på icke-trovärdiga hot, dvs. spelarnas strategier innehåller hot som de inte är beredda att verkställa. Genom att ställa upp ett starkare villkor, perfekt jämvikt, ville Selten garantera jämviktsstrategiernas

trovärdighet. Harsanyi utvecklade en metod för att analysera situationer med ofullständig information, t.ex. om spelarnas målsättningar, vilket har möjliggjort en ökad realism i tillämpningen av spelteorin på ekonomiska fenomen.

1.2.3 Tillämpningar inom biologin

Inom biologi används spelteori främst som metod för analys av de evolutionära konsekvenserna av komplex växelverkan mellan individer, företrädesvis av samma art. Varje individ antas tillämpa en uppsättning regler "en strategi" som är genetiskt betingad men som samtidigt kan vara mer eller mindre flexibel, dvs. olika beteenden kan tillämpas i olika situationer. Kostnader och vinster för alternativa strategier definieras som minskad respektive ökad fitness (framgång i fråga om fortplantning och överlevnad); dessa beror i det typiska fallet av vilka strategier de andra individerna tillämpar. En stabil lösning, så kallad ESS (evolutionärt stabil strategi), som är ett exempel på Nash-jämvikt, är en strategi som inte kan slås ut av alternativa strategier. Exempel på biologiska problem som analyserats spelteoretiskt är kampbeteenden hos djur, uppkomsten av stabila sociala konstellationer (t.ex. fleråriga familjegrupper eller livslångt partnerskap) och evolution av flyttnings- och spridningsbeteenden.

2 Spelteori

2.1 Två-personers noll-summa spel

Det anses generellt att så kallade två-personers noll-summa spel har åstadkommit mycket inom modellformullering och analys. Alla spel vi beskriver här involverar två personer som betecknas P_1 respektive P_2. Termen noll-summa betyder helt enkelt att vad P_1 vinner, förlorar P_2.

2.1.1 Ex 1 , UDDA, JÄMN

Regler.

Bägge spelare väljer antingen talet 1 eller 2 utan att veta den andres val. Då valen offentliggörs så betalar antingen P_1 en krona till P_2 eller betalar P_2 en krona till P_1 beroende på om summan av de valda talen är udda eller jämn.
Eftersom P_1 antingen kan spela på 1 eller på 2, så formulerar vi dessa båda "strategier" som s_1 respektive s_2. För P_2 skriver vi t_1 respektive t_2 och kan således ställa upp vad vi kallar en "vinstmatris",

	t_1	t_2
s_1	1	-1
s_2	-1	1

där talet på raden i och kolonnen j representerar summan som P_1 vinner från P_2 då P_1 spelar s_i och P_2 spelar t_j.
Flera spel, däribland "krona eller klave" ger identisk vinstmatris och betraktas därför som samma spel.

2.1.2 Ex 2 , MORRA

Regler.

Spelarna ska samtidigt visa ett eller två fingrar mot varandra och på samma gång säga varsitt nummer. Om detta nummer för en av spelarna är det samma som summan av båda spelares fingrar, så vinner denna spelare summan av fingrarna. Om bägge spelare gissar rätt/fel så blir det ej någon betalning. Spelarna har således fyra möjliga strategier. Spelaren kan visa ett finger och gissa på två eller tre, eller kan han visa två fingrar och gissa på tre eller fyra.
s_{ij} och t_{ij} får här betydelsen "visa i fingrar och gissa j". Vinstmatrisen blir:

	t_{12}	t_{13}	t_{23}	t_{24}
s_{12}	0	2	-3	0
s_{13}	-2	0	0	3
s_{23}	3	0	0	-4
s_{24}	0	-3	4	0

Vi formulerar och definierar vad vi sett hittills.

2.1.3 Definition

Ett två-personers noll-summa spel Γ består av ett par av mängder S och T och en reell-värd funktion ϕ definierad av par (s,t) där $s \in S$ och $t \in T$.
Terminologi. Elementen s och t av S och T kallas *strategier* för spelarna P_1 respektive P_2. Funktionen ϕ kallas *vinstfunktionen* eller bara *vinsten.*
Tolkning. Talet $\phi(s,t)$ är summan som P_2 betalar P_1 då P_1 spelar strategin s och P_2, strategin t. Termen noll-summa som tidigare nämnts betyder helt enkelt att vad p_1 vinner, förlorar P_2.
Beteckningar. Vi beskriver spelet Γ med strategi-mängder S och T, och vinstfunktion ϕ genom att skriva $(S,T;\phi)$. Om mängderna S och T är ändliga, så kan vinsten ϕ beskrivas av en matris. Dessa spel kallas *matrisspel.*

2.1.4 Ex 3 , BLUFF

Regler.

Man lägger två kort märkta HÖG och LÅG i en hatt och P_1 drar ett kort och tittar på detta. Notera att det endast är P_1 som drar ett kort under hela spelets gång. P_2 vet alltså ej vilket kort han besitter. P_1 kan då antingen lägga sig s_l eller satsa s_s. Då han lägger sig får han betala P_2 summan $a > 0$ om han har LÅG, och då P_1 satsar måste P_2 antingen lägga sig och betala P_1 summan a eller syna t_s. Om P_2 synar får han betala eller ta emot summan $b > a$ från P_1, beroende på vilket kort han har.
Vi kan direkt förkasta en strategi för P_1, nämligen att lägga sig då han får HÖG. Vi illustrerar reglerna något lättare nedan.

- P_1 lägger sig
  - P_2 lägger sig
    - P_1 vinner a (HÖG)
    - P_1 förlorar a (LÅG)
  - P_2 synar
    - P_1 vinner b (HÖG)
    - P_1 förlorar a (LÅG)
- P_1 satsar
  - P_2 lägger sig
    - P_1 vinner alltid a
  - P_2 synar
    - P_1 vinner b (HÖG)
    - P_1 förlorar b (LÅG)

När vi bygger upp vinstfunktionen måste vi ta hänsyn till förväntningar. Om P_1 spelar s_s och P_2 spelar t_s och eftersom P_1 har HÖG eller LÅG med sannolikheten $1/2$ så är förväntade vinsten i det läget noll.
Om P_1 spelar s_s och P_2 spelar t_l vinner P_1 summan a oberoende av korten.
Om P_1 spelar s_l och P_2 spelar t_s så vinner P_1 summan b då han har HÖG,

vilket är hälften av tiden, och förlorar a den andra hälften. Förväntningen blir alltså $\frac{b}{2} - \frac{a}{2} = \frac{b-a}{2}$.

	t_s	t_l
s_s	0	a
s_l	$\frac{b-a}{2}$	0

2.2 Lösningar till spel (Blandade strategier)

Vi fortsätter nu med att lösa våra exempel. Vi betraktar till att börja med en spelare P_1 som ska spela UDDA, JÄMN. Han vet ingenting om i fall P_2 lutar åt att spela en strategi över en annan. Under dessa villkor kan P_1 inte förvänta sig bättre än att gå jämnt upp, och för att göra detta är det lättast för P_1 att bara kasta ett mynt och välja strategin s_1 för krona och s_2 för klave. Det är dock viktigt att kronan har just sannolikheten $1/2$ att visa respektive strategi, ty om P_1 skulle spela s_1 med större sannolikhet än $1/2$, så skulle P_2 erhålla en positiv förväntning genom att spela t_2.
Sammanfattningsvis kan vi säga att det optimala förfarandet för P_1 (och P_2 av symmetriskäl) är att slumpvis välja en strategi, men med lika stor sannolikhet för den andra. Lösningen kan för många anses som trivial, men är bra som grund.
Vi betraktar således spelet MORRA vars lösning ej är lika självklar.
Vinstmatrisen hade följande utseende:

	t_1	t_2	t_3	t_4
s_1	0	2	-3	0
s_2	-2	0	0	3
s_3	3	0	0	-4
s_4	0	-3	4	0

Eftersom spelet är symmetriskt, kan vi bara välja ett tillvägagångssätt som ger P_1 en förväntning av att ej förlora, därför att P_2 bara skulle kunna "kopiera" ett tillvägagångssätt vilket skulle innebära att vinsten blir noll. Vi fastställer att ett sådant sätt för P_1 är att välja mellan s_2 och s_3 med sannolikheten $3/5$ för s_2 och $2/5$ för s_3. Vi tittar nu på vad som händer med denna procedur då P_2 väljer strategi.
Antag att P_2 väljer att spela t_1. Mot denna strategi har P_1 en sannolikhet på $3/5$ att förlora 2 och en sannolikhet på $2/5$ att vinna 3. Vinsten blir således:

$$\frac{3}{5}(-2) + \frac{2}{5}(3) = 0$$

Då P_2 väljer att spela t_2 eller t_3, står det klart att ingen av dem vinner.
Slutligen om P_2 spelar t_4 så vinner P_1 3 med sannolikhet $3/5$ samt förlorar 4 med sannolikhet $2/5$. Förväntad vinst blir:

$$\frac{3}{5}(3) + \frac{2}{5}(-4) = \frac{1}{5}$$

P_1 kan alltså förvänta sig en vinst på minst noll och proceduren blir optimal.

Vi kan faktiskt visa att mängden optimala procedurer är oändligt många.
Låt P_1 välja mellan s_2 och s_3 och sannolikheten för att välja s_2 vara ett tal mellan $4/7$ och $3/5$. Vinsten blir som tidigare noll mot t_2 och t_3. Mot t_1 får vi:

$$p(-2) + (1-p)(3) = -5p + 3 \geq 0$$

ty $p \leq \frac{3}{5}$

och mot t_4:

$$p(3) + (1-p)(-4) = 7p - 4 \geq 0$$

ty $p \geq \frac{4}{7}$

2.2.1 Definition

Låt Γ vara spelet $(S,T;\phi)$. En *blandad strategi* σ för P_1 är en reellvärd funktion definierad över S sådan att:

$$\sigma(s) \geq 0 \tag{1}$$

och

$$\sigma(s) = 0 \tag{2}$$

för alla utom en ändlig summa strategier s av S,
och slutligen

$$\sum_{s \in S} \sigma(s) = 1 \tag{3}$$

Analogt gäller för en blandad strategi τ för P_2. Mängden av alla blandade strategier för P_1 och P_2 skriver vi som <S> respektive <T>.
Tolkning. Talet σ(s) representerar sannolikheten för att P_1 ska spela strategin s. För att verkligen skilja mellan blandade strategier, σ och τ, och vad vi tidigare definierat som strategier, s och t så kallar vi det senare för rena strategier.

Vi utvidgar nu vår vinstfunktion till blandade strategier.

$$\phi(\sigma, \tau) = \sum_{s,t} \sigma(s)\tau(t)\phi(s,t)$$

Uttrycket är det förväntade värdet av vinsten då P_1 spelar σ och P_2 spelar τ. Vi betecknar vår utvidgade funktion på följande sätt:

$$< \Gamma >= (< S >, < T >; \phi)$$

För matris-spel betecknar man gärna blandade strategier på ett speciellt sätt. Om S och T innehåller m respektive n strategier skriver man dessa som $s_1, \cdots, s_m$ och $t_1, \cdots, t_n$,
samt skriver vinstfunktionen $\phi(s_i, t_j)$ som a_{ij}. En blandad strategi för P_1 ges av en m-vektor $x = (\xi_1, \cdots, \xi_m)$, där ξ_i är sannolikheten för s_i. På samma sätt är $y = (\eta_1, \cdots, \eta_n)$ för P_2.
Förväntade vinsten blir:

$$\phi(x,y) = \sum_{i,j} \xi_i a_{ij} \eta_j = xAy$$

där A är vinstmatrisen för spelet.
En lösning till spelet Morra blir nu en blandad strategi

$$\bar{x} = (0, p, 1-p, 0) \ , \ \tfrac{4}{7} \leq p \leq \tfrac{3}{5}$$

Vi visar att detta är de enda blandade strategierna med en förväntning på minst noll oavsett strategi från P_2. Detta gör vi genom att med hjälp av utvalda strategier för P_2, tvinga P_1 till att spela på detta sätt.
Vinstmatrisen för spelet var:

	t_1	t_2	t_3	t_4
s_1	0	2	-3	0
s_2	-2	0	0	3
s_3	3	0	0	-4
s_4	0	-3	4	0

2.2.2 Bevis

Låt $x = (\xi_1, \xi_2, \xi_3, \xi_4)$. Vi börjar med att bestämma ξ_1 och låter $y = (0, \frac{4}{7}, \frac{3}{7}, 0)$. Vinstfunktionen blir:

$$\begin{gathered} \phi(x,y) = \tfrac{4}{7}(2\xi_1 - 3\xi_4) + \tfrac{3}{7}(-3\xi_1 + 4\xi_4) = -\tfrac{1}{7}\xi_1 \\ \Longrightarrow \xi_1 = 0 \\ \text{, ty } \xi_1 \geq 0 \text{ och } \phi(x,y) \geq 0 \end{gathered}$$

På samma sätt för ξ_4, låt $y = (0, \frac{3}{5}, \frac{2}{5}, 0)$. Då får vi:

$$\begin{gathered} \phi(x,y) = \tfrac{3}{5}(2\xi_1 - 3\xi_4) + \tfrac{2}{5}(-3\xi_1 + 4\xi_4) = -\tfrac{1}{5}\xi_4 \\ \Longrightarrow \xi_4 = 0 \\ \text{, ty } \xi_4 \geq 0 \text{ och } \phi(x,y) \geq 0 \end{gathered}$$

Sedan om $\xi_2 > \frac{3}{5}$, låt P_2 spela t_1 vilket ger

$$\phi(x, t_1) = -2\xi_2 + 3(1 - \xi_2) = -5\xi_2 + 3 < 0$$

och slutligen om $\xi_2 < \frac{4}{7}$ låt P_2 spela t_4 vilket ger

$$\phi(x, t_4) = 3\xi_2 - 4(1 - \xi_2) = 7\xi_2 - 4 < 0$$

V.S.B.

2.3 Värdet av ett spel

Vi tittar nu på spelet BLUFF vars matris var

	t_s	t_l
s_s	0	a
s_l	$\frac{b-a}{2}$	0

Eftersom spelet ej är symmetriskt är det heller ej uppenbart vilken vinst P_1 kan förvänta sig. Antag nu att P_1 spelar blandade strategin $\bar{x} = (\frac{b-a}{b+a}, \frac{2a}{b+a})$. Vad blir då hans förväntade vinst?
Vi sätter upp vinstfunktionen och kollar:

$\phi(\bar{x}, y) = \frac{b-a}{b+a}(0 \cdot \eta_1 + a\eta_2) + \frac{2a}{b+a}(\frac{b-a}{2}\eta_1 + 0 \cdot \eta_2) = \frac{a(b-a)}{b+a}\eta_2 + \frac{a(b-a)}{b+a}\eta_1 = \frac{a(b-a)}{b+a}(\eta_2 + \eta 1) = \frac{a(b-a)}{b+a}$

P_1 kan alltså förvänta sig $\omega = \frac{a(b-a)}{b+a}$ oavsett strategi från P_2. Med andra ord kan P_1 försäkra sig om att få minst ω vilket är ett positivt tal, ty $b > a$. Vi ser att det är lönsamt för P_1 att bluffa.
Efter detta kan man visa att P_1 ej kan garantera en högre vinst än ω genom att låta P_2 spela $\bar{y} = (\frac{2a}{b+a}, \frac{b-a}{b+a})$. Om P_2 spelar $\bar{y}$ kommer nämligen P_1 att vinna ω oavsett strategi. Då ω är högsta garanterade vinst för P_1, ser vi att ω spelar samma jämvikts-roll som noll gjorde i de föregående exemplen. Detta kallas spelets värde. Alltså, de blandade strategierna $\bar{x}$ och $\bar{y}$ kallas optimala och ω kallas spelets värde. Tillsammans utgör de lösningen $(\bar{x}, \bar{y}, \omega)$.

2.4 Kombinatorisk spelteori

Ett kombinatoriskt spel är ett spel som spelas mellan två eller flera personer. Spelet är kooperativt till sin natur, d.v.s alla spelarna har information om förutsättningarna när de gör sitt drag, och deltagarna gör inga drag samtidigt. Slumpen är utsluten. Exempel på kombinatoriska spel är schack, othello, nim och "först till tjugo".

2.4.1 Först till tjugo

Först till tjugo spelas mellan två spelare, där den som säger tjugo först vinner spelet. Spelarna räknar i tur och ordning från 1 till 20 och varje spelare har att välja mellan att räkna ett eller två tal framåt. Den förste spelaren kan exempelvis säga "ett" eller "ett, två". Om han skulle säga "ett" kan motspelaren säga "två" eller "två, tre". Om den förste spelaren istället säger "ett, två" kan den andre spelaren säga "tre" eller "tre, fyra", och så vidare. Om någon av spelarna kan säga "sjutton" först så kommer han att ta hem spelet, då måste motspelaren säga antingen "arton" eller "arton, nitton". Ska man vara säker på att vinna spelet måste man hålla redan på talen 2,5,8,11,14,17 och 20. Om man dividerar dessa tal med 3 så ger de rest 2. Alla dessa tal är med andra ord kongruenta med 2 modulo 3. Vårt vinststrategi är att sluta räkna på ett tal kongruent med 20 modulo 3. Om vi istället sätter vinststal till N, och k är maximal antal steg som man kan gå framåt. Vinsttaktiken är då att sluta räkna på tal som är kongruenta med N modulo k+1.

3 Beslutsteori

> *Nytta - Något som har fördelaktig, kvarstående verkan på visst område, för viss person eller verksamhet etc.*
>
> *Nationalencyklopedin*

Utgångspunkten för alla typer av beslutsfattande är att det finns flera olika handlingsmöjligheter. Varje möjlig utfallsmöjlighet värderas. Oftast görs detta med hjälp av referenser i vardagslivet, så som tex pengar eller tid. Då detta inte är möjligt kan en mer teoretisk värderingsskala användas, den så kallade nyttoskalan. I enklare fall kan det vara så att ett visst beslut direkt ger dig en form av nytta, medan det i andra fall kanske krävs att ditt beslutsalternativ i sin tur genererar något som ger dig nytta.

3.1 Beslutsprocess

"Du har bjudit hem två vänner på middag. Under tiden du duschar bränns maten vid. Den är nästan helt förstörd men du lyckas rädda kycklingfiléerna. Potatisen är dock helt förstörd. Gästerna kommer om en halvtimme. Vad gör du?"

A. Du slänger på dig kläderna, kastar dig på cykeln och åker och köper nya potatisar i närmsta affär.
B. Du ringer efter pizza.

Det finns många olika sätt att komma fram till ett beslut. Allra enklast är kanske att singla slant. Men det utfall man då får är kanske inte det man önskar. Då har vi istället ett annat alternativ, vi bestämmer oss för en förutsättning.

"Du vill bestämt ha något färdigt att bjuda på när dina vänner kommer."

Då måste du utesluta alternativ A eftersom detta tar för lång tid.

Detta exempel illustrerar en enkel form av beslutsfattande. Denna modell kan sammanfattas i tre punkter:
1. Konsekvenserna av varje handling betraktas.
2. Konsekvenserna värderas beroende på dess önskvärdhet.
3. Beslutet tags sedan med hjälp av värderingarna.

I första steget av beslutsprocessen värderas beslutets betydelse. Utifrån detta inhämtas och bearbetas information. Både möjligheten och viljan att inhämta information bestäms av hur stor betydelse beslutet har. Detta utgör steg två. Innan denna process ens har börjat har beslutet färgats av beslutsfattarens egna åsikter. Ett beslut kan således aldrig vara helt objektivt. I tredje steget kommer återigen beslutfattarens egna åsikter in i bilden då han eller hon skall välja ut det mest gynsamma alternativet. Här kan även andra metoder så som beslutsträd, diagram, tabeller och simuleringar användas.

För vårt exempel gäller följande:
1. Det tar lång tid att åka och köpa potatis, men det är dyrt att köpa pizzor.

2. Jag vet att mina vänner är hungriga och därför kommer dom uppskatta att maten är färdig. Om jag köper pizza får jag däremot mindre pengar kvar.
3. Jag anser att det är viktigare för mig att mina vänner blir nöjda än att jag har mycket pengar kvar efter middagen.

Hade du istället valt att vara så ekonomisk som möjligt så hade du valt alternativ A, eftersom detta blir mycket billigare och så hade dina vänner helt enkelt fått vänta på sin mat.

3.2 Beslut under säkerhet

Att fatta beslut under säkerhet innebär att alla möjliga utfall är kända. Det återstår därmed enbart för beslutsfattaren att räkna ut någon form av jämförelsetal för de olika utfallen och sedan rangordna dessa efter hans eller hennes egna intressen. Teoretiskt är detta inte svårt, men i praktiken kan det ibland vara svårt om utfallsalternativen är många.

3.3 Beslut under osäkerhet

I exemplet ovan vet vi att vilket alternativ vi än tar så kommer vi att få ett resultat som vi känner till. I andra situationer kan dock resultatet av ett visst val vara okänt. Man kan inte sätta upp utfallssannolikhter. Detta kan bero på att slumpen avgör resultatet av ett visst val, att man till exempel har naturen som motståndare eller att man är beroende av vad en annan person beslutar. En sådan situation kan till exempel vara följande:

"En fånge har rymt från fängelset. För att ta sig hem måste han passera ett bevakat område. Detta område är indelat i två på varandra följande zoner, zon A och zon B. Fången måste alltså ta sig igenom båda zonerna. Polisen spanar efter honom med helikopter. Det finns två helikoptrar i luften som spanar oberoende av varandra och dom bevakar antingen en zon var eller så är båda två inom samma zon. Han har möjlighet att lifta igenom en zon, men måste ta sig till fots igenom den andra. Att lifta innebär att han blir osynlig för poliserna i luften. Om han väljer att ta sig till fots genom en zon där det inte finns någon helikopter så är sannolikheten att bli upptäckt och gripen lika med noll. Är den ena av helikoptrarna i zonen är sannolikheten att bli upptäkt 0,2 i zon A och 0,4 i zon B. Är bägge planen i zon A blir sannolikheten att bli upptäckt 0,2 + 0,2 - 0,2· 0,2 = 0,36. Motsvarande i zon B är 0,4 + 0,4 - 0,4 · 0,4 = 0,64."

Rymlingen sätter upp ett schema över sin sannolikhet att bli gripen i de olika zonerna, beroende på hur helikoptrarna spanar.

	p Rymlingen synlig i zon A	1-p Rymlingen synlig i zon B	rad min
Helikopterspaning i zon A + zon A	0,36	0	0
Helikopterspaning i zon A + zon B	0,2	0,4	0,2
Helikopterspaning i zon B + zon B	0	0,64	0
kolumn max	0,36	0,64	

Med hjälp av detta kan han få fram den handling p, som gynnar honom mest. Detta kan ske antingen genom en grafisk lösningsmetod, eller genom teoretiskt räknande. En grafisk lösningsmetod blir dock inte exakt, så han väljer att genomföra det teoretiskt. Det förväntadet värdet betecknas EV (expected value). Följande beräkningar av EV görs:

EV(AA) = p · 0,36 + (1 - p) · 0 = 0,36p
EV(AB) = p · 0,2 +(1 - p) · 0,4 = 0,4 - 0,2p
EV(AA) = EV(AB) ger 0,36p = 0,4 - 0,2p vilket ger p = 5/7 och
EV(AA) = EV(AB) = 9/35

Rymlingens blandade strategi blir då 5/7 A + 2/7 B och hans förväntade sannolikhet att bli gripen när han tar sig igenom området blir 9/35. Detta innebär att han i snitt bör välja att ta sig till fots genom zon A och lifta igenom zon B.

3.4 Nyttoresonemang

Låt oss nu använda ännu en teori för beslutsfattande, nyttoresonemang. Detta innebär att man tilldelar alla valmöjligheter olika många"nyttopoäng".
I detta exemplet utgår vi från ett företag, vars chef ska fatta ett beslut om varifrån man ska köpa in en viss vara som krävs för att producera en av företagets produkter. Det finns tre olika möjligheter att välja mellan, dessa presenteras här som möjlighet A, B respektive C.

A. Här köps varan in från en fabrik i Kina där de tillverkas av dåligt avlönade barnarbetare. Priset är lågt, 5 kronor per vara och kvaliteten är okej. Frakten till Sverige är lång och man kan räkna med att 20 % av varorna går sönder.

B. Varorna köps in direkt från produktionen i östeuropa där arbetsförhållandena och lönen för arbetarna är dålig. Priset är 10 kronor per vara och kvaliteten är medelhög. Transporten till Sverige är inte lång men ändå går cirka 5 % av varorna sönder.

C. Varan tillverkas i Sverige där arbetsförhållandena håller god klass och för varan får man betala 15 kronor. Kvaliteten är mycket bra och under transpor-

ten går endast cirka 1 % av varorna sönder.

Chefens uppgift är nu att utifrån dessa fakta fatta ett beslut om vilket alternativ som ska användas. Detta gör han genom att sammanställa en tabell. Han värderar de olika aspekterna genom att tilldela dom poäng mellan noll och tio, där tio är det bästa. I exemplet får vi följande tabell och värden:

	A	B	C
Pris	10	7	4
Arbetsförhållande	0	4	10
Frakt	3	5	9
Kvalitet	2	6	9
	15	22	32

Ur tabellen kan man enkelt avläsa att alternativ C är det för företaget mest gynnsamma alternativet utifrån de aspekter man har tagit hänsyn till.
Även om företaget och beslutsfattaren i exemplet har fått ett väldigt klart besked om vilket av alternativen som är mest gynnsamt är det inte säkert att dom följer resultatet. Tvärtom är det mycket vanligt att företagsledare fattar beslut utifrån sin egen "magkänsla" och inte systematiskt.

Referenser

[1] F.N. David. *Games Gods Gambling.* 1962.

[2] John D. Beasly. *The Mathematics of Games.* 1990.

[3] Göran Widén. *Elementär beslutsteori.* 1974.

[4] Irwin D.J Bross. *Besluts planering.* 1966.

[5] Gunnar Blom. *Om spelteori och beslutsteori.* 1961.

[6] David Gale. *The Theory of Linear Economic Models.* 1960.

[7] Murray C. Kemp, Yoshio Kimura. *Introduction to Mathematical Economics.* 1978.

[8] John von Neumann, Oskar Morgenstern. *Theory of Games and Economic Behavior.* 1944.

Informationsteori

Ulf Johansson, Wilhelm Eklund, Benny Klein

10 juni 2005

Innehåll

1 Inledning

Informationsteorins utveckling tog fart 1948 då Claude E. Shannon publicerade sin artikel ‘A mathematical Theory of Communication‘. Den publicerades först i Bell Systems Technical Journal. Shannon definierade ett flertal begrepp inom information samt dess bevis. Shannon förklarade varför ett logaritmiskt system är det bästa då information skall beskrivas matematiskt. Definitionen av entropi som ett kommunikationssystems osäkerhet härrör ur Shannons artikel. Som ett direkt resultat av Shannons arbete har självrättande koder på cd- och dvdskivor utvecklats. All digital teknik som vi använder idag har Shannon att tacka för sin snabba utveckling sedan 50-talet.

2 Information

När ett meddelande skickas från en källa till en mottagare måste meddelandet kodas på något sätt så att det blir lättare att överföras. Hos mottagaren uppstår då ett grundläggande problem, när meddelandet ska avkodas finns det ett stort antal utfall. Vilket av de möjliga utfallen är det som skickades från källan?
Om det är ett ändligt antal utfall i mängden kan det antalet eller en funktion av det antalet anses vara ett mått på informationen som produceras då man väljer utfall i mängden. Av flera anledningar är en logaritmisk funktion den mest användbara. Ofta varierar tid, bandbredd linjärt logaritmiskt med antalet möjligheter. Dessutom känns det logiskt att två identiska kanaler har dubbel kapacitet att sända information. Även matematiskt är det mer användbart då operationerna blir lättare att genomföra logaritmiskt.

3 Kommmunikationssystem

Ett kommunikationssystem består huvudsakligen av fem delar.

1. En informationskälla som producerar ett eller flera meddelande som ska skickas till en destionation. Meddelandet kan vara olika typer, en sekvens med bokstäver, en funktion av tiden (radio) eller en funktion som beror av flera variabler (svart vit TV), flera funktion av flera variabler (färg TV).

2. En sändare som överför meddelandet till en signal som kan sändas över kanalen.

3. En kanal som vidarbefordrar signalen från sändare till mottagare. Det kan vara sladdar, radiofrekvenser, ljud, ljus etc.

4. En mottagare som överför meddelandet från signalform till det ursprungliga meddelandet.

5. Den destination som meddelandet var avsett för.

Då signalen sänds i kanalen kan det finnas störningar som förändrar signalen. Trots det ska mottagaren kunna översätta signalen till det ursprungliga meddelandet.

Shannon valde att dela in kommunikationssystemen i olika kategorier: diskreta, kontinuerliga och och ett tredje som är en blandning av de två första. Det diskreta kommunikationssystemet karakteriseras av att såväl meddelande som signal är en sekvens av diskreta symboler. Ett exempel är telegrafi där det ursprungliga meddelandet är bokstäver och signalen som skickas består av punkter och mellanrum. I den här rapporten behandlar vi uteslutande diskreta kommunikationssystem. Ett kontinuerligt system får

vi då både meddelande och signal utgörs av kontinuerliga funktioner. Vid den tredje kategorin utgörs meddelande och signal av både diskreta symboler och kontinuerliga funktioner.

4 Diskret kanal

Telegrafi är en diskret kanal för att sända information. Shannon definierade en diskret kanal som en sekvens av val från en ändlig mängd $S_1, ...S_n$ kan sändas från en sändare till en mottagare. Varje symbol har en viss utbredning i tid då det sänds över kanalen, det behöver dock inte vara detsamma för alla symboler. I ett system behöver inte alla sekvenser av S_n vara tillåtna. Men hur mäter man då kanalens möjlighet att överföra information?
När det gäller fjärrskrift (som var den första enheten för kommunikation mellan människa och dator) och likt telegrafi är en diskret kanal använder man sig av att att det finns 32 symboler till hands. Varje symbol representeras fem bitar av information om systemet då sänder n symboler per sekund är kanalens maximala kapacitet $5n$ bitar per sekund. Huruvida systemet kommer upp till den kapaciteten beror på informationskällan.
Shannon generaliserade begreppet och definierade kanalkapaciteten av ett system med olika långa symboler och tillåtna sekvenser som

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

där $N(T)$ är antalet tillåtna signaler under tiden T.

En diskret informationskälla kan betraktas som en markovprocess och osäkerheten i systemet kallas för dess entropi. Dessa begrepp ska diskuteras närmre i rapporten nu.

5 Entropi

5.1 Definition av entropi

För att mäta täthet av information börjar vi med att mäta osäkerheten i ett slumpexperiment. Sen visar vi att entropi är en bra mått för information mängden som man kan få genom att utföra experimenten.

5.1.1 Några begrepp inom sannolikhetslära

Anta att vi har ett slump experiment, t.ex. att kasta ett mynt eller en tärning. Experimentet har ett ändligt antal möjliga resultat och i varje försök får vi ett enda resultat. Ett ändligt utfallsrum är mängden av alla dessa (n) utfall. En delmängd av rummet heter en händelse, t.ex. händelsen att få 1 i ett tärningskast ingår i utfallsrummet och är en utfall. Händelsen att få 7 i en 6 sidig tärning är en omöjlig händelse - en tom delmängd i utfallsrummet. Ett delmängd som innehåller ett eller inget utfall kallas elementär händelse. Varje händelse har en sannolikhet. En diskret stokastisk variabel:

$$X \quad = \quad x_1, x_2, .., x_i, .., x_n \;\; 1 \leq i \leq n$$

är en avbildning från utfallsrummet till de reella talen. Detta betyder att man förvandlar händelsen till ett tal. Varje utfall har en sannolikhet, och därför har också varje värde av X en sannolikhet $0 \leq p \leq 1$. Om $X = x_i$ blir $p(x_i)$ sannolikheten för händelsen. Mer allmänt:

$$p(X) \quad = \quad (p(x_1), p(x_2), \ldots, p(x_n)) = (p_1, p_2, \ldots, p_n)$$

Det gäller att summan av sannolikheterna $\sum_{k=1}^{n} p_k = 1$.
Väntevärdet $E(X)$ definieras av:

$$E(X) = \sum_{k=1}^{n} p_k F(X_k) \tag{1}$$

där $F(X_k)$ är en reellvärd funktion. E(X) betyder ett medelvärde av resultaten av experimenten om man använder funktionen F, eller det förväntade värdet av experimenten.

5.1.2 Entropin och dess egenskaper

Titta på följande utfallsrum:

$$A = \begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix}$$

I utfallsrummet A kan vi inte förutse resultatet. Vi kan säga att det finns en stor osäkerhet i experimenten. Men i fallet

$$B = \begin{pmatrix} B_1 & B_2 \\ 0.95 & 0.05 \end{pmatrix}$$

kan vi med stor säkerhet förutse resultatet. Här finns lite osäkerhet. Slutligen i

$$C = \begin{pmatrix} C_1 & C_2 \\ 1.0 & 0.0 \end{pmatrix}$$

så inneholler utfallsrummet C ingen osäkerhet alls, resultaten av experimenten är entydig. C är ett system/experiment som har ett enda läge/resultat.
Man kan visa att ett bra mått på osäkerheten är entropin:

$$H(p_1, p_2, \ldots, p_n) = -\sum_{k=0}^{n} p_k \log(p_k) \tag{2}$$

där vi definierar $p_k \log(p_k) = 0$ om $p_k = 0$. Basen till logaritmen är konstant men godtycklig. I kommunikations sammanhang är det populärt med basen 2 eftersom den ger en bekväm bas för binära/digitala beräkningar. Några exempel:

1. Entropin av mängden A är

   $$H(A) = -0.5\log(0.5) - 0.5\log(0.5) = -\log(\frac{1}{2}) = \log(2)$$

   Om en mängd A innehåller n sannolikheter med konstant sannolikhet $p_k = \frac{1}{n}$ (t.ex en tärning eller ett mynt) och mängden B innehåller också n utfall men med sannolikheter som inte är lika blir $H(A) = \log(n) > H(B)$. Detta ska vi visa allmänt nedan. Uttryckt på ett annat sätt: Om A är en ändlig utfallsrum med n möjliga resultat och n sannolikheter blir $H(A) \leq \log(n)$

2. Entropin för mängd B är

$$H(B) = -0.95\log(0.95) - 0.05\log(0.05)$$

Där $-0.95\log 0.95$ är mycket nära noll, och

$$\begin{aligned} -0.05\log(0.05) &= 0.05\log(\frac{100}{5}) \\ &= 0.05\log(20) = 0.05\log(4) + 0.05\log(5) \\ &= 0.1\log(2) + 0.05\log(5) = 0.15\log(2) + 0.05\log(2.5) \end{aligned}$$

som vi kan jämföra med resultaten för entropin av mängd A.

3. Entropin för mängd C:

$$H(C) = -0\log(0) - 1\log(1) = 0$$

När man utför experimenten eliminerar man osäkerheten och får ett entydigt resultat. Man kan säga att genom utförandet av experimenten erhåller man information. Om osäkerheten i experimenten är liten kan man förutse resultaten och då erhåller man lite information. Om å andra sidan osäkerheten är stor (som i utfallsrummet A) vinner man mycket information, eftersom man inte kan förutse resultaten. Informationen är direkt relaterad till osäkerheten så man använder entropin som en mått på informationen i utfallsrummet.

För ett försök med n möjliga utfall blir entropin störst då alla sannolikheter är lika $p = frac1n$. För att visa detta behöver vi en hjälpsats, IT olikheten:

$$\log(r) \leq (r-1)\log(e). \tag{3}$$

bevis: man flyttar $\log(r)$ till högre delen. sen deriverar man högre delen och visar genom teckenstudium att funktionen alltid större än noll.

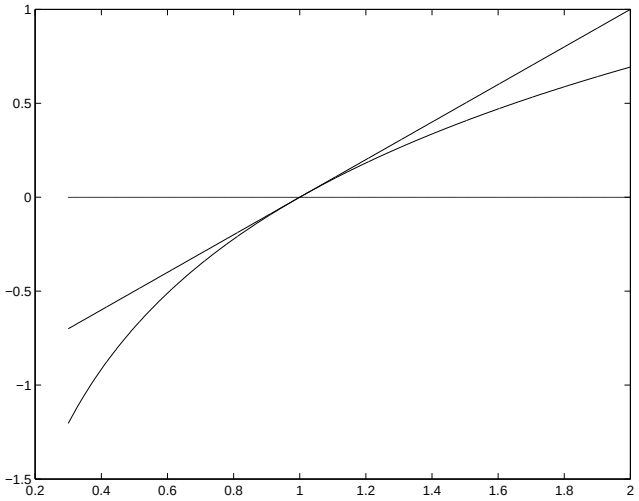
1. basbyte för logaritmer: $\ln(x) = \frac{\log_b(x)}{\log(e)}$.

2. egenskapen $\ln(r) = \ln(r) - \ln(1)$.

3. att använda medelvärdesatsen: $\ln(r) - \ln(1) = \frac{1}{r_0}(a-b)$

Vi ska bevisa att $H(A) \leq \log(n) \Leftrightarrow H(A) - \log(n) \leq 0$ där n är antalet händelser i A.

$$\begin{aligned} H(A) - \log(n) &= -\sum_{k=1}^{n} p_k \log(p_k) - \log(n) \\ &= \sum_{k=1}^{n} p_k \log(\frac{1}{p_k}) - \sum_{k=1}^{n} p_k \log(n) \\ &= \sum_{k=1}^{n} p_k (\log(\frac{1}{p_k}) - \log(n)) \end{aligned}$$



Figur 1: En bild av funktionerna r-1 och ln(r):

$$\begin{aligned} &= \sum_{k=1}^{n} p_k log(\frac{1}{p_k n}) \\ &\leq \sum_{k=1}^{n} p_k (\frac{1}{p_k n} - 1)\log(e) \\ &= \log(e)(\sum_{k=1}^{n} \frac{1}{n} - \sum_{k=1}^{n} p_k) = \log(e)(1-1) = 0. \end{aligned}$$

En annan egenskap hos entropin är att den inte ändras om vi lägger till en eller fler omöjliga händelser och deras sannolikheter (som alltid är noll): Vi lägger till en sex sidig tärning den omöjliga händelsen att få 7. Utfallsrummet blir:

$$A = \begin{pmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \end{pmatrix}$$

$H(A) = -6\frac{1}{6}\log(\frac{1}{6}) - 0\log(0) = \log(6)$ som är lika med entropin utan den omöjliga händelsen.

5.2 Entropi av två utfallsrum

Om det finns två ändliga utfallsrum (A,B) med n respektive m händelser är deras sannolikheter(p_k, q_l).
1) Entropi av oberoende utfallsrum:
Anta att A, B är oberoende av varandra, dvs sannolikheten för att i samma prov av experimenten få händelserna A_k och B_l är $p_k q_l = \pi_{kl}$. Då bildar alla händelserna $A_k B_l$ och deras sannolikheter π_{kl} en ändlig utfallsrum (med *mn* händelser) som vi kallar produkten av A och B. vi får:

$$-H(AB) = \sum_{k=1}^{n}\sum_{l=1}^{m} \pi_{kl}\log(\pi_{kl})$$

$$
\begin{aligned}
&= \sum_{k=1}^{n}\sum_{l=1}^{m} p_k q_l \log(p_k q_l) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{m} p_k q_l (\log(p_k) + \log(q_l)) \\
&= \sum_{k=1}^{n} p_k \log(p_k) \sum_{l=1}^{m} q_l + \sum_{l=1}^{m} q_l \log(q_l) \sum_{k=1}^{n} p_k \\
&= -H(A) - H(B).
\end{aligned}
$$

Entropin av produkten av två oberoende utfallsrum är summan av deras entropier. Ett exempel av en produkt av två utfallsrum är kast med två tärningar.
2) Entropi av beroende utfallsrum:
Anta nu att de två utfallsrum är beroende av varandra . Låt q_{kl} vara sannolikheten att händelsen B_l uppkommer givet att händelsen $A(k)$ uppkommer.

Vi definierar $H_A(B) = \sum_{k=1}^{n} p_k H_k(b)$. Man kan se på $H_A(B)$ som väntevärdet av B om händelsen A_k med sannolikhet p_k har inträffat. (se definitionen av väntevärdet ovan). Det gäller att $H_A(B) \leq H_(B)$. Relationen $H_A(B) \leq H(B)$ betyder för beroende A,B att kännedom av A och dess entropi, kan endast minska entropin av B. Informationsmässigt betyder relationen att vi kan endast minska mängden av informationen som vi får av B. Ett exempel: i engelska ord kommer det alltid ett 'u' efter ett 'q'. Vi kan med säkerhet gissa nästa bokstav, ty i detta fall är u beroende av q. u kan i allmänhet ha en sannolikhet p('u'), men om det inträffar 'q' blir sannolikheten för 'u' p('u')=1 och noll för alla andra bokstäver. Om alfabeten utgör en utfallsrum blir då osäkerheten noll. Men utan kännedom om 'q' har vi en helt annan entropi för nästa bokstav i meningen.
Man definierar I(X,Y) genom

$$
I(X,Y) = H(X) - H_Y(X) \tag{4}
$$

där I betyder den mängd information om Y som finns i X och det gäller:

$$
\begin{aligned}
H(XY) &= H(X) + H_Y(X) && (5)\\
&= H(X) + H(Y) - I(X,Y) && (6)\\
& && (7)
\end{aligned}
$$

vilket är mycket användbart. Detta visar att

$$
\begin{aligned}
I(X,Y) &= I(Y,X) && (8)\\
& && (9)
\end{aligned}
$$

.

5.3 Entropi av markovprocesser

En stokastisk process är en följd av stokastiska variabler som kan vara beroende av varandra, och där beroendet även kan variera). En markovprocess är en stokastisk process där varje variabel är beroende endast av den närmast föregående variablen. Om X,Y,Z är variabler, som kommer i den ordningen, då är $p_Y(Z)$ sannolikheten för Z om Y har inträffat och $p_{XY}(Z)$ är sannolikheten för Z om X och Y har inträffat. För en markovprocess gäller:

$$
p_Y(Z) = p_{XY}(Z) \tag{10}
$$

(anmärkning: p är oberoende av framtiden. Den beror endast av närmast föregående variabel) t.ex i exemplet med engelska kan det endast komma 'u' efter en 'q', men det betyder inte att bara för att det finns ett 'u' måste den föregående symbolen vara ett 'q'. För en markovprocess X,Y,Z gäller följande lemma (data processing lemma):

$$I(X,Z) \leq I(X,Y) \tag{11}$$
$$I(X,Z) \leq I(Y,Z) \tag{12}$$

Betydelsen av lemmat är att informationen som Z innehåller om X är mindre än informationen som Y innehåller om X, och att den informationen som Z innehåller om Y är mindre än informationen som Z innehåller om X. Dvs bearbetning av ett ändligt utfallsrum kan endast minska informationen av det ursprungliga rummet eller lämna mängden av informationen oberörd (med andra ord den kan inte öka informationen). Bildbehandling, t.ex, kan inte tillföra mer information till bilden.
Anta nu att vi har en markov kedja X,Y,Z (vi kan bortse från Z just nu) och en start punkt X. X är en stokastisk variabel som kan anta n olika sannolikheter. Transformen från X till Y är då en funktion från en stokastisk variabel till en annan (Y) som har också n olika sannolikheter q_{kl} ty Y kan bli beroende av X. Om vi har p_k för start sannolikheten av X får vi $\sum_{l=1}^{n} p_k q_{kl} = p_k$. Alla q_{kl} utgör ett ändligt utfallsrum vars entropi är $H_k = -\sum_{l=1}^{n} q_{kl} \log(q_{kl})$ vilket är entropin av ett steg i processen.
Kvantiteten $H = EH_k = -\sum_{k=1} np_k \sum_{l=1}^{m} q_{kl} \log(q_{kl})$ kan ses som det förväntade värdet av entropin när processen tar en steg. Det är klart att samma gäller för nästa steg (med de sannolikheterna som tillhör Z). Om vi tar svenska till exempel: svenska ord kan inte starta med två likadana konsonanter. Om en bokstav är en vokal blir kan den följande bokstav inte bli samma vokal. Sannolikheten för vokalen blir då noll. Detta gäller om ordet inte är ett lånord eller ett sammansatt ord. Detta ger mycket olika entropier för olika ord.

Det som gäller för ett steg kan utvidgas till flera: Varje variabel i kedjan $X_1, X_2, \ldots, X_r$ har n olika utfall, så processen utgör ett ändligt utfallsrum med n^r olika utfall. Jämför detta med två variabler som i exemplet ovan men att de två variabler är avbildningar på samma utfallsrum (möjligtvis med olika sannolikheter). Ett exempel är att kasta en tärning om och om.

H_i^r är entropin av processen med start i $X = x_i$ och av längden r. Entropin för en godtycklig startpunkt blir: $H^r = EH_i^r = \sum_{i=1} np_i H_i^r$ och betyder medelvärdet av informationen som erhålls när processen gått r steg.
När vi väl har ett informationsmått för en process kan vi mäta information per symbol: $M = \frac{H^r}{r}$. T.ex. i engelska har de två processerna "qu" och "q" samma entropi, eftersom det efter ett 'q' endast kan komma ett 'u'. Utfallsrummet för den andra bokstaven har endast ett möjligt utfall, nämligen p(q) =1 vilket ger en entropi av noll. Mängden av information per symbol i "q" är dubbel så stor som i "qu". Om vi till exempel har ett program som kan tillämpa 'q'-regeln på en sekvens kan vi komprimera "enriques que for queueing" med $H = H^{25}$ och $M_1 = H/25$ till "enriqes qe for qeueing" med $M_2 = H/22$. vilket ger $M_2/M_1 = 1.136$ dvs varje symbol i M_2 innehåller i snitt 13 procent mer information än i M_1.

6 Kodning och tillämping av informationsteori

Det finns många tillämpningar utav informationsteori och den mängd information man får ut av en viss företeelse. De flesta av dom ligger inom kommunikationsområdet där man med hjälp av kodning och kunskap om informationsteori kan minimera mängden data som man faktiskt måste skicka.
Det vanligaste sättet att koda information är binärt, dvs. med hjälp av ettor och nollor.

6.1 Analys av text

I ett tal-språk (svenska, engelska etc) finns det regler för hur en mening är uppbyggd. Detta är ej slumpartat och därför ej en Markov-process. Vissa bokstäver förekommer statistiskt oftare än andra och innehåller därför mindre mängd information. Under kursen i matematisk modellering använde vi oss av detta faktum för att systematiskt testa oss fram till vilket ceasarschiffer av en viss text som troligast var klartext.

6.2 Olika tecken innehåller olika mycket information

Inom olika språk är somliga tecken mera ovanliga än andra. Exempelvis är bokstäverna 'x','y','z' mer sällsynta än exempelvis 'a' i det svenska språket. Därför ger ett tecken såsom 'x' i ett ord mer information om hela ordet och meningen än vad ett 'a' hade gett, ty 'a' är en vanligt förekommande bokstav i svenska språket, som därför inte säger lika mycket om hela ordet som en mer utmärkande bokstav 'x' hade gjort. Antag att man vid överföring av ordet "yxa" hade tappat bort exakt ett av tecknen i det. Om det hade varit den första bokstaven 'y' hade några av de möjliga korrekta alternativen varit "axa" och "yxa". Hade det varit 'x' som förlorats hade det funnits många fler möjligheter, exempelvis "yta", "yra", "yla", "yva" samt "yxa". Hade det slutligen varit 'a' som förlorats hade det bara funnits alternativet "yxa" som passat in på svenska. Man ser lätt att det råder mycket större osäkerhet om det är ett 'x' som förlorats medan det vid ett borttappat 'a' knappt råder någon osäkerhet alls. Det är precis detta som menas med att entropin är stor i en bokstav som ett 'x', den innehåller väldigt mycket information och berättar väldigt mycket av meningen.

6.3 Självrättande kod vid enkel/dubbelfel

Inom digital överföring så skickas all information som kombinationer av ettor och nollor. De bildar sviter av en bestämd längd, och översätts sedan till ett tal som symboliserar en bokstav. Exempelvis står 1000001 för $1*2^6+0*2^5+0*2^4+0*2^3+0*2^2+0*2^1+1*2^0=65$ vilket kommer översättas till bokstaven 'A' enligt ASCII-tabellen som ofta används. Om endast en av ettorna eller nollor skulle bytts till det omvända så hade vi fått ett helt annat tal. Om man ger varje möjlig binärkombination en betydelse så kan man inte upptäcka eventuella fel som uppstår vid överföringen. Detta gör att hela sviten av ettor och nollor får en helt annan betydelse än vad som var tänkt. Detta kan man likna vid att en person inte pratar rent exempelvis om en person läspar. Om man istället ordnar det så att vissa sviter saknar betydelse så inser man att det har blivit ett fel om man skulle mottaga en sådan svit. Genom att sprida ut de sviter som har betydelse så kan man direkt se att det blivit ett fel om det bara är ett enda någonstans i sviten.

Exempelvis om någon läspar kan vi människor lägga märke till detta och automatiskt korrigera till det som personen „menar „p.g.a. att det ord vi hör saknar betydelse, och vi antar att det måste ha varit ett annat ord personen i fråga menade. Detta kallas självkorrigerande kod.

Om man sprider ut sviterna med meningen ännu mer så kan man även detektera dubbelfel på detta sätt. Detta leder visserligen till att varje svit som man skickar måste innehålla fler tecken men det gör också att man slipper skicka om informationen om ett enkel eller dubbelfel skulle uppstå.

6.4 Informationsdensitet

När man talar ett språk så förmedlar man information genom de ord man säger. På olika språk så låter det olika, när man förmedlar exakt samma mängd information. Det tar dock olika lång tid och olika många stavelser att säga en mening på två olika språk. I vissa språk måste man använda fler ord för att få fram en viss information än i andra. Man skulle då kunna tala om olika mycket information per tidsenhet (alt. stavelse) på de olika språken, olika informationsdensitet kan man kalla det.

När man lyssnar på någon som talar ett språk som man inte förstår (exempelvis finska) så uppfattar man det ofta som ett brus där man inte hör skillnad på orden utan allt låter ungefär likadant. Ändå lyckas information gå fram genom detta „brus „p.g.a. de små skillnader från „basljudet „som man formar. Man skulle kunna identifiera dessa skillnader och vid överföring enbart skicka skillnaderna. Då skulle man förmedla mer information per del skickad data, och informationsdensiteten hade varit högre. Detta är ofta önskvärt vid överföring eftersom man inte vill skicka mer data än vad som är absolut nödvändigt, och det är detta som är kodning: att komprimera sina data enligt ett känd mönster.

7 Slutsatser

Shannons teori för information är basen till de moderna kommunikationsystemen, från analog radio (AM,FM) till digitala system som GSM, 3G osv. Men teorin går längre. Det är ett mycket kraftfullt verktyg för analys av text, DNA, finansvärlden och kan användas inom många fler områden.

8 Källor

- A Mathematical theory of communication. Claude E. Shannon.
- Mathematical Foundations of Information Theory. A. I. Khinchin.
- Elements Of Information Theory. Thomas M. Cover. Joy A. Thomas.
- Informationsteori - grundvalen för (tele-)kommunikation. Rolf Johannesson
- www.mathworld.com (Information om markovprocesser).

Slumpvandring i $\mathbb{Z}^n$
Projekt i matematisk kommunikation

Olov Karlström, f03ok
Mattias Rydenfelt, d02mry
Kristian Soltesz, f03ks

Handledare: Tobias Rydén

Färdigställd: 2005-05-19

Innehåll

1 Inledning

En slumpvandring är en vandring i vilken man med jämna mellanrum rör sig ett steg. Stegens riktningar är slumpmässig, men följer en viss sannolikhetsfördelning. Vi har valt att studera ett konkret problem relaterat till slumpvandring, nämligen dess *beständighet*. Begreppet beständighet kommer att utredas under resans gång. Något slarvigt uttryckt är beständigheten sannolikheten för återkomst till utgångspunkten. Det kommer visa sig att två saker är avgörande för återkomst, nämligen dimensionen på rummet vi slumpvandrar i och sannolikhetsfördelningen för stegriktningarna.

Intuitivt är det ganska svårt att uppskatta huruvida man återvänder vid en slumpvandring. Av erfarenhet vet de flesta att det kan ta ganska lång tid att återvända när man går vilse i en ny stad. Detta trots att staden är begränsad till två dimensioner. Det är dessutom väldigt sällan vi verkligen slumpvandrar i en stad. (Efter ett tag lär vi oss känna igen platser, och får en känsla för riktning.)

I rummet är intuitionen än svagare. I *Star Trek* tampas Captain Janeway och hennes besättning mot rymdvarelser och meteorsvärmar i sina försök att hitta tillbaks till jorden. De har dessutom federationens stjärnkarta till sitt förfogande.

För att på ett givande sätt diskutera sannolikheten för återkomst krävs det dock att vi först strikt definierar vad vi avser med slumpvandring och framförallt med återvändo.

2 Slumpvandring i $\mathbb{Z}^n$

Vi börjar med slumpvandring i $\mathbb{Z}^n$. Med detta menar vi följande

Def. 1. Slumpvandring i $\mathbb{Z}^n$. Med slumpvandring i $\mathbb{Z}^n$ kan flera saker avses. I denna text avser vi en process som startar i $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{Z}^n$ vid tiden $t = 0$. Mellan $t = m - 1$ och $t = m$ sker en förflyttning $X_m \in \{\pm \mathbf{e}_j; 1 \leq j \leq n\}$, där $\mathbf{e}_j$ är de kanoniska basvektorerna i $\mathbb{Z}^n$. □

Låter vi $P(X_k = \mathbf{e}_j) = P(X_k = -\mathbf{e}_j) = \frac{1}{2n},\ 1 \leq j \leq n$, kallar vi slumpvandringen *symmetrisk*.

Def. 2. Med ovan beteckningar ges processens värde vid tiden $t = m$ av

$$S_m^x = \begin{cases} \mathbf{x} & , m = 0 \\ \mathbf{x} + \sum_{k=1}^{m} X_k & , m \geq 1. \ \square \end{cases} \tag{1}$$

2.1 Slumpvandring i $\mathbb{Z}$

Låt oss nu undersöka vad som händer i det enklaste fallet, nämligen då vi slumpvandrar i $\mathbb{Z}$. Låt $P(X_k = 1) = p,\ P(X_k = -1) = q = 1 - p$. Vi bortser från de triviala fallen då $p \in \{0, 1\}$.

Def. 3. Med $T_y^x = \min_{m \geq 0}(S_m^x = y)$ avses tiden från processens start i x (vid $t = 0$) tills processen når y för första gången. □

Def. 4. Vidare definierar vi $\phi(x) = P(T_d^x < T_c^x),\ c, d \in \mathbb{Z},\ c < d$, som sannolikheten att processen (som startar i x) når d före c. □

Vi fortsätter med en liten undersökning av $\phi(x)$. Med sannolikhetstolkningarna av p och q har vi alltså $\phi(x) = p\phi(x+1) + q\phi(x-1)$. (Då processen befinner sig i x hamnar den efter ett steg i $x+1$ med sannolikhet p och i $x-1$ med sannolikhet q.) Detta medför

$$\begin{cases} \phi(x+1) - \phi(x) &= \frac{q}{p}(\phi(x) - \phi(x-1)) &, c+1 \leq x \leq d-1, \\ \phi(x) &= 0 &, x \leq c, \\ \phi(x) &= 1 &, x \geq d. \end{cases} \tag{2}$$

Notera nu att $\phi(x)$ kan utvecklas i en teleskopsumma som

$$\phi(x) = \sum_{y=c}^{x-1} (\phi(y+1) - \phi(y)), \tag{3}$$

ty $\phi(c) = 0$. Genom rekursivt användade av (2) erhåller vi differensekvationen

$$\begin{aligned} \phi(x+m+1) - \phi(x+m) &= \frac{p}{q}(\phi(x+m) - \phi(x+m-1)) \\ &= \left(\frac{p}{q}\right)^2 (\phi(x+m-1) - \phi(x+m-2)) \\ \ldots &= \left(\frac{p}{q}\right)^{m+1} (\phi(x) - \phi(x-1). \end{aligned} \tag{4}$$

Dessa två observationer ger

$$\begin{aligned} \phi(x) &= \sum_{y=c}^{x-1} \left(\frac{q}{p}\right)^{y-c} (\phi(c+1) - \phi(c)) = \\ &= \phi(c+1) \sum_{y=c}^{x-1} \left(\frac{q}{p}\right)^{y-c} \\ &= \phi(c+1) \frac{1 - (q/p)^{x-c}}{1 - q/p}, \end{aligned} \tag{5}$$

där de två första likheterna följer av (3) och (4). Den tredje likheten följer av att $\phi(c) = 0$ medan den sista är en följd av formeln för geometrisk summa. Väljer vi $x = d$ i (5) erhåller vi

$$\phi(d) = 1 = \phi(c+1) \frac{1 - (q/p)^{d-c}}{1 - q/p} \Rightarrow \phi(c+1) = \frac{1 - q/p}{1 - (q/p)^{d-c}}. \tag{6}$$

Insättning av uttrycket för $\phi(c+1)$ från (6) i (5) ger slutligen

$$\phi(x) = P(T_d^x < T_c^x) = \frac{1 - (q/p)^{x-c}}{1 - (q/p)^{d-c}}, \ c \leq x \leq d, \ p \neq q. \tag{7}$$

Av symmetrin får vi även

$$\psi(x) \equiv P(T_c^x < T_d^x) = \frac{1 - (p/q)^{d-x}}{1 - (p/q)^{d-c}}, \ c \leq x \leq d, \ p \neq q. \tag{8}$$

Från (7) och (8) ser vi att $\phi(x) + \psi(x) = 1,\ c \leq x \leq d$. Låter vi P_0 vara sannolikheten att en process som börjar i $[c, d]$ varken når c eller d då $t \to \infty$ har vi

$$\left.\begin{array}{rcl} \phi(x) + \psi(x) + P_0 & = & 1 \\ \phi(x) + \psi(x) & = & 1 \end{array}\right\} \Rightarrow P_0 = 0 \tag{9}$$

Den första ekvationen i (9) är sann, då något av de tre alternativen (når c innan d, når d innan c, når varken c eller d) måste vara korrekt för processen. $P_0 = 0$ innebär att processen kommer nå antingen c eller d då $p \neq q$, med start i $x \in [c, d]$.

Med $x \in [c, d]$ kan sannolikheten att nå c skrivas som

$$P(T_c^x < \infty) = \lim_{d\to\infty} \psi(x) = \begin{cases} \frac{-\left(\frac{p}{q}\right)^{-x}}{-\left(\frac{p}{q}\right)^{-c}} = \left(\frac{q}{p}\right)^{x-c} & , p > \frac{1}{2}, \\ 1 & , p < \frac{1}{2}. \end{cases} \tag{10}$$

Symmetrin ger

$$P(T_d^x < \infty) = \lim_{d\to\infty} \psi(x) = \begin{cases} 1 & , p > \frac{1}{2}, \\ \left(\frac{p}{q}\right)^{d-x} & , p < \frac{1}{2}. \end{cases} \tag{11}$$

Ytterligare ett intressant resultat får vi genom *starka stora talens lag* (se Appendix)

$$P\left(\lim_{m\to\infty} \frac{S_m^x}{m} = p - q\right) = 1. \tag{12}$$

Med $p < \frac{1}{2}$ kommer processen alltså att *driva* mot $S_m^x = -\infty$. På samma sätt leder $p > q$ till drift mot $+\infty$. Att processen driver i en riktning betyder inte att den inte kan återvända till en punkt. Med utgångspunkt från (10) och (11) i diskussionen ovan, inser man att detta sker med en positiv sannolikhet. En intressant fråga är däremot sannolikheten för att processen ska återvända till x ett oändligt antal gånger.

Def. 5. Med ett *transient tillstånd* menar vi ett x sådant att $P(S_m^x$ oändligt ofta$) = 0,\ p \neq q$. Om processens samtliga tillstånd är transienta, talar vi om en *transient process.* □

(Att processen återvänder till x oändligt ofta innbär att x besöks oändligt många gånger då tiden $\to \infty$.)

Med andra ord är slumpvandringen i $\mathbb{Z}$ transient enligt (12) då $p \neq q$. Vad kan sägas om fallet $p = q$? Insättning av $p = q = \frac{1}{2}$ i (2) ger efter lösning av differensekvationen med tillhörande randvärden att

$$\begin{array}{rcl} \phi(x) & = & P(T_d^x < T_c^x) = \dfrac{x-c}{d-c}, \\ \psi(x) & = & P(T_c^x < T_d^x) = \dfrac{d-x}{d-c} \end{array} , c \leq x \leq d,\ p = q = \frac{1}{2}. \tag{13}$$

Vi ser från (13) att $\phi(x) + \psi(x) = 1$ är uppfyllt även för $p = q$. Dessutom har vi

$$P(S_m^x \text{ når } c) = P(T_c^x < \infty) = \lim_{d\to\infty} P(S_m^x \text{ når } c \text{ före } d) = \lim_{d\to\infty} \frac{d-x}{d-c} = 1. \quad (14)$$

Vid symmetrisk slumpvanding i $\mathbb{Z}$ med start i $x \in [c,d]$ nås alltså garanterat tillståndet c. Detta sker dessutom ett oändligt antal gånger. Vi tänker oss nämligen att processen når c i tiden t_c. Om vi "nollställer" tiden i $t = t_c$, kommer processen garanterat nå c igen enligt (13). Oändlig upprepning av denna procedur visar påståendet. Av symmetriskäl får vi ett analogt resultat för d: $P(T_d^x < \infty) = \lim_{c\to\infty} \frac{x-c}{d-c} = 1$.

Då vi kan välja $[c,d]$ godtyckligt, har vi kommit fram till en intressant slutsats: Vid symmetrisk slumpvandring i $\mathbb{Z}$ kommer varje tillstånd passeras ett oändligt antal gånger om processen låts fortlöpa under oändligt lång tid.

Def. 6. Beständighet. Om $P(S_m^x = y \text{ ett oändligt antal gånger}) = 1$ då $t \to \infty$ kallas tillståndet y *beständigt*. Om en process samtliga tillstånd är *beständiga* är processen en *betändig process*. □

Den symmetriska slumpvandringen i $\mathbb{Z}$ är alltså *beständig*, vilket kan vara bra att veta om man har åkt vilse i tunnelbanan men har gott om tid. Däremot är den *icke*-symmetriska slumpvandringen i $\mathbb{Z}$ *inte beständig*. Intressant är nu (åtminstone för rymdresenärer utan stjärnkarta) att undersöka beständigheten för symmetriska slumpvandringar i $\mathbb{Z}^n, n > 1$. Till detta ägnar vi de kommande avsnitten. I dessa behandlar vi en något modifierad version av ett bevis som först framlades av den ungerske matematikern *Polyá*. (Att icke-symmetriska slumpvandringar ej är beständiga i $\mathbb{Z}^n$ följer direkt ur resultatet för $n = 1$.)

2.2 Slumpvandring i $\mathbb{Z}^2$

Vi fortsätter med det näst enklaste fallet, nämligen slumpvandring i planet av heltalspar. I själva verket hade det varit möjligt att ställa upp det *n-dimensionella* fallet på en gång, men det är lättare att få en överskådlig bild om vi börjar just med $n = 2$.

Vi antar utan inskränkning att processen vi studerar börjar i $\mathbf{x} = (0,0)$; vi kan alltid flytta koordinatsystemet så att $\mathbf{x}$ hamnar i $(0,0)$.

Def. 7. Sannolikheten att processen befinner sig i $(0,0)$ vid tiden $t = m$ betecknar vi $r_m = P(S_m = (0,0))$. □

Vi undersöker nu sannolikheten att processen återvänt till $(0,0)$ vid $t = 2m$. Det är uppenbart att tiden måste vara delbar med två, ty tas l steg i riktning $\mathbf{e}_i$ måste l steg tas i riktgningen $-\mathbf{e}_i$. Att finna den sökta sannolikheten r_{2m} är ett kombinatoriskt problem. Låt antalet steg i riktningen $\mathbf{e}_1$ vara j, vilket även är antalet steg i riktningen $-\mathbf{e}_1$. Antalet steg i riktningen $\mathbf{e}_2$ blir följaktligen samma som antalet steg i riktningen $-\mathbf{e}_2$, nämligen $m - j$. Antalet sätt som processen kan återvända till $(0,0)$ efter $2m$ iterationer är antalet möjliga permutationer av j symboler $\mathbf{e}_1$, j symboler $-\mathbf{e}_1$, $m-j$ symboler $\mathbf{e}_2$ samt $m-j$ symboler $-\mathbf{e}_2$, när vi låter j anta alla värden mellan 0 och m. För ett givet j kan vi tänka oss det hela som en urvalsprocess. Först väljer vi j platser från $2m$ möjliga för $\mathbf{e}_1$. Sedan j platser av de återstående $2m-j$ för $-\mathbf{e}_1$. Därefter väljer

vi $m-j$ platser av de återstående $2m-2j$ för $\mathbf{e}_2$. På de $m-j$ resterande platserna placerar vi $-\mathbf{e}_2$. Med *binomialkoefficienter* (se någon bok i *kombinatorik* eller *diskret matematik*) kan vi skriva detta som

$$\begin{aligned} & \sum_{j=0}^{m} \binom{2m}{j} \binom{2m-j}{j} \binom{2m-2j}{m-j} \binom{m-j}{m-j} \\ = & \sum_{j=0}^{m} \frac{(2m)!}{j!j!(m-j)!(m-j)!} = \sum_{j=0}^{m} \frac{(2m)!}{m!m!} \frac{m!m!}{j!j!(m-j)!(m-j)!} \\ = & \binom{2m}{m} \sum_{j=0}^{m} \binom{m}{j}^2. \end{aligned} \tag{15}$$

Den första likheten följer av definitionen av binomialkoefficienten och förkortning av täljare mot nämnare i det resulterande bråket. Uttrycket under summan visar sig vara kvadraten på en binomialkoefficient. Genom att dra nytta av en *symmetriegenskap* hos binomialkoefficienterna kan vi skriva om uttrycket som

$$\binom{2m}{m} \sum_{j=0}^{m} \binom{m}{j} \binom{m}{m-j} = \binom{2m}{m}^2. \tag{16}$$

Den sista likheten inses genom att göra en urvalstolkning. Antag att vi har en mängd med m olika objekt av sort 1 och m olika objekt av sort 2. På hur många olika sätt kan vi välja m objekt från unionen av dessa mängder? Svaret är att detta kan göras på så många sätt som det går att välja j element av sort 1 och $m-j$ element av sort 2, med $j = 0, 1, \ldots, m$. Men vi vet från kombinatoriken att antalet valmöjligheter är precis $2m$ över m.

I (16) har vi alltså antalet gynnsamma utfall (de utfall då processen återvänder till $(0,0)$). Antalet möjliga utfall är 4^{2m} (vi har $2m$ positioner och på var och en av dessa kan vi välja att placera $\mathbf{e}_1$, $-\mathbf{e}_1$, $\mathbf{e}_2$ eller $-\mathbf{e}_2$). Till slut har vi erhållit den sökta sannolikheten för återvändo till $(0,0)$ vid tiden m som

$$r_{2m} = 4^{-2m} \binom{2m}{m}^2 = 4^{-2m} \frac{(2m)!^2}{m!^4}. \tag{17}$$

Vi använder nu en starkare variant av *Stirlings formel* (18) (se Appendix) för att uppskatta storleken på r_{2m}. Formeln finns visad i [1] och lyder

$$\begin{aligned} 1 \quad \leq & \quad \frac{m!}{(2\pi m)^{1/2} m^m e^{-m}} < \frac{e}{(2\pi)^{1/2}} \\ \Rightarrow & \quad \begin{cases} (2m)!^2 & < & 2 \cdot 16^m m^{4m+1} e^{2-4m} \\ m!^4 & \geq & 4\pi^2 m^{4m+2} e^{-4m} \end{cases} \\ \Rightarrow & \quad r_{2m} > \frac{2 \cdot 16^m m^{4m+1} e^{2-4m}}{4^{2m} 4\pi^2 m^{4m+2} e^{-4m}} = \frac{e^2}{2\pi^2 m}. \end{aligned} \tag{18}$$

Vi är nu närmare svaret än vad man kan tro, ty

$$\sum_{m=1}^{\infty} r_{2m} = \infty. \tag{19}$$

Summan av den angivna serien divergerar. (Detta följer av att $\sum_{m=1}^{\infty} \frac{1}{n} = \infty$). Summan av serien (19) har en viktig tolkning, som kommer att vara viktig även i fallen med högre dimensioner. Att summan *divergerar* är nämligen ekvivalent med att processen är *beständig*. En *konvergent* summa är ekvivalent med en *transient* process.

Bevis. Vi inför A_m så att

$$A_m = \begin{cases} 1 & , \ S_m^{\mathbf{x}} = \mathbf{x}, \\ 0 & , \ \text{annars.} \end{cases} . \tag{20}$$

Väntevärdet (se tex [3])

$$E(A_m), \tag{21}$$

visar sig vara intressant. Av väntevärdesdefinitionen följer direkt att

$$\begin{aligned} E(A_m) &= 1 \cdot P(S_m^{\mathbf{x}} = \mathbf{x}) + 0 \cdot P(S_m^{\mathbf{x}} \neq \mathbf{x}) \\ &= P(S_m^{\mathbf{x}} = \mathbf{x}) = r_m. \end{aligned} \tag{22}$$

Vi har alltså att

$$E\left(\sum_{m=1}^{\infty} A_m\right) = \sum_{m=1}^{\infty} E(A_m) = \sum_{m=1}^{\infty} r_m. \tag{23}$$

Varje gång processen återvänder till i ökar summan i vänsterledet av (23) med 1. Återvänder processen *oändligt ofta* blir summan oändlig, så även väntevärdet. Om processen endast återvänder ett ändligt antal gånger blir summan ändlig. Alltså gäller

$$\begin{aligned} \sum_{n=1}^{\infty} r_m = \infty \quad &\Leftrightarrow \quad \text{beständig process} \\ \sum_{n=1}^{\infty} r_m \neq \infty \quad &\Leftrightarrow \quad \text{transient process. } \square \end{aligned} \tag{24}$$

Vi har nu, tack vare (24), kommit fram till slutsatsen att symmetrisk slumpvandingen i $\mathbb{Z}^2$ är *beständig*. Att implikationerna går åt båda hållen i (24) är dock inte självklart. Betecknar vi sannolikheten att processen återvänder till starttillståndet med $f = P(S_m^{\mathbf{x}} = \mathbf{x})$ och det totala antalet återkomster med $N = \sum_{m>0} A_m$, kan vi skriva $P(N = k) = f^k(1 - f)$. Vi ser att N är geometriskt fördelad med väntevärdet $E(N) = \frac{f}{1-f}$. En beständig process är ekvivalent med $f = 1 \Rightarrow E(N) = \infty$. På samma sätt är en transient process ekvivalent med $f < 1 \Rightarrow E(N) < \infty$. Ovanstående, tillsammans med (23) visar ekvivalenserna i (24).

Givetvis hade vi kunnat använda metoden ovan även för ett bevis i det endimensionella fallet. Anledningen att vi valt två olika angreppsvinklar är i första hand förhoppningen att en av dem uppfattas som naturligare av läsaren.

2.3 Slumpvandring i $\mathbb{Z}^n,\ n \geq 3$

Vi vidareutvecklar idén från det två-dimensionella fallet till $n = 3$, och börjar med att ställa upp ett uttryck för r_{2m}. Låt j beteckna antalet steg i riktningen $\mathbf{e}_1$ resp. $-\mathbf{e}_1$. På samma sätt definierar vi m som antalet steg i riktningen $\mathbf{e}_2$ resp. $-\mathbf{e}_2$. I riktningen $\mathbf{e}_3$ tas $m - j - l$ steg, likaså i riktningen $-\mathbf{e}_3$. Med sex riktningar blir antalet möjliga vandringar om $2m$ steg 6^m. Sannolikheten att återvända till $(0,0,0)$ på $2m$ iterationer blir följaktligen (jämför med det tvådimensionella fallet)

$$\begin{aligned} r_{2m} &= 6^{-2m} \sum_{(j,l);j+l\leq m} \binom{2m}{j} \binom{2m-j}{j} \binom{2m-2j}{l} \\ &\quad \binom{2m-2j-l}{l} \binom{2m-2j-2l}{m-j-l} \binom{m-j-l}{m-j-l}. \end{aligned} \tag{25}$$

Genom att utveckla uttrycket i (26) enligt definitionen på binomialkoefficienten, och förkorta faktorer i täljaren mot nämnaren, erhålles

$$\begin{aligned} &\frac{1}{2^{2m}} \binom{2m}{m} \sum_{(j,l);j+l\leq m} \left(\frac{m!}{j!l!(m-j-l)!} \frac{1}{3^m} \right)^2 \\ = \ &\frac{1}{2^{2m}} \binom{2m}{m} \sum_{(j,l);j+l\leq m} (p_{j,l})^2, \end{aligned} \tag{26}$$

där $p_{j,l}$ är sannolikheter i *trinomialfördelningen*. (För slumpvandring i godtycklig dimension förekommer sannolikheter från motsvarande multinomialfördelning i summan).

Vi gör som i det tvådimensionella fallet och uppskattar storleken på r_{2m}:

$$r_{2m} \leq \frac{1}{2^{2m}} \binom{2m}{m} \sum_{j+l\leq m} (\max_{j,l} p_{j,l}) p_{j,l} = \frac{1}{2^{2m}} \binom{2m}{m} \max_{j,l} p_{j,l}. \tag{27}$$

Anledningen till att summan i (27) ”försvinner” är att summan av sannolikheterna $p_{j,l}$ är 1.

Det gäller att $p_{j,l}$ är som störst då j och l är så nära $m/3$ som möjligt, se Appendix. Följaktligen gäller att

$$r_{2m} \leq \frac{1}{2^{2m}} \binom{2m}{m} \frac{1}{3^m} \frac{m!}{\left(\left[\frac{m}{3}\right]!\right)^3}, \text{ där } [\,] \text{ betecknar heltalsdelen.} \tag{28}$$

Vi använder nu den starkare versionen av *Stirlings formel* från (18) för att uppskatta r_{2m}. För $m \equiv 0 \pmod 3$ kan vi skriva

$$\begin{aligned} r_{2m} &\leq \frac{1}{12^m} \frac{(2m)!}{m! \left(\left(\frac{m}{3}\right)!\right)^3} \leq \frac{\sqrt{2\pi 2m}(2m)^{2m}e^{-2m}}{m^m e^{1-m}\sqrt{m} \cdot \left(\left(\frac{m}{3}\right)^{\frac{m}{3}} e^{1-\frac{m}{3}} \sqrt{\frac{m}{3}}\right)^3} \leq \\ \ldots &\leq \frac{C}{m^{3/2}},\ C \text{ konstant.} \end{aligned} \tag{29}$$

Att $r_{2m} \leq \frac{C}{m^{3/2}}$ gäller även då $m \not\equiv 0 \pmod 3$ visas i Appendix.

$$\sum_{m=1}^{\infty} r_{2m} < \infty, \tag{30}$$

varför slumpvandringen i $\mathbb{Z}^3$ är transient enligt (34). Det är intuitivt självklart att detsamma gäller då dimensionen höjs till $n > 3$. Slutsatsen är att *symmetrisk slumpvandring* i $\mathbb{Z}$ och $\mathbb{Z}^2$ är *beständig*, emedan den är *transient* för $\mathbb{Z}^n$, $n \geq 3$. Det är alltså en dålig idé att ge sig ut på rymdresa (till och med i en diskret rymd) utan ordentlig karta.

3 Appendix

3.1 Starka stora talens lag

Låt $X_1, X_2, \ldots$ vara en oändlig följd av oberoende stokastiska variabler med väntevärdet m. För

$$\overline{X} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{31}$$

gäller enligt starka stora talens lag att

$$P(X = m) = 1. \tag{32}$$

För bevis, se [7].

3.2 Stirlings formel

Med hjälp av Stirlings formel kan man få en approximation på fakulteten av ett tal. Formeln lyder

$$n! \approx \sqrt{2\pi} \frac{n^{n+\frac{1}{2}}}{e^n}. \tag{33}$$

Beviset, vilket återges exempelvis i [6], kräver matematik på högre nivå än vad som lärs ut första året på LTH. Ur formeln ovan kan man relativt lätt härleda (18), se [1].

3.3 Multinomialfördelning

Binomialkoefficienten,

$$\left(\begin{array}{c} n \\ k \end{array} \right), \tag{34}$$

vilken har den kombinatoriska tolkningen ”antalet sätt att välja ut k objekt från n utan hänsyn till ordning”, går att generalisera. Denna generalisering kallas multinomialkoefficient och betecknas

$$\left(\begin{array}{c} n \\ n_1, n_2, \ldots, n_k \end{array} \right) = \frac{n!}{n_1! n_2!, \ldots, n_k!}, \tag{35}$$

där $n_1 + n_2 + \ldots + n_k = n$. Den kombinatoriska tolkning är ”antalet arrangemang av n_1 stycken likadana objekt A_1, n_2 stycken likadana objekt A_2, ..., n_k objekt

A_k". En intressant egenskap hos multinomialkoefficienten är att den antar sitt maximala värde då $n_1, n_2 \ldots, n_k$ ligger så nära $\frac{n}{k}$ som möjligt. Detta kan visas på följande sätt: Antag att $n_1!n_2! \ldots, n_k!$ är minimal. Om det gäller att det finns n_i och n_j så att $n_i - n_j = a \geq 2$, så har vi

$$\begin{aligned}(n_i - 1)!(n_j + 1)! &= (n_j + a - 1)!(n_j + 1)! = (n_j + 1)(n_j + a - 1)!n_j! \\ &< (n_j + a)(n_j + a - 1)!n_j! = (n_j + a)!n_j! = n_i!n_j!, \end{aligned} \quad (36)$$

vilket motsäger att $n_1!n_2! \ldots, n_k!$ skulle vara minimal.

När produkten av fakulteter är minimal är multinomialkoefficienten maximal, vilket visar påståendet.

Med hjälp av multinomialkoefficienterna kan man även definiera en statistisk fördelning, nämligen multinomialfördelningen. Denna definieras enligt följande. Låt $Y = (X_1, X_2, \ldots, X_k)$ vara en stokastisk variabel med $X_1 + X_2 + \ldots + X_k = n$. Den simultana sannolikhetsfunktionen är då

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1!n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \quad (37)$$

där det för summan av sannolikheterna gäller $p_1 + p_2 + \ldots + p_k = 1$.

3.4 Approximation av r_{2m}

I (30) gjordes en approximation av r_{2m} för $m \equiv 0 \pmod 3$. Nedan visar vi att resultatet från approximation är giltigt för alla positiva heltal m. I följande räkningar gäller $m \equiv 0 \pmod 3$. Enligt resonemanget i avsnittet om multinominalfördelning kan vi direkt skriva om (27) som

$$\begin{aligned} r_{2(m+1)} &\leq \frac{1}{12^{m+1}} \frac{(2m)!(2m+1)(2m+2)}{m!(m+1)\left(\left(\frac{m}{3}\right)!\right)^2 \left(\frac{m}{3}+1\right)!} = r_{2m} \frac{1}{12} \frac{(2m+1)(2m+2)}{(m+1)\left(\frac{m}{3}+1\right)} \\ &= r_{2m} \frac{m + \frac{1}{2}}{m+3} \leq r_{2m} \leq \frac{C}{m^{3/2}}, \ C \text{ konstant.} \end{aligned} \quad (38)$$

På samma sätt kan vi skriva

$$\begin{aligned} r_{2(m+2)} &\leq \frac{1}{12^{m+2}} \frac{(2m)!(2m+1)\ldots(2m+4)}{m!(m+1)(m+2)\left(\frac{m}{3}\right)!\left(\left(\frac{m}{3}+1\right)!\right)^2} \\ &= r_{2m} \frac{(m+\frac{1}{2})(m+\frac{3}{2})}{(m+3)^2} \leq \frac{C}{m^{3/2}}, \ C \text{ konstant.} \end{aligned} \quad (39)$$

Alltså gäller $r_{2m} \leq \frac{C}{m^{3/2}}$ $\forall m \in \mathbb{Z}^+$, dvs även då $m \not\equiv 0 \pmod 3$.

4 Källförteckning

Referenser

[1] R. N. Bhattacharya, E. C. Waymire. *Stochastic Processes with Applications.* Wiley, 1990. ISBN: 0-471-84272-9.

[2] T. Rydén, G. Lindgren *Markovkedjor.* LTH, 2000.

[3] G. Blom *Sannolikhetsteori med tillämpningar A och B.* Studentlitteratur, 1984. ISBN: 91-44-04372-4.

[4] S. M. Ross *Introduction to Probability Models.* Academic press, 1985. ISBN: 0-12-598468-5.

[5] U. Körner *Köteori.* Studentlitteratur, 2003. ISBN: 91-44-03103-3.

[6] *http://planetmath.org/encyclopedia/StirlingsApproximation.html.* 2005-04-28

[7] *http://www.cims.nyu.edu/ẽve2/chap2.pdf.* 2005-04-28

Matematisk Kommunikation: Projektarbete

Martin Nilsson
Niklas Nilsson
Ufuk Kirik

8 Maj 2005

Innehåll

1 Introduktion till Brownsk Rörelse: Hur upptäcktes fenomenet?

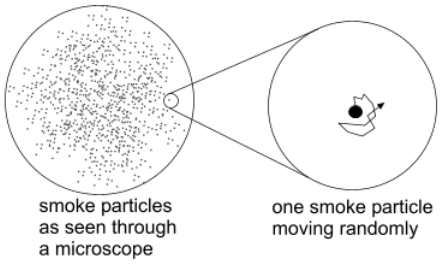
Vad är det egentligen?

Det finns två betydelser av termen Brownsk rörelse:
Den första är det fysiska fenomenet att mycket små partiklar i en vätska[1] rör sig tillsynes slumpmässigt. Rörelsen uppkommer hos partiklar som är så små, att det finns en markant sannolikhet för att mycket färre molekyler från omgivningen stöter emot partikelns ena sida, än på motsatt sida. Resultatet blir att partikeln får en “knuff” i riktning mot den sida med det lägre antalet molekyler. Den andra är de matematiska modellerna som beskriver dessa rörelser.

Hur upptäcktes fenomenet?

Brownsk rörelse har fått sitt namn efter botanisten Robert Brown, även då det inte var Brown som upptäckte fenomenet först - nästan alla som betraktade vatten genom ett mikroskop stötte på det - han nämner själv ett tiotal personer vid namn som råkat på fenomenet i sin skrift år 1829. Han gjorde dock ett exprimentiellt studium av fenomenet som ledde till att många andra kom i kontakt med det. Genom sina betraktelser

[1] Även i gaser, men här domineras rörelsen av turbulens



Figur 1: *Partiklar i rök utgör var för sig brownsk rörelse*

utvecklade han en teori som sade att all materia består av små partiklar, vilka han kallade *levande molekyler* och att rörelsen kom från molekylerna själva, inte vätskan de befann sig i. Även om hans teori visades felaktig så gjorde han ett stort bidrag genom att etablera Brownsk rörelse som ett viktigt fenomen och visa att fenoment förekommer för organiskt så väl som icke organiskt materia. Det senare upptäckte han när han både studerade uppslammade pollenpartiklar i vatten och uppslammade dammpartiklar i vatten och noterade att fenomenet förekom i båda fallen.

Efter Browns artikel studerades *Brownsk rörelse* av många, och olika förklaringar föreslogs (elektriska krafter, lokala temperaturvariationer, växelverkan med ljus). Först 1877 presenterade Delsaux den nu accepterade förklaringen, nämligen att de uppslammade partiklarnas oregelbundna rörelse orsakas av kollisioner med snabba vätskemolekyler.

Genom systematiska studier av bland annat Gouy hade man vid sekelskiftet kommit fram till en hel radattributer hos brownsk rörelse:

1. Rörelsen är väldigt irregulär, bestående av translationer och rotationer och dess bana verkar inte ha någon tangent.
2. Två partiklar verkar röra sig självständigt, även då de närmar sig varandra till ett avstånd mindre än deras diameter.
3. Rörelsen är mer aktiv för mindre partiklar.
4. Uppbyggnad och densitet hos partikeln har ingen påverkan.
5. Rörelsen är aktivare ju lägre viskositet vätskan har.
6. Rörelsen är aktivare vid högre temperatur.
7. Rörelsen avtar aldrig.

Vid sekelskiftet arbetade Louis Bachelier fram en generaliserad funktion för vad som nu är känt som wienerprocessen, vilken kopplade den till diffusionsekvationen[2].

Några år senare tog Einstein sig an problemet utan att egentligen känna till brownsk rörelse genom att tillämpa kinetisk teori för vätskemolekyler, vilket inte motsades av någon av de tidigare upptäckterna. Einstein visade först att de suspenderade partiklarna bör ha samma diffusionsegenskaper som vätskemolekylerna, trots att partiklarna är

[2] *Diffusion* är en lokalt slumpartad process som leder till utjämning av koncentrationsskillnader i gaser och vätskor

så stora att de kan urskiljas i mikroskop. Han kunde sedan härleda ett enkelt samband mellan partiklarnas genomsnittliga förflyttning i en viss riktning (λ_x) under en viss tid t och diffusionskonstanten i vätskan D:

$$\lambda_x = \sqrt{2Dt} \tag{1}$$

Detta plublicerade han sedan i sitt verk år 1905 (*Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen.*). Jean Perrin studerade sedan år 1912 fenomenet och gjorde många experiment som bekräftade att Einsteins teori går att tillämpa på brownsk rörelse.
Teorin för den matematiska Brownska rörelsen lades på fast grund av Norbert Wiener 1923, och utvecklades senare av Paul Lévy.

2 Sannolikhetens Roll

För att få en bättre förståelse av brownsk rörelse så måste man undersöka grunderna i sannolikhetsläran, eftersom rörelsen hos partiklarna är tillsynes slumpmässig. Därför vore det orationellt att utelämna sannolikheten - eller chansens - roll medan vi diskuterar brownsk rörelse.

Till att börja med så kan vi ställa oss frågan vad vi egentligen menar då vi använder ordet "chans". Chanser nämns då vi skall beskriva våra gissningar. Gissningar som görs då man skall göra bedömningar med ofullständig information eller osäker kunskap. Ett exempel skulle kunna vara, *'Det är en 20 procents chans att det blir regn imorgon.'* Vi kan se att detta är ett påstående gjort med begränsad information. Vi vill gissa vad saker är, eller vilka saker som med störst sannolikhet kommer inträffa. I vanliga fall gör vi gissningar eftersom vi måste ta ett beslut. I linje med exemplet ovan, så vore *'Borde jag ta med mig mitt paraply imorgon?'* en korresponderande orsak för att göra en gissning. Som i allt annat så finns det bra gissningar och dåliga gissningar; och sannolikhetsteorin hjälper oss att göra bättre val genom att låta oss prata kvantitativt om en situation som må ha ganska varierande resultat men ändå ett konsekvent medelbeteende.

Ett bra och ganska enkelt exempel på en sådan situation är att singla slant. Om det är ett *rättvist* mynt, så är resultatet helt slumpmässigt mellan två resultat som är lika troliga. Generellt kan vi skatta att sannolikheten för att få ett resultat A från en repeterad händelse är:

$$P(A) = \frac{N_A}{N}$$

där N är antalet repetitioner och att N_A är antalet fall med resultatet A vi fått under dessa N gånger.

När man kommit hit uppkommer en ny viktig punkt, för att kunna prata om sannolikheten för att något kommer att inträffa, så måste det vara ett möjligt resultat av en *repeterbar* observation.
Relationen mellan Brownsk rörelse och sannolikhet kan man hitta i definitionen av Brownsk rörelse. En av de vanliga definitionerna som man kan hitta på nätet[3] refererar till Brownsk rörelse som *"the random walk of small particles suspended in a fluid"*.

[3] tagits från www.mathworld.com

Rörelsen hos varje partikel är slumpmässig, detta innebär att de, mer eller mindre, har lika stor chans att röra sig i vilken riktning som helst de kan röra sig i. Med tänke på hur en sådan rörelse[4] är konstrurerad så kan en sannolikhets fördelning göras och rörelsen kan uttryckas matematiskt.

3 Slumpvandring

Slumpvandring är en slumpmässig förflyttning av ett objekt. Objektet startar i en punkt och låter sen slumpen bestämma åt vilket håll det ska ta ett steg. Denna vandring kan ske i en eller flera dimensioner men vi börjar med att undersöka slumpvandring i en dimension.

Objektet kan då endast röra sig längs en linje och har då bara två möjliga vägar att gå, antingen framåt eller bakåt. För enkelhetens skull studerar vi fallet där sannolikheten att objektet rör sig framåt är lika stor som bakåt. En intressant fråga att ställa sig nu är var objektet kommer att befinna sig efter N steg. Det genomsnittliga värdet av slumpvandringen är noll eftersom det är lika stor chans att objektet går framåt som bakåt. När man undersöker saken närmare upptäcker man att frågan inte har ett fullt så enkelt svar. Det visar sig att det genomsnittliga avståndet till nollan ökar när antalet steg ökar.

Objektets avstånd till startpunkten (nollan) är absolutbeloppet av dess x koordinaten på linjen eftersom ett avstånd inte kan vara negativt. Av praktiska skäl arbetar vi med kvadraten av avståndet istället för absolutbeloppet. Kvadraten av avståndet ger ju också alltid ett positivt resultat. Vi ska nu visa att det genomsnittliga avståndet till nollan i kvadrat efter N steg är just N gånger steglängden.

Vi tänker oss att D_N^2 är kvadratavståndet till noll efter N steg. För att ta reda på genomsnittliga värdet på detta så börjar vi från början. Efter ett steg vet vi att helt säkert att $D_1^2 = 1$. Vi kan beräkna D_N^2 när $N > 1$ genom att använda D_{N-1}. Har vi efter (N-1) steg D_{N-1} så har vi efter N steg $D_N = D_{N-1} + 1$ eller $D_N = D_{N-1} - 1$. Om vi räknar med kvadraten av avståndet istället får vi.

$$D_N^2 = \begin{cases} D_{N-1}^2 + 2D_{N-1}^2 + 1 \\ \\ D_{N-1}^2 - 2D_{N-1}^2 + 1 \end{cases}$$

Det är exakt lika stor chans att vi får det ena resultatet som det andra. Därför tar vi helt enkelt medelvärdet av de båda alternativen:

$$D_N^2 = \frac{D_{N-1}^2 + 2D_{N-1}^2 + 1 + D_{N-1}^2 - 2D_{N-1}^2 + 1}{2} = D_{N-1}^2 + 1 \qquad (2)$$

Vi har tidigare visat att $D_1^2 = 1$, vilket tillsammans med (2) dessa oss den enkla lösningen att medelavståndet från nollan i kvadrat är lika med antalet steg objektet har tagit, $D_N^2 = N$ och därmed att $D_N = \sqrt{N}$.

När vi nu har fastställt detta kan vi göra kvalificerade gissningar om var en partikel hamnar som slumpmässigt vandrar längst en linje. Det skulle nu vara intressant att se

[4] kollisionen med andra molekyler som är mycket mindre

om vi skulle kunna göra likadana *gissningar* om vi utökar slumpvandringen till fler dimensioner.

Nu ska vi undersöka slumpvandring i två dimensioner. Vårt objekt startar i punkten $(0,0)$ och flyttar det sig ett steg med längden 1 i en slumpmässig vald riktning. Detta upprepas med den nya positionen som ny utgångspunkt. Objektets riktning väljs helt slumpmässigt från intervallet $[0;2\pi]$ och alla vinklar är lika sannolika.

I stället för att använda oss av x och y koordinater när vi beskriver objektets punkt använder vi oss av komplexa tal skrivna på polär form. Vi beskriver objektets position efter N steg som talet z som är summan av alla stegen.

$$z = \sum_{j=1}^{N} e^{i\theta_j}.$$

Vi tar nu absolutkvadraten av z som definieras som $|z|^2 = z\bar{z}$ (där $\bar{z}$ är konjugat av z).

$$|z|^2 = z\bar{z} = \sum_{j=1}^{N} e^{i\theta_j} \sum_{k=1}^{N} e^{-i\theta_k} = \sum_{j=1}^{N}\sum_{k=1}^{N} e^{i(\theta_j-\theta_k)} = N + \sum_{k,j=1 k\neq j}^{N} e^{i(\theta_j-\theta_k)}.$$

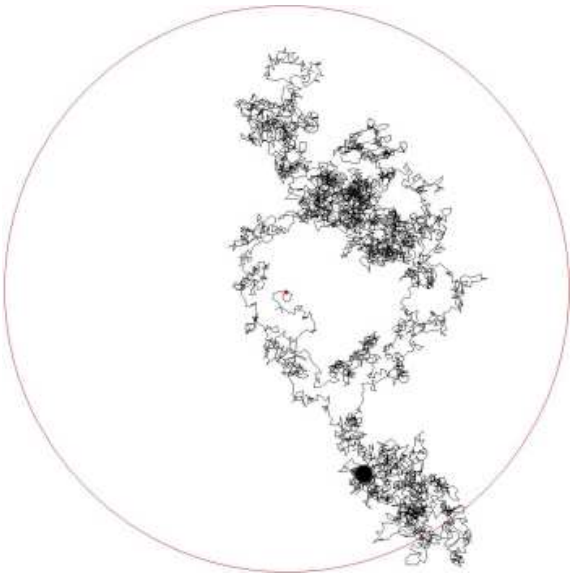
I den sista ekvationen har vi en slumpmässigt vald variabel θ_j minus en annan slumpmässigt vald variabel θ_k. Detta meför att $e^{i(\theta_j-\theta_k)}$ kan ta vilken riktning som helst med lika stor sannolikhet. Medelvärdet av $\sum_{k,j=1 k\neq j}^{N} e^{i(\theta_j-\theta_k)}$ blir då noll. Alltså får vi att medelavståndet i kvadrat är N, $D_N^2 = N$ och därmed att $D_N = \sqrt{N}$.

Nu har vi visat att samma samband mellan medelavståndet till startpunkten och antalet steg gäller för slumpvandring i två dimensioner såväl som i en dimension. I figuren nedan har $\sqrt{N}$ markerats med en cirkel.

Det finns fler saker som slumpvandringar i en respektive två dimensioner har gemensamt. Det gäller nämligen att när antalet tagna steg går mot oändligheten så kommer vi att ha varit i alla punkter. För att ta ett exempel, om en berusad person skulle gå hem från krogen och hans vandring skulle kunna beskrivas som en slumpvandring skulle han alltid komma hem så småningom oavsett var hans hem var beläget. Detta gäller för både en och två dimensioner och bevisades av den ungerske matematikern George Pólya 1921. Däremot gäller inte samma sak för tre eller fler dimensioner, så om en berusad fågel skulle flyga hem från krogen så är det inte alls säkert att någonsin hittar sitt bo.

För att ta ett exempel på vad vi kan använda våra nya kunskaper om slumpvandring till: om vi tänker oss att vi sprider pollenkorn på en vattenyta, pollenkornen börjar krocka med varandra och Brownsk rörelse uppkommer. Om vi tittar på ett pollenkorns rörelse inom detta kaos så skulle vi kunna beskriva den som en slumpvandring. Istället för att slumpgeneratorer bestämmer färdriktning så gör massa slumpmässiga krockar med andra pollenkorn det. Med hjälp av sambandet mellan medelavståndet till startpunkten och antalet tagna steg vi bevisat för slumpvandring i två dimensioner kan vi nu göra kvalificerade gissningar om hur långt pollenkornet har förflyttat sig efter en viss tid. Detta förutsätter att vi kan uppskatta ungefär hur lång tid det tar mellan varje krock och hur långt kornet färdas mellan varje krock.

Det finns stora likheter mellan Brownsk rörelse och slumpvandring och därför är det



Figur 2: *Experimentel observation kan jämföras till förväntade teoretiskt värde*

bra att studera slumpvandring för att förstå Brownsk rörelse bättre. Man skulle kunna säga att Brownsk rörelse är massor av partiklar som slumpvandrar bredvid varandra.

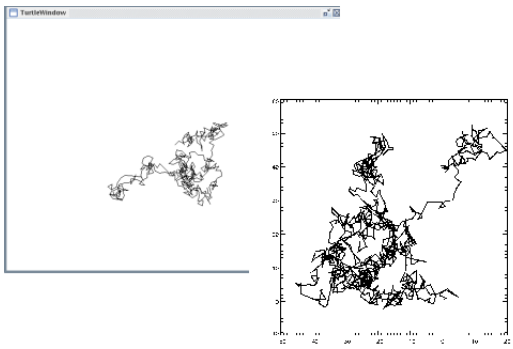
4 Simuleringsmetoder

Det finns ett antal sätt att formulera en simulering av *slumpvandringar* och deras tillämpningar till Brownsk rörelse. Även om olika källor favoriserar och berättar om olika metoder, bygger alla metoder i princip på samma idé. Vi ska försöka att förklara så många av dessa metoder som möjligt i denna sektion.

Enligt Krell Institute[5], finns det fem vanliga metoder att formulera, respektive simulera *slumpvandring* i två dimensioner genom att generera slumpmässigt tagna steg. Den första av de metoderna är byggd på enhetscirkel och trigonometri. I denna metod tar man en slumpmässig vinkel, θ, i intervallet $[0, 2\pi]$. Man sätter sedan $x = \cos\theta$ och $y = \sin\theta$. Då blir *steglängden* normerad till 1 för alla vinklar θ. Vi utvecklade något enkelt program som kör med denna princip. En skärmdump från detta programmet och en bild från *Wikipedia.org* i artikeln om *random walk* i två dimensionella fallet kan ses i figur 3.

I den andra metoden tar man ett slumpmässigt tal i intervallet $[-1, 1]$ som Δx. Sedan beräknar man $\Delta y = \pm\sqrt{1-(\Delta x)^2}$. I denna metod använder man sambandet mellan förändringen i x respektiv y-leden. Vi bestämde oss att använda en liknande metod, som kollar på förändringar på axlarna, inte i det två dimensionella fallet, utan i det tredimensionella fallet. Implementeringar gjordes i *Matlab* och innehåller en algoritm som

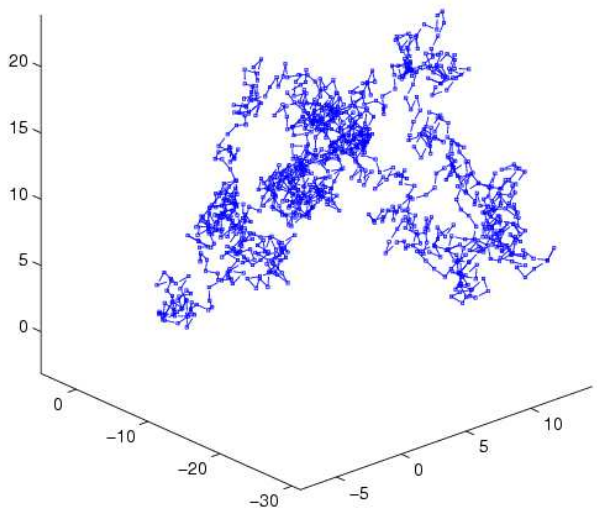
[5]Krell Institute Inc, Ames Indiana USA - http://www.krellinst.org



Figur 3: *Slumpvandring simuleringar på 2D*

genererar ett slumpmässigt tal, $\Delta x = D(1,1)$[6], i intervallet $[0,1]$. Sedan tas ett annat tal, nämligen $\Delta y = D(1,2)$, i intervallet $[0, \sqrt{1-D(1,1)^2}]$. Alltså bestämmer vi förändringar i x, respektive y-led, för vilka måste gälla $y^2 \leq \sqrt{1^2 - x^2}$ eftersom vi har ett Δz också. Naturligtvis blir $\Delta z = \sqrt{1-(\Delta x)^2-(\Delta y)^2}$.

Alltså har vi koordinaterna för en slumpmässig punkt i rummet som har avståndet 1 till origo. Genom att spara koordinaterna för sådana punkter i en annan (mycket större[7]) matris, kan vi registrera rörelsen för vår partikel i rummet. En tredimensionell bild kan ses nedan.



Figur 4: *Slumpmässigt rörelse av en partikel i rummet*

[6]Matris D står för en 1x3 matris

[7]en *n*x3 matrix, då n är antal steg vi vill köra simuleringen

Den tredje metoden innebär att man tar slumpmässiga tal för båda Δx och Δy och sedan normaliserar dem så att *steglängden* blir 1. Det är självklart mycket likt den andra metoden, särskilt om man vill tillämpa metoden till högre dimensioner.

Fjärde metoden innehåller en lite allmännare utgångspunkt. I denna metod väljer man slumpmässigt mellan fyra resultat som är lika troliga, nämligen nord, öst, syd och väst. Läsaren bör observera att den här metoden faktiskt är en förenklad version av den första, som istället hade ett intervall över $[0, 2\pi]$. Även då denna metoden ger snyggare grafer[8] så är den mindre noggrann än metod 1. I denna metod kan vi bara välja riktningar som är multiplar av $\frac{\pi}{2}$. En annan geometrisk tolkning av sådana grafer är att tänka på en funktion definierad från $\mathbf{R}^2$ till $\mathbf{Z}^2$, då steglängden är lika med 1.

I den sista metoden kör man nästan exakt samma princip som i den tredje metoden, men istället för intervallet $[-1, 1]$ använder man ett intervall mellan $[-\sqrt{2}, \sqrt{2}]$ vilket innebär att steglängden inte blir konstant 1 utan närmar sig 1 i medeltal.

Av de fem metoderna favorisera Krell Institute den sista och påstår att den är den enda metod som blir lämplig med stort antal steg. Vi tycker dock inte att det är inte helt uppenbart att den sista metoden är lämpligare för stora stegantal. I princip är metoderna i stort sätt likadana och ingen metod är exakt nog, eftersom det inte finns några slumpmässiga val i datavärlden och siffrorna är mer eller mindre gissningsbara med avancerade data algoritmer. För mer noggrana resultat använder man inmätningsprocedurer då man läser in listor för mättningar av radioaktiv sönderfall från olika ämne. Även då sådana listor antas vara helt slumpmässiga; i princip alla saker omkring oss kan beskrivas som en funktion och har då något mönster. Det är ju hur mycket vi kan förstå av det.

Referenser:

`http://www.doc.ic.ac.uk/~nd/surprise_95/journal/vol4/ykl/report.html`

The Research Goes On ... av Y.K.Lee och Kelvin Hoon, 1995

`http://www.fenews.com/fen42/teach_notes/teaching-notes.html`

Utdrag ur *Essays in Derivatives*, av Don Chance, publiserad i John Wiley & Sons, 1998

http://www.wikipedia.org

http://mathworld.wolfram.com

Dynamical theories of Brownian motion - second edition, av Edward Nelson, Augusti,2001[9]

The Feynman's Lectures on Physics, av Richard P. Feynman

[8] se titelbilden på första sidan

[9] Tillgänglig i elektronisk form på: http://www.math.princeton.edu/ nelson/books.

Kvaternioner och komplexa tal

Niclas Ek, Johan Frisk, Johan Gustafsson och Erik Kronvall

19 maj 2005

1 Komplexa tal

1.1 De komplexa talens historia

De tidigaste referenserna man har till kvadratrötter av negativa tal finns i den grekiske matematikern och uppfinnaren Heron av Alexandrias arbete. Dessa referenser är från någon gång under det första århundradet e.Kr.

De blev mer framträdande under 1500-talet då de italienska matematikerna Niccolo Fontana Tartaglia och Gerolamo Cardano fann formler för rötterna i tredje- och fjärdegradspolynom. Man insåg att dessa formler ibland krävde användning av kvadratrötter av negativa tal. Detta var väldigt förvirrande eftersom inte ens de negativa talen var fullt erkända vid denna tiden. Termen "imaginär" myntades för dessa storheter av René Descartes på 1600-talet. Under 1700-talet gjordes många framsteg av främst Abraham de Moivre och Leonhard Euler som bl.a. tog fram varsin välkänd formel:

- de Moivres formel (1730):

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta, \tag{1}$$

- Euler's formel (1748):

$$\cos\theta + i\sin\theta = e^{i\theta}. \tag{2}$$

Men trots formler och andra framsteg med de komplexa talen så blev de ändå inte helt accepterade förrän deras geometriska tolkning introducerades av Caspar Wessel år 1799. Ett antal år senare återupptog Carl Friedrich Gauss Wessels teorier och gjorde dem allmänt kända vilket ledde till ett uppsving för teorin för de komplexa talen.

År 1804 började Abbé Buée arbeta med en idé om att rötter av negativa tal ska representera en enhetslinje, den imaginära axeln, som är vinkelrät mot de reella talens axel. Av någon anledning publicerades inte Buées arbete förrän 1806, vilket var samma år som en annan matematiker Jean-Robert Argand också slutförde sitt arbete inom samma ämne. Det är till Argands arbete man refererar till nuförtiden då man vill läsa om introduktionen av den geometriska tolkningen av komplexa tal. Trots stora framsteg under början av 1800-talet så var det få som kände till dem. Det var inte förrän Gauss började intressera sig för Argands arbete som han märkte hur relativt okänt det var i matematikens värld. Han utkom med sina mest framstående anteckningar i ämnet 1832 vilket resulterade i att hela den matematiska världen fick upp ögonen för både hans och Argands teorier.

Gauss introducerade sedan skrivsättet $a+bi$ för komplexa tal och förklarade att han använde i som beteckning för $\sqrt{-1}$.

1.2 Definition av komplexa tal

Ett modernt sätt att införa de komplexa talen är att låta dem definieras som ordnade talpar (a, b) där a och b är reella tal. Operationerna addition respektive multiplikation mellan komplexa tal går till på följande vis:

$$(a, b) + (c, d) = (a + c, b + d),$$

$$(a, b) \cdot (c, d) = (ac - bd, bc + ad).$$

Komplexa tal på formen $(a, 0)$ uppför sig som reella tal enligt $(a, 0)+(b, 0) = (a + b, 0)$ och $(a, 0) \cdot (b, 0) = (ab, 0)$ och man identifierar dessa med det reella talet $(a, 0) = a$. Vidare motiverar $(a, 0) \cdot (b, c) = (ab, ac)$ skrivsättet $a(b, c)$ för multiplikation med reella tal. Detta medför att man kan skriva $(a, b) = (a, 0)+(0, b) = a(1, 0)+b(0, 1)$. Här kallar man $(1, 0)$ för 1 och $(0, 1)$ för i, vilket medför beteckningen $(a, b) = a+bi$ för komplexa tal. Formeln för multiplikation ger:

$$i^2 = (0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 1 \cdot 0 + 0 \cdot 1) = (-1, 0) = -(1, 0) = -1.$$

1.3 Geometri

Idéerna med en imaginär och en reell axel resulterade under 1800-talets början i det komplexa talplanet. I detta tvådimensionella koordinatsystem representeras som bekant varje komplext tal av en punkt eller en vektor. x-led motsvarar då den reella delen och y-led den imaginära delen av talet. Detta ligger också till grund för definitionen ovan av komplexa tal som talpar. Precis som vektorer i $\mathbb{R}^2$ kan komplexa tal skrivas på polär form.

$$z = a + bi = r(\cos\theta + i\sin\theta) = re^{i\theta},$$

där den sista likheten, Eulers formel, är ett skrivsätt som motiveras i avsnitt 1.6 nedan. Här är r längden av vektorn (a, b) och benämns absolutbeloppet av z. θ är vinkeln mellan vektorn (a, b) och x-axeln och kallas argumentet av z. Att addera två komplexa tal är detsamma som att addera två vektorer, medan multiplikation med $z = re^{i\theta}$ är detsamma som rotation moturs med vinkeln θ och förlängning med faktorn r. Multiplikation med i ger en rotation med 90 grader moturs, och $i^2 = -1$ ger en rotation med 180 grader, d.v.s. två rotationer med 90 grader vardera.

1.4 Absolutbelopp och konjugat

Byter man tecken på imaginärdelen i ett komplext tal får man ett tal som kallas det komplexa konjugatet. Konjugatet av $z = a + bi$ betecknas $\overline{z}$, och definieras som $\overline{z} = a - bi$. Konjugatet är användbart vid division mellan två komplexa tal. Ett komplext tal $z = a + bi \neq 0$ multiplicerat med sitt konjugat blir alltid ett positivt reellt tal:

$$z\overline{z} = (a + bi)(a - bi) = a^2 - abi + abi - bi^2 = a^2 + b^2.$$

Absolutbeloppet av ett komplext tal $z = a + bi$ betecknas $|z|$ och definieras som

$$|z| = \sqrt{a^2 + b^2} = \sqrt{z\overline{z}}$$

Sambandet $z\overline{z} = |z|^2$ kan då $z \neq 0$ även skrivas $\frac{z\overline{z}}{|z|^2} = 1$ vilket innebär att varje komplext tal $z \neq 0$ har en multiplikativ invers z^{-1} som kan uttryckas:

$$z^{-1} = \frac{\overline{z}}{|z|^2}.$$

1.5 Komplexa tal på matrisform

Komplexa tal kan uttryckas med hjälp av speciella 2x2 matriser. Man kan låta det komplexa talet $z = a + bi$ representeras av en matris på formen

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Varje matris på denna form är en linjärkombination av matriserna

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

vilket innebär att matrisen Z som motsvarar det komplexa talet $z = a + bi$ kan skrivas:

$$Z = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = aE + bI.$$

E och I är basmatriser som motsvarar 1 och i, och uppfyller sambandet

$$I^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -E.$$

För det komplexa talet z representerat av matrisen Z inses lätt att det komplexa konjugatet $\overline{z}$ motsvaras av den transponerade matrisen Z^T. Dessutom gäller att $|z|^2 = \det Z$, d.v.s. absolutbeloppet av z är lika med kvadratroten ur determinanten av Z:

$$\det Z = \begin{vmatrix} a & -b \\ b & a \end{vmatrix} = a^2 + b^2 = |z|^2.$$

Räknelagarna för matriser gör att addition och multiplikation fungerar exakt likadant när de komplexa talen representeras av matriser som annars. Detta kan visas med två godtyckliga komplexa tal $z_1 = a + bi$ och $z_2 = c + di$.

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} + \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} a+c & -(b+d) \\ b+d & a+c \end{pmatrix}$$

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} ac-bd & -(ad+bc) \\ ad+bc & ac-bd \end{pmatrix}$$

att jämföras med

$$(a, b) + (c, d) = (a + b, c + d)$$

respektive

$$(a, b) \cdot (c, d) = (ac - bd, bc + ad).$$

1.6 Komplexa funktioner

I grundläggande analys behandlas funktioner som har de reella talen som definitions- och värdemängd. $f : \mathbb{R} \to \mathbb{R}$. En komplex funktion har de komplexa talen $\mathbb{C}$ som definitions- och värdemängd. $g : \mathbb{C} \to \mathbb{C}$. Ibland talar man om en komplexvärd funktion, vilket endast talar om att funktionens värdemängd är en delmängd av de komplexa talen.

En komplexvärd funktion är bestämd av sin real- och imaginärdel: $g = u+iv$ där u och v är reellvärda funktioner av det komplexa talet z. Eftersom z bestäms av två reella parametrar x och y kan man se funktionerna u och v som reellvärda funktioner av två reella variabler. En komplexvärd funktion g kan då skrivas: $g(x, y) = u(x, y) + iv(x, y)$. Komplex analys handlar om att ge reella funktioner så som $\sin x$, $\cos x$, $\ln x$, e^x, x^a o.s.v. en tolkning för komplexa argument och att ta fram teori för komplexa funktioner och deriverbarhet.

Detta sysselsatte 1700-talets matematiker. Den första att studera komplexa logaritmer var matematikern Johann Bernoulli som nämnde dessa i en artikel 1702. Vidare noterade engelsmannen Roger Cotes 1714 att det verkade finnas ett samband mellan komplexa logaritmer och de trigonometriska funktionerna sinus och cosinus. Resultatet av deras arbete sammanfattades av Leonard Euler då han 1748 i en artikel publicerade formeln (2). Euler kom fram till formeln genom att använda sig av serieutvecklingarna:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \tag{3}$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \dots \tag{4}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \dots \tag{5}$$

Euler utgick från (3) och ersatte x med ix vilket enligt räkneregler för imaginära tal ger:

$$e^{ix} = 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} - \frac{x^6}{6!} - i\frac{x^7}{7!} + \frac{x^8}{8!} + \dots$$

Uppdelning i real och imaginärdel ger:

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \dots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \dots\right)$$

som med (4) och (5) ger formeln (2). Detta är dock bara formella räkningar , formeln (3) är ju härledd med förutsättningen att $x \in \mathbb{R}$. Ett fullständigt bevis måste innehålla en konvergensanalys.

Insättning av $\theta = \pi$ i Eulers formel ger det märkliga sambandet:

$$e^{i\pi} + 1 = 0,$$

som mycket kort sammanfattar tusentals år av matematisk utveckling. Kopplingen mellan imaginära tal och periodiska funktioner genom Eulers formel kan användas för att härleda många trigonometriska formler. Om vi antar att formeln $e^{a+b} = e^a \cdot e^b$ även gäller för imaginära exponenter kan additionsreglerna för sinus och cosinus härledas: $e^{ix+iy} = e^{i(x+y)} = \cos(x+y)+i\sin(x+y) = e^{ix} \cdot e^{iy} = (\cos x+i\sin x)(\cos y+i\sin y) = \cos x\cos y-\sin x\sin y+i(\cos x\sin y+\sin x\cos y)$. Identifiering av realdel och imaginärdel ger nu additionsformlerna för både sinus och cosinus.

1.7 Komplexa logaritmer

Den naturliga logaritmen för ett reellt tal $x > 0$, $\ln x$ uppfyller: $e^{\ln x} = x$. Den naturliga logaritmen kan alltså härledas ur funktionen e^x. Eftersom $e^{ix} = \cos x + i \sin x$ kan vi ge en tolkning av logaritmen för ett komplext tal z. Om $z = i = e^{i\frac{\pi}{2}}$ kan vi tolka $\log z$ som $i\frac{\pi}{2}$ då detta tal uppfyller relationen $e^{\log z} = z$. Vi ser dock att logaritmen för ett komplext tal ej är entydigt bestämd eftersom: $i = e^{i\frac{\pi}{2}} = e^{i\frac{\pi}{2}+2k\pi}, k \in \mathbb{Z}$. Detta är en följd av periodiciteten hos sinus och cosinus. Vid räkningar med komplexa logaritmer måste man naturligtvis kräva att funktionen $\log z$ är entydigt bestämd. Detta löses genom att man föreskriver logaritmens värden till ett visst område. Något förenklat kan man säga att man bestämmer sig för viljet värde på k som skall gälla. En konsekvens av detta är att vissa logaritmlagar ej gäller för komplexa logaritmer.

2 Kvaternioner

2.1 En utvidgning av de komplexa talen

Den framgångsrika idén med att låta varje komplext tal representeras av en punkt i ett plan gjorde att man börja söka efter en utvidning till tre dimensioner, där man kunde låta varje tal i ett nytt talsystem åskådliggöras i något slags komplext rum. Det visade sig vara svårt att göra en sådan utökning av de komplexa talen och samtidigt behålla de regler och egenskaper som gör ett talsystem meningsfullt.

Ett enkelt försök att introducera tredimensionella komplexa tal skulle kunna bestå i att låta talen skrivas på formen $t = a + bi + cj$, där a, b och c är reella tal medan i och j är något slags imaginära enheter som uppfyller speciella villkor. Det är enkelt att införa en naturlig definition av addition mellan två tal $t_1 = a_1 + b_1 i + c_1 j$ och $t_2 = a_2 + b_2 i + c_2 j$ på denna form:

$$t_1 + t_2 = (a_1 + a_2) + (b_1 + b_2)i + (c_1 + c_2)j.$$

En definition av multiplikation skulle kunna se ut såhär:

$$t_1 \cdot t_2 = (a_1 + b_1 i + c_1 j)(a_2 + b_2 i + c_2 j) =$$

$$= a_1 a_2 + a_1 b_2 i + a_1 c_2 j + b_1 a_2 i + b_1 b_2 i^2 + b_1 c_2 ij + c_1 a_2 j + c_1 b_2 ji + c_1 c_2 j^2 =$$

$$= a_1 a_2 + (a_1 b_2 + b_1 a_2)i + (a_1 c_2 + c_1 a_2)j + b_1 c_2 ij + c_1 b_2 ij + b_1 b_2 i^2 + c_1 c_2 j^2.$$

Men denna definition är inte fullständig så länge man inte bestämt vad produkterna i^2, j^2, ij och ji ska betyda. Och faktum är att det är omöjligt att definiera dessa produkter så att talsystemet får alla egenskaper som man kan anse vara nödvändiga för att det ska vara användbart.

2.2 William Rowan Hamilton

En av de personer som skulle komma att ägna stor tid och möda på att hitta något slags generalisering av de komplexa talen till fler dimensioner var irländaren William Rowan Hamilton. Han föddes den 4:e Augusti 1805 i Dublin. Båda hans föräldrar dog då han fortfarande var en ung pojke. Hans farbror, som var

en anglikansk präst, utbildade honom. Geniet i Hamilton yttrade sig till en början genom hans förmåga att hantera språk. Redan vid fem års ålder läste han grekiska, hebreiska och latin. Vid tio års ålder hade han lärt sig ett halv dussin orientaliska språk utöver de han redan behärskade.

Hamilton gick på en av en liten men fin skola för matematiker sammankopplad med Trinity College i Dublin. Det var här han tillbringade större delen av sitt liv. Redan vid knappa 22 års ålder blev han utnämnd till professor i astronomi. Hamiltons matematiska studier skötte han helt på egen hand. När han sedan började på universitetet var han bäst på alla kurser.

De komplexa talen fascinerade Hamilton och det var han som först definierade komplexa tal som talpar av reella tal. Han var också som sagt mycket engagerad i sökandet efter en utvidgning av dessa. Detta problem kämpade han med länge, men till slut hittade han en något oväntad lösning. Han kunde inte utvidga de komplexa talen till tre dimensioner, men väl till fyra. Enligt historien ska Hamilton år 1843 ha vandrat med sin fru längs Royal Canal i Dublin då han plötsligt insåg det fundamentala villkor som gjorde utvidgningen möjlig:

$$i^2 = j^2 = k^2 = ijk = -1.$$

Detta ska han genast ha ristat in i ena sidan av Broughambron. Ristningen finns inte kvar, men istället finns en minnesplatta uppsatt som berättar om Hamiltons plötsliga insikt. Den nya fyrdimensionella algebra han hade upptäckt kallade han för kvaternioner, och han ägnade stor del av sitt fortsatta liv till försöka marknadsföra dem på olika sätt.

Hamilton hann även bidra till utvecklingen inom optiken, dynamiken och algebran med sitt arbete. En del av hans forskning har även blivit signifikativ för utvecklingen av kvantmekanik. Han dog den 2:e september 1865.

2.3 Definition av kvaternioner

Det Hamilton visade när han upptäckte kvaternionerna att man genom att använda sig av tre imaginära enheter kunde skapa ett system som äger de flesta av de värdefulla egenskaper man önskar av ett talsystem. Kvaternionerna definieras som tal på formen $q = a + bi + cj + dk$, där a, b, c och d är reella tal medan i, j och k är imaginära enheter som uppfyller det fundamentala villkoret $i^2 = j^2 = k^2 = ijk = -1$.

Genom att använda sig av ekvationen $ijk = -1$ kan man enkelt härleda regler för multiplikation av två olika imaginära enheter. Exempelvis ger multiplikation med i att $iijk = -i$, vilket tillsammans med villkoret $i^2 = -1$ medför att $-jk = -i$ som måste vara ekvivalent med $jk = i$. Multiplicerar man nu med j ser man att $jjk = ji$, som enligt $j^2 = -1$ innebär att $ji = -k$. Men om sambandet $ijk = -1$ multipliceras med k visar det sig att $ijkk = -k$, vilket tillsammans med $k^2 = -1$ betyder att $-ij = -k$, eller ekvivalent $ij = k$. Eftersom $ij \neq ji$ kan vi dra slutsatsen att multiplikation mellan två kvaternioner q_1 och q_2 inte är kommutativ, d.v.s att $q_1 \cdot q_2$ i allmänhet inte är lika med $q_2 \cdot q_1$.

Med samma metoder som ovan kan man visa produkterna mellan samtliga par av imaginära enheter, och ställa upp dessa i en multiplikationstabell:

$\cdot$	1	i	j	k
1	1	i	j	k
i	i	-1	k	$-j$
j	j	$-k$	-1	i
k	k	j	$-i$	-1

2.4 Addition och multiplikation

En naturlig additionsregel för kvaternioner kan enkelt införas om man använder samma princip som för komplexa tal, och helt enkelt adderar real- och de olika imaginärdelarna var för sig. Summan mellan två kvaternioner $q_1 = a_1 + b_1 i + c_1 j + d_1 k$ och $q_2 = a_2 + b_2 + c_2 + d_2$ definieras som:

$$q_1 + q_2 = (a_1 + a_2) + (b_1 + b_2)i + (c_1 + c_2)j + (d_1 + d_2)k.$$

Addition mellan kvaternioner är alltså inget mer än addition mellan reella tal. Att regeln ovan är både associativ, så att $(q_1 + q_2) + q_3 = q_1 + (q_2 + q_3)$, och kommutativ, så att $q_1 + q_2 = q_2 + q_1$, följer därför direkt av att detsamma gäller för addition mellan reella tal.

Nästan lika enkelt kan man införa en allmän multiplikationsregel för kvaternioner. Regeln kan naturligt resoneras fram om man förutsätter att $ai = ia$ för alla reella tal a, och att motsvarande gäller om i ersätts med någon annan av de imaginära enheterna. För samma kvaternioner q_1 och q_2 som användes i additionsregeln borde produkten av dessa kunna tänkas se ut som följer:

$$q_1 \cdot q_2 = (a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k) =$$

$$= a_1 a_2 + a_1 b_2 i + a_1 c_2 j + a_1 d_2 k + b_1 a_2 + b_1 b_2 i^2 + b_1 c_2 ij + b_1 d_2 ik +$$

$$+ c_1 a_2 j + c_1 b_2 ji + c_1 c_2 j^2 + c_1 d_2 jk + d_1 a_2 k + d_1 b_2 ki + d_1 c_2 kj + d_1 d_2 k^2.$$

Använder man sig av multiplikationstabellen ovan kan man sortera termerna:

$$q_1 \cdot q_2 = (a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2) + (a_1 b_2 + b_1 a_2 + c_1 d_2 - d_1 c_2)i +$$

$$+ (a_1 c_2 + c_1 a_2 + d_1 b_2 - b_1 d_2)j + (a_1 d_2 + d_1 a_2 + b_1 c_2 - c_1 b_2)k.$$

2.5 Vektor- och skalärprodukt

Hamilton tänkte sig att en kvaternion $q = a + bi + cj + dk$ bestod av två delar. Talet a kallade han för skalärdel, medan de resterande tre termerna tillsammans utgjorde vad Hamilton kallade en vektor. Om man ser på en kvaternion q som uppdelad i en skalärdel a och en vektordel, som kan betecknas $\mathbf{v} = (b, c, d)$, är det möjligt att skriva produkten mellan två kvaternioner på ett enklare sätt: $q = a + \mathbf{v}$. Här kan man betrakta skalärdelen a av q som dess realdel, och vektorn $\mathbf{v}$ som dess imaginärdel.

Genom att sortera termerna i definitionen av multiplikation på ett något annorlunda sätt än ovan upptäcker man några intressanta samband:

$$q_1 \cdot q_2 = (a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2) + (a_1 b_2 + b_1 a_2 + c_1 d_2 - d_1 c_2)i +$$

$$+ (a_1 c_2 + c_1 a_2 + d_1 b_2 - b_1 d_2)j + (a_1 d_2 + d_1 a_2 + b_1 c_2 - c_1 b_2)k =$$

$$= a_1a_2 - (b_1b_2 + c_1c_2 + d_1d_2) + a_1(b_2i + c_2j + d_2k) + a_2(b_1i + c_1j + d_1k) +$$
$$+(c_1d_2 - d_1c_2)i + (d_1b_2 - b_1d_2)j + (b_1c_2 - c_1b_2)k.$$

Till att börja med kan man nu notera att $b_1b_2 + c_1c_2 + d_1d_2$ är lika med det tal vi kallar skalärprodukten av vektorerna $\mathbf{v}_1 = (b_1, c_1, d_1)$ och $\mathbf{v}_2 = (b_2, c_2, d_2)$. Vidare så är vektorn $(c_1d_2 - d_1c_2, d_1b_2 - b_1d_2, b_1c_2 - c_1b_2)$, som återfinns i slutet av uttrycket ovan, just den vektor vi känner som vektorprodukten av $\mathbf{v}_1$ och $\mathbf{v}_2$. Detta innebär att produkten av två kvaternioner $q_1 = a_1 + \mathbf{v}_1$ och $q_2 = a_2 + \mathbf{v}_2$ kan skrivas på ett enklare sätt med hjälp av symbolerna för skalär- och vektorprodukt:

$$q_1 \cdot q_2 = a_1a_2 - \mathbf{v}_1 \cdot \mathbf{v}_2 + a_1\mathbf{v}_2 + a_2\mathbf{v}_2 + \mathbf{v}_1 \times \mathbf{v}_2.$$

Värt att lägga märke till är att vektorprodukten $\mathbf{v}_1 \times \mathbf{v}_2$ är det enda i skrivsättet ovan som inte kommuterar. För denna gäller att $\mathbf{v}_1 \times \mathbf{v}_2 = -\mathbf{v}_2 \times \mathbf{v}_1$, vilket betyder att skillnaden mellan produkten q_1q_2 och den omvända produkten q_2q_1 är $q_1q_2 - q_2q_1 = 2(\mathbf{v}_1 \times \mathbf{v}_2)$. Om produkten mellan två specifika kvaternioner är kommutativ måste detta innebära att vektorprodukten mellan deras vektordelar är noll. Reglerna för vektorprodukt säger då att vektorerna är parallella.

2.6 Konjugat och absolutbelopp

Precis som man till varje komplext tal $z = a + bi$ kan tilldela ett unikt tal $\overline{z} = a - bi$ som kallas det komplexa konjugatet, kan man definiera ett konjugat för kvaternioner. Detta gör man i analogi med de komplexa talen genom att låta imaginärdelen, i detta fall samtliga av de tre koefficienterna för i, j och k, byta tecken. Konjugatet till kvaternionen $q = a+bi+cj+dk = a+\mathbf{v}$ definieras alltså som $\overline{q} = a-bi-cj-dk = a-\mathbf{v}$. För två kvaternioner $q_1 = a_1+b_1i+c_1j+d_1k = a_1 + \mathbf{v}_1$och $q_2 = a_2 + b_2i + c_2j + d_2k = a_2 + \mathbf{v}_2$ gäller att:

$$\overline{q_1 + q_2} = \overline{(a_1 + a_2) + (b_1 + b_2)i + (c_1 + c_2)j + (d_1 + d_2)k} =$$

$$= (a_1 + a_2) - (b_1 + b_2)i - (c_1 + c_2)j - (d_1 + d_2)k = \overline{q_1} + \overline{q_2}.$$

Ovanstående samband har exakt samma utseende som motsvarande för komplexa tal. För komplexa tal gäller dessutom att $\overline{z_1z_2} = \overline{z}_1\overline{z}_2$, vilket däremot inte är sant för kvaternioner. Eftersom produkten av två komplexa tal är kommutativ gäller dock även att $\overline{z_1z_2} = \overline{z}_2\overline{z}_1$. Detta samband är sant också för kvaternioner, vilket kan verifieras genom att utveckla $\overline{q_1q_2}$ och $\overline{q}_2\overline{q}_1$ var för sig:

$$\overline{q_1q_2} = \overline{a_1a_2 - \mathbf{v}_1 \cdot \mathbf{v}_2 + a_1\mathbf{v}_2 + a_2\mathbf{v}_1 + \mathbf{v}_1 \times \mathbf{v}_2} =$$

$$a_1a_2 - \mathbf{v}_1 \cdot \mathbf{v}_2 - a_1\mathbf{v}_2 - a_2\mathbf{v}_1 - \mathbf{v}_1 \times \mathbf{v}_2.$$

$$\overline{q}_2\overline{q}_1 = (a_2 + (-\mathbf{v}_2))(a_1 + (-\mathbf{v}_1)) =$$

$$= a_2a_1 - (-\mathbf{v}_2) \cdot (-\mathbf{v}_1) + a_2(-\mathbf{v}_1) + a_1(-\mathbf{v}_2) + (-\mathbf{v}_2) \times (-\mathbf{v}_1) =$$

$$= a_1a_2 - \mathbf{v}_1 \cdot \mathbf{v}_2 - a_1\mathbf{v}_2 - a_2\mathbf{v}_1 - \mathbf{v}_1 \times \mathbf{v}_2.$$

Liksom för komplexa tal är produkten av en kvaternion och dess konjugat alltid ett reellt tal. Varje kvaternion har också ett reellt absolutbelopp $|q|$ vilket i likhet

med komplexa tal och även vektorer definieras som $|q| = \sqrt{a^2+b^2+c^2+d^2}$. För absolutbeloppet gäller följande samband:

$$q\overline{q} = (a+\mathbf{v})(a+(-\mathbf{v})) = a^2 - \mathbf{v}\cdot(-\mathbf{v}) + a\mathbf{v} + a(-\mathbf{v}) + \mathbf{v}\times(-\mathbf{v}) = a^2 + \mathbf{v_1}\cdot\mathbf{v_2} =$$

$$= a^2 + b^2 + c^2 + d^2 = |q|^2.$$

Ekvationen $q\overline{q} = |q|^2$ ger att varje kvaternion $q \neq 0$, i analogi med de komplexa talen, har en entydigt bestämd multiplikativ invers

$$q^{-1} = \frac{\overline{q}}{|q|^2}.$$

2.7 Matrisrepresentation

Liksom för komplexa tal går det utmärkt att låta kvaternionerna representeras av matriser. Detta kan göras på två sätt. Antingen med 2x2-matriser där varje element är ett komplext tal, eller med 4x4-matriser där alla element är reella. Representationen med komplexa 2x2-matriser ser för en godtycklig kvaternion $q = a + bi + cj + dk$ ut på följande vis:

$$\begin{pmatrix} a-di & -b+ci \\ b+ci & a+di \end{pmatrix}. \tag{6}$$

Exempelvis skulle kvaternionen $1+2i+3j+4k$ få följande utseende på en sådan matrisform:

$$\begin{pmatrix} 1-4i & -2+3i \\ 2+3i & 1+4i \end{pmatrix}.$$

Varje matris på denna form är en linjärkombination av basmatriserna

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad K = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}$$

som motsvarar 1, i, j respektive k och uppfyller villkoret

$$I^2 = J^2 = K^2 = IJK = -E.$$

Låter man nu räknelagarna för komplexa tal och matriser gälla kan man enkelt visa att addition och multiplikation mellan kvaternioner fungerar på samma sätt även när dessa man låter kvaternionerna representeras av matriser på formen (6). Om $q_1 = a_1 + b_1 i + c_1 j + d_1 k$ och $q_2 = a_2 + b_2 i + c_2 j + d_2 k$ skulle deras summa se ut såhär:

$$\begin{pmatrix} a_1-d_1i & -b_1+c_1i \\ b_1+c_1i & a_1+d_1i \end{pmatrix} + \begin{pmatrix} a_2-d_2i & -b_2+c_2i \\ b_2+c_2i & a_2+d_2i \end{pmatrix} =$$

$$= \begin{pmatrix} a_1-d_1i+a_2-d_2i & -b_1+c_1i-b_2+c_2i \\ b_1+c_1i+b_2+c_2i & a_1+d_1i+a_2+a_2i \end{pmatrix} =$$

$$= \begin{pmatrix} (a_1+a_2)-(d_1+d_2)i & -(b_1+b_2)+(c_1+c_2)i \\ (b_1+b_2)+(c_1+c_2)i & (a_1+a_2)+(d_1+d_2)i \end{pmatrix}.$$

Den erhållna matrisen är som synes exakt den som motsvarar kvaternionen (q_1+q_2). Multiplikationsregeln kan kontrolleras lika enkelt med hjälp av reglerna för matrismultiplikation.

För de 2x2-matriser som kan representera komplexa tal konstaterades tidigare att absolutbeloppet av ett komplext tal var detsamma som kvadratroten av determinanten av motsvarande matris. Även matriser på formen (6) som representerar kvaternioner har denna egenskap, och för kvaternionen $q = a + bi + cj + dk$ representerad av 2x2-matrisen Q gäller därför att $|q|^2 = \det Q$:

$$\det Q = \begin{vmatrix} a - di & -b + ci \\ b + ci & a + di \end{vmatrix} =$$

$$= (a - di)(a + di) - (-b + ci)(b + ci) = a^2 + b^2 + c^2 + d^2 = |q|^2.$$

Ett komplext tal representerat av en 2x2-matris har som bekant ett konjugat som motsvaras av den transponerade matrisen. Detta är inte sant för kvaternioner och deras respektive matriser på formen (6). Däremot gäller för kvaternionen q och motsvarande 2x2-matris att konjugatet $\overline{q}$ representeras av det konjugerade transponatet av samma matris. Med det konjugerade transponatet menas den matris som erhålls om en ursprungsmatris transponeras efter att alla dess element bytts ut mot respektive komplexa konjugat.

Det andra sättet att representera kvaternioner med matriser består som sagt i att använda 4x4-matriser där varje element är ett reellt tal. Kvaternionen $q = a + bi + cj + dk$ representeras då av matrisen:

$$\begin{pmatrix} a & -b & d & -c \\ b & a & -c & -d \\ -d & c & a & -b \\ c & d & b & a \end{pmatrix}. \tag{7}$$

De matriser som på denna form motsvarar 1 respektive de imaginära enheterna är:

$$E = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad I = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$J = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \qquad K = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Precis som i fallet med 2x2-matriser är varje matris på formen (7) en linjärkombination av E, I, J, K som givetvis uppfyller det fundamentala villkoret

$$I^2 = J^2 = K^2 = IJK = -E.$$

Att additions- och multiplikationsreglerna gäller även för matriser på denna form är inte svårt att kontrollera. En kvaternion $q = a + bi + ci + dk$, representerad av en matris Q på formen (7), har ett konjugat $\overline{q}$ som representeras av transponatet av Q:

$$Q^T = \begin{pmatrix} a & b & -d & c \\ -b & a & c & d \\ d & -c & a & b \\ -c & -d & -b & a \end{pmatrix}.$$

2.8 Kvaternionernas betydelse

När Hamilton spred läran om sina kvaternioner i mitten av 1800-talet var de något kontroversiellt, främst på grund av att de inte uppfyllde den kommutativa lagen för multiplikation. Kvaternionerna vann dock viss respekt hos matematiker, men de blev ändå aldrig det revolutionerande verktyg som Hamilton trott och hoppats på. Mot slutet av 1800-talet hade den mer allmängiltiga vektoralgebran tagit över det mesta av den matematik och fysik som kunde beskrivas med hjälp av kvaternioner, och idag tillämpas Hamiltons upptäckt i mycket få sammanhang.

Men även om kvaternionerna idag inte används i särskilt stor utsträckning har de haft stor historisk betydelse för viktiga delar av matematikens utveckling. Vektoralgebran har i själva verket hela sitt ursprung i Hamiltons kvaternioner, och det var Hamilton som introducerade begreppen skalär och vektor. Det var också Hamilton som genom multiplikationsregeln för kvaternionerna först stötte på de samband som idag kallas skalär- och vektorprodukt. I viss litterlatur finns i, j, k fortfarande kvar som beteckning för basvektorer.

Vektoralgebran som från början var ett sidospår ur kvaternionerna utvecklades och kom på sikt att nästan helt ersätta dem. Men kvaternionerna dyker fortfarande upp inom vissa tillämpningar. Precis som de komplexa talen kan användas för att beskriva vridningar i planet, kan kvaternionerna användas för att beskriva vridningar i rummet. Detta låter sig emellertid göras även med matriser, varför kvaternionerna kan verka överflödiga. En viss fördel finns dock med att använda kvaternioner, eftersom de erbjuder ett kompakt och smidigt skrivsätt. En vektor $\mathbf{x}$ roteras med hjälp av följande funktion:

$$\mathbf{y} = q\mathbf{x}q^{-1}.$$

Här är q en kvaternion med $|q| = 1$, och $q^{-1} = \overline{q}$ är dess invers. $\mathbf{y}$ är den vektor som erhålls efter rotation av $\mathbf{x}$. Sambandet får ett snarlikt utseende om rotationen utförs med matriser, men den matris som används är betydligt mer komplicerad uttryckt i de variabler som beskriver rotationen. Matrisen i fråga är en ortogonal 3x3-matris, och eftersom en sådan består av nio olika tal medan en kvaternion bara består av fyra krävs det också betydligt fler beräkningar om rotationen utförs med hjälp av matrismultiplikation än om den utförs genom multiplikation av kvaternioner. Detta gör att det oftast går både snabbare och smidigare att använda kvaternioner. I de moderna datorspelens 3d-grafik måste beräkningar av vridningar gå mycket snabbt, och därför föredrar man ofta kvaternioner framför matriser. Andra områden där kvaternionerna tillämpas är kontrollteori och robotteknik.

3 Källförtäckning

3.1 Tryckta källor

Boyer, Carl och Merbach, Uta, A History of Mathematics, New York (1989).

Johansson, Bo Göran, Matematisk historia, Studentlitteratur, Lund (2004).

Kantor, I.L. och Solodovnikov, A.S., Hypercomplex Numbers, Originaltitel: Giperkompleksnye chisla, Översättning: A. Shenitzer, Springer-Verlag, New York (1989).

Sparr, Gunnar, Linjär algebra, andra upplagan, Studentlitteratur, Lund (1994).

Stillwell, John, Mathematics and its History, Second edition, Springer-Verlag, New York (2002).

3.2 Övriga källor

http://www.mai.liu.se/ thkar/kurser/TATM87/Sem.Kvaternionerv03.pdf, författare: Keijo Hildén.

http://www.mathworld.com

http://en.wikipedia.org

http://sv.wikipedia.org

Kvaternioner

Gustaf Sörnmo och Tor Gillberg $\pi 04$

20 maj 2005

1 Historik

William Rowland Hamilton föddes vid midnatt mellan den 4 och 5 augusti 1805, vilket har lett till viss oklarhet kring födelsedatumet. Han hade en farbror som var lingvist och ska ha lärt William 14 språk under hans uppväxt. Hamiltons matematikintresse började väldigt tidigt och helt utan någon lärares inblandning. Vid tio års ålder kom han över en bok om Euklidisk geometri som han studerade med iver. Som tolvåring började han studera Newtons *Arithmetica Universalis* och som sjuttonåring upptäckte han ett fel i Laplaces *Celestial Mechanics*. Att Hamilton inte hade någon lärare gjorde att hans matematiska språk fått en egen touch. Under sin collegeperiod fick Hamilton mängder av utmärkelser. Han var alltid först från tentorna och hade alltid bäst resultat. Hamilton slutförde inte sina studier vid colleget sedan han 1827 blivit utsedd till Royal Astronomer of Ireland och han fick som uppgift att sköta ett observatorium i nordvästra Dublin. Hamilton kunde mer än de flesta om teorin kring astronomi men hade ingen erfarenhet av det praktiska arbetet. Ganska snart tröttnade han på att sitta vaken och göra observationer och hyrde in tre av sina systrar. Istället ägnade han sig åt att skriva poesi. Detta var dock inte en av Hamiltons stora begåvningar och han blev snart rådd att inte skriva mer. Hamilton studerade även optik och dynamik och skrev många böcker kring bägge ämnena.

Hamilton var mycket intresserad av de imaginära talen. Den moderna definitionen av de komplexa talen som ett par av reella tal infördes av Hamilton. Länge försökte han hitta ett liknande talsystem för tre dimensioner. Genom att använda talpar av komplexa tal kom han istället fram med en teori för fyra dimensioner. Under en promenad över Brougham "Broom" bron i Dublin den 16 oktober 1843 fick Hamilton idén om reglerna; $i^2 = j^2 = k^2 = ijk = -1$. Genast sprang han in under bron för att där karva in sin geniala tanke. Han döpte de fyrdimensionella talen till kvaternioner och skrev dem på formen $p = a + bi + cj + dk$, där i, j, k är imaginärdelarna. Om kvaternionens tillkomst har han själv sagt att

"*The quaternion was born, as a curious offspring of a quaternion of parents, say of geometry, algebra, metaphysics, and poetry.... I have never been able to give a clearer statement of their nature and their aim than I have done in two lines of a sonnet addressed to Sir John Herschel: 'And how the One of Time, of Space the Three Might in the Chain of Symbols girdled be'*.".

Hamilton satt ofta djupt försjunken i sina egna tankegångar, vilket inte helt sällan ledde till att han glömde bort att äta och sköta andra jordliga ting. Eftersom han var oerhört noggrann med allt han företog sig, blev bara en bråkdel

av hans totala arbete publicerat. Mycket av det han skrev finns inte längre. Sina sista 6 år ägnade han åt boken *Elements of Quaternions*, vilken blev färdig en eller två dagar innan han dog, den 2 september 1865.

2 Räkneregler

De kvaternionska räknereglerna påminner till stor del om räknereglerna för reella tal. Den huvudsakliga skillnaden är att kvaternioner är icke-kommutativa. Här är några exempel på regler som är gemensamma, där p, q och r betecknar kvaternioner.

$$p + q = q + p, \tag{1}$$

$$p + (q + r) = (p + q) + r, \tag{2}$$

$$p(qr) = (pq)r, \tag{3}$$

$$p(q + r) = pq + pr, \tag{4}$$

där (1) är den kommutativa lagen för addition, (2) och (3) de associativa lagarna och (4) den distributiva lagen.

En kvaternion skrivs alltså som $p = a + bi + cj + dk$. Vid räkneoperationer med kvaternionens tre imaginärdelar används

$$i^2 = j^2 = k^2 = ijk = -1. \tag{5}$$

Vi kan nu härleda de icke-kommutativa räknereglerna utifrån (1). Om vi först tittar på $ijk = -1$ kan vi multiplicera med k i båda leden, vilket alltså ger $k = ij$ (eftersom $k^2 = -1$). Detta bevisas på samma sätt för i och j. Vidare ser vi att det även är möjligt att skriva ij på ett annat sätt; $i = jk$ multiplicerat med $-j$ från vänster ger $k = -ji = ij$. Alltså kan vi konstatera att räknelagarna är icke-kommutativa (detta kan även ses ur de kvaternionska matrisrepresentationerna, se avsnitt 2.2.2). Kvaternionernas icke-kommutativa egenskaper illustreras effektivt med följande tabell

.	*1*	*i*	*j*	*k*
1	1	i	j	k
i	i	-1	k	-j
j	j	-k	-1	i
k	k	j	-i	-1

Hamilton presenterade kvaternionerna som en realdel och en vektor (vi väljer att använda två kvaternioner p och q, då de båda kommer nyttjas vid senare uträkningar och bevis)

$$p = a + \vec{u}, \tag{6}$$

$$q = t + \vec{v}. \tag{7}$$

Vektorn består alltså av de tre imaginärdelarna

$$\vec{u} = bi + cj + dk,$$

$$\vec{v} = xi + yj + zk.$$

Likt de komplexa talen har kvaternioner ett konjugat, definierat som

$$\bar{p} = a - \vec{u},$$

$$\bar{q} = t - \vec{v}.$$

Multiplikation för kvaternioner är även definierad, men då icke-kommutativitet råder är mulitiplikationsordningen av betydelse. Ett enkelt exempel följer. Nu passar vi även på att härleda en formel för multiplikation med kvaternionen uppställd som en realdel plus en vektor, se (6) och (7).

$$\begin{aligned} pq = (a + \vec{u})(t + \vec{v}) &= (a + bi + cj + dk)(t + xi + yj + zk) \\ &= at - bx - cy - dz + i(ax + bt + cz - dy) \\ &\quad + j(ay + ct + dx - bz) + k(az + dt + by - cx) \\ &= at - \vec{u} \cdot \vec{v} + t\vec{u} + a\vec{v} \\ &\quad + i(cz - dy) + j(dx - bz) + k(by - cx). \end{aligned}$$

Vi nöjer oss med att visa multiplikation åt ett håll, men produkten kommer att ändra tecken beroende på ordningen. Som vi tidigare har nämnt, infördes vektorprodukten i samband med kvaternionen och vi kan nu konstatera att vår okända produkt blir just vektorprodukten

$$\vec{u} \times \vec{v} = -\vec{v} \times \vec{u} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ b & c & d \\ x & y & z \end{vmatrix} = i(cz - dy) + j(dx - bz) + k(by - cx).$$

En enkel formel för att räkna multiplikation med kvaternioner blir således

$$pq = at - \vec{u} \cdot \vec{v} + a\vec{v} + t\vec{u} + \vec{u} \times \vec{v}. \tag{8}$$

Konjugatet kan användas för att reducera bort eventuella imaginärdelar. Detta används främst vid division

$$p^{-1} = \frac{1}{p\bar{p}}\bar{p}. \tag{9}$$

Absolutbeloppet av en kvaternion brukar kallas dess norm $N(p)$, definierad som

$$\begin{aligned} |p| = N(p) &= \sqrt{p\bar{p}} = \sqrt{\bar{p}p} \\ &= \sqrt{(a + \vec{u})(a - \vec{u})} \\ &= \sqrt{a^2 - \vec{u} \cdot \vec{u}} \\ &= \sqrt{a^2 + b^2 + c^2 + d^2}. \end{aligned} \tag{10}$$

2.1 Skalärprodukt

I kvaternionska sammanhang kallas skalärprodukten ofta för inre produkt. Den inre produkten kan användas för att isolera separata element ur kvaternionen. Man kan även räkna ut kvaternionens yttre produkt, även känd som hakprodukt (skrivs matematiskt som $\wedge$).

$$p \cdot q = \frac{\bar{p}q + \bar{q}p}{2}. \tag{11}$$

Om vi beräknar (11) genom att använda de grundläggande räkneregler vi formulerade i början av kapitlet får vi

$$\begin{aligned}\frac{1}{2}[(a-\vec{v})(t+\vec{u}) + (t-\vec{u})(a+\vec{v})] &= \frac{1}{2}[at - \vec{v}\cdot\vec{u} + a\vec{u} - \vec{v}t + ta - \vec{u}a + t\vec{v} - \vec{u}\cdot\vec{v}] \\ &= \frac{1}{2}[2at - 2\vec{u}\cdot\vec{v}] \\ &= at - \vec{u}\cdot\vec{v} \\ &= at + bx + cy + dz.\end{aligned}$$

Ett mer demonstrativt exempel på den inre produktens isolerande egenskaper ges här

$$\begin{aligned}p \cdot i = \frac{1}{2}[(a-\vec{v})(i) + (-i)(a+\vec{v})] &= \frac{1}{2}[ai - \vec{v}i - ia - i\vec{v}] \\ &= \frac{1}{2}[-\vec{v}i - i\vec{v}] \\ &= \frac{1}{2}[b - cji - dki + b - cij - dik] \\ &= b.\end{aligned}$$

Ur kvaternionen p har vi alltså isolerat b, den reella konstant som ursprungligen tillhörde imaginärdelen i.

Hakproduktens formulering är mycket lik den för skalärprodukt

$$p \wedge q = \frac{\bar{p}q - \bar{q}p}{2}. \tag{12}$$

Ett exempel på hakproduktens egenskaper

$$\begin{aligned}p \wedge i = \frac{1}{2}[(a-\vec{v})(i) - (-i)(a+\vec{v})] &= \frac{1}{2}[ai - \vec{v}i + ia + i\vec{v}] \\ &= \frac{1}{2}[2ai + b - cji - dki - b + cij + dik] \\ &= ai.\end{aligned}$$

Vi ska inte gå in på djupet med vektorprodukten, men den är vid kvaternionräkningar relaterad till skalär- och hakprodukten. Vektorprodukten skrivs som

$$p \times q = \frac{pq - qp}{2}. \tag{13}$$

2.2 Matrisdefinitioner

Som vi tidigare har nämnt går det att utnyttja matriser för att representera kvaternioner. Matrisernas icke-kommutativa struktur visar sig fungera utmärkt för att beskriva imaginärdelarnas egenskaper.

2.2.1 Komplexa tal

Ett komplext tal $w = aE + bI$ kan på matrisform definieras som

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{14}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = E,$$

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = I.$$

Matrisen E med den reella konstanten a definieras som det komplexa talets realdel (enhetsmatrisen). Matrisen I svarar mot det komplexa talets imaginärdel. Det visar sig att de komplexa talens räkneregler går att utvinna ur matriserna, t.ex får vi att $E^2 = E$ och $I^2 = -E$.

2.2.2 Kvaternioner

Vi beskriver två former för att representera kvaternioner med matriser. I de båda formerna svarar kvaternionsk addition och multiplikation mot addition och multiplikation med matriser. Första formen består av en 4×4 matris där kvaternionen p blir

$$\begin{pmatrix} a & -b & d & -c \\ b & a & -c & -d \\ -d & c & a & -b \\ c & d & b & a \end{pmatrix}. \tag{15}$$

Här motsvarar matrisens transponat kvaternionens konjugat.

Den andra formen, som vi huvudsakligen kommer att jobba med är en utveckling av de komplexa talens 2×2 matrisform i (14)

$$\begin{pmatrix} a - di & -b + ci \\ b + ci & a + di \end{pmatrix}. \tag{16}$$

Vi skriver här kvaternionen som $z = aE + bI + cJ + dK$, där E och I symboliseras av matriser med samma utseende som för komplexa tal (se avsnitt 2.2.1). Med enkel utveckling av (16), finner vi de resterande imaginärdelarnas matrisrepresentanter ger

$$\begin{pmatrix} a - di & -b + ci \\ b + ci & a + di \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} + i \begin{pmatrix} -d & c \\ c & d \end{pmatrix}$$

$$= a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + i\left[c\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + d\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

$$= a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + c\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} + d\begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}.$$

$$\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = J,$$

$$\begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = K.$$

2.2.3 Exempel med matriser

Vi ska nu undersöka hur många kvaternionska lösningar det finns för en viss polynomekvation, till exempel

$$z^2 + 1 = 0.$$

Vi använder oss av matrisdefinitionen (16) för att lösa ekvationen.

$$\begin{aligned} z^2 &= \begin{pmatrix} a - di & -b + ci \\ b + ci & a + di \end{pmatrix}^2 \\ &= \begin{pmatrix} a^2 - b^2 - c^2 - d^2 - 2adi & -2ab + 2aci \\ 2ab + 2aci & a^2 - b^2 - c^2 - d^2 + 2adi \end{pmatrix} \\ &= \begin{pmatrix} a^2 - b^2 - c^2 - d^2 & 2ab \\ 2ab & a^2 - b^2 - c^2 - d^2 \end{pmatrix} + 2i\begin{pmatrix} -ad & ac \\ ac & ad \end{pmatrix} \\ &= (a^2 - b^2 - c^2 - d^2)E + 2abI + 2aJ + 2adK = -1. \end{aligned}$$

Eftersom de imaginära komponenterna I, J och K inte kan uppfylla högerledet i uträkningens nedre rad, måste $a = 0$ vilket ger oss en ekvation med enbart reella tal.

$$\Rightarrow \quad -b^2 - c^2 - d^2 = -1 \quad \Leftrightarrow \quad b^2 + c^2 + d^2 = 1.$$

Här ser man direkt att det måste finnas oändligt många kvaternionska lösningar, eftersom a, b och c är reella tal.

Man kan givetvis göra en betydligt mer djuplodande analys av hur antalet kvaternionska lösningarna förhåller sig till olika polynomekvationer, men vi nöjer oss med att konstatera att det i många fall rör sig om ett oändligt antal lösningar. Som nästa steg skulle man, exempelvis, kunna undersöka om $z^2 = q$ alltid har lösningar (där q är en kvaternion).

2.3 Dimensionella möjligheter

Kvaternioner är ett tal i fyra dimensioner; man kan fråga sig varför det inte finns någon motsvarighet i exempelvis tre dimensioner. Om vi nu antar att vi, istället för tre, har två imaginära delar som betecknas i och j. Varje tal kan då skrivas $w = x + yi + zj$ där x, y och z är reella konstanter.

För att rummet av våra tal skall vara slutet under multiplikation, måste vi definiera ij som $ij = x + yi + zj$.

Men vi vill samtidigt att associativitet skall gälla. Vi får då att

$$i \cdot (ij) = (ii) \cdot j = (-1) \cdot j = -j. \tag{17}$$

Enligt ovan får vi nu

$$\begin{aligned} i \cdot (ij) &= i(x + yi + zj) \\ &= xi - y + zij \\ &= xi - y + z(x + yi + zj) \\ &= (zx - y) + (x + zy)i + z^2 j \end{aligned}$$

där (17) ger att

$$-j = (zx - y) + (x + zy)i + z^2 j.$$

För att vänster- och högerleden skall vara lika måste vi ha

$$\begin{cases} zx - y = 0 \\ x + zy = 0 \\ z^2 = -1. \end{cases}$$

Men eftersom x, y och z är reella konstanter kan den sista likheten inte gälla. Alltså kan vi konkludera att komplexa tal inte förekommer i tre dimensioner.

Med ett mer allmänt bevis, som är uppbyggt på liknande sätt, kan man visa att det bara finns komplexa tal i 2^n dimensioner (för högre dimensioner är samlingsnamnet hyperkomplexa tal), med en reell komponent och $n - 1$ imaginära komponenter.

3 Tillämpningar

När Hamilton upptäckte kvaternionerna 1843 trodde han och hans samtida matematiker att upptäckten skulle innebära en mindre matematisk revolution. Man hoppades att kvaternionerna skulle vara lika användningsbara som de komplexa talen. Det visade sig dock snabbt att deras användningsområde är avsevärt mycket mindre än de komplexa talen. Ett exempel på detta är i samband med analytiska funktioner, som för kvaternioner är ytterst begränsade och till stor del oanvändbara.

Trots att kvaternionerna i många matematiska sammanhang avfärdades som totalt irrelevanta, kan man säga att de indirekt banade väg för vektorer och olika vektoroperation, t.ex skalär- och vektorprodukt. En tidigt variant av Maxwells vågekvationer skrevs faktiskt mha kvaternioner.

Kvaternionernas särställning har inte förändrats märkbart i dagens teoretiska matematik. Däremot används de flitigt i diverse programmeringssammanhang, t.ex grafik, kontrollteori och signalbehandling, framför allt för att talen kan representera rotationer och orienteringar. Matriser kan även användas till detta ändamål, men kvaternionerna ger färre räkneoperationer. Detta är en högst önskvärd egenskap och alltså föredrar man här att arbeta med kvaternioner.

4 Källor

http://en.wikipedia.org/wiki/Quaternion
http://mathworld.wolfram.com/Quaternion.html
www.mai.liu.se/ kehil/XXX.pdf
http://www.daviddarling.info/encyclopedia/Q/quaternion.html

Dimension och fraktaler

Hampus Sahlin
Petter Strandmark
Daniel Svensson
Karl Wernersson

9 maj 2005

> ”År 1953 insåg jag att den räta linjen för till mänskligheten undergång. Men den räta linjen har blivit rena tyranniet. Den räta linjen dras fegt med linjal utan tanke eller känsla; det är den linje som inte existerar i naturen. Och den linjen är det ruttna fundamentet för vår dödsdömda civilisation. Trots att man på vissa håll medgivit att den leder i fördärvet fortsätter man att följa den. [...] Varje mönster som leder till den räta linjen är dödfött. Idag befinner vi oss i rationalismens triumf och finner oss samtidigt att stå inför tomheten. Ett estetiskt tomrum, en öken av enformighet, kriminell sterilitet, förlorad skaparkraft. Till och med kreativiteten är prefabricerad. Vi har blivit maktlösa. Vi kan inte skapa längre. Det är vår verkliga analfabetism.”

-Freidensreich Hundertwasser, Österrikisk målare

Ordet fraktal myntades inte förrän så sent som 1975 av Benoit Mandelbrot men människan har jobbat med fraktalliknade problem sedan 1100-talet – antagligen tidigare. Den ordentliga föregångaren till fraktalforskningen startades inte förrän under 1870 då man lyckades skapa kurvor som fyllde upp kvadrater, kurvor som korsade sig skälva i varje punkt, kurvor med oändlig längd och ändlig area samt kurvor utan längd. Nästan alla dessa exemplen kom från mängdläran.

Först ut var Karl Weierstrass då han 1872 lyckades bevisa att det fanns kontinuerliga funktioner som saknar derivata. Detta vände uppochner på den klassiska analysen, vilket gjorde de flesta matematiker ganska förbannade. 1883 uppfann Georg Cantor cantormängden i viken man efter oändligt med iterationer har tagit bort hela sträckan men lämnat oändligt många punkter kvar. Se figur 3.

Nästa genombrott kom när Giuseppe Peano 1890 skapade en kurva som korsade varje punkt i en godtycklig kvadrat. Detta komplicerade dimensionsdefinitionen eftersom man nu kunde beskriva en tvådimensionell yta med en variabel, se figur 1.

De följdes 1906 upp av Helge von Kochs snöflinga, vilket är en geometrisk figur som precis som i Weierstrass funktion saknar derivata. Den har även en annan intressant egenskap då den har en ändlig area men en oändlig omkrets. Se figur 4.

Figur 1: Peanokurvan är en kurva som fyller en yta.

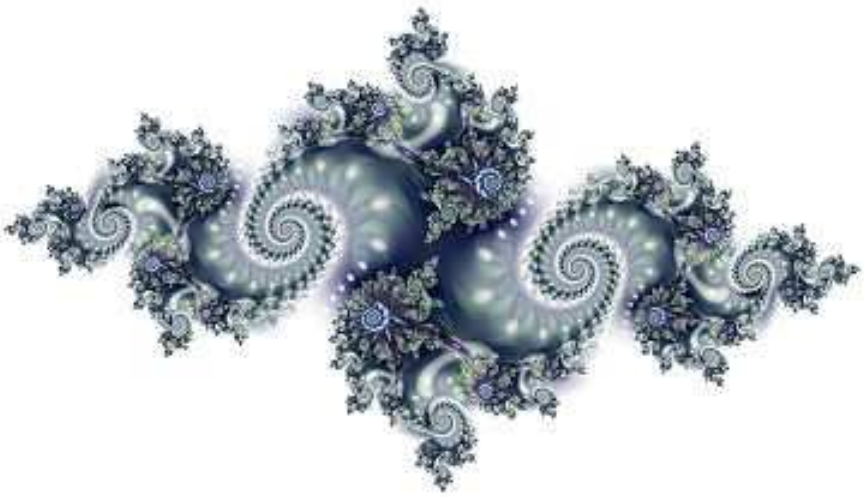
1918 publicerade Gaston Julia Journal de Mathematic Pure et Appliqué i vilken han behandlar Juliamängder vilket är iterering av rationella funktioner i det komplexa planet t.ex. Mandelbrotsmängden $z_{n+1} = z_n^2 + C$ som stannar inom en sluten mängd då n går mot oändligheten. Detta gav honom le Grand Prix de l'Académie des Sciences vilket gjorde honom världskänd under 20-talet.

Många matematiker ogillade dessa nya funktioner Poincaré kallade dem ett galleri av monster och Hermite skrev om hur han "vände sig i skräck och rädsla ifrån den hemska plågan av funktioner utan derivata" och han försökte stoppa Lebesgue att publicera en artikel om icke-differentierbara ytor. De flesta matematiker hade vid den här tiden samma åsikt som Hermite viket gjorde att ofta när Lebesgue försökte delta i en matematisk var det någon analysprofessor som sa "detta intresserar inte dig, vi talar om deriverbara funktioner". Detta ledde till att mycket av dessa funktioner och figurer glömdes bort.

1945 introducerades Benoit Mandelbrot av sin farbror för Juliamängden som ansåg att det var ett mycket intressant problem. Mandelbrot var däremot inte alls intresserad och började jobba med vad som kan tyckas vara helt andra saker men 1970 ledde hans forskning tillbaka till Juliamängderna och med hjälp av datorer kunde man nu rita upp dem på ett effektivt sätt.

> "Hur kommer det sig att geometrin ofta beskrivs som kall och torr? En orsak ligger i dess oförmåga att beskriva formen på ett moln, ett berg, en kust eller ett träd. Moln är inga sfärer, berg är inga koner, kuster är inga cirklar, bark är inte slät och blixtar beskriver inga räta linjer. Mer generellt vill jag hävda att många av naturens mönster är så oregelbundna och fragmenterade ett naturens mönster inte bara är högre än den euklidiska geometrin utan av en helt annan art."

- Benoit B. Mandelbrot [5]



Figur 2: Exempel på en Juliamängd.

Dimension

Vad menar vi egentligen med begreppet dimension? Det finns flera tänkbara sätt att definera dimension, och hur man gör det beror på vad man vill åstadkomma.

I denna uppsats kommer vi att vilja behandla olika geometriska figurer, eller punktmängder som vi hädanefter skall kalla dem. Ett möjligt sätt är att införa ett kordinatsystem för att se hur många variabler som krävs för beskriva mängden. I ett plan kan alla punkter beskrivas med två variabler och vi kallar det därför två-dimensionellt [1]. Dimension hänger också samman med hur man mäter mängder och hur måttet skalar, en kvadrat har ett tvådimensionellt mått som vi kallar area, när vi dubblar sidan kommer arean att öka med en faktor 2^2.

Här vill vi även kunna beräkna dimensioner för fraktaler, tex mänger som saknar area men har oändlig längd. Med detta i åtanke använder vi Hausdorffdimension uppkallat efter Felix Hausdorff.

Hausdorffdimension

Låt F vara en vanlig kurva i $\mathbb{R}^3$ med definerad längd som man vill beräkna längden på. Ett sätt är att täcka kurvan med klot med diameter δ. Då är längden summan av klotens diametrar.

$$L(F) = \sum_{i=1} \delta_i$$

Eftersom kloten fortfarande kan överlappa varandra så säger vi att vi tar minsta möjliga antal klot som krävs för att täcka kurvan.

$$L(F) = \inf \left\{ \sum_{i=1} \delta_i \right\}$$

För att förbättra approximationen låter vi $\delta \to 0$.

$$L(F) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} \delta_i \right\} \right\}$$

Låt nu F vara en vanligt ytstycke med bestämd area som man vill beräkna. Om vi använder samma metod som ovan kan vi beräkna arean. $\gamma = \frac{\pi}{4}$.

$$A(F) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} (\delta_i^2 \gamma) \right\} \right\}$$

Mer allmänt kan det d-dimensionella måttet bestämmas med

$$\mathscr{H}^d(F) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} (\delta_i^d \gamma) \right\} \right\} \tag{1}$$

[1] Det kan beskrivas med en variabel (t.ex. med en Peanokurva), men inte på ett särskilt meningsfullt sätt

Detta mått kallas Hausdorffmåttet för d.

Vad händer om vi försöker mäta längden på den en yta?

$$L(Y) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} \delta_i \right\} \right\}$$

Vi kan se att detta gränsvärde divergerar då $\delta \to 0$, en given yta kan inte täckas av ett ändligt antal linjestycken.

Vi försöker även mäta ytans volym:

$$V(Y) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} (\delta_i^3 \gamma) \right\} \right\}$$

Detta gränsvärde går mot 0 då $\delta \to 0$. Det enda meningsfulla måttet för en yta är det tvådimensionella – arean.

Vi kan definera dimensionen för en mängd som den kritiska dimension där måttet ändras från oändligheten till 0.

Sats 1 *För varje mängd $F \neq \emptyset$ finns ett D sådant att.*

$$\mathscr{H}^d(F) = \lim_{\delta \to 0} \left\{ \inf \left\{ \sum_{i=1} (\delta_i^d \gamma) \right\} \right\} = \begin{cases} \infty, & d < D \\ 0, & d > D \end{cases} \tag{2}$$

Detta D kallas Hausdorffdimensionen för mängden F. Detta sätt att definera dimension är teoretiskt tillräckligt då det fungerar för alla mängder, men i praktiska sammanhang behövs andra metoder att räkna ut eller approximera dimensioner. Obsververa även att D inte behöver vara ett heltal, vi kommer att få se flera exempel på detta.

Sats 2 *Om $F \in \mathbb{R}^n$ och $\lambda > 0$ så gäller att*

$$\mathscr{H}^d(\lambda F) = \lambda^d \mathscr{H}^d(F)$$

där $\lambda F = \{\lambda x : x \in F\}$, dvs. mängden F skalad med en faktor λ.

För bevis, se [6], sida 27.

Exempel

Här följer några exempel på fraktaler samt hur man räknar ut deras Hausdorffdimensioner.

Cantormängden

Tag en linje med längden ett, dela den i tre delar, tag bort den mittersta delen. Tag nu de två återstående delarna och dela var och en i tre delar, tag bort den mittersta delen på var och en osv. Det du får är Cantormängden. Se figur 3.

Cantormängdens dimension

Dela in Cantormängden F i två delar, $F_v = F \cap [0, \frac{1}{3}]$ och $F_h = F \cap [\frac{2}{3}, 1]$. Vi ser då att både F_v och F_h geometriskt liknar F men är skalade 1:3, och att $F = F_v \cup F_h$. Om vi använder (2) så får vi för något D:

$$\mathscr{H}^D(F) = \mathscr{H}^D(F_v) + \mathscr{H}^D(F_h) = \Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F) + \Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F) = 2\Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F)$$

Vi söker D, Cantormängdens dimension. För denna kan vi anta att $0 < \mathscr{H}^D(F) < \infty$, vilket gör att vi kan dividera med $\mathscr{H}^D(F)$.

$$1 = 2\Big(\frac{1}{3}\Big)^D \iff D = \frac{\log 2}{\log 3}$$

Cantormängdens dimension blir då $D \approx 0{,}6309$

Von Kochs kurva

Tag en linje med längden ett, dela den i tre delar, tag bort den mittersta och lägg till två nya (se figur 4). Repetera samma sak för varje ny del så får du von Kochs kurva.

Dimension på von Kochs kurva

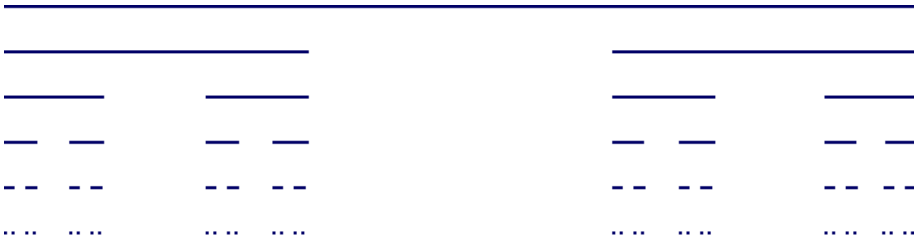
Dela in kurvan F i tre delar, F_v, F_m och F_h. F_v och F_h liknar geometriskt F men är skalade 1:3, medan F_m består av två mot varandra lutande F skalade 1:3. Vi får då att $F = F_v \cup F_m \cup F_h$, tillsammans med (2) för något D får vi

$$\mathscr{H}^D(F) = \mathscr{H}^D(F_v) + \mathscr{H}^D(F_m) + \mathscr{H}^D(F_h) =$$
$$= \Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F) + 2\Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F) + \Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F) = 4\Big(\frac{1}{3}\Big)^D \mathscr{H}^D(F)$$

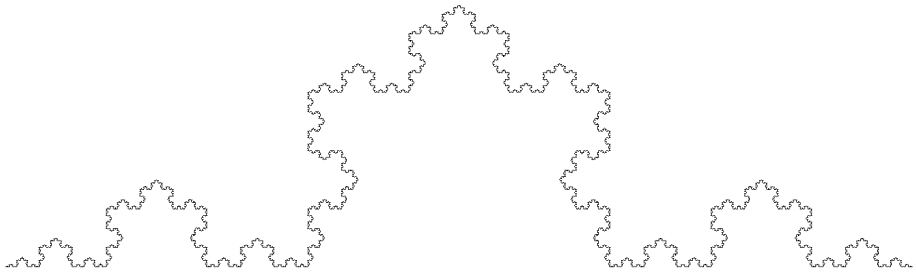
Om vi antar att kurvan har ett mått $0 < \mathscr{H}^D(F) < \infty$ så kan vi dividera med $\mathscr{H}^D(F)$.

$$1 = 4\Big(\frac{1}{3}\Big)^D \iff D = \frac{\log 4}{\log 3}$$

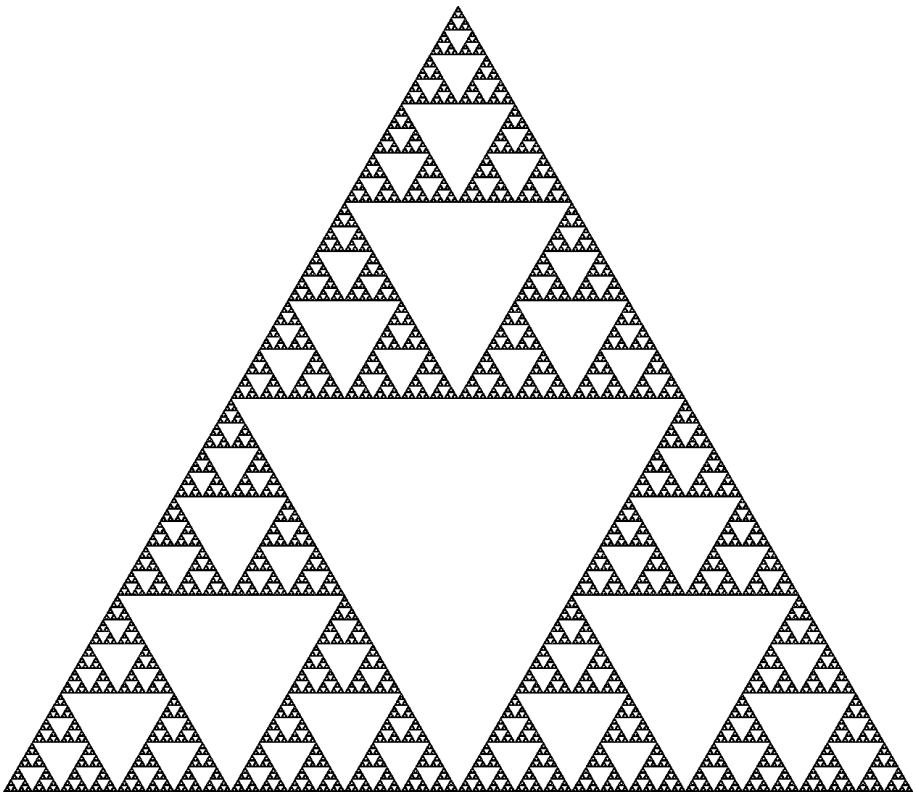
Dimensionen på von Kochs kurva blir då $D \approx 1{,}2619$.



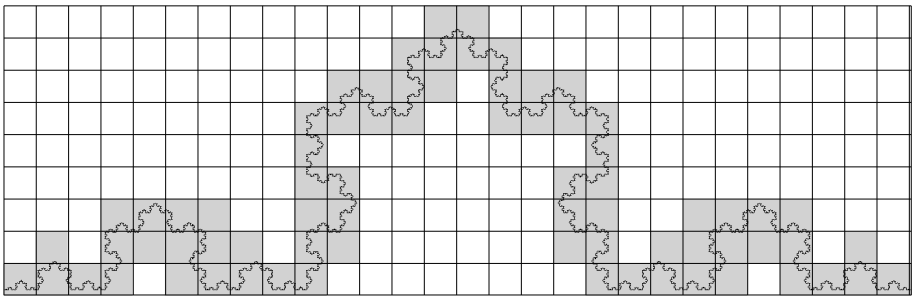
Figur 3: Cantormängden har en dimension på $\log 2/\log 3 \approx 0{,}6309$. Detta betyder att den innehåller ett oändligt antal punkter ($\mathscr{H}^0 = \infty$), men har ingen längd ($\mathscr{H}^1 = 0$).



Figur 4: Von Kochs kurva. Dimensionen är $\log 4/\log 3 \approx 1{,}2619$.



Figur 5: Sierpinskitriangelen har en dimension på $\log 3/\log 2 \approx 1{,}5850$.



Figur 6: Lådräkning på von Kochs kurva. $\delta = 50$ bildpunker och $N_\delta(F) = 73$.

Numerisk beräkning av dimension

För att räkna ut Hausdorffdimensionen täcker vi över en mängd med små mängder med en viss diameter. Dessa små mängder kan se ut hur som helst och även överlappa varandra, vilket gör det besvärligt att bestämma Hausdorff-dimensionen, utom i vissa specialfall vi har sett exempel på ovan.

För att numeriskt bestämma dimensionen hos en mängd ställer vi större krav på dessa små mängder: de skall vara rutor (eller n-dimensionella lådor) i ett kvadratiskt rutnät som mängden ifråga täcks över med. Detta gör, som vi skall se, det lätt att beräkna en dimension av t.ex. en bild på en fraktal med hjälp av en dator.

Lådräkningsdimension[2]

Tag en mängd F i ett rektangulärt område. Lägg ett rutnät på detta område med bredden δ_0 på rutorna. Kalla nu antalet rutor som innehåller någon del av F för $N_{\delta_0}(F)$. Se figur 6.

Vad händer då vi gör rutnätet mindre? Hur förändras $N(F)$? Om F är en linje, som är endimensionell, kommer $N(F)$ att fördubblas om rutnätet görs dubbelt så fint. Är istället F en tvådimensionell mängd, t.ex. en rektangel, kommer $N(F)$ öka med en faktor $2^2 = 4$. För en tredimensionell kub blir ökningen istället $2^3 = 8$ osv.

Om föregående stycke generaliseras till en godtycklig dimension D och rutnätets upplösning förändras från δ_0 till δ erhålls uttrycket

$$N_\delta(F) = \left(\frac{\delta_0}{\delta}\right)^D N_{\delta_0}(F),$$

då $\delta \to 0$, vilket är samma resultat som precis resonerades fram. Nu tar vi logaritmen av båda leden:

$$\log N_\delta(F) = D \log\left(\frac{\delta_0}{\delta}\right) + \log N_{\delta_0}(F),$$

[2]Eng. box-counting dimension

som efter omskrivning får formen

$$\log N_\delta(F) = D \cdot -\log \delta + \log \delta_0 N_{\delta_0}(F). \tag{3}$$

Sambandet mellan $-\log \delta$ och $\log N_\delta(F)$ är alltså en rät linje, med riktningskoefficienten D. Detta gör att lådräkningsdimensionen är mycket lätt att beräkna numeriskt utifrån bilder.

Det bör påpekas att eftersom $N_\delta(F) \to \infty$ om $\delta \to 0$ och $D > 0$, så kan (3) skrivas som

$$D = \lim_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta}, \tag{4}$$

vilket är den definition som brukar hittas i litteratur.

Detta dimensionsbegrepp heter egentligen Minkowski-Bouligand-dimension och hänger starkt samman med Hausdorff-dimensionen. För alla mängder är $D_{\text{Haus}} \leq D_{\text{Låd}}$ och likhet råder för många fraktaler. Ett exempel där likhet inte råder är den uppräkningsbara mängden $\left\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}$, som har lådräkningsdimension $\frac{1}{2}$, men Hausdorffdimension lika med, kanske lite mer intuitivt, 0.

Dimension av bilder

Om man har en bild på en fraktal och vill ta reda på dess dimension kan man med fördel använda sig av (3). Genom att beräkna N_δ för olika värden på δ och anpassa en linje till dessa värden erhålls D som lutningen på denna linje. Figur 7 visar beräknade dimensioner på fyra olika mängder. Observera att linjens position i höjdled är ointressant, eftersom den beror på vilken enhet δ mäts i och inte har något att göra med bildens dimension.

Självklart kan denna metod generaliseras till delmängder av $\mathbb{R}^3$ och $\mathbb{R}^n$, men antalet lådor som skall räknas ökar förstås då exponentiellt.

Kuststräckor

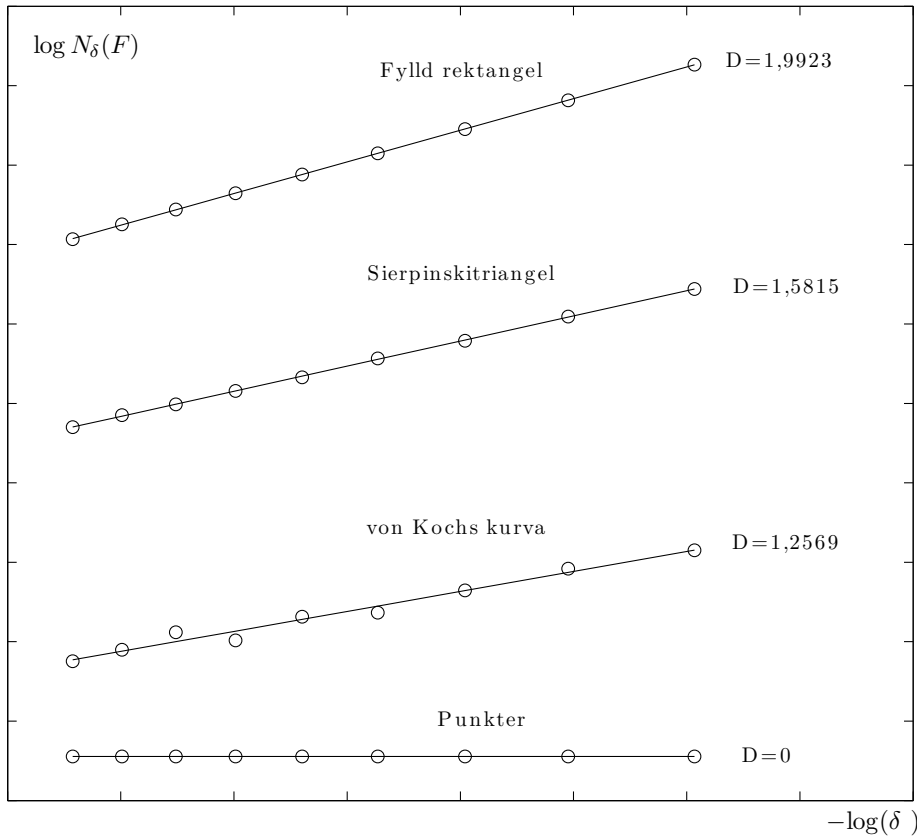
Hur lång är Skånes västkust mellan Malmö och Helsingborg? Lägger man ett snöre längs på en sverigekarta och mäter sträckan kommer man kanske att få ett värde på 7 mil. Men vad händer om man istället mäter på en karta med högre upplösning, skulle då inte kuststräckan bli avsevärt mycket längre? Jo, och skulle någon få för sig att gå längs hela kuststräckan med ett måttband skulle kusten mellan de båda städerna bli oerhört mycket längre än så. Ju mer man förminskar skalan, desto längre blir sträckan.

Kuststräckor har alltså ett endimensionellt mått som är lika med oändligheten, men beränsas av ett ändligt ytområde. Följdaktligen borde den ha en fraktal dimension > 1 och med metoden som introducerades ovan borde den även gå att beräkna. Med hjälp av högupplösta kartor från GIS-centrum vid Lunds Universitet har dimensionen på Skånes västkust kunnat beräknas till ungefär 1,1. Se figur 8.

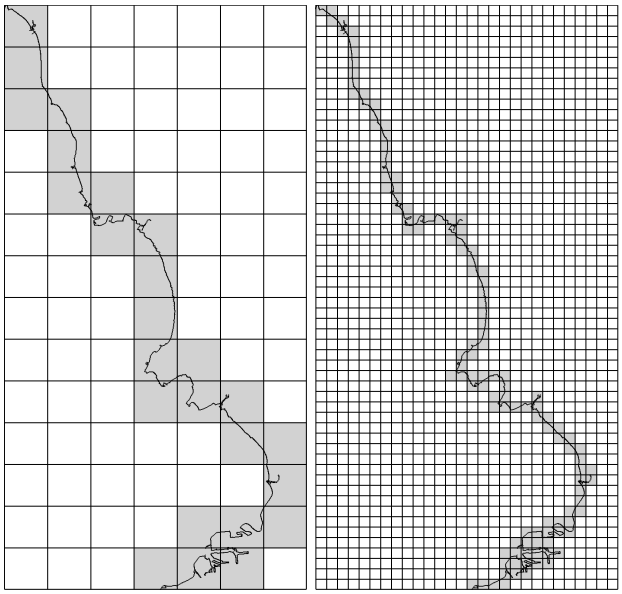
Fraktaler är ofta mycket bra modeller för naturliga företeelser, såsom kuststräckor, landskap och moln. När man inom datorgrafiken vill skapa slumpmässiga

landskap är fraktaler ett viktigt redskap. Figur 9 visar en slumpmässigt skapad fraktal med samma dimension som Skånes västkust. Visst kan kan man se likheter med en naturlig kustlinje?

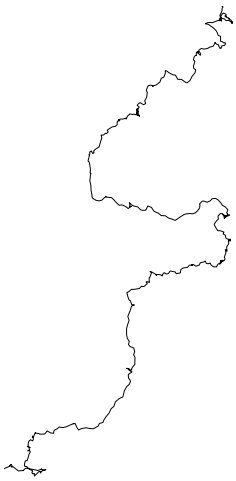
Beräkningar på Norges kust gjorda av Jens Feder [1] uppskattar dimensionen på Norges kust till $1{,}52 \pm 0{,}01$. Detta kan vi tro på, eftersom fraktalens dimension kan ses som ett mått på dess "sönderbrutenhet", vilket medför att Norges kust sannolikt har en högre dimension än Skånes.



Figur 7: Dimension på olika fraktaler: teoretiska värden är 2, 1,5850, 1,2619 och 0. Noggrannheten beror helt på bildens upplösning.



Figur 8: Beräkning av dimensionen av Skånes västkust från Helsingborg till Malmö. Skånes kust är ganska slät och har en fraktal dimension på omkring 1,1.



Figur 9: Slumpmässigt skapad fraktal med samma dimension som Skånes västkust (ungefär 1,1).

MATLAB-kod

Koden som användes för att beräkna bilders dimension i MATLAB finns här utskriven. Exempel på användning:

Exempel

```
D = boxDimension( imread('fraktal.bmp') );
```

Bilden skall vara svartvit, med 1-bits färgdjup. Fraktalen antas vara svart.

boxDimension.m

```
% Bilden som skickas till funktionen skall vara svartvit (1-bit)
% Mängden, vars dimension skall räknas ut, har färgen svart (=0).
%
% Skrivet av Petter Strandmark 2005
%
function D = boxDimension(I, deltaRange)

   if nargin < 2
      %Dessa storlekar på boxar skall vi räkna
      deltaRange = 4:12;
   end

   %Räkna boxar för dessa storlekar
   for n = 1:length(deltaRange)
      delta = deltaRange(n);
      logN(n)     = log(N(delta,I));
      logDelta(n) = -log(delta);
   end

   %Hitta en linje som passar till datan
   lineP = polyfit(logDelta,logN,1);
   %Dimensionen är lutningen på denna linje
   D = lineP(1);

   %Om utargument inte önskas, rita en graf istället
   if nargout == 0
      plot(logDelta,logN,'o');
      hold on
      %Rita dimensionslinjen som jämförelse
      plot(logDelta,polyval(lineP,logDelta),'-');
   end
end

%Ger 1 om I innehåller någon pixel fylld (del av mängden)
%annars 0 om I är helt tom
function n = check(I)
   n = min(1, length( find( I == 0) ));
end

function boxCount = N(delta,F)
   [width height] = size(F);
   %Räkna boxarna som skär mängden F
   boxCount = 0;
```

```
   for row = 1:delta:width
      for col = 1:delta:height
         boxCount = boxCount + check(F(row: min(row+delta-1,width),...
                                  col:min(col+delta-1,height)) );
      end
   end
end
```

Referenser

[1] Feder, Jens. *Fractals.* Plenum Press, 1988.

[2] Peitgen, Heinz-Otto och Saupe, Dietmar. *The Science of Fractal Images.* Springer Verlag, 1988.

[3] Peitgen, Heinz-Otto et al. *Fractals for the Classroom.* Springer Verlag, 1992.

[4] Mandelbrot, Benoit B. *Fractals – Form, Chance and Dimension.* 1977.

[5] Mandelbrot, Benoit B. *The Fractal Geometry of Nature.* 1983.

[6] Kenneth Falconer. *Fractal Geometry – Mathematical Foundations and Applicatons.* 1983.

Projekt i Matematisk Kommunikation
Principalkomponentanalys

Sebastian Nilsson, Jonas Olsson och Björn Samvik

Handledare: Johan Råde

18 maj 2005

Innehåll

1 Inledning

Principalkomponentanalys (PCA) innebär att man reducerar antalet dimensioner i en datamängd till ett greppbart antal (2 eller 3) på ett sådant sätt att variansen är så stor som möjligt. Förhoppningsvis innebär detta att mycket av informationen är bevarad. PCA (från engelskans Principal Component Analysis) är ett bra verktyg för att analysera data. Det lättaste exemplet av PCA är projektionen från rummet till planet där man går från tre till två dimensioner. PCA är alltså ett statistiskt verktyg som har stor användbarhet inom många områden. Exempel på användning av PCA är bildkompression, bildjämförelser (fingeravtryck, iris, ansikten), m.m. Ytterligare ett användningsområde är att analysera gendata, i vårt fall vilka gener som är aktiva vid olika cancerdiagnoser.

2 Projektionsformeln

2.1 Sats (Projektionsformeln)

Låt $e_1, e_2, ..., e_n$ vara en ortonormerad bas i R^n. Då kan vektorn $\mathbf{u}$ skrivas som

$$\mathbf{u} = (\mathbf{u} \cdot e_1)e_1 + (\mathbf{u} \cdot e_2)e_2 + ... + (\mathbf{u} \cdot e_n)e_n$$

2.2 Bevis

Eftersom $e_1, e_2, ..., e_n$ utgör en ON-bas existerar $C_1, C_2, ..., C_n$ så att

$$\mathbf{u} = C_1 \cdot e_1 + C_2 \cdot e_2 + ... + C_n \cdot e_n$$

Genom att skalärmultiplicera med lämplig riktning fås

$$\mathbf{u} \cdot e_i = C_1 e_1 e_i + C_2 e_2 e_i + ... + C_n e_n e_i = C_i$$

2.3 Projektion på en linje

I rummet R^3 så kan $\mathbf{u}$ skrivas som

$$\mathbf{u} = (\mathbf{u} \cdot e_1)e_1 + (\mathbf{u} \cdot e_2)e_2 + (\mathbf{u} \cdot e_3)e_3$$

Exempelvis fås projektionen av $\mathbf{u}$ på e_2 genom skalärmultiplikation med e_2.

$$\mathbf{u} \cdot e_2 = (\mathbf{u} \cdot e_1)e_2 + (\mathbf{u} \cdot e_2)e_2 + (\mathbf{u} \cdot e_3)e_2 = 0 + (\mathbf{u} \cdot e_2)e_2 + 0$$

2.4 Projektion på ett plan

Anta att vi har ett plan π som går genom origo, där e_1 och e_2 spänner upp hela planet, och att e_1, e_2, e_3 är en ON bas för R^3. Det medför att e_3 är vinkelrätt mot π.

$\mathbf{u} = (\mathbf{u} \cdot e_1)e_1 + (\mathbf{u} \cdot e_2)e_2 + (\mathbf{u} \cdot e_3)e_3$ där alltså $(\mathbf{u} \cdot e1)e_1$ och $(\mathbf{u} \cdot e_2)e_2 \| \pi$ och $(\mathbf{u} \cdot e_3)e_3 \perp \pi$.

U:s projektion på planet π är alltså $(\mathbf{u} \cdot e_1)e_1 + (\mathbf{u} \cdot e_2)e_2$.

2.5 Projektion på ett valfritt delrum

Om vi har ett rum R^{13} och vill projicera på ett femdimensionellt delrum A gör vi precis som tidigare.

e_1, e_2, e_3, e_4, e_5 är alltså ON-bas för vårt delrum och $e_6, e_7, ..., e_{13}$ är alla vinkelräta mot vårt delrum.

U:s projektion på delrummet A blir då

$$\mathbf{u} = (\mathbf{u} \cdot e_1)e_1 + (\mathbf{u} \cdot e_2)e_2 + (\mathbf{u} \cdot e_3)e_3 + (\mathbf{u} \cdot e_4)e_4 + (\mathbf{u} \cdot e_5)e_5$$

2.6 Flera delrum

Anta att vi har R^n uppdelat i två delrum, M och L. Dessa är vinkelräta mot varandra och kan skrivas som:

$R^n = L \bigoplus M$

där $e_1, e_2, ..., e_n$ är bas för R^n och L rummet har $e_1...e_k$ som ON-bas. M-rummet har $e_{k+1}, e_{k+2}, ..., e_n$ som ON-bas.

$\mathbf{U} = (e_1, e_2...e_k)$ ($e_1, ..., e_k$ är kolonnvektorer) och vektor $\mathbf{X} \in R^n$.

Vi kan nu skriva om detta som

$$\mathbf{U}^T = \begin{pmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_k^T \end{pmatrix}$$

$$\Longleftrightarrow$$

$$\mathbf{U}^T\mathbf{X} = \begin{pmatrix} e_1^T\mathbf{X} \\ e_2^T\mathbf{X} \\ \vdots \\ e_k^T\mathbf{X} \end{pmatrix}$$

Detta är koordinaterna för $\mathbf{X}$ i L med avseende på basen $e_1, e_2...e_k$.

3 Egenvärdesdekomposition

3.1 Sats (Spektralsatsen)

Varje symmetrisk matris $\mathbf{A}$ kan skrivas som $\mathbf{USU}^T$ där $\mathbf{U}$ är en ortogonalmatris och $\mathbf{S}$ är en diagonalmatris.

3.2 Diskussion

Matrisen $\mathbf{S}$ ser ut som följer

$$S = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

U är en ortogonalmatris och kan skrivas som

$$\mathbf{U} = (u_1, u_2, ..., u_n)$$

där $u_1, u_2, ..., u_n$ är en ON-bas. **A** kan uttryckas som

$$\mathbf{A} = \mathbf{USU}^T$$

Genom matrismultiplikation fås

$$\mathbf{AU} = \mathbf{US}$$

Med ovanstående formler för **U** och **S** kan sambandet skrivas om som

$$\mathbf{A}(u_1, u_2, ..., u_n) = (u_1, u_2, ..., u_n)\mathbf{S}$$

$$\Longleftrightarrow$$

$$\mathbf{A}u_i = \lambda_i u_i \tag{1}$$

λ_i är med andra ord egenvärden till **A** och u_i egenvektorer.

4 Singulärvärdesdekomposition

4.1 Sats

Varje $m \times n$ matris **A** kan skrivas som $\mathbf{USV}^T$. **U** är en $m \times k$ matris med ortonomerade kolonnvektorer och $k = min(m, n)$. **S** är en $k \times k$ diagonalmatris med icke negativa diagonalelement och **V** är en $n \times k$ matris med ortonomerade kolonnvektorer.

4.2 Bevis

Eftersom **A** inte har förutsats vara en symmetrisk matris kan vi inte utföra spektraldekomposition direkt. Däremot kan vi istället titta på $\mathbf{AA}^T$ som är symmetrisk. Från sats 3.1 följer

$$\mathbf{AA}^T = \mathbf{URU}^T \tag{2}$$

där **R** är en diagonalmatris och U en ortogonalmatris. Matrisen **R** innehåller inte några negativa värden. Detta kan inses genom följande resonemang.

Ekvation (1) ger

$$\mathbf{AA}^T u_i = \lambda_i u_i$$

Genom skalärmultiplikation med vektorn u_i får man följande samband

$$\lambda_i = \lambda_i |u_i|^2 = u_i \cdot \mathbf{AA}^T u_i = \mathbf{A}^T u_i \cdot \mathbf{A}^T u_i = |\mathbf{A}^T u_i|^2 \geq 0$$

Detta betyder att alla egenvärden måste vara positiva. Låt $\mu_i = \sqrt{\lambda_i}$. Låt **S** vara en diagonalmatris med $\mu_1 ... \mu_n$ som diagonalelement. Då är $\mathbf{R} = \mathbf{S}^2$.

Insättning i (2) ger

$$\mathbf{AA}^T = \mathbf{US}^2\mathbf{U}^T$$

Om alla egenvärden är skilda från noll kan man sätta $\mathbf{V} = \mathbf{A}^T\mathbf{U}\mathbf{S}^{-1}$. Då är

$$\mathbf{V}^T\mathbf{V} = \mathbf{S}^{-1}\mathbf{U}^T\mathbf{A}\mathbf{A}^T\mathbf{U}\mathbf{S}^{-1}$$

Eftersom $\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$ är

$$\mathbf{V}^T\mathbf{V} = \mathbf{S}^{-1}\mathbf{U}^T\mathbf{U}\mathbf{S}^2\mathbf{U}^T\mathbf{U}\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{I}\mathbf{S}^2\mathbf{I}\mathbf{S}^{-1} = \mathbf{I}$$

och

$$\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{U}\mathbf{S}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{A} = \mathbf{A}$$

Om ett eller flera egenvärden är noll (vilket gör att $\mathbf{S}^{-1}$ inte kan bildas) kan man anta att $\mu_1, ..., \mu_l > 0$ och att $\mu_{l+1}, ..., \mu_k = 0$, för något l, $0 \leq l \leq m$. Nya beteckningar måste då införas.

Låt $\tilde{\mathbf{S}}$ vara diagonalmatrisen $\mathbf{S}$ där alla egenvärden som är noll har plockats bort. $\tilde{\mathbf{S}}$ blir en $l \times l$ matris sådan att

$$\tilde{\mathbf{S}} = \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_l \end{pmatrix}$$

och

$$\tilde{\mathbf{U}} = (u_1, u_2, ..., u_l)$$
$$\mathbf{U}_0 = (u_{l+1}, u_{l+2}, ..., u_m)$$

Då är

$$\mathbf{S} = \begin{pmatrix} \tilde{\mathbf{S}} & 0 \\ 0 & 0 \end{pmatrix}$$

och

$$\mathbf{U} = \begin{pmatrix} \tilde{\mathbf{U}} & \mathbf{U_0} \end{pmatrix} \tag{3}$$

Följande egenskaper gäller

$$\mathbf{A}\mathbf{A}^\mathbf{T} = \mathbf{U}\mathbf{S}^\mathbf{2}\mathbf{U}^\mathbf{T} = \begin{pmatrix} \tilde{\mathbf{U}} & \mathbf{U_0} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{S}^2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{U}}^T \\ \mathbf{U_0} \end{pmatrix} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}^\mathbf{2}\tilde{\mathbf{U}}^\mathbf{T} \tag{4}$$

$$\mathbf{I} = \mathbf{U}\mathbf{U}^\mathbf{T} = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\mathbf{T} + \mathbf{U_0}\mathbf{U_0^T} \tag{5}$$

$$\mathbf{I} = \mathbf{U}^\mathbf{T}\mathbf{U} = \begin{pmatrix} \tilde{\mathbf{U}}^T\tilde{\mathbf{U}} & 0 \\ 0 & \mathbf{U_0^T}\mathbf{U_0} \end{pmatrix}$$

Detta medför att

$$\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \mathbf{U_0^T}\mathbf{U_0} = \mathbf{I} \tag{6}$$

Vidare har vi att

$$(\mathbf{U_0^T}\mathbf{A})(\mathbf{U_0^T}\mathbf{A})^\mathbf{T} = \mathbf{U_0^T}\mathbf{A}\mathbf{A}^\mathbf{T}\mathbf{U_0} = \mathbf{U_0^T}\tilde{\mathbf{U}}\mathbf{S}^\mathbf{2}\tilde{\mathbf{U}}\mathbf{U_0} = \mathbf{0}$$

Enligt Lemma 5.2 följer att

$$\mathbf{U_0^T A = 0} \tag{7}$$

$\mathbf{V}$ kan definieras som

$$\mathbf{V = A^T \tilde{U} \tilde{S}^{-1}}$$

Då följer det från ekvationerna (4) och (6) att

$$\mathbf{V^T V = \tilde{S}^{-1} \tilde{U}^T A A^T \tilde{U} \tilde{S}^{-1} = \tilde{S}^{-1} \tilde{U}^T \tilde{U} \tilde{S}^2 \tilde{U}^T \tilde{U} \tilde{S}^{-1} = I} \tag{8}$$

Och från ekvationerna (5) och (7)

$$\mathbf{\tilde{U} \tilde{S} V^T = \tilde{U} \tilde{S} \tilde{S}^{-1} \tilde{U}^T A = \tilde{U} \tilde{U}^T A = (\tilde{U} \tilde{U}^T + U_0 U_0^T) A = A}$$

Detta bevisar sats 4.1 för $m \leq n$. På samma sätt kan också fallet $n < m$ visas genom att titta på $\mathbf{A}^T\mathbf{A}$ istället för $\mathbf{A}\mathbf{A}^T$.

□

4.3 Diskussion

Matrisen $\mathbf{S}$ ser ut som följer

$$S = \begin{pmatrix} \mu_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \mu_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & \mu_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \mu_n \end{pmatrix}$$

$\mathbf{U}$ och $\mathbf{V}$ är matriser med ortonormerade kolonnvektorer. Dessa kan skrivas som

$$\mathbf{U} = (u_1, u_2, ..., u_k)$$

$$\mathbf{V} = (v_1, v_2, ..., v_k)$$

A kan uttryckas som

$$\mathbf{A = USV}^T$$

Genom matrismultiplikation fås enligt (8)

$$\mathbf{AV = US}$$

Med ovanstående formler för $\mathbf{U}$ och $\mathbf{S}$ kan sambandet skrivas om som

$$A(v_1, v_2, ..., v_n) = (u_1, u_2, ..., u_n)S$$

$$\Longleftrightarrow$$

$$\mathbf{A}v_i = \mu_i u_i$$

μ_i kallas för singulärvärden och v_i för singulärvektorer till $\mathbf{A}$.

5 Att välja projektion

Vi vill hitta en projektion från $R^n \longrightarrow R^3$ så att projektionen av punkterna $a_1, ..., a_n$ har så stor varians som möjligt, för att lättast möjligt se samband mellan datapunkterna.

5.1 Spår

För att förstå hur man gör det bästa dimensionsvalet behöver vi definiera begreppet spår (eng trace).

Låt $\mathbf{A}$ vara en kvadratisk matris. Då gäller att spåret av $\mathbf{A}$ är summan av diagonalelementen. Detta skrivs som

$$Tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

En viktig räkneregel är följande

$$Tr(\mathbf{XY}) = Tr(\mathbf{YX}) = \sum_{i,j} \mathbf{X}_{ij}\mathbf{Y}_{ji}$$

vilken kommer användas för att visa vilka dimsnsioner som bör väljas, då variansen mellan kolonnerna i matrisen A definieras som $Tr(\mathbf{AA}^T)$.

5.2 Lemma

Låt $\mathbf{A}$ vara en matris. Om $\mathbf{A}^T\mathbf{A} = \mathbf{0}$ är $\mathbf{A} = \mathbf{0}$

Bevis

$$Tr(\mathbf{AA}^T) = \sum_{i,j} \mathbf{A}_{ij}\mathbf{A}^T_{ji} = \mathbf{0}$$

5.3 Sats

Låt $\mathbf{A} = \mathbf{USV}^T$. Låt $\mathbf{B} = \mathbf{W}^T\mathbf{A}$ där $\mathbf{W}$ är en $m \times 3$ matris med ortonomerade kolonnvektorer samt att $\mathbf{S}$ är en diagonalmatris med diagonalelementen $\mu_1 ... \mu_k$ ordnade så att $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \geq 0$. Då är

$$Tr(\mathbf{B}^T\mathbf{B}) \leq \mu_1^2 + \mu_2^2 + \mu_3^2$$

5.4 Bevis

Enligt sats 4.1 kan $\mathbf{A}$ skrivas som

$$\mathbf{A} = \mathbf{USV}^T$$

Låt

$$\mathbf{B} = \mathbf{W}^T\mathbf{A} = \mathbf{W}^T\mathbf{USV}^T$$

Låt $\mathbf{Z}^T = \mathbf{W}^T\mathbf{U}$ vilket ger

$$\mathbf{B} = \mathbf{W}^T\mathbf{A} = \mathbf{Z}^T\mathbf{SV}^T$$

Eftersom $\mathbf{Z}$ har ortonormerade kolonnvektorer är $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$.

$$\mathbf{Z} = \begin{pmatrix} z_1 & z_2 & z_3 \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix} z_1' \\ \vdots \\ z_k' \end{pmatrix}$$

Vi vill maximera variansen som kan uttryckas som

$$Tr(\mathbf{B}^T\mathbf{B}) = Tr(\mathbf{VSZZ}^T\mathbf{SV}^T) = Tr(\mathbf{ZZ}^T\mathbf{SV}^T\mathbf{VS}) = Tr(\mathbf{ZZ}^T\mathbf{S}^2) = \sum_{i=1}^{k} |\mathbf{z}_i'|^2\mu_i^2 \tag{9}$$

där $|\mathbf{z}_i'|^2 \leq 1$ och

$$\sum_{i=1}^{k} |\mathbf{z}_i'|^2 = \sum_{i=1}^{3} |\mathbf{z}_i|^2 = 3.$$

Vi har koefficienterna $|\mathbf{z}_1|, |\mathbf{z}_2|, \ldots, |\mathbf{z}_i|$ som alla är ≤ 1. Vi kan välja $|\mathbf{z}_1| = |\mathbf{z}_2| = |\mathbf{z}_3| = 1$ vilket också maximerar funktionen $\mathbf{B}^T\mathbf{B}$ då $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \geq 0$.

Lemma 1.

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$$
$$0 \leq t_i \leq 1$$
$$t_1 + t_2 + \ldots + t_n = 3$$
$$\sum t_i\mu_i \leq \mu_1 + \mu_2 + \mu_3$$

Bevis av lemma 1.

$$\begin{aligned} & t_1\mu_1 + \ldots + t_n\mu_n \\ & \leq t_1\mu_1 + t_2\mu_2 + t_3\mu_3 + t_4\mu_4 + t_5\mu_4 + \ldots + t_n\mu_4 \\ & = t_1\mu_1 + t_2\mu_2 + t_3\mu_3 + (t_4 + \ldots + t_n)\mu_4 \\ & = t_1\mu_1 + t_2\mu_2 + t_3\mu_3 + (3 - t_1 - t_2 - t_3)\mu_4 \\ & = 3\mu_4 + t_1(\mu 1 - \mu 4) + t_2(\mu 2 - \mu 4) + t_3(\mu 3 - \mu 4) \\ & \leq 3\mu_4 + (\mu 1 - \mu 4) + (\mu 2 - \mu 4) + (\mu 3 - \mu 4) \\ & = \mu_1 + \mu_2 + \mu_3 \end{aligned}$$

Det följer från lemma 1 att

$$Tr(\mathbf{B}^T\mathbf{B}) \leq \mu_1 + \mu_2 + \mu_3$$

Vi bör alltså använda de kolonner som motsvarar $\mu_1 + \mu_2 + \mu_3$ Då är

$$\mathbf{W} = \begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix}$$

och

$$\mathbf{W^T A} = \begin{pmatrix} S_1V_1^T \\ S_2V_1^T \\ S_3V_1^T \end{pmatrix}$$

dvs $\mathbf{W^T A}$ är de tre första raderna i $\mathbf{SV^T}$.

□

6 Gener och patienter, en praktisk tillämpning av PCA

En av många användningsområden är att analysera gendata för att kunna se samband.

Om vi har en matris **A** med k (gener) × n (patienter), motsvarar varje kolonn en patient och varje rad en gen.

A kan enligt sats 4.1 skrivas som $\mathbf{A} = \mathbf{USV^T}$.

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \mu_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \mu_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & \mu_3 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \mu_n \end{pmatrix} \mathbf{V^T}$$

där μ_1 till μ_n är singulära värden som vi ordnar så att $\mu_1 \geq \mu_2 \geq \mu_3 ... \geq \mu_k > 0$

Man får den bästa projektionen genom att ta de tre första raderna i $\mathbf{SV}^T$.

A Programkod

Detta är en liten del ur programmet som har skrivits till projektet. Programmet är skrivet med biblioteken QT, OpenGL, GLUT, MKL (Intel Math Kernel Library) och ett matrisbibliotek skrivet av Johan Råde.

Programkoden nedan läser in data från en fil och bearbetar den till punkter i rummet med hjälp av PCA.

Hela programet kan laddas ner från http://www.net-bear.com/pca.tar.gz

```
// Read Size
ifstream f(path.c_str());
f >> height_ >> width_;

// Read data
matrix<double> data(height_, width_);
for(int i = 0; i < height_; ++i)
  for(int j = 0; j < width_; ++j)
    f >> data(i,j);

// Calculate mean & standard deviation
matrix<double> mu, sigma;
tie(mu, sigma) = statistics2(data);
matrix<double> stdDev = sqrt(sigma);

// Apply mean and stdDev
MATRIX_LOOP(data, i, j) {
  data(i,j) -= mu(i,0);
  if(stdDev(i,0) != 0)
  data(i,j) /= stdDev(i,0);
}
```

```
// Singular value decomposition
matrix<double> U,S,V;
tie(U,S,V) = svd(data);

// Calculate coordinates
posdata_ = ((diag(S)*V.t()).row(0,3)).t();
```

Minsta kvadratmetoden och dess släktingar

Stefan Adalbjörnsson
Simon Burgess
Matias Quiroz

9 maj 2005

Innehåll

1 Inledning

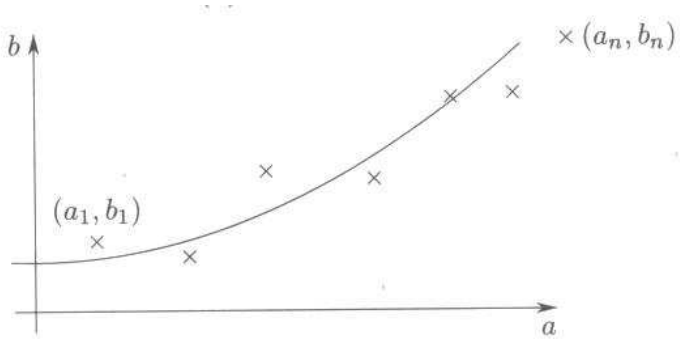
Minsta kvadratmetoden är en algoritm skapad för att lösa minimeringsproblem. Det finns en del vidareutvecklingar som bygger på samma idé: att minimera en viss enhet. Användningsområden är bl a statistik, datorteknik, numerik, biologi och kemi. På grund av sin kraftfullhet har minsta kvadratmetoden blivit ett viktigt matematiskt verktyg i näringslivet.

Metoden publicerades och namngavs 1805 av *Adrien-Marie Legendre* (1752-1833), som använde den för beräkning av kometbanor. *Carl Friedrich Gauss* (1777-1855) anmärkte året därpå att han själv använt metoden i många år, och 1809 publicerade han en härledning utifrån mätvärden med normalfördelade fel. År 1808 förespråkade även amerikanen *Robert Adrain* (1775-1843) metoden, oberoende av Legendre och Gauss. Sedan dess har metoden växt i sitt användningsområde, och inspirerat till nya områden inom matematiken. Bland dessa kan nämnas variationskalkyl och finita elementmetoden.

2 Kurvanpassning till överbestämda ekvationssystem

2.1 Inledning

Vi börjar med att se hur minsta kvadratmetoden kan användas för att anpassa ett polynom till ett antal givna mätpunkter (se Figur 1). Även om vi själva ofta kan rita en kurva i ett diagram som vi tycker ger en god beskrivning av sambandet för ett experiment, vill man ha en exakt algoritm för att beräkna kurvan. Kurvanpassningar av högre ordning blir mycket subjektiva för hand.



Figur 1: Kurvanpassning till ett givet antal mätpunkter (a_n,b_n)

Minsta kvadratmetoden tillhandahåller en algoritm för att plocka fram en kurva till ett antal mätpunkter, som inte nödvändigtvis kan placeras exakt på en kurvan (se figur 1): Om sambandet för experimentet misstänks vara ett polynom av andra graden, går det oftast inte att göra en kurva så att alla mätpunkter ligger på kurvan. Mer generellt går det oftast inte att exakt anpassa ett polynom av graden n till mer än $n + 1$ mätpunkter (Med två mätpunkter kan vi anpassa en linje exakt, men med tre mätpunkter är det i allmänhet inte möjligt).

Exempel 1: På ett havregrynspaket av ett känt märke finns följande specifikationer om ingredienser för olika stora portioner:

portioner	*vatten*	*havregryn*
1	1,5 dl	0,5 dl
2	2 dl	1 dl
3	3 dl	1,5 dl

Tabell 1: Portionsbeskrivning för havregrynsgröt.

Om vi avsätter mängden vatten i dl på y-axeln, och mängden havregryn i dl på x-axeln i ett diagram får vi tre punkter, $(0,5;1,5)$, $(1;2)$ och $(1,5;3)$. Dessa tre punkter kan omöjligt placeras på en och samma linje. Vad händer om en större mängd havregryn tillagas? I vilka proportioner ska då havregryn och vatten tillsättas? Ett rimligt antagande är att vatten och havregryn i verkligheten följer ett linjärt samband. En linjeanpassning bör göras, och med minsta kvadratmetoden ska vi göra detta längre fram.

2.2 Problembeskrivning

Vi vill anpassa ett godtyckligt polynom av ordningen m

$$b = b(a) = x_1 + x_2 a^2 + \ldots + x_{m+1} a^m \tag{1}$$

till en uppsättning mätdata $(a_1, b_1), \ldots, (a_n, b_n)$. Om de n antal mätpunkterna ska anpassas till ett polynom av ordningen m bildar vi $AX = Y$ på matrisform:

$$Y = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad A = \begin{pmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^m \\ 1 & a_2 & a_2^2 & \ldots & a_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^m \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_{m+1} \end{pmatrix}.$$

Ekvationen $y = Ax$ innebär då n stycken ekvationer $b_i = b_i(a) = x_1 + x_2 a_i^2 + \ldots + x_{m+1} a_i^m$. Beroende på hur många mätpunkter vi har i jämförelse med hur hög grad vi vill ha på det anpassade polynomet, kan flera situationer uppträda:

Jämförelse	*situation*
n<m+1	Det finns oändligt antal sätt att anpassa polynomet så att mätpunkterna ligger exakt på kurvan
n=m+1	AX=Y går att lösa, och har oftast exakt en lösning.
n>m+1	AX=Y är överbestämt och har ingen lösning.

Tabell 2: Tabellen visar vilka situationer som kan inträffa då olika antal mätdata n finns att tillgå när ett polynom av ordningen m skall genomlöpa mätpunkterna.

Det är i det sista fallet minsta kvadratmetoden kommer till nytta. Med minsta kvadratmetoden kan vi anpassa en kurva som är så bra som möjligt.

Detta är oftast fallet vid riktiga experiment, exempelvis ett linjärt samband med 10 mätdata.

Exempel 2: Vi fortsätter på vårt tidigare exempel om havregrynstillagning. Då vi vill göra en linjeanpassning till våra mätdata, $b(a) = x_1 + x_2 a$ där b är vattenmängd och a är mängd havregryn, bildar vi följande matriser:

$$Y = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \; A = \begin{pmatrix} 1 & a_1 \\ 1 & a_2 \\ 1 & a_3 \end{pmatrix}, \; X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

som med insatta värden enligt Exempel 1 blir

$$Y = \begin{pmatrix} 1,5 \\ 2 \\ 3 \end{pmatrix}, \; A = \begin{pmatrix} 1 & 0,5 \\ 1 & 1 \\ 1 & 1,5 \end{pmatrix}, \; X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Om det bara funnits en mätpunkt hade situationen varit som i fall ett (Se Tabell 2). Då det bara finns en punkt som en linje ska gå genom finns det oändligt många olika linjer som uppfyller detta. Om det funnits två mätpunkter finns bara en given linje som uppfyller kraven. Med tre eller mer mätdata som i vårt fall behövs minsta kvadratmetoden.

2.3 Lösning med Minsta kvadratmetoden

När vi befinner oss i det sistnämnda fallet (se Tabell 2) används minsta kvadratmetoden för att få en bra kurvanpassning. Vi ska nu bilda normalekvationen, $A^T AX = A^T Y$, (bevis följer i avsnitt 3.3) för att få ut de optimala parametrarna. $A^T A$ blir här

$$A^T A = \begin{pmatrix} n & \sum a_i & \sum a_i^2 & \dots & \sum a_i^m \\ \sum a_i & \sum a_i^2 & \sum a_i^m & \dots & \sum a_i^{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum a_i^m & \sum a_i^{m+1} & \sum a_i^{m+2} & \dots & \sum a_i^{2m} \end{pmatrix}$$

och $A^T Y$ blir

$$A^T Y = \begin{pmatrix} \sum b_i \\ \sum a_i b_i \\ \vdots \\ \sum a_i^m b_i \end{pmatrix}.$$

Genom att lösa ekvationssystemet

$$\begin{pmatrix} n & \sum a_i & \sum a_i^2 & \dots & \sum a_i^m \\ \sum a_i & \sum a_i^2 & \sum a_i^m & \dots & \sum a_i^{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum a_i^m & \sum a_i^{m+1} & \sum a_i^{m+2} & \dots & \sum a_i^{2m} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{m+1} \end{pmatrix} = \begin{pmatrix} \sum b_i \\ \sum a_i b_i \\ \vdots \\ \sum a_i^m b_i \end{pmatrix}$$

bestämmer vi parametrarna i (1). $A^T A$ går att invertera om $a_1, a_2, \dots, a_n$ inte alla är identiska. En kurva med en ekvation av graden m har nu anpassats till n stycken mätpunkter, $n > m+1$, så att det sammanlagda avståndet från punkterna till linjen blir minimalt. (se Figur 1).

Exempel 3: Detta exempel fortsätter där Exempel 2 slutade. Ett linjärt samband mellan vattenmängd och havregrynsmängd ska nu tas fram. Lösningen på problemet fås med hjälp av minsta kvadratmetodens normalekvation (se ekv. (3)), som ger:

$$X = (A^T A)^{-1} A^T Y \tag{2}$$

där X innehåller parametrarna x_1 och x_2, som bestämmer linjeanpassningen $b(a) = x_1 + x_2 a$, där b är vattenmängd och a är havremängd. $A^T A$ blir en 2x2-matris

$$A^T A = \begin{pmatrix} 3 & 3 \\ 3 & 3,5 \end{pmatrix}$$

och dess invers blir

$$(A^T A)^{-1} = \begin{pmatrix} 7/3 & -2 \\ -2 & 2 \end{pmatrix}.$$

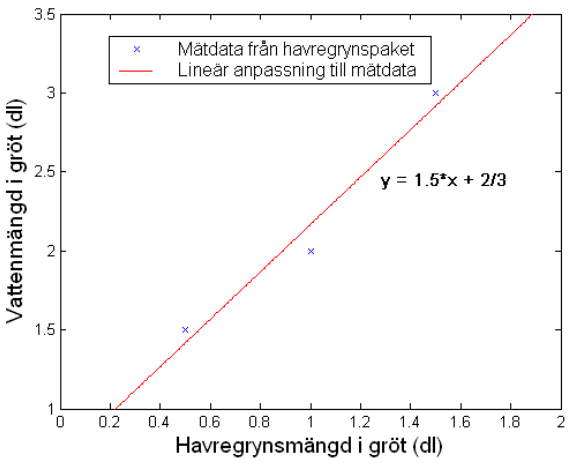
$A^T Y$ blir

$$A^T Y = \begin{pmatrix} 6,5 \\ 7,25 \end{pmatrix}$$

som slutligen ger X enligt ekvation (2)

$$X = \begin{pmatrix} 2/3 \\ 3/2 \end{pmatrix}.$$

Detta ger att mängden vatten (b) beror på mängden havregryn (a) med följande linjära aprroximation: $b(a) = \frac{3}{2}a + \frac{2}{3}$.



Figur 2: Linjär anpassning till tre mätpunkter

Vi har nu tagit fram en linjeanpassning (Se Figur 2) till ett problem som innehåller ett överbestämt system med hjälp av minsta kvadratmetoden. Som bilden visar stämmer approximationen bra. Anmärkning: Modellen stämmer dock inte bra för värden nära a=0. Modellen antyder här att man kan tillaga god havregrynsgröt med endast vatten som ingrediens. Dock bör den stämma bra för större värden på a.
Avslutningsvis kan det nämnas att man inte vill välja för hög grad på polynomet. Man kan tex frestas att välja så hög grad på polynomet att alla punkter genomlöps, men anledningen att man mätt många punkter är för att få bort det mätfel som metoden medför. Genom att genomlöpa alla punkter arbetar man under förutsättningen att alla mätvärden är perfekta. Detta skulle ge en helt fel bild av sambandet.

3 Härledning och bevis för minsta kvadratmetoden

3.1 Definitioner

För att kunna förstå härledningen och beviset behöver vi först ta nytta av följande definitioner.

Definition 1: Ett *linjärt rum* över R är en mängd H där 2 räkneoperationer är definerade: *addition:* $u \in H, v \in H \longrightarrow u + v \in H$
multiplikation med en skalär: $\lambda \in R, u \in H \longrightarrow \lambda u \in H$, sådana att de vanliga räknereglerna gäller.

Definition 2: Med ett *Euklidisktrum* (pre-Hilbertrum) menas ett linjärt rum försett med en skalärprodukt (se definition 4).
Definition 3: Med ett *Hilbertrum* menas ett Euklidisktrum som är fullständigt. Med fullständighet menas, vagt uttryckt, att följder som vill konvergera gör det.
Definition 4: Låt H vara ett linjärt rum över $\mathbf{K}$. En skalärprodukt på H är en funktion $(x, y) \mapsto (x|y)$ från $H \times H$ till K med egenskaperna

- $(x|x) \geq 0$ för alla x, $(x|x) = 0 \Longleftrightarrow x = 0$,
- $(x|c_1y_1 + c_2y_2) = c_1(x|y_1) + c_2(x|y_2)$,
- $(x|y) = \overline{(y|x)}$.

Då skalärprodukten mellan två vektorer är noll sägs de vara *ortogonala*.

Definition 5: Med *rangen* av en matris A menas antal linjärt oberoende kolonner i matrisen A. Med *full rang* menas en $m \times n$ matris med rang n.

3.2 Problembeskrivning

Antag att man ska lösa ekvationssystemet $Ax = b$ då A är en $m \times n$ matris. Om $m > n$ är systemet överbestämt och saknar då i allmänhet lösning. Minsta kvadrat metoden går ut på att finna en lösning som på något sätt är optimal. Men vad är det som ska optimeras? Ekvationen som vi vill lösa är en vektor ekvation och det verkar logiskt att vilja minimera längden på vektorn $Ax - b$ eftersom längden är 0 precis då x är en lösning. Detta leder till att vi vill minimera:

$$|Ax - b| = \sqrt{([Ax]_1 - b_1)^2 + ([Ax]_2 - b_2)^2 + \cdots + ([Ax]_n - b_n)^2}$$

Eftersom rotfunktionen är monotont växande räcker det att minimera $|Ax - b|^2$. Nästa steg är då att hitta detta minimum.

3.3 Bevis för Normalekvationen

Låt $f(x) = |Ax - b|^2$ och låt x_0 vara en minstakvadratlösning. Då är $f(x_0)$ det minsta värdet f kan anta vilket vi utnyttjar för att härleda x_0. Vi ser att $f(x_0+ty)-f(x_0) \geq 0$, för alla vektorer y och alla reella tal t, eftersom $f(x_0)$ är minsta värdet. Om vi använder skalärprodukten i motsvarande Hilbertrum fås att

$$\begin{aligned} f(x_0 + ty) - f(x_0) &= |A(x_0 + ty) - b|^2 - |Ax_0 - b|^2 \\ &= (Ax_0 - b + tAy|Ax_0 - b + tAy) + (Ax_0 - b|Ax_0 - b) \\ &= (Ax_0 - b|tAy) + (tAy|Ax_0 - b) + (tAy|tAy) \\ &= 2tRe(Ax_0 - b|Ay) + t^2|Ay|^2 \geq 0. \end{aligned}$$

För små t är t^2 mycket mindre än t och då vi kan välja valfritt tecken på t. Men eftersom olikheten måste gälla för alla t så blir $(Ax_0 - b|Ay) = 0$. Detta innebär att $Ax_0 - b$ är vinkelrät mot Ay för alla y. Alltså kan vi mha reglerna för skalärprodukt skriva $(A^*(Ax_0 - b)|y) = 0$ (A^* *innebär konjugat och transponat av* A) och eftersom detta ska gälla för alla vektorer y måste vi ha $A^*Ax_0 = A^*b$. Sammanfattningsvis har vi en minsta kvadratlösning då x_0 löser *normalekvationen*

$$A^*Ax = A^*b,$$

vilket skulle bevisas. Man kan visa att ekvationssystemet alltid är lösbart eftersom högerledet ligger i $V(A^*)$ och det gäller att $V(A^*A) = V(A^*)$.
Om A^*A är inverterbar, det vill säga om A har maximal rang, så ges minstakvadratlösningen av $x_0 = (A^*A)^{-1}A^*b$. För reella matriser är $A^* = A^T$ och vi får

$$x_0 = (A^TA)^{-1}A^Tb. \tag{3}$$

4 Minsta kvadratmetoden i allmänna Hilbertrum

4.1 Introduktion

En vanlig problemställning, som genomsyrar hela detta arbete, är att man vill approximera en given vektor med någon klass av vektorer. Hur bra en approximation är avgörs av hur nära den givna och den approximerade vektorn ligger. Avståndet kallas för norm och kan mätas på olika sätt. Exempelvis kan en norm tolkas som ett avstånd mellan två vektorer (funktioner). Man vill givetvis inte inskränka sig i att endast behandla vektorer i den mening som är vanligast. Med hjälp av teori för Hilbertrum kan minsta kvadratmetoden generaliseras till givna funktioner i rummet. Nedan förklaras Hilbertteorin mycket kortfattat och vi bygger vidare på samma idéer vi introducerade i avsnitt 3.1.

I Hilbertrumsteorin så behandlas funktioner som vektorer. Detta är ett mycket abstrakt begrepp, dock kan man i vissa fall illustrera med bilder av vektorer, men oftast får man nöja sig med formella räkningar. Med hjälp av en *skalärprodukt* i Hilbertrummet så kan man studera väsentliga egenskaper, såsom ortogonalitet. Som ett exempel på ett hilbertrum kan vi nämna $L_2(w, I)$ där skalärprodukten ges av

$$(u|v) = \int_I \overline{u(x)}v(x)w(x)\,dx.$$

u och v är vektorer (funktioner), och $w(x)$ kallas viktfunktion och är alltid given. I är intervallet man arbetar över. Vidare kan nämnas att L_2 normen har många fysikaliska tillämpningar och är enkel att hantera räknemässigt. För minimeringsproblem behövs även projektionssatsen som ges nedan utan

bevis, det bör nämnas att beviset går till på samma sätt som för matriser.

4.2 Projektionssatsen

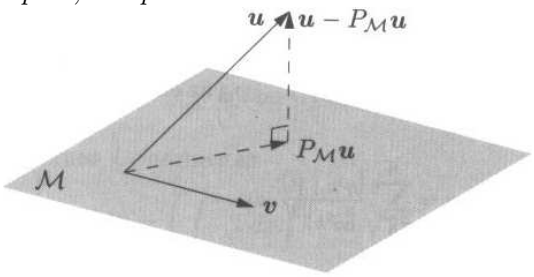
Antag att $\varphi_1, \varphi_2, \ldots, \varphi_n$ är en ortogonal bas (i tillhörande skalärprodukt) i H. Varje $u \in H$ kan då skrivas $u = \sum_{k=1}^{n} \frac{(\varphi_k|u)}{||\varphi_k||^2}\varphi_k$ och vi betecknar högerledet med $P_M u$ samt kallar den för projektionen av u på det linjära höljet $M = [\varphi_1, \ldots, \varphi_n]$. Se figur 3 för en geometrisk åskådning. För varje $u, v \in H$ gäller

- $P_M u \in M$
- $u - P_M u \perp M$
- $\min ||u - v||^2 = ||u - P_M u||^2 = ||u||^2 - \sum_{k=1}^{n} \frac{(\varphi|u)}{||\varphi_k||^2}\varphi_k$

När vi nu har dessa begrepp så faller det naturligt att studera avstånd mellan funktioner. Med hjälp av projektionssatsen kan vi nu projicera funktioner på varandra och därmed avgöra vilka polynom som approximerar en given funktion bäst i L_2 normen. Det handlar alltså om att minimera integralen

$$\int_I |u(x) - av(x)|^2 \, dx$$

där a är en konstant. Detta löses genom att med projektionsformeln projicera u på funktionen v. För den nyfikne läsaren rekommenderas boken *Kontinuerliga system: G Sparr, A Sparr.*



Figur 3: geometrisk förklaring av projektionssatsen

4.3 Konkreta exempel

I fysikaliska sammanhang så betecknar integralen

$$\int_I |u(x)|^2 dx$$

kostnaden (kan t.ex vara energi). Även här vill man minimera integralen. Det som berörts i detta avsnitt är alltså en generalisering som dock bygger på samma idé: att vi vill minimera ett avstånd, oavsett vilken norm vi mäter i. Vi kan illustrera detta med ett exempel:

Exempel 2: Bestäm de tal a och b som minimerar integralen

$$\int_0^\pi (x - a\sin x - b\sin 2x)^2\, dx$$

Vi ska alltså minimera $||u - v||^2$ i $L_2[0,\pi]$, där $u = x$ och $v \in M = [\sin x, \sin 2x]$. Vi ser genom enkla beräkningar att $\sin x$ och $\sin 2x$ är ortogonala i L_2:

$$\int_0^\pi \sin x \sin 2x\, dx = \frac{-1}{4}\int_0^\pi (e^{ix} - e^{-ix})(e^{i2x} - e^{-2ix})\, dx = 0.$$

Väljer vi $v = P_M u$ så har vi minimerat avståndet. Talen a och b beräknas alltså med hjälp av projektionsformeln, dvs

$$a = \frac{(\sin x|x)}{(\sin x|\sin x)}, \quad b = \frac{(\sin 2x|x)}{(\sin 2x|\sin 2x)}$$

där alla skalärprodukter tas i $L_2[0,\pi]$. Beräkning av integralerna ger sedan att $a = 2$ samt $b = -1$.

5 Variationskalkyl

5.1 Historik

Variationskalkylen är en gren av matematiken som behandlar problemet att minimera eller maximera storheter som beror av en obekant funktion. Under 1700-talet utvecklade Euler och Lagrange systematiska metoder för variationsproblem. Genom att betrakta små variationer av den obekanta funktionen fann man för vissa allmänna optimeringsproblem en differentialekvation, Eulerekvationen (även kallad *Euler-Lagranges* differentialekvation), som måste satisfieras av en funktion som ger maximum eller minimum. Vi kommer att se detta alldeles strax.

5.2 Euler-Lagrange ekvationen

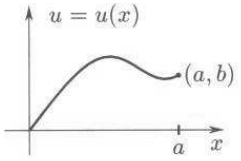
Fysikaliska lagar är ofta formulerade som variationsprinciper. Ett exempel på detta är Hamiltons princip: Låt K stå för kinetisk energi och U för potentiell energi. Då rör sig en partikel i ett konservativt fält så att

$$\int_\gamma (K - U)\, ds$$

är minimal.
Exempel 3: Låt A vara mängden $\{u \in C^1(a,b), u(a) = u_0, u(b) = u_1\}$. Hitta minimum till funktionalen

$$J(u) = \int_a^b \sqrt{1 + (u'(x))^2}\, dx, \text{ u} \in A.$$

Geometrisk kan vi tolka funktionalen som en beskrivning av kurvlängden i nedanstående figur.



Figur 4: kurvlängden för funktionalen i Exempel 3.

I detta triviala exemplet vet vi vilken som är den kortaste vägen. Dock leder sättet att beskriva det till ett intressant resultat. Säg att vi har en funktional

$$J(u) = \int_a^b f(x, u(x), u'(x))\, dx$$

som uppfyller vissa randvillkor. Det svåra i minimeringsproblem är att definitionsmängden utgörs av funktioner och inte punkter. Antag att minumum antas för $u = \tilde{u}(x)$. Tag $\varphi(x) \in C^1$ med randvillkoren $\varphi(a) = \varphi(b) = 0$. Då uppfyller $\tilde{u}(x) + t\varphi$ randvillkoren och det gäller att

$$J(\tilde{u} + t\varphi) \geq J(\tilde{u})$$

för varje t. För fixt φ på vänster sida i olikheten står en funktion av t, som har sitt minimum för $t = 0$. Alltså är funktionens derivata (förutsatt att den existerar) lika med noll i punkten $t = 0$. Detta ger

$$0 = \frac{d}{dt} J(\tilde{u} + t\varphi) = \int_a^b \frac{d}{dt} f(x, \tilde{u} + t\varphi, \tilde{u'} + t\varphi')\, dx = \int_a^b (\varphi \frac{\partial}{\partial u} f + \varphi' \frac{\partial}{\partial u'} f)\, dx$$

$$= \int_a^b \varphi \frac{\partial}{\partial u} f\, dx + \int_a^b \varphi' \frac{\partial}{\partial u'}\, dx = \int_a^b \varphi (\frac{\partial}{\partial u} f - \frac{d}{dx} \frac{\partial}{\partial u'} f)\, dx$$

där vi använt partialintegration för att komma fram till svaret. Och eftersom denna likhet gäller för alla φ så följer det att minimallösningen $u = \tilde{u}$ satisfierar ekvationen

$$\frac{\partial}{\partial u} f - \frac{d}{dx} \frac{\partial}{\partial u'} f = 0.$$

Denna ekvation kallas för *Euler-Lagranges* differentialekvation.
Vi ska nu se hur det föregående exemplet kan lösas med denna metod: Sätt $f(x, u(x), u'(x)) = \sqrt{1 + (u'(x))^2}$. Insättning i Euler-Lagrange ekvationen ger:

$$\frac{\partial}{\partial u}\sqrt{1 + (u'(x))^2} - \frac{d}{dx}\frac{\partial}{\partial u'}\sqrt{1 + (u'(x))^2} = 0 - \frac{d}{dx}\frac{1}{\sqrt{(1 + u'(x)^2)}}u'(x) = 0.$$

Detta betyder att

$$\frac{u'(x)}{\sqrt{1 + (u'(x))^2}} = C, \text{ där C är en konstant.}$$

Alltså följer det att $u'(x)$ måste vara en konstant, och då erhålles lösningen $u(x) = Cx + D$ *(en linje som väntat)*. Vi kan även bestämma konstanterna C och D genom att utnyttja givna randvillkor.

6 Källförteckning

Böcker
Källén, A. (1997), *Kompendium för matematik för naturvetare II*, Studentlitteratur, Lund
Petterson, P. (2004), Utlämnat material i kursen matristeori, Lund
Sparr, G. & Sparr, A. (2000), *Kontinuerliga System*, Studentlitteratur, Lund

Internet
`http://ceee.rice.edu/Books/LA/index.html`
`http://www.sm.luth.se/~johanb/applmath/chap3`
`http://www.ne.se/jps/search/article.jsp?i_art_id=256691&i_word=minsta%20kvadratmetoden`

Projektarbete i Matematisk Kommunikation
Varför är en såpbubbla rund?
– En introduktion i variationskalkyl

Ebba Brink
Martin Larsson
Daniel Persson
Martin Strömqvist
π-04

19 maj 2005

Sammanfattning

I detta arbete formuleras och härleds några grundläggande resultat inom variationskalkylen. Vi kommer att lösa några problem som visserligen är tämligen enkla, men som trots det har haft stor betydelse för variationskalkylens utveckling. Vi kommer också att belysa några praktiska tillämpningar och redogöra for ämnets historiska bakgrund.

Innehåll

1 Inledning: Vad är variationskalkyl?

Hur kommer det sig att en såpbubbla är sfärisk? Varför rör sig himlakropparna som de gör? Vad är det som bestämmer formen på ett membran uppspänt över en ram? Vilken väg mellan två punkter färdas en ljusstråle?

Svaren på sådana frågor är inte alltid uppenbara. Dessa exempel, liksom många andra liknande problem, har dock en gemensam nämnare: i samtliga fall ordnar sig naturen på ett sådant sätt att någon fysikalisk storhet minimeras. Såpbubblan antar den form som innebär lägst ytspänning. Himlakropparna rör sig på ett sådant sätt att den totala verkan, det vill säga skillnaden mellan kinetisk och potentiell energi, minimeras. Membranet formar sig så att den potentiella spänningsenergin minimeras, och ljusstrålen tar den snabbaste vägen mellan två punkter.

Detta är i grund och botten ett sätt att formulera olika naturlagar — nämligen som principer, såsom att "energin minimeras", "gångtiden minimeras", etc. Det visar sig att detta sätt att formulera naturlagarna ofta är väldigt generellt. Till exempel följer Newtons kraftlag direkt av principen om minsta verkan, dessutom förblir denna princip korrekt även i situationer där kraftlagen inte längre gäller. Denna typ av principer kallas *variationsprinciper*, och kan användas för att härleda lösningar på de matematiska problem som uppkommer ur de fysikaliska frågeställningarna.

För att komma vidare inför vi ett nytt begrepp.

Definition[1]**:** En *funktional* $J[y]$ är en regel som till varje funktion y i någon lämpligt vald klass av funktioner ordnar ett reellt tal. Om y ger utseendet på membranet i exemplet ovan skulle $J[y]$ kunna beteckna spänningsenergin.

I fallet med membranet innebär motsvarande variationsprincip att bland alla möjliga former y antas den som minimerar energin $J[y]$. För att ta reda på hur denna form ser ut måste vi alltså finna den *funktion* y som minimerar *funktionalen* $J[y]$. Hur detta går till i praktiken kommer vi att se längre fram.

En vanligt förekommande klass av funktionaler är ett integraluttryck där integranden beror av den sökta funktionen y och dess derivator. Vi betraktar här fallet då y är en reellvärd funktion av en reell variabel x. Funktionalen kan då skrivas

$$J[y] = \int_a^b f(x, y, y')dx, \tag{1}$$

där $f(x, y, y')$, som kallas *Lagrangianen*, är ett uttryck som innehåller x, y och y'. Här har vi antagit att endast förstaderivatan ingår, men definitionen kan generaliseras till att även omfatta derivator av högre ordning. I det fall y beror av flera variabler kan integralen i (2) komma att ersättas av en multipelintegral.

De exempel på variationsproblem som givits så här långt har alla haft en mycket konkret fysikalisk bakgrund. Under variationskalkylens utveckling har givetvis

[1] Denna definition är något oprecis, men den räcker för behoven i detta arbete.

en mängd andra problem, som ofta kan kännas något krystade eller konstruerade, haft betydelse. Det har dock visat sig, som så ofta inom matematiken, att de metoder som arbetats fram har kunnat tillämpas på högst påtagliga och verklighetsnära problem.

I detta arbete formuleras och härleds några grundläggande resultat inom variationskalkylen. Vi kommer att lösa några problem som visserligen är tämligen enkla, men som trots det har haft stor betydelse för variationskalkylens utveckling. Vi kommer också att belysa några praktiska tillämpningar och redogöra for ämnets historiska bakgrund.

2 Historik

Frågeställningarna i inledningen är på intet sätt nya, utan har fängslat människor i alla tider. Isoperimetriska problem, det vill säga problem som innebär att man vill maximera ett område med en given omkrets, studerades redan i antikens Grekland. På senare tid har flera av historiens mest framstående matematiker och fysiker ägnat sig åt att korrekt försöka beskriva och lösa dem, däribland Newton, Euler, Lagrange och Hilbert.

Men, låt oss börja med Fermat, som på sätt och vis var den som satte variationskalkylens hjul i rullning.

2.1 Fermats princip

Pierre de Fermat var en fransk advokat som utövade matematik på fritiden. Han är speciellt känd för att skriva ner sina matematiska resultat i marginaler på böcker eller i brev till sina vänner. År 1662 skickade han två artiklar till en kollega, i vilka han skrivit om hur ljusstrålar bryts när de passerar från ett optiskt medium till ett annat. Detta kom att bli början till något mycket mer omfattande än bara det optiska problemet med ljusets brytning.

Det var i dessa artiklar som Fermat först redogjorde sin princip om minsta tid, som är en grundläggande föreställning inom variationskalkylen. Principen kan enkelt uttryckas som att "naturen verkar på de sätt som är snabbast och lättast". I enlighet med detta förutsatte alltså Fermat att ljuset rör sig på så sätt att det genomkorsar mediet på kortast möjliga tid.

Fermats resultat kan ses som banbrytande då han var den första inom området som använde sig av analys istället för geometriska metoder.

2.2 Newtons kroppar

Efter Fermats artiklar gjordes inga framsteg inom området förrän 1685, då Isaac Newton formulerade och löste variationkalkylens första djupa problem.

Vid den här tiden arbetade Newton med att analysera olika kroppars rörelser genom en sorts trög vätska, eller som han själv uttryckte det; "ett sällsynt medium av mycket små partiklar i vila, vilka är av samma mått och fritt fördelade

på samma avstånd från varandra". Hans tanke var att resultatet skulle kunna användas inom konstruktionen av skepp.

Ett av problemen innefattade att hitta formen på den rotationskropp som rör sig längs sin egen rotationsaxel i detta medium med minsta möjliga motstånd. Tidigare hade han studerat t.ex. konformiga kroppar och då nyttjat den vanliga teorin för maxima och minima i sina beräkningar. För att lösa det nya problemet var Newton dock tvungen att introducera ett mer kraftfullt verktyg: det som senare skulle kallas variationskalkyl.

Det finns åtskilliga anledningar till att just det här problemet är av stor relevans för variationskalkylen, förutom det faktum att det var först. För det första skulle tekniken som Newton använde i sin lösning komma att användas både av Newton själv och av Jakob Bernoulli vid lösandet av det så kallade brachistokron-problemet. Ännu senare skulle samma idéer användas och systematiseras av Euler. För det andra är problemet intressant i sig, då det är av ovanlig sort och kan ge lösningskurvor med hörn - icke kontinuerliga punkter - och kan därför sakna lösning i klassisk form. För det tredje är problemet ett av de mest komplexa specialfall i hela teorin, och kroppar med minimalt motstånd kan utgöras av inte mindre än tre olika typer av kurvor.

När Newton publicerade sina lösningar förekom det inga direkta förklaringar till de nya metoderna. Resultaten förbryllade den matematiska societeten, och många förstod inte vidden av Newtons arbete förrän långt senare.

2.3 Bröderna Bernoulli och brachistokron-problemet

Johann och Jakob Bernoulli var schweiziska matematiker och bröder, som tillsammans med Newton och Leibniz kan ses som analysens huvudgrundare. Trots sina framgångar led de av syskonrivalitet och hade bittra gräl om kvaliteten på varandras arbete.

År 1696 utmanade Johann Bernoulli den matematiska världen att lösa vad han döpt till brachistokron-problemet (från *brachystos*, kortast och *chronos*, tid). Detta problem, som först formulerats av Galileo Galilei 1638, gick ut på att hitta kurvan mellan två punkter i ett vertikalt plan längs vilken en friktionsfri pärla under gravitationskraftens inverkan kunde glida på kortast möjliga tid.

Galilei hade publicerat en slutsats i sitt verk "Två Nya Vetenskaper" som utkom fyra år innan hans död. Här bevisade han korrekt att den snabbaste vägen för pärlan inte var densamma som den kortaste (det vill säga en rät linje). Han hade samtidigt påstått att den sökta vägen var en del av en cirkel - men detta var felaktigt.

Ironiskt nog var lösningen på problemet en kurva som Galilei själv hade studerat och namngett redan 1599 - nämligen cykloidbågen, som genereras om man följer en fix punkt på periferin av ett rullande hjul. Cykloidbågen har andra intressanta egenskaper, varav en är *tautokronism* ("samma tid"), vilket kan illustreras som att oberoende av från vilken höjd en pärla börjar glida längs med kurvan, så kommer den att nå slutpunkten på samma tid. Som en fotnot kan nämnas att cykloiden, på grund av alla oenigheter den skapade mellan matematiker på

1600-talet, skämtsamt kallats “Geometrins Helena” (efter den bortrövade skönheten från Troja).

Gottfried Leibniz, tysk matematiker, fysiker och filosof, var en av de första som löste problemet. Han hade fått det beskrivet i ett privat brev från Bernoulli den 9 juni 1696, och en vecka senare postade han sitt svar.

Johann Bernoullis egen lösning var ett skickligt nyttjande av Fermats princip. Han delade upp området mellan de två punkterna i smala horisontella skikt och antog att den fallande partikeln bröts vid varje gräns som om den vore en ljusstråle, så att den totala falltiden minimerades.

Vad gäller Newton sägs det att han läste Johann Bernoullis utmaning och skrev sedan klart sin lösning innan han gick till sängs samma kväll. Klart är det i alla fall att han den 30 januari 1697 publicerade en artikel med ett kort och kärnfullt svar på spörsmålet. Det är rimligt att anta att han genast insett likheterna mellan brachistokron-problemet och hans tidigare arbete på området, och dragit nytta av detta. En intressant anmärkning är att fastän Newton skickade in sin artikel anonymt, kunde Bernoulli se vem som skrivit den - han påstod att han “kände igen lejonet på dess avtryck”.

Andra samtida vetenskapsmän som löste problemet var Johanns bror Jakob och densammes elev Guillaume de l’Hôpital.

Brachistokron-problemet kom att fungera som en sporre för bröderna Bernoulli, och de kommande åren formulerade och löste de en rad andra, mer generella problem och började på så sätt bygga upp detta nya matematiska område.

2.4 Eulers generaliseringar

En annan av Johann Bernoullis elever som skulle komma att betyda mycket för variationskalkylens utveckling var Leonhard Euler.

Euler var en schweizisk matematiker, som även skulle komma att bidraga stort inom optik, mekanik, ellära och magnetism. Han sades ha ett förbluffande minne och kunde utföra mycket komplicerade beräkningar i huvudet, vilket kom honom väl till gagn då han från och med 1766 var helt blind. Han fortsatte publicera resultat under hela sin livstid, och kan räknas som en av historiens flitigaste matematiska skribenter med nära 900 författade artiklar.

Även om Euler rimligtvis leddes in på området av sin läromästare Johann Bernoulli tror man det var långt senare hans intresse för variationskalkyl tändes på riktigt. Under 1720- och 30-talen snuddade han då och då vid ämnet - som vid denna tidpunkt fortfarande inte hade något namn - i sin efterforskning, och skrev bland annat om geodesiska kurvor.

Den stora sensationen kom dock 1744 då Euler gav ut sitt mästerliga opus om “metoder att hitta kurvor i planet med maximi- eller minimiegenskaper”. Här tog han bland annat upp 100 exempelproblem som han inte bara löste utan använde för att visa på en mer allomfattande teori. Det kanske allra viktigaste han gjorde i denna bok var dels att sätta upp ett generellt redskap för att skriva ner de så kallade Euler-ekvationerna, och dels att bringa fram och diskutera lagen om

minsta verkan som han upptäckt i april 1743. Det var första gången denna lag publicerades. Av någon anledning tillskrivs den nu vanligen Maupertuis, fastän dennes arbete på samma område inte presenterades förrän omkring ett år senare.

En annan viktig sak Euler gjorde för variationskalkylen var att ändra diskussionen från att i grund och botten ha rört sig om specialfall till att behandla en väldigt generell klass av problem, även med moderna mått mätt.

2.5 Maupertuis och beviset för Guds existens

Pierre de Maupertuis var en fransk fysiker som också studerat under Johann Bernoulli. År 1736 ledde han ett franskt forskarlag på en expedition till Lappland där de gjorde undersökningar om jordens radie. Trots svåra insektsattacker på sommaren, en odrägligt kall vinter och ett smärre skeppsbrott i Baltiska havet på väg hem lyckades de 1737 presentera för den parisiska vetenskapsakademin sin slutsats att jordklotet har en oblat form (hoptryckt i axiell riktning). Expeditionen gjorde deltagarna berömda.

Maupertuis formulerade lagen om minsta verkan oberoende av Euler och framlade sitt resultat 23 april 1744. Som kuriosa kan nämnas att han hävdade att den var ett metafysiskt bevis för Guds existens.

2.6 Lagranges variationer

Den 12 augusti 1755 mottog Euler ett brev från 19-årige Guiseppe Lodovico Lagrangia - mer känd under sitt franska namn, Joseph-Louis Lagrange.

Brevet innehöll matematiska detaljer av en mycket tilltalande och revolutionär idé som Lagrange utarbetat i samband med att han studerat tautokronproblemet (där en partikel ska glida friktionsfritt längs en kurva och nå den nedre ändpunkten på samma tid oavsett från vilken höjd den startar - lösningskurvan är i analogi med brachistokron-problemet en cykloidbåge).

Det den unge matematikern hade upptäckt var ett sätt att använda Eulers metoder från 1744 utan krav på någon geometrisk inblick i problemet - istället kunde han göra motsvarande beräkningar på analytisk väg och erhålla ett nästan automatiskt svar. Finessen var att införa något som kom att kallas för *variation*, en liten störning som gör det möjligt att finna en lösning till maximi- eller minimiproblem.

Allt detta banade väg för en ny epok inom variationskalkylen. Efter att Euler hade sett Lagranges metod anslöt han sig till den och övergav han sin egen. Som en eloge till Lagrange gav han också ämnesområdet namnet "variationskalkyl".

Lagrange brevväxlade tämligen flitigt med Euler innan han publicerade sitt resultat 1760. I slutet av samma decennium kom hans andra stora verk, där han fortsatte att utveckla sina idéer till att gälla i mer komplexa situationer.

2.7 Legendres andra variation

Nästa betydelsefulla framsteg gjordes av fransmannen Adrien-Marie Legendre. Legendre kom från en rik familj och hade fått en toppkvalitativ universitetsutbildning i matematik och fysik. 1782 deltog han i en pristävling som utlystes i Berlin, där uppgiften bestod i att matematiskt bestämma banan för kanonkulor eller bomber, medräknat luftmotståndet. Hans uppsats vann priset och ledde in honom på en forskningskarriär.

Han blev intresserad av problemet att bestämma huruvida en given extrem är en minimerande eller maximerande båge. För att göra det införde han något han kallade en "andra variation". Hans resultat, som presenterades för den parisiska vetenskapsakademin 1786, var anmärkningsvärt men innehöll en del felaktigheter. Till exempel lyckades han inte ge något gällande bevis. Han mottog kritik från bland andra Lagrange.

2.8 Senare upptäckter

Efter Legendres upptäckt skulle det dröja 50 år innan nästa stora steg i variationskalkylens utveckling togs, denna gång av Carl Jacobi. Jacobi var son till en judisk bankman och visade tidigt ett stort intresse för studier, speciellt i matematik. Han kom sedermera att bli en mycket uppskattad professor på universitet i Berlin. För att få tillåtelse att undervisa där hade han dock varit tvungen att avsäga sig sin judiska tro och bli kristen.

Det var 1836 som hans berömda artikel kom ut. Den var kort, och innehöll nästan inga förklaringar eller bevis - detta kan ha berott på att han ville ge ut den snabbt innan någon annan publicerade liknande material. Men den gav upphov till en hel skola av kommentatorer som arbetade för att fylla i luckorna (däribland Clebsh, Hesse och V.A. Lebesque).

En annan som bidragit till området är Karl Weierstrass, som var verksam under andra hälften av 1800-talet. Han sysslade bland annat med den "andra variationen" och parametrisering av problem, men formulerade också villkor som är nödvändiga för variationskalkylen.

Ytterligare vetenskapsmän som skulle komma att föra ämnet framåt var Alfred Clebsch, Adolph Mayer, David Hilbert och Gilbert Bliss.

3 Grundläggande begrepp och teori

Som vi sett går problemen inom variationskalkylen ut på att minimera (eller maximera) en funktional,

$$J[y] = \int_a^b f(x, y, y')dx, \tag{2}$$

Eftersom y' ingår i integraluttrycket måste y förutsättas vara deriverbar i intervallet (a, b). Om man vill att $J[y]$ ska anta ett (ändligt) reellt värde för alla

tillåtna funktioner y måste man givetvis också kräva att integralen är ändlig; om den är generaliserad i till exempel någon av ändpunkterna måste alltså y vara sådan att integralen konvergerar. Det finns alltså en mängd kriterier som definitionsmängden, det vill säga mängden av alla tillåtna funktioner y, måste uppfylla. Det är inte alltid på förhand givet hur denna klass av funktioner ska väljas, något vi kommer att illustrera längre fram.

Ett bivillkor som ofta dyker upp i olika problemställningar är att funktionsgrafen till y ska börja och sluta i bestämda punkter. Man kräver alltså att $y(a) = A$ och $y(b) = B$, med A och B givna. Hur ska man nu göra för att finna den minimerande funktionen? Lösningsidén, som först introducerades av Lagrange vid mitten av 1700-talet, innebär väsentligen att man varierar funktionen y lite grand och undersöker hur det påverkar $J[y]$. Då man har funnit ett y sådant att $J[y]$ växer oavsett hur y ändras har man en lösning på problemet, det vill säga en funktion som minimerar $J[y]$.

För att genomföra detta i praktiken inför vi *variationen* av $J[y]$. Detta begrepp är analogt med begreppet differential från en- och flervariabelanalysen. Vi tar därför avstamp i detta senare begrepp.

3.1 Inkrement och variation

Varje oändligt många gånger deriverbar funktion $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ av två variabler kan Taylorutvecklas, till exempel kring punkten (x, y). Man får då:

$$g(x + \Delta x, y + \Delta y) = g(x, y) + \frac{\partial g}{\partial x}\Delta x + \frac{\partial g}{\partial y}\Delta y + R$$

$$\Longleftrightarrow$$

$$\Delta g = g(x + \Delta x, y + \Delta y) - g(x, y) = \frac{\partial g}{\partial x}\Delta x + \frac{\partial g}{\partial y}\Delta y + R,$$

där de partiella derivatorna ska tas i punkten (x, y). Δx och Δy anger avståndet från (x, y), och R är högre ordningens termer (resttermen). Differentialen dg av g definieras som den linjära delen av Δg, det vill säga den del som endast utgörs av första ordningens termer:

$$dg(\Delta x, \Delta y) = \frac{\partial g}{\partial x}\Delta x + \frac{\partial g}{\partial y}\Delta y$$

Observera att definitionen endast förutsätter att g är minst en gång deriverbar.

Vi överför nu resonemanget till funktionalen $J[y]$. Antag att vi varierar y genom att addera en deriverbar funktion $h(x)$, vars belopp $|h|$ är litet överallt och som uppfyller $h(a) = h(b) = 0$. Taylorutvecklar vi $J[y+h] = \int_a^b f(x, y+h, y'+h')dx$ kring "punkten" (y, y') får vi på liknande sätt som ovan

$$J[y+h] = \int_a^b f(x, y, y')dx + \frac{\partial}{\partial y}\int_a^b f(x, y, y')dx \cdot h + \frac{\partial}{\partial y'}\int_a^b f(x, y, y')dx \cdot h' + R$$

$$= J[y] + \int_a^b h\frac{\partial f}{\partial y}(x,y,y')dx + \int_a^b h'\frac{\partial f}{\partial y'}(x,y,y')dx + R.$$

Med $\Delta J = J[y+h] - J[y]$ får vi alltså till slut

$$\Delta J[h] = \int_a^b \left(h\frac{\partial f}{\partial y}(x,y,y') + h'\frac{\partial f}{\partial y'}(x,y,y') \right) dx + R. \tag{3}$$

Storheten ΔJ kallas *inkrementet* av J. Variationen δJ av funktionalen J definieras, i analogi med differentialen av en vanlig funktion, som den linjära delen av inkrementet, dvs.

$$\delta J[h] = \int_a^b \left(h\frac{\partial f}{\partial y}(x,y,y') + h'\frac{\partial f}{\partial y'}(x,y,y') \right) dx \tag{4}$$

Från den vanliga analysen vet vi att ett nödvändigt villkor för en extrempunkt är att gradienten är noll, vilket är ekvivalent med att differentialen är noll för alla Δx och Δy. Det visar sig att ett analogt resultat gäller i variationskalkylen.

Sats 1: Variationen av $J[y]$ blir noll då y är den minimerande funktionen. Detta kan även formuleras som att ett nödvändigt villkor för att $J[y]$ ska ha ett minimum för $y = \hat{y}$, är att $\delta J[\hat{y}] = 0$.

Bevis: Antag att $\hat{y}$ minimerar $J[y]$. För varje deriverbar funktion $h(x)$, med $h(a) = h(b) = 0$, gäller att

$$J[\hat{y}+h] \geq J[\hat{y}] \Longleftrightarrow \Delta J \geq 0, \tag{5}$$

eftersom $J[\hat{y}]$ är det minsta värde $J[y]$ kan anta. Inkrementet är alltså icke-negativt för varje tillåtet $h(x)$. I enlighet med (3) och (4) skriver vi inkrementet som

$$\Delta J[h] = \delta J[h] + R,$$

där R liksom tidigare betecknar de återstående, högre ordningens termer. Då h är litet kommer tecknet på ΔJ att bestämmas av tecknet på δJ, eftersom övriga termer då är försumbara. Alltså gäller att $\delta J \geq 0$; annars skulle $\Delta J < 0$, vilket strider mot (5).

Antag nu att variationen $\delta J[h]$ är positiv för något $h = h_0$. Då har vi för varje konstant $a > 0$, att

$$\begin{aligned}
\delta J[-ah_0] &= \int_a^b \left((-ah)\frac{\partial f}{\partial y}(x,y,y') + (-ah)'\frac{\partial f}{\partial y'}(x,y,y') \right) dx = \\
&= -a\int_a^b \left(h\frac{\partial f}{\partial y}(x,y,y') + h'\frac{\partial f}{\partial y'}(x,y,y') \right) dx = \\
&= -a\delta J[h_0] < 0
\end{aligned}$$

Men detta innebär att även inkrementet $\Delta J < 0$, vilket strider mot olikheten (5). Vårt antagande att variationen var strängt positiv ledde alltså till en motsägelse, varför den enda återstående möjligheten är att variationen är exakt noll. □

3.2 Euler-Lagranges differentialekvation

Vi går nu vidare med problemet att finna den minimerande funktionen y. Sats 1 kombinerat med ekvation (4) ger att ett nödvändigt villkor är

$$0 = \int_a^b \left(h\frac{\partial f}{\partial y} + h'\frac{\partial f}{\partial y'} \right) dx = \int_a^b h\frac{\partial f}{\partial y}dx + \int_a^b h'\frac{\partial f}{\partial y'}dx$$

Partialintegration av den andra termen ger

$$0 = \int_a^b h\frac{\partial f}{\partial y}dx + \int_a^b h'\frac{\partial f}{\partial y'}dx = \int_a^b h\frac{\partial f}{\partial y}dx + \left[h\frac{\partial f}{\partial y'} \right]_a^b - \int_a^b h\frac{d}{dx}\frac{\partial f}{\partial y'}dx =$$

$$= \int_a^b h\left(\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} \right) dx + \left[h\frac{\partial f}{\partial y'} \right]_a^b$$

Förutsättningen att $h(a) = h(b) = 0$ medför att den sista termen är lika med noll. Vi har alltså

$$\int_a^b h\left(\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} \right) dx = 0$$

Eftersom detta ska gälla för alla $h(x)$ måste den andra faktorn i integranden vara identiskt lika med noll. Vi har därmed härlett *Euler-Lagranges differentialekvation*:

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} = 0$$

Genom härledningen har vi också bevisat följande sats:

Sats 2: Varje funktion som minimerar $J[y]$ är en lösning till Euler-Lagranges differentialekvation.

Observera att detta endast är ett nödvändigt, inte tillräckligt, villkor för minimum. Det finns funktioner som uppfyller ekvationen, men som ändå inte löser minimeringsproblemet. Det finns dessutom situationer då problemet helt saknar lösning; existensfrågan har vi än så länge över huvud taget inte behandlat. Det är inte alltid uppenbart huruvida en lösning existerar, vilket vi kommer att se exempel på längre fram. I många fall kan dock lösningen på ett variationsproblem fås genom att man löser Euler-Lagrange-ekvationen.

4 Några karakteristiska problem

4.1 Kortaste sträckan mellan två punkter

Vi söker den kortaste kurvan mellan punkterna (a, y_0) och (b, y_1). Det första problemet man stöter på är att välja vilka kurvor som får vara med. Vi begränsar oss här till funktionskurvor av typen $y = y(x)$, och låter därför A (mängden av tillåtna funktioner) vara mängden av funktioner så att

$$A = \{y \in C^1[a, b], \quad y(a) = y_0, \quad y(b) = y_1\}$$

Vi söker minimum till funktionalen

$$J[y] = \int_{x_0}^{x_1} \sqrt{1 + y'(x)^2}dx, \quad y \in A.$$

Lagrangianen till denna funktional är som synes

$$f(y') = \sqrt{1 + y'^2}.$$

Insättning i Euler-Lagrange-ekvationen ger

$$0 - \frac{d}{dx}(\frac{y'}{\sqrt{1 + y'^2}}) = 0,$$

varefter integration av bägge led ger

$$C = \frac{y'}{\sqrt{1 + y'^2}} \Longleftrightarrow y' = \frac{C}{\sqrt{1 - C^2}} = K,$$

där K är konstant. Ytterligare en integration ger

$$y = Kx + L$$

Inte helt oväntat visar sig alltså lösningen, om den existerar, vara en rät linje. Konstanterna K och L bestäms med hjälp av randvillkoren $y(a) = y_0$ och $y(b) = y_1$.

4.2 Brachistokronproblemet

Låt en partikel med en viss massa m starta från (x_0, y_0) där partikeln befinner sig i vila. Låt partikeln under påverkan av gravitationskraften glida friktionsfritt till punkten (x_1, y_1) längs en kurva $y = y(x)$. Vilken sådan kurva minimerar falltiden?

Partikelns hastighet längs kurvan kan tecknas $v = \frac{ds}{dt}$, det vill säga $dt = \frac{1}{v}ds$. Linjeelementet ds definieras som $ds = \sqrt{1 + y'(x)^2}dx$.

Detta ger oss den totala falltiden som

$$T(y) = \int_{x_1}^{x_2} \frac{1}{v} ds = \int_{x_1}^{x_2} \frac{\sqrt{1 + y'(x)^2}}{v} dx$$

För att klara av detta problem så måste vi uttrycka v som en funktion av y och/eller x. Eftersom partikeln glider friktionsfritt medför varje förändring i potentiell energi en lika stor förändring i kinetisk energi. Detta ger oss den kinetiska energin då partikeln förflyttat sig från den ursprungliga höjden y_0 till höjden $y(x)$. Vi får

$$\frac{mv^2}{2} = mg(y_0 - y) \Longleftrightarrow v = \sqrt{2g(y_0 - y)}$$

Vi ska således minimera funktionalen

$$T[y] = \int_{x_1}^{x_2} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2g(y_0 - y)}} dx$$

Lagrangianen är alltså

$$f(y, y') = \frac{1}{\sqrt{2g}} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{y_0 - y}}$$

Efter lite räknande med i huvudsak kedjeregeln övergår Euler-Lagranges differentialekvation i

$$f(y, y') - y' \frac{\partial f}{\partial y'}(y, y') = C$$

för någon godtycklig konstant C. Sätter vi in uttrycket för $f(y, y')$ får vi

$$\frac{\sqrt{1 + y'(x)^2}}{\sqrt{y_0 - y(x)}} - y'(x)^2 \frac{(1 + y'(x)^2)^{-\frac{1}{2}}}{\sqrt{y_0 - y(x)}} = C,$$

vilket kan förenklas till

$$y'(x)^2 = \left(\frac{dy}{dx}\right)^2 = \frac{1 - C^2(y_0 - y(x))}{C^2(y_0 - y(x))}$$

Detta kan vi skriva om till

$$dx = \pm \frac{\sqrt{y_0 - y(x)}}{\sqrt{C_1 - (y_0 - y(x))}} dy, \quad C_1 = C^{-2} \tag{6}$$

Eftersom vi "vet" att partikeln rör sig nedåt längs bankurvan måste $\frac{dy}{dx} \leq 0$. Vi väljer därför minustecknet i uttrycket för dx.

För att komma vidare gör vi ett variabelbyte:

$$y_0 - y(x) = C_1 \sin^2(\frac{\theta}{2}) = \frac{C_1}{2}(1 - \cos(\theta)),$$

varav

$$\frac{dy}{d\theta} = -C_1 \sin(\frac{\theta}{2})\cos(\frac{\theta}{2}) \Longleftrightarrow dy = -C_1 \sin(\frac{\theta}{2})\cos(\frac{\theta}{2})d\theta$$

Insättning i ekvation (6) ger

$$dx = C_1 \sin^2(\frac{\theta}{2})d\theta = \frac{C_1}{2}(1 - \cos(\theta))d\theta$$

Integration av denna likhet ger

$$x = \frac{C_1}{2}(\theta - \sin(\theta)) + C_2,$$

vilket tillsammans med

$$y = y_0 - \frac{C_1}{2}(1 - \cos(\theta))$$

utgör lösningen i form av en cykloid framställd på parameterform. Konstanterna C_1 och C_2 bestäms med hjälp av de två randvillkoren $y(x_0) = y_0$ och $y(x_1) = y_1$.

Antag nu att man istället för en partikel som glider längs banan betraktar ett hjul som rullar. Då blir man tvungen att ta hänsyn till hjulets rotationsenergi, vilken måste läggas till rörelseenergin i räkningarna ovan. Den totala kinetiska energin blir i det fallet $7mv^2/10$ istället för $mv^2/2$. Det fantastiska är att detta leder till samma differentialekvation och samma lösning som i fallet med en glidande partikel. Lösningskurvan blir alltså fortfarande en cykloid.

En närmare analys visar att cykloiden inte bara löser brachistokronproblemet. Kurvan löser även det så kallade *tautokronproblemet*, som går ut på att finna den kurva som ger samma falltid, oavsett var på kurvan partikeln startar. Cykloiden har alltså den märkliga egenskapen att två partiklar som börjar glida (eller rulla) på två olika platser längs kurvan når slutpunkten (x_1, y_1) samtidigt!

5 Existerar lösningarna?

I alla de fall vi behandlat så här långt har vi undvikit att behandla existensfrågan — har det ställda problemet verkligen en lösning? Det kan lätt förefalla självklart eller sannolikt att en lösning existerar, om inte annat på grund av de fysikaliska frågeställningar som ligger bakom många av problemen. Vi ska ska nu se ett exempel på att det inte alls behöver vara uppenbart. Vi betraktar problemet att finna den minimerande funktionen till funktionalen

$$J[y] = \int_0^1 |y'(x)|^{1/2} dx \tag{7}$$

under bivillkoren att $y(0) = 1$ och $y(1) = 0$. Som definitionsmängd $\mathcal{A}$ väljer vi alla kontinuerliga, styckvis deriverbara funktioner som uppfyller bivillkoren.

För att visa att det inte finns någon lösning $y \in \mathcal{A}$ går vi till väga på följande sätt. Vi visar först att en eventuell lösning $\hat{y}$ måste uppfylla $J[\hat{y}] = 0$. Därefter visar vi att det inte finns någon funktion i $\mathcal{A}$ som har denna egenskap.

Att $J[y]$ inte kan vara negativ framgår av (7); integranden utgörs av kvadratroten ur ett absolutbelopp och är alltså hela tiden icke-negativ. Följaktligen är även hela integralen större än eller lika med noll.

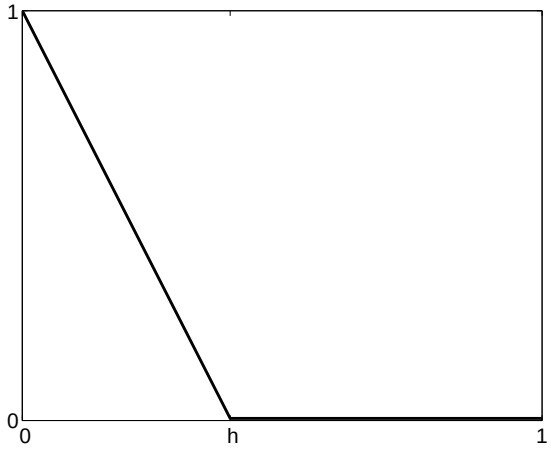
Betrakta nu funktionen

$$y(x) = \begin{cases} 1 - \frac{x}{h}, & 0 \leq x < h; \\ 0, & h \leq x \leq 1; \end{cases}$$

med $0 < h < 1$. Funktionsgrafen finns avbildad i figuren.

Derivation ger att

$$y'(x) = \begin{cases} -\frac{1}{h}, & 0 \leq x < h; \\ 0, & h \leq x \leq 1; \end{cases}$$



Figur 1: Funktionsgraf till $y = y(x)$

Insatt i (7) erhålls

$$J[y] = \int_0^h \left(\frac{1}{h}\right)^{1/2} dx + \int_h^1 0 dx = \left(\frac{1}{h}\right)^{1/2} [x]_0^h = h^{1/2}$$

Tydligen gäller det att $J[y] \to 0$ då $h \to 0$, vilket visar att den minimerande funktionen $\hat{y}$ måste vara sådan att $J[\hat{y}] = 0$. Den funktion vi betraktat här duger dock inte; så länge h inte är exakt noll går det att hitta en "bättre" funktion genom att minska h. Men om h blir noll är $y = 0$ på hela intervallet [0,1], och uppfyller således inte randvillkoren.

Men kan det inte finnas någon annan funktion som både tillhör definitionsmängden $\mathcal{A}$ och får funktionalen att bli noll? Svaret är nej, och för att inse detta börjar vi med att konstatera att $J[y] = 0$ endast om integranden i (7) är noll, eftersom denna inte kan bli negativ. Detta innebär att derivatan $y'(x) = 0$ för alla $x \in [0, 1]$. Men då är $y(x)$ styckvis konstant och kan inte samtidigt uppfylla $y(0) = 1$, $y(1) = 0$ och kontinuitetsvillkoret.

Om vi skulle ändra vår definitionsmängd $\mathcal{A}$ till att omfatta även icke-kontinuerliga funtioner skulle till exempel $y(x) = \{1, x = 0;\ 0, 0 < x \leq 1\}$ i någon mening kunna anses vara en lösning. Valet av definitionsmängd kan alltså vara direkt avgörande för huruvida en lösning existerar eller inte.

Värd att notera är även exponentens betydelse i integranden i (7). Om den skulle ändras från 1/2 till exempelvis 1 visar det sig att funktionalens värde blir konstant lika med ett, oberoende av h.

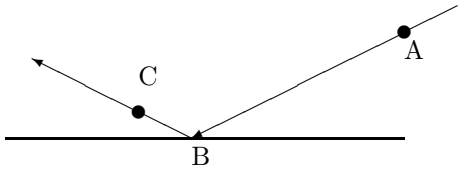
Även om exemplet i detta avsnitt kan verka konstruerat illustrerar det hur oförutsägbara resultaten kan vara beträffande existensen av lösningar.

6 Tillämpningar av variationskalkyl

6.1 Principen om minsta verkan

Då Maupertuis formulerade principen om minsta verkan på 1740-talet, grundade han den på känslan att universum var alltför perfekt och välordnat för att slösa med det vi idag kallar energi. Rörelser i naturen måste vara beskaffade på så sätt att att någon storhet minimeras, och han kom fram till att detta var den kinetiska energin. Senare vidareutvecklades teorin av bland andra Euler, som fann att kroppar i vila anordnas så att den potentiella energin minimeras.

Variationskalkyl handlar till stor del om att hitta det minimala eller maximala utfallet av en process, eller det effektivaste sättet att göra något, och kan förklara många fenomen i naturen. Ett exempel vi tagit upp är problemet att hitta det kortaste avståndet mellan två punkter. Något mer avancerat är att hitta den kortaste vägen som sammanbinder två punkter, A och C, och en linje.



Det visar sig att den kortaste vägen beskrivs av en ljusstråle som går genom A, reflekteras någonstans på linjen, B, för att sedan gå genom C. Här kan linjen lämpligtvis ses som en spegel vinkelrät mot pappersplanet. Även en oskruvad biljardboll som som går genom A, studsar mot på linjen och sedan går genom C beskriver denna väg. Tydligen rör sig ljus på så sätt att det väljer den tidseffektivaste vägen mellan två punkter, eller tre som i fallet ovan. Detta kallas *Fermats princip om minsta tid* . När ljuset stöter på ett material där det färdas långsammare tar det en genväg för att komma igenom materialet så fort som

möjligt, för att sedan rikta sig mot punkten igen. Detta leder visserligen till att ljuset färdas en längre sträcka, men under kortare tid. Allt det här återspeglar principen om att naturen alltid försöker minimera någonting, använda minsta möjliga energi.

Inom geodesin kan man hitta många ytterligare exempel på principen om minsta verkan. Geodesi är ett område där man utgår från principen om minsta möjliga verkan, man säger att fysikaliska processer och förändringar tar den enklaste eller minimala vägen. Varför tar rinnande vatten den väg som är brantast? Svaret är att det är det effektivaste sättet att bli av med potentiell energi och ta sig från en punkt till en annan.

Något som verkligen kopplar ihop principen om minsta verkan med variationskalkyl är Hamiltons princip. Principen går ut på att beskriva den väg en partiken rör sig med hjälp av integralen över skillnaden mellan kinetisk och potentiell energi, det vi kallar verkan. I ekvationen nedan är $L = V - T$, där L är den kinetiska energin och T den potentiella.

$$\int_{t_1}^{t_2} L(x, x', t)dt$$

(Jämför med denna integral med funktionaluttrycket i inledningen). Det visar sig att den korrekta beskrivningen av partikelns rörelse uppenbarar sig då integralen för verkan har ett stationärt värde, helt i enighet med principen om minsta verkan.

6.2 Optimal reglering

Optimal reglering är ett område inom reglertekniken där variationskalkyl används flitigt och har varit till stor nytta inom till exempel rymdforskning och militär verksamhet. Ett problem där optimal kontroll ska användas kan formuleras på följande sätt. Ett systems initiala tillstånd är givet i matematiska termer. Målet, eller det som ska uppnås genom någon process är också givet. Det gäller att utföra processen, ta sig från ett tillstånd till ett annat, på ett så effektivt sätt som möjligt. Ett klassiskt exempel är Goddards raketproblem. Goddard var en pionjär inom rymdforskningen och arbetade med hur man ska kunna skjuta upp en raket så långt som möjligt. Hur ska man variera rakettrycket under uppskjutningen för att nå så högt som möjligt? Problemet kan formuleras med följande ekvation.

$$v' = \frac{1}{m}(u - D(v, h)) - g$$

Här står m för massan, som minskar då bränsle förbrukas, u står för motorkraften, g för gravitationskonstanten och D för det aerodynamiska motståndet, som beror av hastigheten v och höjden h. I det här uttrycket är det mycket som varierar med tiden, variationskalkyl är alltså ett mycket bra verktyg för att hantera problemet. Tar man inte hänsyn till det aerodynamiska motståndet är lösningen att använda full motorkraft hela tiden, eftersom man måste pressa upp en större massa om man inte använder lika mycket bränsle. Problemet med aerodynamiskt motstånd löstes med hjälp av variationskalkyl.

6.3 Såpfilmens minimalyta

Hur kommer det sig då att en såpfilm uppför sig som den gör, och varför är en såpbubbla rund?

Svaret på frågan är den minimala energiprincipen. Det som ska minimeras är ytspänningen, och den är proportionell mot ytans area. När en såpyta spänns upp med ståltråd bildas den minsta yta som har ståltråden som rand. För en given volym är den minsta möjliga begränsningsytan en sfär. Därför är såpbubblor sfäriska.

7 Källförteckning

Byström, Johan, (2002)
Introduktion till variationskalkyl,
http://www.sm.luth.se/ johanb/applmath/chap3/.

Davis, Chris, (1998)
Idle Theory: Least Action Principles,
http://ourworld.compuserve.com/homepages/cuius/idle/evolution/ref/leastact.html.

Gelfand, I. M. och Fomin, S. V., (1963)
Calculus of Variations,
Prentice-Hall

Goldstine, Herman H., (1980)
A History of the Calculus of Variations,
New York, Springer-Verlag New York Inc.

Hjorteland, Trond, (1999)
The Action Variational Principle in Cosmology
http://trond.hjorteland.com/thesis/hoved.html.

School of Mathematics and Statistics, University of St Andrews, Scotland (2004)
The MacTutor History of Mathematics archive: Biographies,
http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Euler.html.

Sparr, G. och Sparr, A., (2000)
Kontinuerliga system, andra upplagan,
Lund, Studentlitteratur

Weisstein, Eric W., (1999)
Brachistochrone Problem,
http://mathworld.wolfram.com/BrachistochroneProblem.html.
Cycloid,
http://mathworld.wolfram.com/Cycloid.html.
Eric Weisstein's World of Biography: Mathematicians,
http://scienceworld.wolfram.com/biography/topics/Mathematicians.html.

[Material från internetkällor är hämtat i slutet av april 2005.]

De platonska kropparna

Christian Lundtofte och Ola Svensson

19 maj 2005

Inledning

De platonska kropparna är regulära polyedrar. En polyeder är en kropp som begränsas av plana ytor. Regulär är den om sidorna är kongurenta, regelbundna månghörningar och om det i varje hörn möts precis lika många ytor. Det finns exakt fem stycken regulära polyedrar: tetraedern, hexaedern, oktaedern, dodekaedern och ikosaedern (figur 1). Anledningen till att de kallas platonska är dialogen Timaios där Platon sammanbinder tetraeden, hexaedern, ikosaedern, oktaedern med de fyra elementen eld, jord, vatten och luft. Dodekaedern lät Platon symbolisera universum.

Figur 1: De platonska kropparna

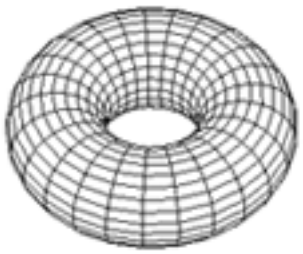
Hur kommer det sig att det endast kan existera fem olika typer av sådana här kroppar? Med hjälp av det matematiska område som kallas Eulerkaraktäristik kan en förklaring till denna frågeställning ges på ett enkelt sätt.

Eulers formel

För samtliga enkla polyedrar, dvs. kroppar som ej innehåller hål likt figur 2 och vars yta kan omformas till en sfär skapade Leonard Euler (1707-1783) följande samband:

$$V - E + F = 2, \tag{1}$$

där V är antalet hörn, E antalet kanter och F är antalet sidor på polyedern.



Figur 2: Torus

Då sambandet upptäcktes, 1750 var Euler oerhört förvånad att en så uppenbar observation ej gjorts tidigare under de två millenium som polyedern känts till och skrev till sin vän Christian Goldback i ett brev samma år följande:

"I find it surprising that these general result in solid geometry have not been noticed by anyone, so far as i am aware..."

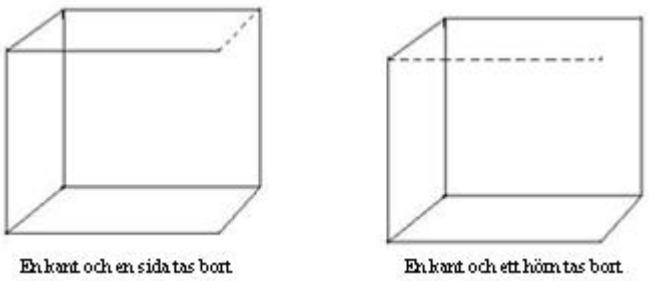
Dock var han ej kapabel till att konstruera ett fullständigt bevis för sitt samband. Det första generellt accepterade bevis gjordes av Adriene-Marie Legendre 1794 och kan beskrivas på följande sätt. Om man från början har en hel kropp som vi sedan bryter ner steg för steg till inget är kvar, då har vi från början

$$V - E + F = 2.$$

om vi tar bort en kant kommer två stycken sidor att försvinna vilket ger

$$V_1 - E_1 + F_1 = V - (E - 1) + (F - 2) = 1.$$

Nu fortsätter vi från det hörnet och tar bort en kant till. Det kan ske på två olika sätt antingen om kanten är ensam om att binda i det hörnet då försvinner även hörnet eller så kommer kanten att tillhöra en hel polygon och då försvinner en sida istället(figur 3). Om vi fortsätter så här kommer vi tillslut endast att ha ett hörn kvar.



Figur 3:

Om vi tar bort en kant och ett hörn får vi

$$V_2 + E_2 + F_2 = (V_1 - 1) - (E_1 - 1) + F_1 = 1.$$

Om vi istället tog bort en kant och en sida får vi

$$V_3 + E_3 + F_3 = V_1 + E_1 + F_1 = V - (E - 1) + (F - 1) = 1.$$

När vi sedan endast har ett hörn kvar så har vi

$$V_4 + E_4 + F_4 = 1 + 0 + 0 = 1.$$

Eulers formel är bevisad.

Eulerkaraktäristik

För kroppar utöver de enkla polyedrarna har en generalisering av Eulers samband $V - E + F$ visat sig gälla för alla typer av kroppar. Generaliseringen ser ut på följande vis:

$$V - E + F = X(g)$$

där $X(g) = 2 - 2g$ och g står för genus som grovt sagt är antalet av den typ av hål som figur 2 visar. Därmed genererar kroppar som cylinder, torus, dubbel torus och sfären följande värde på $X(g)$:

Kropp	*X(g)*
cylinder	0
torus	0
dubbel torus	-2
sfär	2

De platonska kropparna

Då förståelse skapats för Eulers formel (1), kan ett bevis för att de platonska kropparna är precis fem göras efter ett antal observationer. Vi antar att varje sidoyta på kroppen är en regelbunden n-hörning och att i dess hörn sammanfaller m stycken kanter. Då kan följande samband observeras:

$$nF = 2E, \tag{2}$$

och

$$mV = 2E. \tag{3}$$

Anledningen till (2) är att varje kant räknas två gånger, och eftersom varje kant går mellan två hörn på kropparna gäller (3). Med hjälp av sambanden ovan ger (1) det nya sambandet:

$$\frac{2E}{m} - E + \frac{2E}{n} = 2,$$

som kan skrivas

$$\frac{1}{m} - \frac{1}{2} + \frac{1}{n} = \frac{1}{E},$$

eller

$$\frac{1}{m} + \frac{1}{n} = \frac{1}{2} + \frac{1}{E}. \tag{4}$$

Två nödvändiga villkor för (4) är dels att $m \geq 3$ eftersom minst tre kanter sammanfaller i ett hörn, och dels att $n \geq 3$ på grund av att varje sidoyta har minst tre hörn. Dock kan ej både m och n vara större än 3. Ty låt oss säga att m och $n \geq 4$ och observera följande:

$$\frac{1}{m} + \frac{1}{n} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} < \frac{1}{2} + \frac{1}{E},$$

vilket strider mot (4). Om $n = 3$ ger (4):

$$\frac{1}{m} - \frac{1}{6} = \frac{1}{E},$$

eller

$$E = \frac{6m}{6 - m}.$$

Detta innebär att m måste vara mindre än 6, dvs. $m = 3, 4$ eller 5 då $n = 3$. Detta ger med hjälp av (2) och (3) följande värden på V, E och F: (V,E,F) = (4,6,4), (6,12,8) eller (12,30,20). Dessa värden motsvarar tetraedern, oktaedern och ikosaedern. Då istället $m = 3$ kan vi på analogt sätt konstatera att begränsningen av E blir

$$E = \frac{6n}{6 - n},$$

och med hjälp av (2) och (3) fås nu värdena (4,6,4), (8,12,6) eller (20,30,12). Dessa värden motsvarar tetraedern, kuben och dodekaedern. Därmed har vi visat att det endast existerar fem reguljära polyedrar.

De arkimediska kropparna

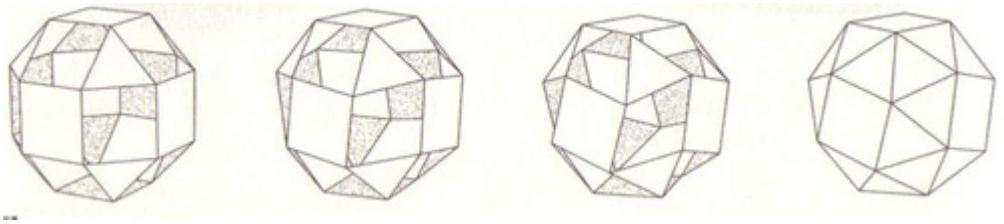
Tidigare såg vi att då en kropp endast består av kongruenta och liksidiga sidoytor kan endast utseendet variera mellan fem olika alternativ. Om kroppen istället skulle bestå av två eller flera typer av regelbundna sidoytor, dock fortfarande regelbundna så att alla hörnen består av samma antal av varje polygon, kommer antalet kombinationer att vara begränsat till tretton stycken.
Ett sätt att bilda arkimediska kroppar på är genom stympning av de platonska kropparna, det går till så att man skär bort hörnen på polyedern och på så sätt bildar en ny polyeder (figur 4). På detta sättet bildas sju av de arkimediska kropparna (stympad tetraeder, stympad hexaeder, stypad oktader, stympad dodekaeder, stympad ikosaeder, kuboctaeder och ikosidodekaeder)

Figur 4: Stympning

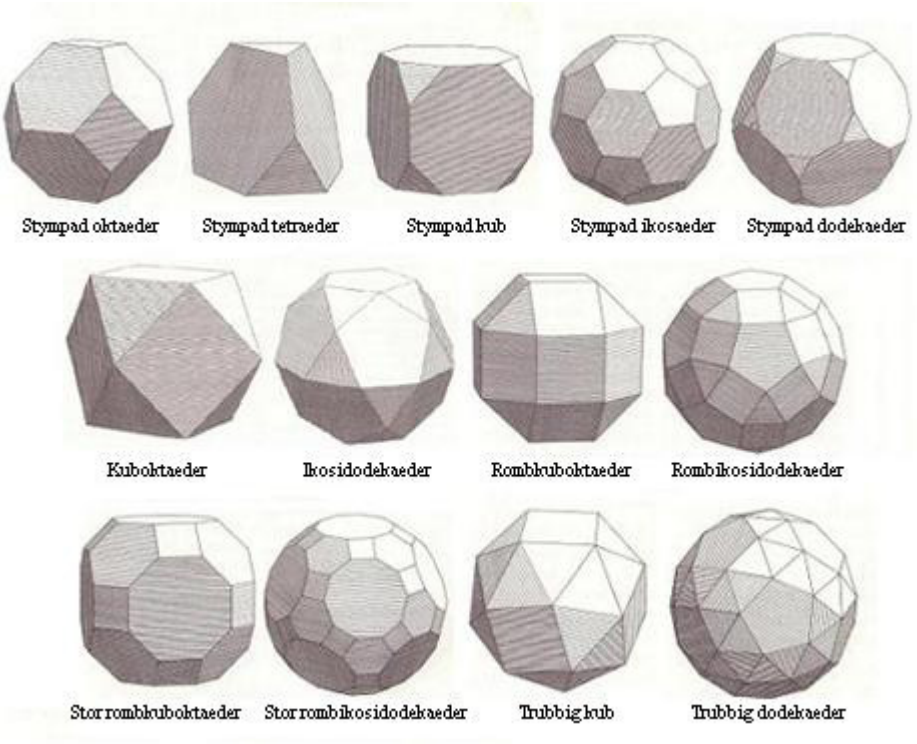
Ytterligare två polyedrar (rombkuboktaeder och rombikosidodekaeder) skapas genom utvigdning av de platonska kropparna och ytterligare två kroppar (stor rombkuboktaeder och stor rombikosidodekaeder) kan fås genom utvigdning av de tidigare nio arkimediska kropparna. Utvigdning går till så att man flyttar sidorna ut från centrum så de är på en kantlängds avstånd från varandra (figur 5).

Figur 5: Utvigdning

De två sista arkimediska kropparna (trubbig hexaeder och trubbig dodekaeder) bildas genom en listig vridning, först gör man en utvigdning som i föregående bildning, genom att sedan vrida den figuren kan de sista kropparna bildas (figur 6).



Figur 6: Vridning

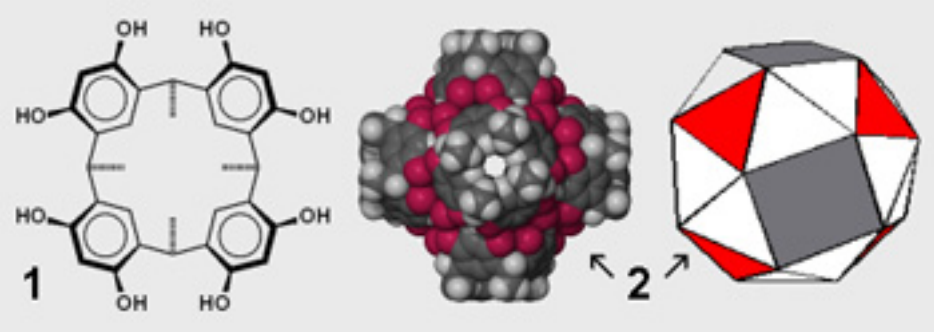


Figur 7: De arkimediska kropparna

Tillämpningar och intresse i dagsläget

Kemisk struktur

Eftersom de platonska kropparna är kända sedan över 2000 år tillbaka i tiden finns det i dagsläget ej någon större vetenskapligt forskning kring dessa. Nämnvärt är dock att man i naturen kan de platonska kropparna framträda i form av bl.a. kristaller och mineraler. På University of Missouri - Columbia, USA leder professor Jerry L. Atwood ett forskningsteam med inriktning på kemiska strukturer utöver de molekylära. Området är känt som supramolekylär kemi och Atwoods studier visar starka samband mellan platonska och arkimediska kroppar och byggstenarna i de supramolekylära strukturerna.



Figur 8: resorcin[4]-arener

Ett exempel är resorcin[4]-arenerna som spontant kan slå sig samman och bilda strukturer likt den arkimediska kropp som figur 8 visar. Denna upptäckt gjorde det möjligt att genom ett perspektiv utifrån matematisk struktur klassificera dessa supramolekylära sammansättningar, vilket ej var lika uppenbart utifrån ett kemiskt perspektiv. Detta visar att de platonska och arkimediska kropparna utmärkta som modeller för att beskriva olika typer av kemiska strukturer likt de supramolekylära.

Topologi

Detta matematiska område dök upp på mitten av 1800-talet och sågs som ett oerhört banbrytande tankesätt inom geometri. Den gav insikten i proportioner som bevaras efter deformering av kroppar. Cirkeln är topologiskt ekvivalent med en ellips och samma ekvivalens gäller för sfären och ellipsoiden, eftersom topologi endast tar hänsyn till hur kroppar och plana nätverk är system av sammanlänkade punkter och ytor. I verkligheten existerar det ej en perfekt ritad cirkel utan avvikelser från idéernas värld och dessa avvikelser som naturen skapar är med andra ord inga hinder för ett topologiskt samband. Med detta tankesätt har olika typer av topologi utformats, såsom algebraisk-, differential- och lågdimensionell topologi. Bernhard Riemann (1826-1866) uppfattade topologin som nyckeln till förståelse av de djupaste proportioner av analytiska funktioner av en komplex variabel. Riemanns funktionsteorier innehar därmed fundamentala topologiska koncept. Eftersom Eulers samband $V - E + F = 2$ för enkla polyedrar, som upptäcktes efter studier av de platonska kropparna, gäller oavsett form på kroppen sägs detta samband vara grundläggande inom topologin. Det var nämligen inom topologin som detta samband dök upp på nytt då generaliseringen $V - E + F = X(g)$ skapades av Poincaré. Denna generalisering gjorde nu det möjligt att topologiskt beskriva alla typer av kroppar och ytor i naturen utan begränsningar. I dagens läge anses fortfarande topologi som ett grundläggande och viktigt kunskapsområde att behärska då analys görs av naturens egna geometriska förhållanden.

Källförteckning

Richard Courant, Herbert Robbins "What is mathematics?", Oxford University Press, Inc., 1996

Coxeter, "Introduction to Geometry", Braun-Brumfield, Inc., 1989

Daud Sutton, "Kuber, klot och andra geometriska kroppar", Svenska förlaget, 2002

http://www.chem.missouri.edu/faculty/Atwood/research.html
http://mathworld.wolfram.com/EulerCharacteristic.html
http://mathworld.wolfram.com/Topology.html
http://mathworld.wolfram.com/ArchimedeanSolid.html

Matematisk Kommunikation π 04
Omöjliga figurer

Emma Blom
Lova Brodin
Petter Cederholm
Susanne Olsson
Handledare Gunnar Sparr

19 maj 2005

Innehåll

1 Inledning

Bilderna på framsidan är sk omöjliga figurer. Detta arbete består av två delar, i den första delen visas hur man rent matematiskt kan bestämma om en figur är en bild av ett omöjligt objekt i tre dimensioner. Den andra delen tar upp hur den mänslkiga hjärnan uppfattar – eller försöker uppfatta – bilder av den typen. Inledningsvis kommer historien om dessa figurer att tas upp och speciellt de mest berömda personerna i detta sammanhang, däribland M.C. Escher och Oscar Reutersvärd. Därefter kommer teorin för hur man räknar på möjliga figurer att klargöras. Utifrån denna härleds ett bevis för vad som är en omöjlig figur. Avslutningsvis räknas ett exempel med Reutersvärds trappa.

I den perceptionspsykologiska delen görs ett försök att förklara varför hjärnan (trots att de inte kan existera) kan uppfatta omöjliga figurer som rumsliga objekt.

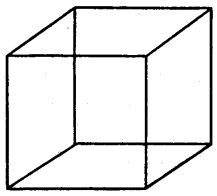
2 Historik

Bilder som lurar ögat och sinnena har länge fascinerat människan, omöjliga figurer är exempel på detta.

Det är framförallt tre namn som man stöter på när man läser om eller studerar omöjliga figurer. *M.C. Esher* är ett av dessa. Escher föddes 1898 i Leeuwarden (Nederländerna). 1919 började han på *Haarlem School of Architecture and Decorative arts*. Escher studerade först arkitektur men övergick till dekorativ konst och det var här han kom in på vägen till de omöjliga figurerna.

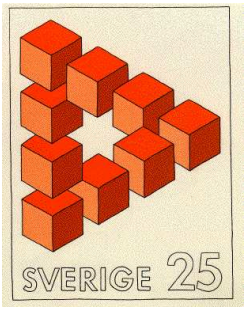
Två av Eschers mest kända verk är *Drawing Hands*, som visar två händer som ritar varandra och *Sky and water* som med ljus och skuggors hjälp omvandlar en fisk i vattnet till en fågel i himmeln.

Eschers arbete har även en stark matematisk koppling och många av de världar som han ritat är byggda runt omöjliga figurer som senare kommer att kallas *Neckers kub* och *Penrose's triangel*.



Eschers verk är omtyckta hos forskare, speciellt matematiker, av den anledningen att han på ett intressant sätt använder polyedrar och geometriska desorienteringar. Matt Groening, skaparen av bl.a *The Simpsons*, har använt en referens från Escher i ett avsnitt *Life in Hell*, där kaniner ramlade ner för trappor från omöjliga vinklar. Groening använde också detta skämt i ett avsnitt av *Futurama*
Escher dog 1972 i Laren (Nederländerna) efter att ha bott i fyra olika länder, se [1].

Oscar Reutersvärd föddes 1915 och var konstnär och konsthistoriker vid *Lunds Universitet*. Reutersvärds triangel, återgiven på frimärket nedan, skapar en paradoxal figur som uppstår ur konflikten mellan yta och djup, se [2]. Reutersvärd har även gett namn åt det sk Reutersvärd index som senare i detta arbete kommer ha en central roll för att visa om en figur är omöjlig.



Roger Penrose är ett namn man stöter på inom många delar av matematiken. Penrose föddes 1931 och är en brittisk matematisk fysiker som har varit professor vid *Birkbeck College* (1967-73) i London och sedan 1973 är professor vid *Oxford*. Penrose har gjort viktiga insatser inom flera grenar av den matematiska fysiken. Inom ramen för klassisk allmän relativitetsteori har han, delvis i samarbete med Stephen Hawking, löst generella singularitetsproblem.

Penrose har även arbetat med omöjliga figurer. Det mest kända är hans omöjliga triangel som är väldigt lik *Oscar Reutervärds* triangel. Penrose triangel, se figur på sidan 5, är även känd under namnet *tribar*, se [3].

Även om Reutersvärds triangel och Penrose's tribar vid ett ytligt betraktande ser ut att bygga på samma idé, såg Oscar Reutersvärd stora skillnader. Så här skriver han i ett brev till Gunnar Spar, [5].

(Jag har ett litet klargörande att komma med. I litteraturen förekommer beteckningen "reutersvärdtriangeln." Den används för att skilja min omöjliga, triangulära figur från Roger Penroses "tribar". Min figur är japansk – alltså axonometriskt konstruerad. Den består ursprungligen av ett antal glest grupperade kuber, därefter av kompakt hopfogade sådana, för att slutligen utgöra en monolitisk kropp. Penroses figur blev en kolossal "förbättring" av min. Min ingår i "betraktelserummet." "Tribar" ingår i "upplevelserummet". Den består av tre balkar, sedda i försvinningsperspektiv, som medsols suger betraktaren bortåt i evigt kretslopp och motsols evigt hitåt.)

3 Olika typer av omöjliga figurer

Omöjliga figurer kan delas upp i äkta och oäkta. Även illusioner räknas ibland hit, men de är egentligen inte omöjliga utan kan uppfattas på olika sätt beroende på hur man tittar på dem. Exempel på detta är tex Neckers kub, se figur på

föregående sida. Den kan uppfattas som en konvex eller konkav kub, dvs gå ut från pappersplanet eller inåt bakom pappersplanet. Andra klassiska exempel på illusioner är vas/två ansikten och gammal/ung kvinna. Äkta omöjliga figurer uppfattas av hjärnan som slutna geometriska figurer, medan de oäkta tycks lösas upp när man följer begränsningslinjerna, se bild nedan. Se även [4]. I fortsättningen används termen "omöjlig figur" för "äkta omöjlig figur".

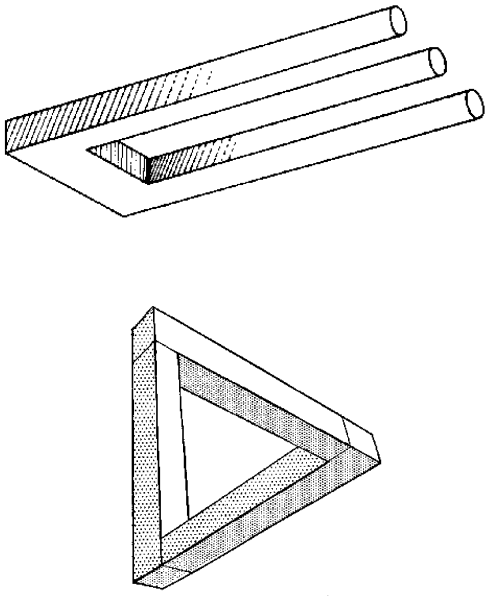


Fig. 1. Perspective drawing of impossible structure.

4 Definition av omöjlig figur

Låt Y vara en punktkonfiguration i 2D uppbyggd av polygonområden. *Om Y ej kan vara en projektion av en sann 3D konfiguration X med plana begränsningsytor, sägs den vara en omöjlig figur.*

Märk att en konfiguration kan ha delfigurer som utgör möjliga delobjekt utan att hela konfigurationen är möjlig. Omvänt om man bygger på en omöjlig figur med tex en rektangulär yta så fås en sammansatt figur som inte är omöjlig. Sådana undantagsfall betraktas inte här, jämför [8].

5 Formteori

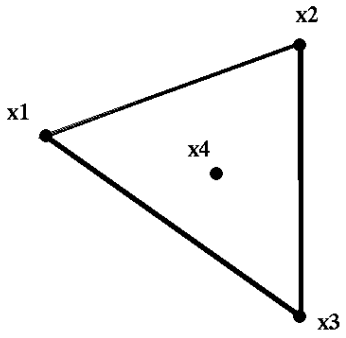
För att kunna utföra beräkningar på omöjliga figurer måste grunderna för möjliga figurer klargöras. Med utgångspunkt från generella metoder för formteori formuleras en teori för att avgöra om en figur är omöjlig eller ej. För en utförligare behandling se [8].

5.1 Shape

Inledningsvis görs några illustrerande exempel för punktkonfigurationer. Låt X^1, X^2, X^3 vara tre punkter i ett plan och låt x^1, x^2, x^3 vara dess koordinatre-

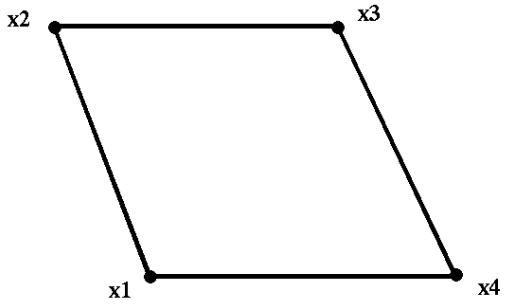
presentation i någon viss bas. Från linjär algebra är det känt att masscentrum i en triangel med hörnen x^1, x^2, x^3 ges av

$$x^4 = \frac{1}{3}(x^1 + x^2 + x^3) \Longleftrightarrow x^1 + x^2 + x^3 - 3x^4 = 0$$



För hörnen x^1, x^2, x^3, x^4 i ett parallellogram gäller att

$$\overline{x^1x^3} = \overline{x^1x^2} + \overline{x^1x^4}$$
$$\Longleftrightarrow (x^3 - x^1) = (x^2 - x^1) + (x^4 - x^1)$$
$$\Longleftrightarrow x^1 - x^2 + x^3 - x^4 = 0$$



Koefficienterna $(1, 1, 1, -3)$ för triangeln med masscentrum och $(1, -1, 1, -1)$ för parallellogrammet gäller för alla sådana konfigurationer. Dessa kallas för ξ–vektorer. Mängden av möjliga sådana koefficienter kallas *Shape*. För alla punktkonfigurationer gäller att summan av elementen i shapevektorn är noll.

Inför beteckningen $X = (X^1, ..., X^4)$ för en punktkonfiguration. Låt $s(X)$ beteckna shape av X. För det första exemplet, där $X = (X^1, X^2, X^3, X^4)$ med masscentrum x^4 för triangeln x^1, x^2, x^3, gäller att

$$s(X) = t(1, 1, 1, -3),\ t \in R.$$

Om X är hörnen i ett parallellogram (x^1, x^2, x^3, x^4) numrerade enligt exemplet gäller

$$s(X) = t(1, -1, 1, -1),\ t \in R.$$

På liknade sätt kan man behandla punktkonfigurationer på en linje (x^1, x^2, x^3). Låt X^1, X^2, X^3 vara punkter på en linje där X^2 är mittpunkten. Då gäller

$$s(X) = t(1, -2, 1),\ t \in R.$$

Mer komplicerade konfigurationer delas lättast upp i flera delfigurer. Koefficienterna framför de punkter som inte tillhör delfiguren sätts då till noll då $s(X)$ beräknas.

Låt nu $X = (X^1, X^2, ..., X^n)$ vara en godtycklig punktkonfiguration. $(X^1, X^2, ..., X^n)$ står för de fysiska punkterna, medan $(x^1, x^2, ..., x^n)$ är punkterna med koordinatrepresentation.

Punkterna kan beskrivas i olika baser, där sambanden mellan olika baser kan skrivas på formen $x'^k = Ax^k + b,\ k = 1, 2, ..., n$, där x och x' är koordinater för samma punkt i olika koordinatsystem. För att matematiskt definiera formen av X används följande lemma.

Lemma. Låt x^k och $x'^k,\ k = 1, ..., n$, vara koordinatrepresentationer för punkterna $X^1, ..., X^n$ i två olika baser, dvs $x'^k = Ax^k + b,\ k = 1, 2, ..., n$. Då gäller

$$\{\xi | \sum \xi_k x^k = 0, \sum \xi_k = 0\} = \{\xi | \sum \xi_k x'^k = 0, \sum \xi_k = 0\}. \tag{1}$$

Detta innebär speciellt att alla vektorer ξ är sådana att summan av dess elementen är noll. För ett parallellogram är $\xi = (1, -1, 1, -1)$, enl exemplet ovan. Lemmat uttrycker att oavsett bas kommer alltid $\sum \xi_k x^k$ vara noll.
Bevis : Om $\sum \xi_k = 0$, $\sum \xi_k\ x^k = 0$ så följer ur

$$\sum \xi_k x'^k = \sum \xi_k (Ax^k + b) = A(\sum \xi_k x^k) + (\sum \xi_k) b$$

att $\sum \xi_k\ x'^k = 0$. För beviset i andra riktningen se [7] och [8]. Lemmat säger att $s(X)$ är oberoende av koordinatrepresentation. Ett annat sätt att uttrycka detta är att $s(X)$ är en fullständig affin invariant, dvs

$$X' = \text{ Aff}(X) \Longleftrightarrow s(X') = s(X)$$

Detta inspirerar till följande definition.

Definition. Formen av X definieras av

$$s(X) = \{\xi | \sum \xi_k x^k = 0, \sum \xi_k = 0\} \tag{2}$$

5.2 Egenskaper

Ett annat sätt att uttrycka definitionen är att $s(X)$ är mängden av lösningar (nollrummet) till det homogena ekvationsystemet på matrisform

$$\begin{pmatrix} x^1 & x^2 & \dots & x^n \\ 1 & 1 & \dots & 1 \end{pmatrix} \xi = 0,$$

där ξ är kolonnmatrisen

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}.$$

Innebörden av detta är att

$$\sum \xi_k x^k = 0, \ \sum \xi_k = 0.$$

Av definitionen följer att $s(X)$ är ett linjärt underrum i R^n. Dimensionen av ett linjärt underrum är antalet oberoende basvektorer som behövs för att entydigt representera alla vektorer i rummet, se [9]. Observera att i fallen med triangeln och parallellogrammet är dim $s(X) = 1$, likaså för exemplet med den räta linjen. I det allmänna fallet gäller

Sats 1. Låt $X = (X^1, ...X^n)$. Då gäller

$$\dim s(X) = \begin{cases} n-4 & \text{om} \quad X \quad \text{är 3D, ej 2D} \\ n-3 & \text{om} \quad X \quad \text{är 2D, ej 1D} \\ n-2 & \text{om} \quad X \quad \text{är 1D} \end{cases} \tag{3}$$

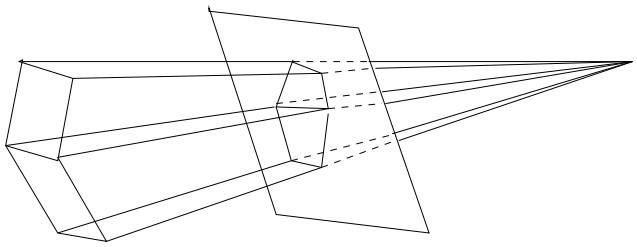
För bevis se [6], [7].

5.3 Perspektivavbildningar

För projektion av en punktkonfiguraton på ett plan används perspektivavbildningar. Sådana beskrivs av funktioner P av formen

$$P : X \to Y, \ \overline{\phi X} = \alpha \overline{\phi Y}.$$

För utförligare diskussion se [7].



Punkten ϕ är strålarnas utgångspunkt, fokalpunkten. $X^1, ..., X^n$ är en punktkonfiguration i rymden och $Y^1, ...Y^n$ är dess projektion i planet. Talet α_k kallas för djupet och uttrycker hur X^k och Y^k förhåller sig till varandra. Hela vektorn $\alpha = (\alpha_1, \ldots, \alpha_n)$ är djupvektorn för konfigurationen X med avseende på Y. Ur definitionerna ovan följer att

$$\xi \in s(X) \Longrightarrow 0 = \sum \xi_k \overline{\phi X^k} = \sum \alpha_i \xi_i \overline{\phi Y^k} = \sum \eta_i \overline{\phi Y^k}$$

Här är per definition $\alpha_i \xi_i = \eta_i$. Ur detta följer att $\eta \in s(Y)$. Relationen $\alpha_i \xi_i = \eta_i$ visar sambandet mellan djupet α och formen för X och Y. Speciellt för fyrpunktskonfigurationen gäller

$$s(X) = (\xi_1, \xi_2, \xi_3, \xi_4),$$

$$s(Y) = (\eta_1, \eta_2, \eta_3, \eta_4)$$

$$\Longrightarrow \alpha = (\frac{\eta_1}{\xi_1}, \frac{\eta_2}{\xi_2}, \frac{\eta_3}{\xi_3}, \frac{\eta_4}{\xi_4}).$$

För ett parallellogram är $s(X) = (1, -1, 1, -1)$ och därmed

$$\alpha = (|\eta_1|, |\eta_2|, |\eta_3|, |\eta_4|).$$

Resonemangen sammanfattas i sats 2:

Sats 2

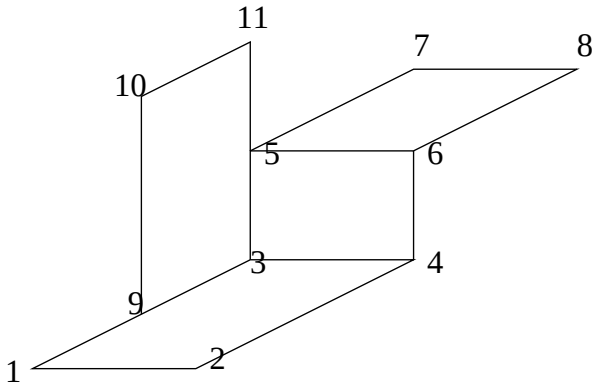
$$3D \to 2D \; P : X \longrightarrow Y, \text{ med djupet } \alpha \Longleftrightarrow \alpha s(X) \subset s(Y)$$

$$2D \to 2D \; P : X \longrightarrow Y, \text{ med djupet } \alpha \Longleftrightarrow \alpha s(X) = s(Y)$$

Detta är hjälpmedlet för att avgöra om en 2D figur är omöjlig eller ej.

5.4 Sammansatta konfigurationer

De klassiska omöjliga objekten är uppbyggda av polyedrar. Konfigurationen X nedan används för att illustrera principen för att bestämma formen $s(X)$, av sådana polyedrar.



Bilda en matris C^X vars kolonner tillhör $s(X)$. Eftersom X är uppbygd av delkonfigurationer så kopplas kolonnerna i C^X till delkonfigurationerna. Raderna i matrisen motsvarar en numrerad punkt i den totala bilden. Kolonnerna beskriver hur punkter och delfigurer hänger samman. Nollorna representerar de koordinater som inte ingår i respektive konfiguration. Låt tex det första parallellogrammet med punkterna 1, 2, 3 och 4 motsvaras av den första kolonnen. Fortsätter man på samma sätt erhålles den så kallade kompositmatrisen för X:

$$C^X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 1 & 1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & -2 & -1 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix} \tag{4}$$

De fyra första kolonnerna i matrisen beskriver alltså delfigurerna, dvs parallellogrammen. Kolonn 5 och 6 representerar de linjer som är parallella i hela figuren och kolonn 7 säger att sträckorna 1-3 och 5-7 är parallella. Genom beräkning (tex med matlab) visas att denna matris har rang $C^X = 7$. Detta stämmer med sats 1 (3) som säger att i detta fall är dim $s(X) = 11 - 4 = 7$, ty antalet hörnpunkter i exemplet är 11. [8]

6 Analys av omöjliga figurer

Låt enligt avsnitt 5.3 $P : X \longrightarrow Y$ vara en perspektivavbildning med djupvektorn α. För sammansatta konfigurationer som i avsnitt 5.4, gäller då att

$$\text{diag } (\alpha)C^X = C^Y \text{diag } (c), \tag{5}$$

där c är den vektor som måste finnas pga den obekanta proportionalitetskonstanten t som finns i varje ξ-vektor. Detta kallas *konsistensekvationen*. Om C^X och C^Y är kända så ger konsistensekvationen ett linjärt ekvationssystem i α och c. Detta ekvationssystem ger djupet för punkterna som ingår i X.

Enligt (5) gäller att rang C^X = rang C^Y. Definitionen av shape ger att rang $C^X \leq$ dim $s(X)$, rang $C^Y \leq$ dim $s(X)$. Om nu rang $C^Y = n - 3$ så följer att rang $C^X = n - 3$ och att dim $s(X) = n - 3$. Från sats 1 (3) följer att X är en 2D-konfiguration. Detta visar att rang C^Y avgör om en figur är omöjlig eller ej.

6.1 Reutersvärdsindex

Med beteckningar enligt ovan görs följande definition.

Definition. Med Reutersvärdsindex för Y menas talet

$$\rho = n - \text{rang } C^Y \tag{6}$$

För sammansatta figurer som i exemplet ovan, kan ρ också uttryckas som

$$\rho = n - f - l - p \tag{7}$$

där n är antalet hörnpunkter och f är det maximala antalet oberoende sidytor. l är det maximala antalet oberoende lineära incidenser och p är det maximala

antalet oberoende plana incidenser. Detta följer av att $f + l + p = \text{rang } C^X$.

En punkt och en rät linje (eller ett plan) sägs vara incidenta, om punkten ligger på linjen (i planet); en rät linje och ett plan sägs vara incidenta om linjen ligger i planet. Definitionen leder fram till två möjliga formuleringar av Huvudsatsen, se [8].

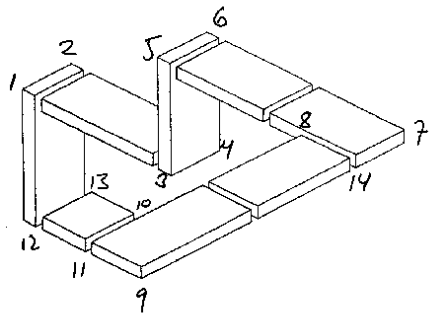
7 Huvudsatsen

$$Y \text{är en omöjlig figur} \iff \text{rang } C^Y = n - 3. \tag{8}$$

$$Y \text{är en omöjlig figur} \iff \text{Reutersvärdindex } = 3 \tag{9}$$

8 Reutersvärds trappa

I detta exempel beräknas ρ för Reutersvärds trappa:



(10)

Denna figur har kompositmatrisen

$$C^X = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 & -2 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & -2 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} \tag{11}$$

Det finns naturligtvis fler matriser som beskriver samma figur. Detta är dock inte intressant för beräkning av Reutersvärdsindex, där hänsyn endast tas till *antalet* villkor som krävs för att göra figuren entydig. Om rangen för matrisen beräknas i en dator spelar det därför ingen roll om man tar med för många villkor. Kolonnerna blir då linjärkombinationer av varandra och påverkar alltså inte matrisens rang.

Kort beskrivning av matrisens innehåll:
De första 6 kolonnerna beskriver precis som i föregående matris delfigurerna. Kolonn 7 säger att punkterna 5, 8 och 14 ligger på en linje. Kolonn 8 och 9 beskriver analogt att 8, 10 och 11 resp 9, 11 och 12 ligger på en linje. Kolonn 10 säger att punkterna 1, 3, 9 och 12 utgör ett parallellogram. Kolonn 11 uttrycker att punkterna 5, 9, 12 och 14 utgör en parallelltrapets.

Med utgångspunkt från detta blir i exemplet $n = 14$, $f = 6$, $l = 3$, $p = 2$.
Enligt definitionen av Reutersvärdsindexet (7) blir

$$\rho = 14 - 6 - 3 - 2 = 3.$$

Enligt huvudsatsen är alltså Reutersvärds trappa en omöjlig figur.
Den slutsats som hjärnan redan dragit är nu också matematiskt bevisad.

9 Perceptionspsykologi

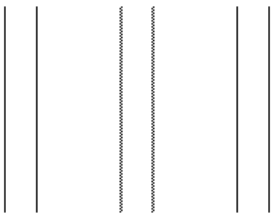
När ögat betraktar linjer i ett plan försöker det genast översätta dessa till rumsliga objekt. Information om avstånd och vinklar behandlas i hjärnan som försöker framställa en tredimensionell bild av linjerna i planet. Detta för att den är inställd på att uppfatta plana figurer som objekt i rummet. Det sker genom att hjärnan utför beräkningar, som i vissa fall kan ta lång tid. När det handlar om omöjliga figurer kommer hjärnan till slutsatsen att detta föremål inte kan existera rumsligt.

Hjärnan får alltså två separata signaler som motsäger varandra, en som säger att det man ser är ett föremål och en som säger att detta föremål inte kan existera.

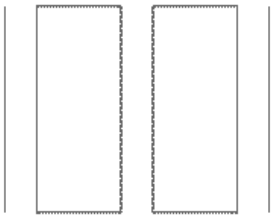
Trots detta så fortsätter ögat att uppfatta bilden som ett föremål, ögat har alltså skapat en rumslig absurditet. Det är detta som gör omöjliga figurer så fascinerande; ögat formar ett objekt som aldrig går att konstruera.

9.1 Gestaltspsykologi

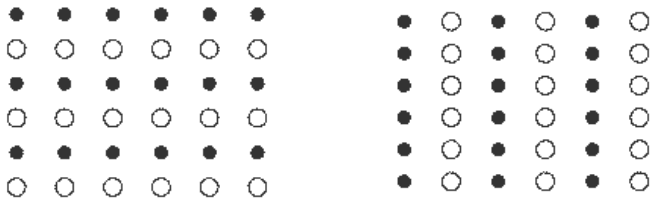
Gestaltpsykologi är den del av psykologin som beskriver hur människor försöker förstå och beskriva saker som helheter. Detta innebär att hjärnan försöker referera till sådant den känner igen. Gestaltpsykologin kan sammanfattas som allmängiltiga lagar. Dessa är det naturliga sättet för hjärnan att uppfatta gestalter, därför är de samma för alla människor och ingen behöver lära sig dem. Den viktigaste lagen är **Närhetslagen** som säger att delar med litet avstånd kopplas samman.



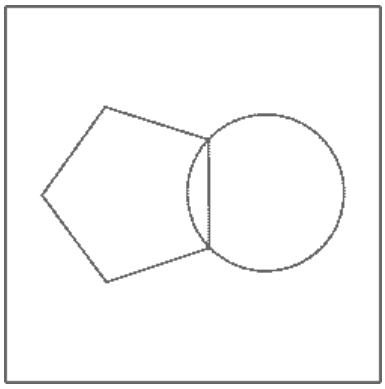
Nästa lag är **Slutenhets- eller konturlagen** där hjärnan uppfattar slutna linjer som ytor. Denna lag verkar starkare än närhetslagen vilket illustreras tydligt av bilden nedan. Linjernas avstånd är detsamma, men nu uppfattar hjärnan inte längre att linjerna med kortast avstånd hänger samman.



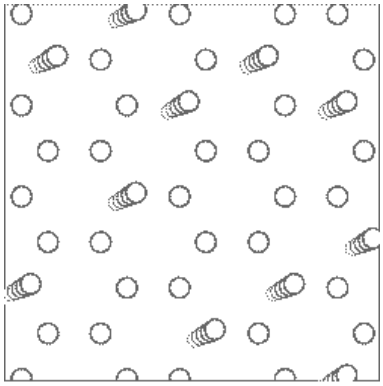
Likhetslagen sammanfogar de delar i ett system som liknar varandra. Därför uppfattas figurerna som "randiga". Denna lag är jämnstark med närhetslagen vilket gör att de konkurrerar.



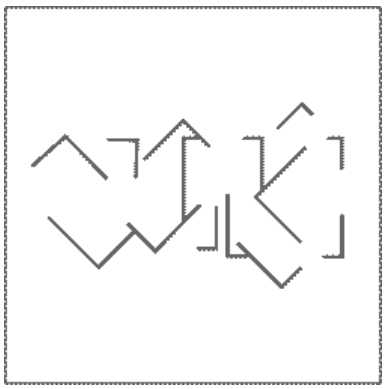
Härefter följer **Den goda kurvans lag** som säger att när välkända figurer sätts samman motverkas närhetslagen. Alltså vill hjärnan hellre uppfatta figurer den känner igen.



Den gemensamma rörelsens lag upphäver de tidigare lagarna och innebär att hjärnan först ser till de rörliga elementen i ett dynamiskt system. I denna figur uppfattas först cirklarna i rörelse och sedan mönstret av större cirklar som bildas av de små cirklarna.



Den sista lagen är **Erfarenhetens lag** som liknar den goda kurvans lag, dvs hjärnan kopplar till tidigare erfarenheter. I exemplet nedan uppfattas bokstaven L därför att vår erfaren säger det. Liksom den goda kurvans lag kan erfarenhetens lag motverka de tidigare.



9.1.1 Gestaltpsykologisk syn på omöjliga figurer

När man försöker applicera gestaltpsykologi på omöjliga figurer inser man att alla delar kan förklaras till en upplevd helhet, men helheten är den illusion som bilden verkar genom. Gestaltpsykologin kommer alltså inte åt helheten, helt enkelt för att figuren är omöjlig, se [10].

Kanske hade det varit möjligt att beskriva varför hjärnan kan uppfatta omöjliga figurer med hjälp av denna psykologiska inriktning om den fortsatt att utvecklas. Tyvärr har ingen större forskning bedrivits inom detta område sedan 30-talet och de lagar som presenterats utgör i stort sett hela inriktningen. De som forskade inom gestaltspsykologi under nittonhundratalets första årtionden var till största del judar. När andra världskriget började splittrades grupperna och forskningen dog ut. För närmare information, se [11].

9.2 Hur uppfattas illusioner?

Illusioner är egentligen inte omöjliga i bemärkelsen att de inte går att konstruera. Motsättningen består här istället i att hjärnan inte kan bestämma sig för vad den ska uppfatta.

Ett vanligt exempel på en illusion är Necker's kub, som kan uppfattas både med konvex och konkav orientering. Konvex innebär att kuben ser ut att gå ut från pappersplanet och konkav är motsatsen, alltså in i planet.

De flesta människor uppfattar konvexiteten som dominant, men detta kan variera från person till person. Konkret innebär detta att hjärnans viloläge ställer in sig på den komplexa figuren, medan en ansträngning krävs för att övergå till den konkava. Dominansen upphävs om man kompletterar med detaljer och skuggor. Om man ritar "inredning", tex ett fönster eller en lampa i kuben blir istället konkaviteten dominant. Ritar man istället prickar, som en tärning, blir det i princip omöjligt för hjärnan att uppfatta kuben som konkav, jämför [4].

9.3 En omöjlig beräkning

En intressant parentes är att det faktiskt går att beräkna volymen på en omöjlig figur, tex en trebalk.

Till att börja med delas varje balk upp i kuber. Om varje balk består av fem kuber blir det sammanlagda antalet 12 stycken. Antag nu att varje kub har sidan 1 dm, alltså att varje kub har volymen 1 dm^3. Detta ger att den totala volymen är 12 dm^3. Även arean kan beräknas på liknande sätt och blir i detta fall 48 dm^2.

10 Hur uppfattar hjärnan omöjliga figurer?

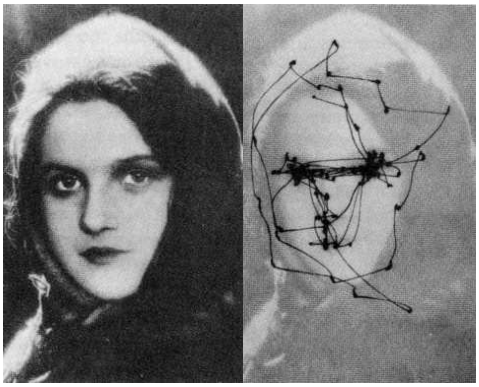
Människans synfält omfattar ca 210°, men det är bara i 2-3° av dessa som man ser riktigt skarp. Detta område motsvaras ungefär storleken av en nagel på armlängds avstånd. Trots detta uppfattar de flesta att de ser skarpt i större delen av sitt synfält. Anledningen till detta är att ögat hela tiden flyttar fokus för att skapa en helhet. Förflyttningen sker ca 4-5 ggr per sekund vilket är tillräckligt snabbt för att man ska uppfatta att man ser skarpt i ett mycket större område än de 2-3 graderna.

Det finns mycket noggrant beskrivet hur varje bild behandlas av hjärnan, men forskarna har inte ännu kunnat visa hur sekvenserna kopplas ihop till en helhetsbild.

En annan faktor som påverkar hur hjärnan uppfattar bilder är att människans "bildnärminne" inte är alls så bra som man kan tro. När ögat går från en fokusering till en annan förloras genast mycket information om den förra. Den information som den aktuella fokuseringen ger dominerar över de tidigare sekvenserna. I vardagen spelar detta ingen större roll, därför att de olika sekvenserna oftast inte motsäger varandra. Däremot när man betraktar en omöjlig figur är detta avgörande för att man ska kunna uppfatta den som ett rumsligt objekt.

Den ryska vetenskapsamannen Yarbus studerade på 1960-talet hur blicken vandrar över en bild, var den stannar och fokuserar. Hans studier visar att man aldrig fokuserar på hela bilden, utan bara på de intressanta punkterna, dvs de punkter där det händer något. Fixeringen pågår vanligen mellan två och åtta tiondelar av en sekund och ögonrörelsen mellan fixeringarna tar från en till åtta

hundradelars sekunder. Människan tittar alltså normalt på en bild genom ett stort antal fixeringar i snabb följd. Dessa fixeringspunkter kan vara hörn, delar med skarp färgkontrast eller skuggor. Betraktar man ett mänskligt ansikte är de intressantaste punkerna tex ögonen och munnen, se [11].



(12)

Diskussion

Då det aldrig forskats kring hur hjärnan uppfattar omöjliga figurer är det svårt att ge en exakt bild av hur detta går till. Med hjälp av de aspekter som framlagts ges här ett försök till en förklaring. Med hjäp av den givna bakgrunden kan man nu försöka förstå varför man trots motsättningarna kan uppfatta omöjliga figurer som faktiska föremål.

För det första analyserar hjärnan bara en liten del av bilden i taget. De delar som blicken koncentreras till är här oftast hörnpunkter i figuren.

Ser man på delfigurerna i en omöjlig figur var för sig är de fullt möjliga. När man sedan förflyttar fokus till nästa del har inte hjärnan kvar tillräckligt mycket information för att avgöra om de sista bitarna kan hänga ihop. Därför kan hjärnan utan problem framställa en tredimensionell figur.

Betrakar man tex Reutersvärds trappa (10) utgörs dess fixpunkter av varje trappsteg. Följer man trappan med blicken så uppfattar man alla detaljer som korrekta, och eftersom man har hunnit tappa information under tiden uppstår ingen motsättning mellan punkterna. Det är först när man börjar analysera och jämföra olika delar med varandra som man inser att figuren faktiskt är omöjlig.

Referenser

[1] *http : //en.wikipedia.org/wiki/M.C.Escher*

[2] Nationalencyklopedin *1994, "Reutersvärd, Oscar".*

[3] *http : //en.wikipedia.org/wiki/Roger_Penrose*

[4] Bruno Ernst *Äventyr med omöjliga figurer.* New Interlitho SpA, Italien 1986.

[5] Del av Oscar Reutersvärds brev till Gunnar Sparr, Lund 1999-04-02.

[6] Gunnar Sparr *On the "reconstruction" of impossible objects.* Proc. SSAB 1992.

[7] Gunnar Sparr *Depth computations from polyhedral images.* Image and vision computing. vol10 no 10, 1992.

[8] Gunnar Sparr och Oscar Reutersvärd *Omöjliga figurer. Japanska perspektiv inom konst, matematik och datorseende.* Föredragsmanuskript, Lunds Matematiska Sällskap, mars 2000.

[9] Gunnar Sparr *Linjär algebra.* Studentlitteratur, Lund, 1994.

[10] *www.all2know.com/sv/wikipedia/g/ge/gestaltpsykologi.html*

[11] Christian Balkenius, inst för kognitionsforskning, Lunds universitet, muntlig källa.

Matematik
Matematikcentrum
Lunds universitet
Box 118, 221 00 Lund

http://www.maths.lth.se/