Stockholms matematiska cirkel

Matematik och AI

Siobhán Correnty
Linus Bergkvist

Institutionen för matematik, KTH och
Matematiska institutionen, Stockholms universitet
2021–2022

Stockholms matematiska cirkel genom tiderna
(tidigare KTH:s matematiska cirkel)

2021-2022	Matematik och AI
2020-2021	Musik och matematik
2019-2020	Datorernas matematik
2018-2019	Grafteori med inriktning på färgläggning
2017-2018	Geometriska konstruktioner
2016-2017	Vad är ett tal?
2015-2016	Fraktaler
2014-2015	Polytoper
2013-2014	Grupper, mönster och symmetrier
2012-2013	Den matematiska analysens grunder
2011-2012	Diofantiska ekvationer
2010-2011	Polynom
2009-2010	Hyperbolisk geometri
2008-2009	Talteori
2007-2008	Sannolikhetsteori
2006-2007	Gruppteori
2005-2006	Vad är ett tal?
2004-2005	Integraler
2003-2004	Linjär algebra och bioinformatik
2002-2003	Algebra och kryptografi
2001-2002	Analysens grunder
2000-2001	Talföljder, rekursioner och iterationer
1999-2000	Linjära avbildningar

Innehåll

Lista över grekiska alfabetet **v**

Några ord på vägen **vi**

1 Matematik **1**

1.1 Definition, axiom, sats och bevis 1

1.2 Mängder 2

1.3 Funktioner 6

1.4 Bevistekniker 8

1.5 Olika satser och hur man bevisar dem 11

2 Euklidiska rum **17**

2.1 Rummet $\mathbb{R}^n$ 17

2.2 Några viktiga olikheter 20

2.3 Delmängder av $\mathbb{R}^n$ 23

3 Optimering **27**

3.1 Optimering för funktioner av en variabel 27

3.2 Partiella derivator 29

3.3 Differentierbarhet 32

3.4 Gradient 33

3.5 Riktningsderivata 34

3.6 Optimering för funktioner av flera variabler 37

4 Sannolikhetsteorins grunder **39**

4.1 Händelser och utfallsrum 39

4.2 Sannolikheter på utfallsrum 40

4.3 Betingning och oberoende 42

4.4 Slumpvariabler 45

5 Mått på slumpvariabler och normalfördelningar **49**
5.1 Diskreta slumpvariabler och sannolikhetsfunktioner 49
5.2 Fördelningsfunktioner 50
5.3 Väntevärde 53
5.4 Varians och standardavvikelse 54
5.5 Normalfördelning 55

6 Maskininlärning **58**
6.1 Vad är maskininlärning? 58
6.2 Vad är djupinlärning? 58
6.3 Ett djupt neuralt nätverk 59
6.3.1 Att beräkna det andra lagret 59
6.3.2 Mer om noder, vikter och bias 61
6.3.3 Andra aktiveringsfunktioner 64
6.3.4 Förlustfunktionen 65
6.3.5 Att träna modellen 69

7 Att approximera en funktion från datapunkter **73**
7.1 Modellen och träningen 74
7.2 Att evaluera ett neuralt nätverk 75
7.3 Andra viktiga begrepp inom djupinlärning 76

8 Att använda Keras och TensorFlow **79**
8.1 En guidad implementering 79
8.2 Resultat 81
8.3 Sammanfattning 82

Lösningar till udda övningsuppgifter **85**

Referenser och förslag till vidare läsning **98**

Lista över grekiska alfabetet

A	α	alfa
B	β	beta
Γ	γ	gamma
Δ	δ	delta
E	ε	epsilon
Z	ζ	zeta
H	η	eta
Θ	θ	theta
I	ι	iota
K	κ	kappa
Λ	λ	lambda
M	μ	my
N	ν	ny
Ξ	ξ	xi
O	o	omikron
Π	π	pi
P	ρ	rho
Σ	σ	sigma
T	τ	tau
Υ	υ	ypsilon
Φ	ϕ	fi
X	χ	chi
Ψ	ψ	psi
Ω	ω	omega

Några ord på vägen

Detta kompendium är skrivet för att användas som kurslitteratur till Stockholms matematiska cirkel under läsåret 2021–2022 och består av åtta kapitel.

Kompendiet är inte tänkt att läsas enbart på egen hand, utan ska ses som ett skriftligt komplement till undervisningen. Alla elever rekommenderas att läsa igenom varje kapitel själv innan föreläsningen. Det är inte nödvändigt att förstå alla detaljer vid den första genomläsningen.

Som de flesta matematiska skrifter på högre nivå är kompendiet kompakt skrivet. Detta innebär att man i allmänhet inte kan läsa det som en vanlig bok. Istället bör man pröva nya satser och definitioner genom att på egen hand exemplifiera. Därmed uppnår man oftast en mycket bättre förståelse av vad dessa satser och deras bevis går ut på.

Till varje kapitel finns ett antal övningsuppgifter. De udda övningarna har lösningar längst bak i kompendiet. Syftet med dessa är att eleverna ska kunna lösa dem och på egen hand kontrollera att de förstått materialet. Övningar med jämna nummer saknar facit och kan användas som examination. Det rekommenderas dock att man försöker lösa dessa uppgifter även om man inte examineras på dem.

Om man kör fast kan man alltid fråga en kompis, en lärare på sin skola eller någon av författarna. Under årets gång kommer det att finnas övningstillfällen där eleverna kan jobba med uppgifterna, själva eller i grupp, och få hjälp av oss.

De övningsuppgifter som är något svårare markeras med en stjärna ($\star$). Uppgifter som är extra utmanande markeras med två stjärnor ($\star\star$).

Vissa övningar kan ha flera lösningar och det som står i facit bör i detta fall endast ses som ett förslag.

Målet med årets kurs är att introducera den övergripande idén om hur neurala nätverk fungerar. Vi gör detta genom att först studera de grundläggande principerna i matematik i Kapitel 1. I Kapitel 2 studerar vi Euklidiska rum så att vi bättre förstår de underliggande mängderna i matematisk analys. Kapitel 3 går igenom grunderna i optimering, vilket kommer att användas vid träning av neurala nätverk, och Kapitel 4 och 5 behandlar sannolikhetsteori, vilket är ett viktigt ämne inom maskininlärning eftersom vi arbetar med slumpmässighet och vill kunna förutse nya utfall. I Kapitel 6 studerar vi ett neuralt nätverk för att klassificera handskrivna siffror i syfte att förstå alla delar av det neurala nätverket. Slutligen så bygger vi i Kapitel 7 och 8 ett neuralt nätverk för att approximera en kvadratisk funktion med hjälp av alla verktyg vi har lärt oss under kursen.

Några ord om cirkeln

Stockholms matematiska cirkel är en kurs för matematikintresserade gymnasieelever, som arrangeras av Kungliga Tekniska högskolan och Stockholms universitet. Cirkeln startade 1999. Vid starten hade den namnet KTH:s matematiska cirkel och hölls i KTH:s ensamma regi. Ambitionen med cirkeln är att sprida kunskap om matematiken och dess användningsområden utöver vad eleverna får genom gymnasiekurser, och att etablera ett närmare samarbete mellan gymnasieskolan och högskolan. Cirkeln ska särskilt stimulera elevernas matematikintresse och inspirera dem till fortsatta naturvetenskapliga och matematiska studier.

Till varje kurs skrivs ett kompendium som distribueras gratis till eleverna. Detta material, föreläsningsschema och övrig information om Stockholms matematiska cirkel finns tillgängligt på

`www.math-stockholm.se/cirkel`

Cirkeln godkänns ofta som en gymnasiekurs eller som matematisk breddning på gymnasieskolorna. Det är upp till varje skola att godkänna cirkeln som en kurs och det är lärarna från varje skola som sätter betyg på kursen. Lärarna är självklart också välkomna till cirkeln och många har kommit överens med sin egen skola om att få cirkeln godkänd som fortbildning eller som undervisning.

Vi vill gärna understryka att föreläsningarna är öppna för alla gymnasieelever, lärare eller andra matematikintresserade.

Vi har avsiktligt valt materialet för att ge eleverna en inblick i matematisk teori och tankesätt och presenterar därför både några huvudsatser inom varje område och bevisen för dessa resultat. Vi har också som målsättning att bevisa alla satser som används om de inte kan förutsättas bekanta av elever från gymnasiet.

Författarna, sommaren 2021

1 Matematik

Temat för årets matematiska cirkel är maskininlärning och AI.

Kursen består mestadels av teori, men i slutet av kursen kommer vi fokusera på hur man kan tillämpa teorin i praktiken. Detta kommer involvera programmering i Python.

Detta kapital är en introduktion i den matematiska metoden och ett antal grundbegrepp som vi kommer använda oss av i kursen.

1.1 Definition, axiom, sats och bevis

I detta avsnitt ska vi beskriva den matematiska metoden utifrån fyra begrepp: *definition*, *sats*, *bevis* och *axiom*.

En *definition* bestämmer vad en term betyder så att man kan arbeta matematiskt med den. Till exempel kan vi definiera udda och jämna tal på följande sätt.

Definition 1.1.1. Ett heltal n är *udda* om det finns ett heltal k som uppfyller att $n = 2k + 1$. △

Definition 1.1.2. Ett heltal n är *jämnt* om det finns ett heltal k som uppfyller att $n = 2k$. △

Ofta har man en intuition om vad en term betyder redan innan man definierar den. Läsaren hade till exempel säkert en uppfattning om vad udda och jämna tal är innan vi definierade. Syftet med en definition är att precisera detta.

När definitionen är gjord, så överger man sina tidigare uppfattningar om vad termen betyder och utgår endast ifrån definitionen. Man säger att definitionen är *stipulativ*. En definition är alltså inte rätt eller fel, utan bara mer eller mindre användbar och intuitiv.

Definitioner bygger ofta på begrepp som läsaren är bekant med. Till exempel utgår Definition 1.1.1 och 1.1.2 från att läsaren redan vet vad ett heltal är.

En *sats* är ett påstående som bevisats vara sant. Varje sats hör samman med ett *bevis*: ett argument för att påståendet är sant.

Sats 1.1.3. *Om n är udda, så är $n + 1$ jämnt.*

Bevis. Om n är udda så finns det ett heltal k så att $n = 2k + 1$. Då gäller att

$$n + 1 = 2k + 1 + 1 = 2k + 2 = 2(k + 1).$$

Eftersom $k + 1$ är ett heltal, så är $n + 1$ ett jämnt tal. □

Bevisen kombinerar definitioner och olika logiska slutledningsregler för att nå den önskade slutsatsen. Sats 1.1.3 har en syskonsats. Beviset är mer eller mindre identiskt, och lämnas som övning.

Sats 1.1.4. *Om n är jämnt, så är $n + 1$ udda.*

En sats vars främsta syfte är att användas i beviset av en annan sats kallas för en *hjälpsats* eller ett *lemma*. En sats som följer omedelbart ur en annan sats, till exempel som ett specialfall, kallas för en *följdsats* eller ett *korollarium*.

Ett påstående måste vara bevisat för att få kallas för en sats. Om man har goda skäl att tro att ett påstående är sant men inte formellt bevisat det kallas påståendet för en *förmodan*, eller *hypotes*. Två exempel är *Riemannhypotesen* och *primtalstvillingsförmodan*.

En förmodan kan förbli obevisad i hundratals år. Ett berömt exempel är *Fermats sista sats*, som formulerade av Pierre de Fermat (1607–1665) år 1637 men bevisades först av Andrew Wiles år 1995. Riemannhypotesen, som ännu är obevisad, formulerades 1859 av Bernard Riemann (1826–1866.)

Eftersom bevisen utgår ifrån definitionen, och inte vår intuition, så behöver man ibland bevisa saker som känns uppenbara. Läsaren vet till exempel att

(i) alla tal antingen är udda eller jämna, och

(ii) ett tal kan inte vara udda och jämnt samtidigt.

Men om man läser Definition 1.1.1 och 1.1.2 så är detta inte självklart. Kan man inte tänka sig tal som varken är udda eller jämnt? Eller tal som är båda?

Bevis bygger på antaganden. Dessa antaganden måste dock bevisas innan de kan anses giltiga. Men dessa bevis måste också bygga på antaganden, som också måste bevisas, och så vidare.

För att undvika en oändlig kedja av bevis, eller ett cirkulärt bevis (ett bevis som använder sig av det man försöker bevisa) så måste man göra grundantaganden som inte behöver bevisa. Dessa kallas för *axiom*.

1.2 Mängder

En *mängd* är en samling objekt. Man kan samla nästan vad man vill i en mängd: tal, katter, och andra mängder.[1] Det viktiga är att man alltid kan avgöra ifall ett objekt tillhör mängden eller inte. De objekt som ligger i mängden kallas för *element*.

Det lättaste sättet att beskriva en mängd är att räkna upp element som ingår i den. För att markera att objekten ligger i en mängd, så omger man listan med *mängdklamrar* { och }. Mängden som innehåller 1, 2 och 3 skrivs alltså som

$$\{1, 2, 3\}.$$

Två mängder är lika om de innehåller samma element, vilket skrivs $A = B$. Det spelar ingen roll i vilken ordning som man skriver elementen eller hur många gånger de listas. Därför gäller att

$$\{1, 1, 2, 3\} = \{1, 2, 3\} = \{2, 3, 1\}.$$

[1] Vi skriver *nästan* av en anledning. Det finns samlingar av objekt som kan beskrivas men som inte utgör en mängd. Detta kallas *Russells paradox*, efter Bertrand Russell (1872-1970). Russells exempel är samlingen av alla mängder som inte innehåller sig själva.

Om ett element x tillhör en mängd A kan man skriva $x \in A$, vilket uttalas som x tillhör A. Om x inte tillhör A skriver man $x \notin A$. Mängden som inte innehåller några element alls kallas den *tomma mängden*, och betecknas med $\emptyset$.

En mängd kan innehålla andra mängder som element. Mängden

$$A = \{\{1, 2\}, 3\}$$

har två element: mängden $\{1, 2\}$ och talet 3. Mängden $\{1, 2\}$ innehåller i sin tur elementen 1 och 2. Däremot innehåller A varken 1 eller 2.

Att mängder kan innehålla andra mängder kan ha paradoxala konsekvenser. Till exempel kan vi lägga den tomma mängden i en mängd, och bilda mängden av den tomma mängden.

$$A = \{\emptyset\} = \{\{\}\}$$

Mängden A innehåller ett element, den tomma mängden, och är därför inte tom. Mängden av den tomma mängden är alltså inte lika med den tomma mängden.

Detta verkar motsägelsefullt. Den tomma mängden är ju tom, så mängden av den tomma mängden borde ju också vara tom? Tricket är att skilja på mängden och elementen i mängden. Den tomma mängden är ju ett element i sig, även om den inte innehåller några element, precis som att 0 är ett tal, trots att representerar ett antal som inte finns.

Det finns ingen begränsning på hur stor en mängd kan vara, och de flesta mängder man studerar innehåller oändligt många element. Dessa mängder kan naturligtvis inte skrivas ut som en lista. Istället beskriver man dem med *mängdbyggaren*, som har följande allmänna form.

$$\{x \mid \text{villkor på } x\}$$

Den här mängden består av alla element som uppfyller villkoret. Ett exempel är mängden

$$\{n \mid n \text{ är jämnt}\} = \{\ldots, -4, -2, 0, 2, 4, \ldots\}$$

som innehåller alla jämna tal.

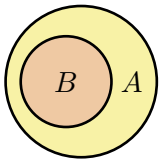
En mängd B är en *delmängd* av en mängd A om alla element i B är ett element i A. Man skriver detta som $B \subset A$. Till exempel så är $\{1, 2\}$ en delmängd av $\{1, 2, 3\}$, eftersom 1 och 2 är element i båda mängderna. Om två mängder är delmängder av varandra så är de lika.

En mängd har alltid minst två delmängder: sig själv och den tomma mängden. En delmängd B av A är *äkta* om B varken är den tomma mängden eller A.

Det är lätt att blanda ihop element och delmängder. Det beror på att mängder kan innehålla andra mängder, så att en delmängd av en mängd kan vara ett element i mängden. Mängden $A = \{\emptyset\}$ är ett bra exempel. Den tomma mängden är både ett element i och en delmängd av A.

I mängden $A = \{1, 2, \{1, 2\}\}$ är $\{1, 2\}$ både en delmängd och ett element. Däremot så är $\{1\}$ enbart en delmängd av A, medan 1 enbart är ett element.

För att illustrera mängder använder man *Venndiagram*, efter matematikern John Venn (1834–1923). Där representeras mängder som enkla former, oftast cirklar, och formernas förhållanden till varandra motsvarar mängdernas. Till exempel kan man illustrera att B är en delmängd av A genom att rita dem som två cirklar, där B ligger inuti A.



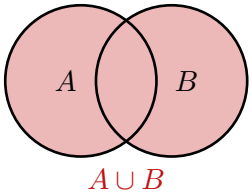
Figur 1.1: Venndiagram för $B \subset A$.

Ur gamla mängder kan vi skapa nya genom de så kallade *mängdoperationerna*.

(i) *Unionen* av mängderna A och B är mängden som består av alla element som ligger i A *eller* i B. Den betecknas med $A \cup B$ och definieras som

$$A \cup B = \{x \mid x \in A \text{ eller } x \in B\}.$$

Ett exempel är $\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}$.

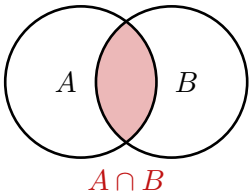


Figur 1.2: Venndiagram för $A \cup B$.

(ii) *Snittet* av mängderna A och B är mängden som består av alla element som ligger i A *och* i B. Den betecknas med $A \cap B$ och definieras som

$$A \cap B = \{x \mid x \in A \text{ och } x \in B\}.$$

Ett exempel är $\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}$.

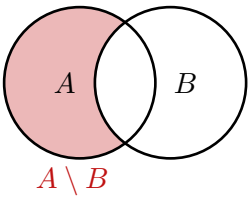


Figur 1.3: Venndiagram för $A \cap B$.

(iii) *Differensen* av en mängd A och B är mängden som består av alla element som ligger i A men inte i B. Den betecknas med $A \setminus B$ och definieras som

$$A \setminus B = \{x \mid x \in A \text{ och } x \notin B\}.$$

Ett exempel är $\{1, 2, 3\} \setminus \{2, 3, 4\} = \{1\}$.



Figur 1.4: Venndiagram för $A \setminus B$.

Notera att $A \setminus B$ inte är lika med $B \setminus A$, exempelvis gäller $\{2, 3, 4\} \setminus \{1, 2, 3\} = \{4\}$.

Två mängder A och B är *disjunkta* om de inte har några gemensamma element, det vill säga om $A \cap B = \emptyset$.

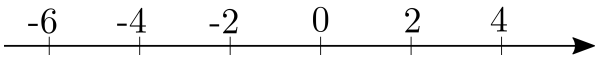
Om alla uppträdande mängder är delmängder av en viss mer eller mindre underförstådd grundmängd M talar man ofta om $M \setminus A$ som *komplementet* till A (med avseende på M). Vi kommer att beteckna komplementet till A med A^c.

Om vi till exempel pratar om mängder av heltal, och vi till exempel betraktar mängden $A = \{1, 2, 3\}$, så avser komplementet till A mängden av alla heltal *förutom* 1, 2 och 3.

De olika talsystemen kan ses som mängder av tal, och har fått egna beteckningar. De *naturliga talen* betecknas med $\mathbb{N}$ och består av talen 0, 1, 2, 3, och så vidare.[2]

Naturliga tal kan adderas och multipliceras utan problem. Resultatet är alltid ett nytt naturligt tal. För att subtrahera behöver vi införa de negativa talen -1, -2, och så vidare. De naturliga talen tillsammans med de negativa talen kallas för *heltalen*, och betecknas med $\mathbb{Z}$ (av tyskans *Zahl* = tal).

Heltal kan adderas, subtraheras och multipliceras. Man kan däremot inte dividera dem med varandra. För detta krävs *rationella tal*. De definieras som alla kvoter a/b, där a och b är heltal och b är skilt från 0. Mängden av alla rationella tal betecknas med $\mathbb{Q}$.

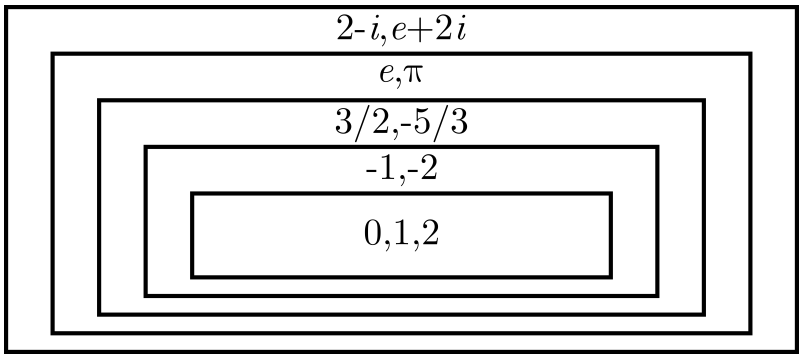


Figur 1.5: Tallinjen runt 0.

De rationella talen ligger på den så kallade *tallinjen*, som går från negativa tal till vänster och till positiva tal till höger (se Figur 1.5). Det finns dock tal som inte är rationella, men som ändå ligger på tallinjen. Ett exempel är $\sqrt{2}$, som är längden på diagonalen i en kvadrat med sidan 1. Lägger man till dessa tal får

[2]Vissa exkluderar 0 från de naturliga talen. Att inkludera 0 har dock fördelar. Om man börjar räkna från 0 och går ett steg i taget kommer man ha gått n steg när man räknat till n. Exempel: om vi räknar till 3 från 0 så får vi $0 \to 1 \to 2 \to 3$, vilket är 3 steg. Om vi börjar från 1 får vi istället $1 \to 2 \to 3$, vilket är 2 steg.

de *reella talen*, som betecknas med $\mathbb{R}$.[3] Reella tal som inte är rationella kallas för *irrationella*.



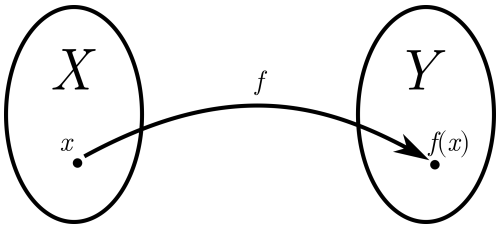
Figur 1.6: De olika talsystemen från $\mathbb{N}$ till $\mathbb{C}$.

De reella talen kan utvidgas ytterligare till de *komplexa talen*, som betecknas med $\mathbb{C}$, genom att lägga till ett tal i som uppfyller $i^2 = -1$.

1.3 Funktioner

En funktion $f : X \to Y$ parar ihop element i en mängd X med element i en mängd Y. Mängden X kallas för *definitionsmängd* och mängden Y kallas för *målmängd*. Man kan se f som en process som tar ett element i mängden X och avger ett element som ligger i mängden Y. När man tillämpar en funktion på ett element x i X så kallas x för funktionens *argument*.

Två funktioner är lika när de har samma definitionsmängd, samma målmängd och de är lika på alla element i definitionsmängden. Definitions- och målmängden är alltså en del av funktionen.



Figur 1.7: En funktion f från X till Y.

Funktioner beskrivs ofta med formler. Exempelvis så kan funktionen $f : \mathbb{N} \to \mathbb{N}$ som tar ett naturligt tal och returnerar dess kvadrat beskrivas som $f(n) = n^2$. Alla polynom kan ses som en funktion från $\mathbb{R}$ till $\mathbb{R}$, som beräknas genom att man sätter in talet x i uttrycket.

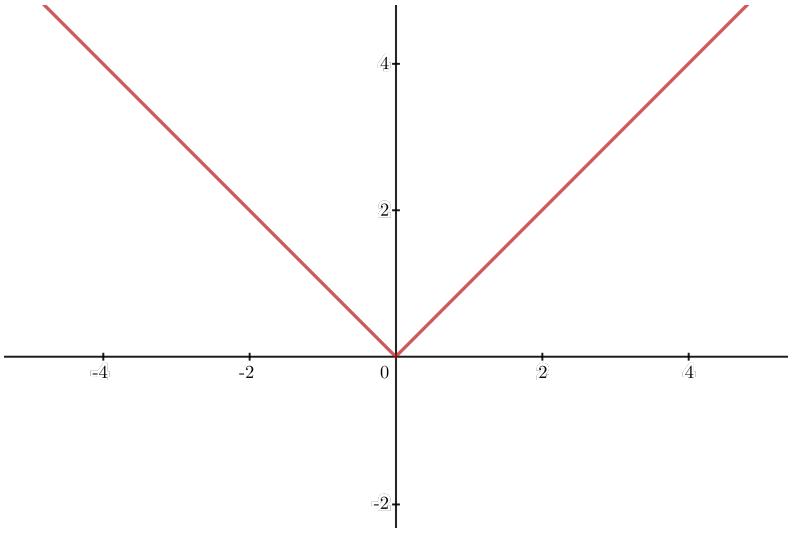
[3] Reella tal är mycket mystiska. Den matematiska cirkeln 2016-2017, *Vad är ett tal?*, handlade om hur man kan definiera dem i termer av rationella tal. Den intresserade läsaren uppmanas att söka upp kompendiet på Cirkelns hemsida: `https://www.math-stockholm.se/samverkan/cirkel/`

En funktion måste dock inte ges av en formel. Det enda som krävs är att funktionen är definierad för alla element i X, och att den ger alltid samma svar. Ett exempel är absolutbeloppet $|x|$ av ett reellt tal x, som definieras som avståndet från x till origo på tallinjen. Man kan beräkna det genom att man tar bort eventuella minustecken framför talet, det vill säga

$$|x| = \begin{cases} x & \text{om } x \geqslant 0 \\ -x & \text{om } x < 0. \end{cases}$$

Exempelvis så gäller $|-3| = -(-3) = 3$ och $|2| = 2$.

En funktion $f : \mathbb{R} \to \mathbb{R}$ kan beskrivas genom sin *graf*, som definieras som mängden av punkter i planet på formen $(x, f(x))$.



Figur 1.8: Grafen av funktionen $f(x) = |x|$.

Om x är ett reellt tal så är $\lfloor x \rfloor$ det största heltalet som är mindre än eller lika med x. Till exempel så är $\lfloor 5/2 \rfloor = 2$ och $\lfloor -\pi \rfloor = -4$. Man kan se det som att $\lfloor x \rfloor$ är avrundningen av x, förutsatt att man alltid avrundar neråt.

Definition 1.3.1. *Golvfunktionen* är funktionen som avbildar reella tal x på $\lfloor x \rfloor$. △

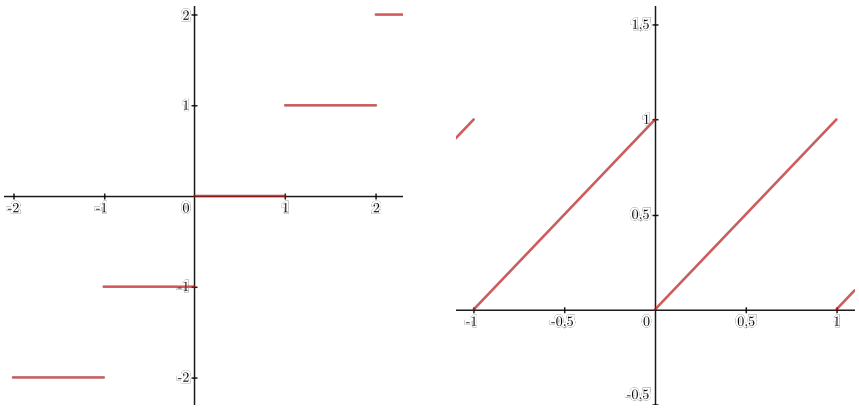
Ett annat sätt att se golvfunktionen är att den avbildar x på det största heltalet n som uppfyller olikheterna

$$n \leqslant x < n + 1.$$

Definition 1.3.2. *Fraktionsfunktionen* $\text{frac} : \mathbb{R} \to \mathbb{R}$ är funktionen som avbildar reella tal x på $x - \lfloor x \rfloor$. △

Värdet $\text{frac}(x)$ ligger alltid mellan 0 och 1, och $\lfloor x \rfloor + \text{frac}(x) = x$. Det senare följer direkt av definitionen, eftersom

$$\lfloor x \rfloor + \text{frac}(x) = \lfloor x \rfloor + x - \lfloor x \rfloor = x.$$



(a) Graf av $\lfloor x \rfloor$ **(b)** Graf av $\text{frac}(x)$

Figur 1.9: Golv- och fraktionsfunktionen

Talet $\text{frac}(x)$ kallas för *fraktionsdelen* av x. Observera att $\text{frac}(x) = 0$ när x är ett heltal. Två andra användbara samband är att $\text{frac}(x+n) = \text{frac}(x)$ och $\lfloor x+n \rfloor = \lfloor x \rfloor + n$ för alla heltal n.

1.4 Bevistekniker

Ett bevis för en sats är ett argument som förklarar varför satsen är sann. Vi har redan sett ett exempel när vi bevisade Sats 1.1.3. I detta avsnitt ska vi gå igenom tre tekniker för att bevisa matematiska satser: direkta bevis, motsägelsebevis och induktionsbevis.[4]

Ett *direkt bevis* utgår ifrån satsens antaganden och definitioner och bevisar satsen rakt på, så att säga. Beviset av Sats 1.1.3 är ett exempel på direkt bevis. Ett annat är följande sats.

Sats 1.4.1. *Antalet funktioner från en mängd A med n element till en mängd B med m element är m^n.*

Bevis. Varje funktion från A till B kan beskrivas som en tabell där varje element i A motsvaras av precis ett element i B. Listan innehåller totalt n platser, och på varje plats kan vi välja bland m element att välja bland. Alltså finns det totalt

$$\underbrace{m \cdot m \cdots m \cdot m}_{n \text{ stycken}} = m^n$$

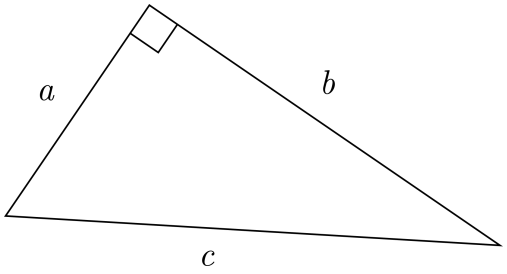
olika funktioner. □

Ibland går det inte att använda direkta bevis, till exempel när man ska bevisa att något inte är fallet. Då kan det vara enklare att anta att det man vill bevisa är falskt, och visa att detta leder till en motsägelse. Om alla steg i beviset är korrekta så måste det ursprungliga antagandet vara fel. Detta kallas för ett *motsägelsebevis*.

[4] Ibland förekommer termen *indirekt bevis*. Vissa använder det som synonym till motsägelsebevis, andra som en synonym till bevisregeln *modus tollens*. Vi undviker den helt.

Sats 1.4.2. *Längden av hypotenusan i en rätvinklig triangel är alltid mindre än summan av längderna av kateterna.*

Bevis. Låt a och b vara längderna av kateterna i triangeln och c längden av hypotenusan.



Figur 1.10: En rätvinklig triangel.

Enligt Pytagoras sats är $a^2+b^2=c^2$. Antag motsatsen till satsen, det vill säga att $a+b \leqslant c$. Då gäller att

$$(a+b)^2 \leqslant c^2 \implies a^2+2ab+b^2 \leqslant c^2 \implies 2ab \leqslant 0.$$

Eftersom a och b är längderna i en rätvinklig triangel så är a och b både större än 0. Eftersom produkter av positiva tal är positiva, så är detta en motsägelse. Alltså måste motsatsen till $a+b \leqslant c$ gälla, det vill säga $a+b>c$. □

Sats 1.4.3. *Ett tal kan inte vara udda och jämnt samtidigt.*

Bevis. Antag att n är ett tal som är båda udda och jämnt. Då finns det två tal, k och l, så att $n=2k$ och $n=2l+1$. Då gäller att

$$2k=n=2l+1 \implies 2k-2l=1 \implies 2(k-l)=1.$$

Med andra ord finns det heltal $m=k-l$ så att $2m=1$. Kan det finns ett sådant tal? Det finns två fall.

(i) Om $m \leqslant 0$, så är $1=2m \leqslant 0$. Motsägelse!

(ii) Om $m \geqslant 1$ så är $2m \geqslant 2>1$. Motsägelse! □

Ett berömt motsägelsebevis är följande.

Sats 1.4.4. *Talet $\sqrt{2}$ är irrationellt.*

Bevis. Antag motsatsen, det vill säga att $\sqrt{2}=a/b$ för några heltal a och b. Antag att a och b är förkortade så långt som möjligt. Då kan endast en av a eller b vara jämn, eftersom om båda är jämna kan vi skriva

$$\sqrt{2}=\frac{a}{b}=\frac{2c}{2d}=\frac{c}{d}$$

och då var inte a och b förkortade så långt som möjligt.

Av definitionen av $\sqrt{2}$ får vi att

$$\sqrt{2}^2 = 2 = \frac{a^2}{b^2} \implies 2b^2 = a^2.$$

Den sista ekvationen säger att a^2 är jämn. Eftersom kvadrater av udda tal är udda (se Övning 1.14), så måste a vara ett jämnt tal, det vill säga $a = 2k$ för något heltal k. Då får vi att

$$2b^2 = (2k)^2 = 4k^2 \implies b^2 = 2k^2.$$

Eftersom b^2 är jämnt, så måste b vara jämnt. Men nu har vi bevisat både a och b är jämna, vilket var omöjligt eftersom vi hade förkortat bråket så långt som möjligt. Detta är en motsägelse. □

Bevisen av Sats 1.4.3 och 1.4.2 bygger båda på ett antagande som vi rent formellt inte bevisat (kan du se vilket?). Att bevis inkluderar sådana antaganden är snarare regel än undantag. Ifall man bevisade precis vartenda antagande utifrån axiomen skulle bevisen bli väldigt långa och komplicerade. Läsaren förväntas själv fylla i de luckor som uppstår.

Det händer dock att uppenbara antaganden är mycket svåra, till och med omöjliga, att bevisa utifrån definitionerna. Historien är fylld av matematiker som gjort till synes självklara antaganden som sedan visat sig vara svåra att bevisa.

Beviset av Sats 1.4.4 är ett exempel på det. Vi antar att ett bråk kan förkortas så långt som möjligt. Detta är inte självklart, utan bygger i själva verket på aritmetikens fundamentalsats, en sats som vi återkommer till senare i kursen.

Den tredje bevistekniken som vi kommer gå igenom kallas för *induktionsbevis*. Säg att du har en följd av påståenden P_0, P_1, P_2 och så vidare, ett för varje naturligt tal. För att bevisa att alla dessa påståenden gäller, så kan man göra det i två steg:

(i) **Basfall**: Bevisa att P_0 är sann.

(ii) **Induktionssteget:** Bevisa att om P_n är sant för något naturligt tal n så är P_{n+1} sant.

Idén är att om båda dessa gäller, så är P_m sant för alla naturliga tal m. Man kan motivera det på följande sätt. Säg att vi undrar ifall P_m är sant. Enligt basfallet är P_0 sant, och genom att tillämpa induktionssteget med $n = 0$ så kan vi dra slutsatsen att P_1 är sant. Vi kan nu sätta $n = 1$, och induktionssteget säger då att P_2 är sant. Genom att upprepa detta m gånger, kan vi bevisa att P_m är sant. Vi kan illustrera det hela som en följd av implikationer:

$$P_0 \implies P_1 \implies P_2 \implies \cdots \implies P_{m-1} \implies P_m \implies \cdots.$$

Ett exempel förtydligar hur det fungerar.

Sats 1.4.5. *Alla naturliga tal är udda eller jämna.*

Bevis. Låt P_n vara påståendet att n är antingen udda eller jämnt. Då är påståendet att alla naturliga tal är udda eller jämna samma som att P_n är sant för alla n.

(i) **Basfall**: När $n = 0$ säger satsen att 0 är udda eller jämnt, vilket är sant eftersom $0 = 2 \cdot 0$ är ett jämnt tal.

(ii) **Induktionssteg**: Antag att P_m är sant för något m, det vill säga m är antingen udda eller jämnt. Vi ska bevisa att $m + 1$ är udda eller jämnt. Det gör vi i två fall.

- Om m är udda så är $m + 1$ jämnt enligt Sats 1.1.3.
- Om m är jämnt så är $m + 1$ udda enligt Sats 1.1.4.

Detta bevisar att $m + 1$ antingen är udda eller jämnt, det vill säga att P_{m+1} är sant. □

Induktionsbevis reducerar satser som handlar om alla tal till satser som handlar om specifika tal. Dessa kan vara enklare att bevisa, eftersom du istället får ett konkret tal att arbeta med.

Induktionsbevis kan vara klurigt i början. Ett organiserat sätt att göra dem är följande sätt.

(i) Omformulera satsen så att den blir av typen: P_n är sann för alla n.

(ii) Bevisa P_0.

(iii) Anta att det finns ett tal n så att P_n är sant, och bevisa att P_{n+1} är sant.

Ett induktionsbevis bygger på att fallet $n + 1$ kan beskrivas i termer av fallet n. Ifall man inte kan hitta en sådan reduktion, kommer ett induktionsbevis vara svårt att använda.

1.5 Olika satser och hur man bevisar dem

I föregående avsnitt diskuterade vi olika bevistekniker. Men vilka tekniker är lämpliga för vilka typer av satser?

- **Implikation:** Man säger att P implicerar Q om Q är sant när P är det. Ett exempel är Sats 1.1.4, som säger att om ett tal n är jämnt, så är talet $n + 1$ udda. Man brukar beteckna implikationer med en tjock pil $\Longrightarrow$, så att P medför Q skrivs

$$P \implies Q.$$

En implikation kan bevisas med ett direkt bevis. Då antar man att P är sant, och sedan visar man att Q också måste vara sant (det är så vi bevisar Sats 1.1.4). Man kan också använda ett motsägelsebevis. Då antar man att P är sann och att Q är falsk, och bevisar en motsägelse.

Ett tredje sätt att bevisa att P implicerar Q är att bevisa att om Q är falsk, så är P falsk. Detta kallas för *omvändningen* av en implikation.

- **Ekvivalens:** En ekvivalens är när två påståenden P och Q implicerar varandra, alltså att om P så Q, och om Q så P. Man brukar använda frasen P om och endast om Q. Man använder tjocka dubbelpilar för att beteckna ekvivalenser, så att P om och endast om Q skrivs som

$$P \iff Q.$$

  Ekvivalenser bevisas genom att första visa att P implicerar Q, och sedan att Q implicerar P.

- **Universalsats:** En universalsats säger att alla n i en mängd M uppfyller något villkor P. Universalsatser kan bevisas som implikationer, genom att omformulera universalsatsen som att om n ligger i mängden M, så uppfyller n villkoret P, det vill säga

$$n \in M \implies n \text{ uppfyller } P$$

  Ifall M är mängden av alla naturliga tal kan man också använda ett induktionsbevis, som vi beskrev ovan.

  Man kan även bevisa en universalsats genom ett motsägelsebevis. Då antar man att det finns ett n i M som inte uppfyller P, och bevisar att det är omöjligt.

- **Existenssats:** En existenssats säger att finns ett objekt n som har egenskapen P. Den typiska existenssatsen är ekvationslösning. Att $x^2 = 3$ har en lösning är en existenssats, och kan omformuleras som att det finns ett tal x så att $x^2 = 3$.

  Ett sätt att bevisa en existenssats är att konstruera det sökta objektet utifrån objekt man redan vet finns. Till exempel så kan man bevisa att det finns ett udda kvadrattal genom att notera att $3^2 = 9$ är udda och ett kvadrattal.

  Man kan också använda ett motsägelsebevis. Då antar man att det inte existerar någon objekt med egenskapen P och visar att det leder till en motsägelse. Dessa bevis har fördelen att vi inte behöver beskriva hur objektet konstrueras. I gengäld kan bevisen vara mycket komplicerade.

En variant på universalsatsen är att inget n i M uppfyller P. Den kan omformuleras som att alla n i M saknar egenskapen P. För dessa typer av satser är motsägelsebevis ofta smidiga: man antar att det finns ett n i M som uppfyller P och härleder en motsägelse.

Universal- och existenssatser är duala till varandra, i bemärkelsen att om du ska bevisa en existenssats med hjälp av ett motsägelsebevis så antar du en universalsats, och vice versa, se bevisen av Sats 1.4.2, 1.4.3 och 1.4.4

Övningar

Övning 1.1. Lista elementen i följande mängder.

(i) $A = \{n \in \mathbb{N} \mid k < 5\}$.

(ii) $B = \{1, 2, \{2, 3\}\}$.

(iii) $C = \{k \in \mathbb{Z} \mid k^2 < 16\}$.

(iv) $A \cap B$.

(v) $(C \setminus A) \cup B$.

Övning 1.2. Lista elementen i följande mängder.

(i) $A = \{x \in \mathbb{Q} \mid x^2 = 2\}$.

(ii) $B = \{0, 1, 2, 3\}$.

(iii) $C = \{p/q \mid 0 \leqslant p < 3 \text{ och } 1 \leqslant q < 3\}$.

(iv) $(B \cup A) \cap C$.

(v) $(C \cap B) \setminus A$.

Övning 1.3. För nedanstående par av mängder A och B, avgör om A och B är lika, disjunkta, någon av dem är en äkta delmängd av den andra eller ingetdera.

(i) $A = \{1, 2, 3\}$ och $B = \{1, 1, 2\}$.

(ii) $A = \{0, 1, 2\}$ och $B = \{n \in \mathbb{N} \mid n^2 < 9\}$.

(iii) $A = \{\{\}\}$ och $B = \{\{x \in \mathbb{N} \mid 2x = -2\}\}$.

(iv) $A = \{x \in \mathbb{R} \mid |x| < 1\}$ och $B = \{x \in \mathbb{R} \mid |x - 1| < 1\}$.

(v) $A = \{x \in \mathbb{Q} \mid x^2 = 2\}$ och $B = \{x \in \mathbb{R} \mid x^2 = 2\}$.

Övning 1.4. För nedanstående par av mängder A och B, avgör om A och B är lika, disjunkta eller någon av dem är en äkta delmängd av den andra.

(i) $A = \{-2, 0, 2\}$ och $B = \{x \in \mathbb{Z} \mid |x| < 3 \text{ och } x \text{ är jämnt}\}$.

(ii) $A = \{x \in \mathbb{R} \mid x^2 < 2\}$ och $B = \{x \in \mathbb{Q} \mid x^2 \geqslant 2\}$.

(iii) $A = \{x \in \mathbb{Z} \mid x \text{ är jämnt}\}$ och $B = \{x \in \mathbb{Z} \mid x \text{ är kvadrattal}\}$.

(iv) $A = \{x \in \mathbb{Z} \mid 2x = -2\}$ och $B = \{x \in \mathbb{N} \mid 2x = 2\}$.

(v) $A = \{\emptyset, \{\emptyset\}\}$ och $B = \{\emptyset\}$.

Övning 1.5. Använd mängdbyggaren för att definiera följande mängder.

(i) Mängden av jämna, positiva heltal.

(ii) Mängden av rationella tal r så att $2r$ är ett heltal.

(iii) Mängden av irrationella tal som ligger inom avstånd 1 från origo.

Övning 1.6. Använd mängdbyggaren för att definiera följande mängder.

(i) Mängden av alla kvadrattal som är större än 2.

(ii) Mängden av rationella lösningar till $x^4 + x^2 - 1 = 0$.

(iii) Mängden av rationella tal som är volymen av en kub med rationella sidor.

Övning 1.7. Ange möjlig definitionsmängd och målmängd för följande funktioner.

(i) Funktionen som ger det n:te kvadrattalet.

(ii) Funktionen som beräknar arean av triangel.

(iii) Funktionen beräknar derivatan av ett andragradspolynom.

Övning 1.8. Ange möjlig definitionsmängd och målmängd för följande funktioner.

(i) Funktionen som ger arean av cirkel med radie r.

(ii) Funktionen som ger avståndet mellan 1 och ett tal r på tallinjen.

(iii) Funktionen som ger de rationella nollställena till ett förstagradspolynom med rationella koefficienter.

Övning 1.9. Är följande funktioner eller inte? Om inte, motivera varför.

(i) $f : \mathbb{R} \to \mathbb{R}$ där
$$f(x) = \begin{cases} 1 \text{ om } x \in \mathbb{Q} \\ 0 \text{ om } x \notin \mathbb{Q}. \end{cases}$$

(ii) $f : \mathbb{N} \to \mathbb{Q}$ där $f(n) = \sqrt{n}$.

(iii) $f : \mathbb{R} \to \mathbb{R}$ så att $f(x) = 0$ med sannolikhet $1/2$ och $f(x) = 1$ med sannolikhet $1/2$.

(iv) $f : \{0\} \to \mathbb{R}$ där $f(0) = 1$ om ordet Balkong börjar på B.

Övning 1.10. Är följande funktioner eller inte? Om inte, motivera varför.

(i) $f : \mathbb{Z} \to \mathbb{N}$, där $f(n)$ är siffersumman i det vanliga (decimala) talsystemet.

(ii) $f : \mathbb{R} \to \mathbb{Q}$, där $f(x) = x/2$.

(iii) $f : \mathbb{N} \to \mathbb{R}$, där $f(n) = \sqrt[n+1]{n+1}$.

(iv) $f : \mathbb{Q} \to \mathbb{Z}$, där $f(p/q) = p$.

Övning 1.11. Avgör om följande funktioner är lika eller inte? Motivera varför.

(i) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$ och $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \sqrt{x^2}$.

(ii) $f : \mathbb{Z} \to \mathbb{Q}$, $f(n) = 1/n$ och $g : \mathbb{N} \to \mathbb{Q}$, $g(m) = 1/m$.

(iii) $f : \mathbb{N} \to \mathbb{Q}$, $f(n) = n/(n+1)$ och $g : \mathbb{N} \to \mathbb{R}$, $g(z) = z/(z+1)$.

Övning 1.12. Avgör om följande funktioner är lika eller inte? Om inte, motivera varför.

(i) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 + 2x + 1$ och $g : \mathbb{R} \to \mathbb{R}$, $g(x) = (x+1)^2$.

(ii) $f : \mathbb{Z} \to \mathbb{Z}$, $f(x) = x^2$ och $g : \mathbb{Z} \to \mathbb{Z}$, $g(x) = |x|^2$.

(iii) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \mathrm{frac}(x)$ och $g : \mathbb{Q} \to \mathbb{R}$, $g(x) = x - \lfloor x \rfloor$.

Övning 1.13. Använd ett direkt bevis för att bevisa att om n är ett jämnt tal så är $n + 1$ ett udda tal (detta är Sats 1.1.4).

Övning 1.14. Använd ett direkt bevis för att bevisa att om n är udda så är n^2 udda.

Övning 1.15. Använd ett motsägelsebevis för att bevisa att summan av ett irrationellt tal och ett rationellt tal är irrationellt.

Övning 1.16. Använd ett motsägelsebevis för att bevisa att om $a + b \geqslant c$ så är antingen $a \geqslant c/2$ eller $b \geqslant c/2$

Övning 1.17. Använd induktion för att bevisa att summan av de n första udda naturliga talen är n^2.

Övning 1.18. Använd induktion för att bevisa att summan av de n första naturliga talen är $n(n+1)/2$.

Satsen har också ett direkt bevis. Försök att hitta det.

Övning 1.19. Bevisa att om $ab = c$ så är $a \geqslant \sqrt{c}$ och $b \leqslant \sqrt{c}$, eller tvärtom.

Övning 1.20 ($\star$)**.** Bevisa att vinkelsumman av en n-hörning är $180(n-2)$ grader

(i) med ett direkt bevis.

(ii) med induktion.

Övning 1.21. Ett sätt att definiera ordningen $\geqslant$ på naturliga tal är följande.

Ett naturligt tal n är *större eller lika med* ett naturligt tal m om det finns ett naturligt tal k så att $n = m + k$. Vi skriver detta som $n \geqslant m$.

Bevisa följande satser:

(i) $n \geqslant 0$ och $n \geqslant n$ för alla n.

(ii) Om $n \geqslant m$ och $m \geqslant k$ så gäller $n \geqslant k$.

(iii) Om $n \geqslant m$ och $m \geqslant n$, så är $n = m$.

(iv) Om n och m uppfyller $n \geqslant m$ så gäller $n + k \geqslant m + k$ för alla $k \geqslant 0$

(v) Om n och m uppfyller $n \geqslant m$ så gäller $nk \geqslant mk$ för alla $k \geqslant 0$.

Övning 1.22 ($\star$)**.** Bevisa att det finns två irrationella tal a och b så att a^b är rationellt. Tips: använd att $\sqrt{2}$ är irrationellt.

Övning 1.23. Bevisa att

$$1^2 + 2^2 + \cdots + n^2 < n^3$$

för alla $n \geqslant 2$.

Övning 1.24 ($\star$)**.** Bevisa att det finns alla rationella tal kan skrivas som en produkt av två irrationella tal.

Övning 1.25 ($\star\star$)**.** Ett *minsta element* i en delmängd S av $\mathbb{N}$ ett tal $n \in S$ med egenskapen att $n \leqslant m$ för alla $m \in S$. Bevisa att alla icke-tomma delmängder av $\mathbb{N}$ har ett minsta element.

Ledtråd: Omformulera satsen och kombinera ett induktionsbevis med motsägelsebevis.

Övning 1.26 ($\star$)**.** Följande påståenden är alla sanna. Vilka bör vara satser och vilka bör vara definitioner? Motivera hur du tänker.

(i) Det finns både udda och jämna kvadrattal.

(ii) Ett tal n är ett kvadrattal om det finns en kvadrat vars area är n.

(iii) Ett tal n är ett kvadrattal om det finns ett heltal k så att $n = k^2$.

(iv) Det n:te kvadrattalet är summan av de n första udda talen.

Övning 1.27 ($\star$)**.** Följande påståenden är alla sanna. Vilka bör vara satser och vilka bör vara definitioner? Motivera hur du tänker.

(i) En rätvinklig triangel har vinkelsumma på 180 grader.

(ii) En rätvinklig triangel har två vinklar vars summa är 90 grader.

(iii) En rätvinklig triangel bildas när man drar en diagonal i en kvadrat.

(iv) En triangel är rätvinklig om den har en rät vinkel.

2 Euklidiska rum

När man för första gången introduceras till funktioner i matematik pratar man nästan enbart om funktioner av en variabel. I praktiken är dock detta ofta ett enkelt specialfall, då de flesta relationer "i verkligheten" beror på mer än en sak. Till exempel beror hur mycket pengar ett stort företag omsätter under ett år på näst intill oändligt många variabler, vilket gör att man med nödvändighet måste analysera en förenklad modell för att kunna dra några slutsatser. Men även en förenklad modell lär bero på väldigt många variabler, som till exempel antalet sålda produkter, pris på produkterna, antalet anställda, momssatser, arbetsgivaravgifter, inkomstskatter etc. Även enklare formler från fysik innehåller ofta mer än en variabel. Till exempel ges den kinetiska energin för en punktformig kropp med massan m och hastigheten v av $mv^2/2$.

I den här kursen kommer vi behöva analysera funktioner som beror av flera variabler, som till exempel $f(x, y, z) = e^{xy}/z$ som då är en funktion av 3 variabler. För vilka värden på (x, y, z) är funktionen definierad? I vilken punkt antar $f(x, y, z)$ sitt största respektive minsta värde? Finns det ens något största respektive minsta värde? Vad är funktionens maximum, om det nu existerar, om vi antar att $1/2 \leqslant x \leqslant y \leqslant z \leqslant 1$. Om vi börjar i punkten $(x, y, z) = (1, 1, 1)$, åt vilket håll ska vi då röra oss för att funktionen ska växa så snabbt som möjligt?

För att kunna analysera funktioner av flera variabler ordentligt och besvara frågorna ovan måste vi till att börja med förstå mängderna på vilka de är definierade.

2.1 Rummet $\mathbb{R}^n$

I den här kursen, och i *flervariabelanalys* i allmänhet, betraktar vi huvudsakligen funktioner som tar en *reell n-tupel* som argument, det vill säga argumentet är på formen $(x_1, \ldots, x_n)$, där var och en av de n komponenterna x_j, $j = 1, 2, \ldots, n$ är ett reellt tal, det vill säga $x_j \in \mathbb{R}$.

Vi använder beteckningen $\mathbb{R}^n$ för att beteckna mängden

$$\{(x_1, \ldots, x_n) : x_j \in \mathbb{R}, \quad j = 1, \ldots n\}.$$

Normalt skriver vi dock bara $\mathbb{R}$ istället för $\mathbb{R}^1$.

Ibland kommer vi även använda fet stil för att beteckna punkter i $\mathbb{R}^n$. Då är det underförstått att den j:te komponenten i $\mathbf{x}$ är x_j, och att x_j då är ett reellt tal. Ibland skriver vi ändå ut $\mathbf{x} = (x_1, \ldots, x_n)$ för att vara extra tydliga.

Däremot är inte alla funktioner i n variabler definierade på hela $\mathbb{R}^n$. Till exempel är funktionen från vårt föregående exempel, $f(x, y, z) = e^{xy}/z$, inte definierad om $z = 0$. Precis som tidigare är *definitionsmängden*, som vi vanligtvis kommer beteckna med D_f, en viktig del av funktionen, och en funktions egenskaper beror i hög grad på såväl definitionsmängden som på själva regeln.

I nästa kapitel då vi mer noggrant undersöker funktioner av flera variabler och deras egenskaper kommer vi att definiera kontinuitet ordentligt, men följande

exempel är förhoppningsvis tydligt ändå, även om vi tills vidare nöjer oss med en intuitiv känlsa av vad kontinuitet innebär.

Exempel 2.1.1. Om vi till exempel betraktar funktionen

$$f(x,y) = \begin{cases} \frac{x}{y} \text{ om } y \neq 0 \\ 0 \text{ om } y = 0 \end{cases}$$

så ser vi att den är kontinuerlig för en mängd som inte innehåller någon punkt där $y = 0$, det vill säga någon punkt på formen $(x, 0)$. Till exempel är funktionen kontinuerlig på mängden

$$D_f = \{(x,y) : 1 < x < 2, 1 < y < 2\}.$$

Däremot är funktionen inte kontinuerlig i alla punkter på mängden

$$D_f = \{(x,y) : -2 < x < 2, -2 < y < 2\}.$$

Till exempel är ju

$$\lim_{y\to 0} f(1,y) = \lim_{y\to 0} \frac{1}{y} \neq 0 = f(1,0).$$

Huruvida en funktion är kontinuerlig eller ej beror alltså till viss del på definitionsmängden. ▲

Såväl $\mathbb{R}^n$ som delmängder till $\mathbb{R}^n$ är i grunden bara mängder, det vill säga samlingar med punkter. Men väldigt ofta väljer vi att lägga till ytterligare *struktur* på dessa mängder. På samma sätt som vi har matematiska operationer på reella tal — som att vi till exempel kan addera reella tal med varandra och få ett nytt reellt tal, och multiplicera reella tal med varandra och få ett nytt reellt tal — kan vi *definiera* operationer på n-tuplar i $\mathbb{R}^n$.

Till att börja med definierar vi tre operationer i $\mathbb{R}^n$, nämligen *addition*, *multiplikation* med ett reellt tal $\lambda \in \mathbb{R}$, samt *skalärmultiplikation*.

Om $\mathbf{x} = (x_1, \ldots, x_n)$ och $\mathbf{y} = (y_1, \ldots, y_n)$ definierar vi

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n).$$

$$\lambda\mathbf{x} = (\lambda x_1, \ldots, \lambda x_n).$$

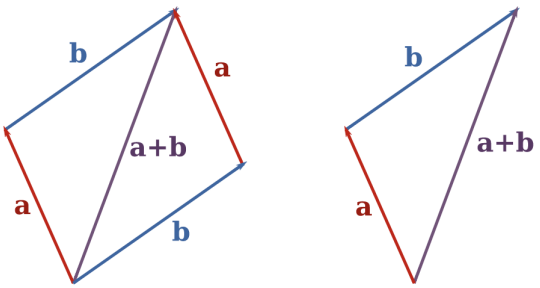
samt

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \ldots + x_n y_n.$$

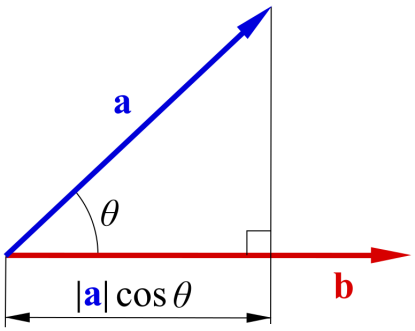
Notera att addition och multiplikation med ett reellt tal ger nya element i $\mathbb{R}^n$, medan skalärprodukten av två element i $\mathbb{R}^n$ blir ett tal i $\mathbb{R}$ (det är därför operationen kallas just *skalär*produkt). En bild som visar hur det ser ut när vi adderar 2 vektorer finns i Figur 2.1 och en bild som visar vad skalärprodukten representerar finns i Figur 2.2.

Det är värt att verifiera att vanliga formler för multiplikation och addition av reella tal också gäller med operationerna ovan. Till exempel är

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}, \quad \lambda(\mathbf{x} + \mathbf{y}) = \lambda\mathbf{x} + \lambda\mathbf{y}, \quad \mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}.$$



Figur 2.1: Addition av 2 vektorer, $a, b \in \mathbb{R}^2$, bild från Wikipedia



Figur 2.2: Skalärprodukten av 2 vektorer, $a, b \in \mathbb{R}^2$, bild från Wikipedia

I två och tre variabler kan punkter i $\mathbb{R}^n$ realiseras geometriskt som pilar i planet respektive rummet. Med denna geometriska tolkning kan addition av två element i $\mathbb{R}^n$ tolkas som att man helt enkelt klistrar fast den andra pilens början på den första pilens spets (se Figur 2.1), och multiplikation med ett positivt reellt tal skalar bara om längden på pilen. Till exempel kommer $\mathbf{x}$ och $2\mathbf{x}$ båda vara pilar som pekar åt samma håll, bara att $2\mathbf{x}$ är dubbelt så lång. Multiplicerar vi istället med ett negativt tal λ kommer vi att få en pil som pekar i motsats riktning, och vars längd har blivit omskalad proportionerligt med absolutbeloppet av λ.

Vi behåller dessutom terminologin från $\mathbb{R}^2$ och $\mathbb{R}^3$ och säger att två element $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ är *parallella* om det finns ett reellt tal λ så att

$$\lambda \mathbf{x} = \mathbf{y}.$$

Om $\lambda \geqslant 0$ säger vi dessutom att $\mathbf{x}$ och $\mathbf{y}$ har *samma riktning* eller är *lika riktade*.

Vanligtvis kommer vi bara skriva 0 för *nollvektorn* i $\mathbb{R}^n$, alltså $(0, \ldots, 0)$. Det bör alltid vara uppenbart från sammanhanget om 0 syftar på det reella talet 0 eller en $n-$tupel vars alla komponenter är nollor. Notera att med definitionen ovan är alla element i $\mathbb{R}^n$ parallella och lika riktade med 0. Vissa väljer dock att i definitionen kräva att två parallella vektorer båda ska vara nollskilda just för att undvika detta.

Skalärproduktens geometriska tolkning är inte lika uppenbar, och den kommer inte riktigt behövas i den här kursen. Men det kan vara värt att veta att

skalärprodukten är relaterad till vinkeln mellan två vektorer och deras längder enligt formeln $\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}||\mathbf{y}| \cos(\theta)$, där θ är vinkeln mellan vektorerna $\mathbf{x}$ och $\mathbf{y}$.

Ett viktigt specialfall från den geometriska tolkningen med pilar i två och tre variabler är att

$$\mathbf{x} \cdot \mathbf{x} = x_1^2 + x_2^2 (+x_3^2),$$

vilket vi känner igen från Pythagoras sats som kvadraten av längden av pilen.

Med detta som motivation definierar vi nu *längden* eller *beloppet* av ett element i $\mathbf{x} \in \mathbb{R}^n$ som

$$|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + \ldots + x_n^2}.$$

Notera att skalärprodukten av ett element $\mathbf{x}$ med sig själv alltid är större än eller lika med 0, med likhet om och endast om $\mathbf{x} = 0$. Däremot är skalärprodukten av två olika element inte nödvändigtvis positiv.

Exempel 2.1.2. Låt $\mathbf{x} = (1, 2, 3, 4)$ och $\mathbf{y} = (-1, -1, -1, -1)$. Båda dessa vektorer är uppenbarligen nollskilda, men de är inte parallella. Vidare är

$$|\mathbf{x}| = \sqrt{1^2 + 2^2 + 3^2 + 4^2} = \sqrt{1 + 4 + 9 + 16} = \sqrt{30},$$

och

$$|\mathbf{y}| = \sqrt{(-1)^2 + (-1)^2 + (-1)^2 + (-1)^2} = \sqrt{4} = 2.$$

Slutligen är

$$\mathbf{x} \cdot \mathbf{y} = -1 - 2 - 3 - 4 = -10.$$

▲

Slutligen definierar vi avståndet mellan två punkter $\mathbf{x}$ och $\mathbf{y}$ som

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

Notera att detta överensstämmer med hur vi mäter avstånd i $\mathbb{R}^2$ och $\mathbb{R}^3$ med hjälp av Pythagoras sats.

När vi tänker på element i $\mathbb{R}^n$ utrustade med *strukturen* ovan — det vill säga att vi kan mäta deras längd, vi kan addera dem, vi kan multiplicera dem med reella tal, och vi kan ta skalärprodukter — så säger vi att dessa element är *Euklidiska vektorer*, eller bara *vektorer*. En vektor är alltså inte bara en punkt i rummet, utan den har mer struktur, som till exempel en längd och en riktning.

2.2 Några viktiga olikheter

En oerhört viktig olikhet, som vi bland annat kommer att behöva för att kunna avgöra i vilken riktining en funktion växer snabbast, är *Cauchy–Schwarz' olikhet*.

Sats 2.2.1. *(Cauchy–Schwarz' olikhet)*

I $\mathbb{R}^n$ *gäller*

$$|\mathbf{x} \cdot \mathbf{y}| \leqslant |\mathbf{x}||\mathbf{y}|, \tag{2.1}$$

eller utskrivet

$$|x_1y_1 + \ldots + x_ny_n| \leqslant \sqrt{x_1^2 + \ldots + x_n^2}\sqrt{y_1^2 + \ldots + y_n^2}.$$

Likhet inträffar om och endast om $\mathbf{x}$ *och* $\mathbf{y}$ *är parallella.*

Bevis. Vi kan anta att $\mathbf{x} \neq 0$, för om $\mathbf{x} = 0$ är båda leden i (2.1) lika med 0, och $\mathbf{x}$ och $\mathbf{y}$ är dessutom parallella.

Den viktiga insikten som beviset bygger på är det enkla faktum att skalärprodukten av *vilken vektor som helst* med sig själv är större än eller lika med 0.

Låt t vara ett reellt tal och betrakta vektorn $t\mathbf{x} + \mathbf{y}$. Enligt räknereglerna för skalärprodukten är

$$0 \leqslant (t\mathbf{x} + \mathbf{y}) \cdot (t\mathbf{x} + \mathbf{y}) = t^2|\mathbf{x}|^2 + 2t\mathbf{x} \cdot \mathbf{y} + |\mathbf{y}|^2.$$

Eftersom $|\mathbf{x}|^2 \neq 0$ kan vi bryta ut $|\mathbf{x}|^2$ och kvadratkomplettera med avseende på t, vilket ger

$$0 \leqslant |\mathbf{x}|^2 \left(t^2 + 2\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2}t + \frac{|\mathbf{y}|^2}{|\mathbf{x}|^2} \right) = |\mathbf{x}|^2 \left(t + \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2} \right)^2 + |\mathbf{y}|^2 - \frac{(\mathbf{x} \cdot \mathbf{y})^2}{|\mathbf{x}|^2}.$$

Notera att denna olikhet håller för *alla* reella värden på t. Om vi nu väljer ett lämpligt värde på t, nämligen

$$t = -\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2},$$

så försvinner den första termen i uttrycket ovan och vi får då

$$0 \leqslant |\mathbf{y}|^2 - \frac{(\mathbf{x} \cdot \mathbf{y})^2}{|\mathbf{x}|^2},$$

Vilket kan skrivas om till (2.1).

Från beviset framgår det även att vi har likhet i (2.1) precis då skalärprodukten av $(t\mathbf{x} + \mathbf{y})$ med sig själv är noll. Men det händer bara om $(t\mathbf{x} + \mathbf{y}) = 0$, vilket innebär att

$$\mathbf{y} = -t\mathbf{x}.$$

Så $\mathbf{x}$ och $\mathbf{y}$ måste alltså vara parallella om vi har likhet. □

En annan väldigt viktig olikhet i $\mathbb{R}^n$ är följande generalisering av Sats (1.4.2) som säger att längden av hypotenusan i en rätvinklig triangel är kortare än summan av kateternas längder.

Sats 2.2.2. *(Triangelolikheten i* $\mathbb{R}^n$*)*

I $\mathbb{R}^n$ *gäller*

$$|\mathbf{x} + \mathbf{y}| \leqslant |\mathbf{x}| + |\mathbf{y}|,$$

med likhet precis när $\mathbf{x}$ *och* $\mathbf{y}$ *är parallella och lika riktade.*

Bevis. Genom att använda räknereglerna för skalärprodukten ser vi att

$$|\mathbf{x}+\mathbf{y}|^2 = (\mathbf{x}+\mathbf{y})\cdot(\mathbf{x}+\mathbf{y}) = |\mathbf{x}|^2 + 2\mathbf{x}\cdot\mathbf{y} + |\mathbf{y}|^2. \tag{2.2}$$

Genom att använda Cauchy-Scwarz' olikhet ser vi att

$$\mathbf{x}\cdot\mathbf{y} \leqslant |\mathbf{x}\cdot\mathbf{y}| \leqslant |\mathbf{x}||\mathbf{y}|,$$

Insättning i (2.2) ger nu att

$$|\mathbf{x}+\mathbf{y}|^2 \leqslant |\mathbf{x}|^2 + 2|\mathbf{x}||\mathbf{y}| + |\mathbf{y}|^2 = (|\mathbf{x}|+|\mathbf{y}|)^2,$$

vilket är ekvivalent med den sökta olikheten.

Vidare har vi likhet i triangelolikheten om och endast om vi har likhet i Cauchy-Scwarz' olikhet och

$$\mathbf{x}\cdot\mathbf{y} = |\mathbf{x}\cdot\mathbf{y}|.$$

Vi vet att vi har likhet i Cauchy-Schwarz' olikhet exakt då $\mathbf{x}$ och $\mathbf{y}$ är parallella, det vill säga då $\mathbf{x} = \lambda\mathbf{y}$ för något $\lambda \in \mathbf{R}$.

Villkoret att

$$\mathbf{x}\cdot\mathbf{y} = |\mathbf{x}\cdot\mathbf{y}|$$

betyder alltså att

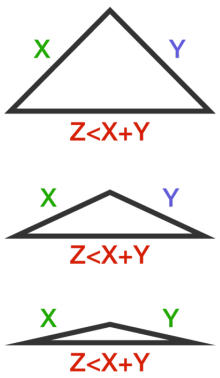
$$\mathbf{x}\cdot\mathbf{y} = \lambda\mathbf{y}\cdot\mathbf{y} = \lambda|\mathbf{y}|^2 = |\lambda||\mathbf{y}|^2 = |\lambda|\mathbf{y}|^2| = |\mathbf{x}\cdot\mathbf{y}|,$$

vilket håller precis då $\lambda = |\lambda|$, alltså när $\lambda \geqslant 0$.

Vi har alltså likhet i traingelolikheten precis då

$$\mathbf{x} = \lambda\mathbf{y}, \quad \lambda \geqslant 0,$$

det vill säga då $\mathbf{x}$ och $\mathbf{y}$ är parallella och lika riktade. □



Figur 2.3: Triangelolikheten för $x, y \in \mathbb{R}^2$, bild från `brilliant.org`

Figur 2.3 visar triangelolikheten med 2 vektorer x och y. Ett viktigt specialfall av triangelolikheten är den så kallade *omvända triangelolikheten*.

Eftersom

$$|\mathbf{x}| = |(\mathbf{x}+\mathbf{y})-\mathbf{y}| \leqslant |\mathbf{x}+\mathbf{y}| + |\mathbf{y}|$$

är

$$|\mathbf{x}| - |\mathbf{y}| \leqslant |\mathbf{x} + \mathbf{y}|.$$

På samma sätt visas att

$$|\mathbf{y}| - |\mathbf{x}| \leqslant |\mathbf{x} + \mathbf{y}|,$$

och dessa två olikheter ger tillsammans att

$$||\mathbf{x}| - |\mathbf{y}|| \leqslant |\mathbf{x} + \mathbf{y}|. \tag{2.3}$$

Olikheten (2.3) kallas ofta för den *omvända triangelolikheten*

Med hjälp av induktion kan dessutom Sats (2.2.2) generaliseras till fler än två termer:

$$|\mathbf{x_1} + \ldots + \mathbf{x_n}| \leqslant |\mathbf{x_1}| + \ldots + |\mathbf{x_n}|.$$

Vi lämnar beviset av denna generalisering som en övningsuppgift.

2.3 Delmängder av $\mathbb{R}^n$

I många fall är man intresserade av att analysera en *delmängd* av $\mathbb{R}^n$. Till exempel kan man vara intresserad av att hitta den kortaste vägen att gå för att ta sig över ett bergspass, och i det fallet är den naturliga mängden att utföra sina analyser på den tvådimensionella delmängden till $\mathbb{R}^3$ som beskriver bergytan av bergspasset.

I den här kursen kommer vi dock huvudsakligen vara intresserade av delmängder till $\mathbb{R}^n$ som är naturliga definitionsmängder för någon funktion vi är intresserade av.

Exempel 2.3.1. Betrakta funktionen

$$f(x, y) = \frac{1}{x(x-1)y(y-1)}.$$

Denna funktion är inte definierad för $x = 0, x = 1, y = 0$ och $y = 1$. Vi skulle till exempel kunna välja att som definitionsmängd ha hela $\mathbb{R}^2$, men att vi tar bort de fyra problemlinjerna $x = 0, x = 1, y = 0, y = 1$.

Den mängden är dock inte *sammanhängande* (notera att vi inte har definierat detta koncept ordentligt, så här får vi förlita oss på vår intuition om vad en sammanhängande mängd bör vara). Om vi vill ha en definitionsmängd som är sammanhängande skulle vi istället kunna betrakta mängden

$$D_f = \{(x, y) : 1 < x < 2, 1 < y < 2\}.$$

Notera att denna definitionsmängd inte är maximal på så sätt att vi kan *utvidga* definitionsmängden och få en strikt större definitionsmängd som också är sammanhängande. Till exempel kan vi betrakta

$$D_f = \{(x, y) : 1 < x < 3, 1 < y < 3\}.$$

Ytterligare en alternativ definitionsmängd är

$$D_f = \{(x, y) : 0 < x < 1, 0 < y < 1\}.$$

Denna definitionsmängd är dock maximal på så sätt att $f(x, y)$ inte är väldefinierad överallt på någon sammanhängande mängd som innehåller D_f. Problemet är att en sådan mängd med nödvändighet måste innehålla någon punkt vars $x-$ eler $y-$koordinat är antingen 0 eller 1.

Det är dock viktigt att vara medveten om att vi *inte* har bevisat påståendet ovan. Vi har ju inte ens definierat vad vi menar med sammanhängande mängd ordentligt! ▲

Några av de vanligaste mängderna man intresserar sig för är *sfärer* och *klot*. Definitionen av dessa motiveras av att en cirkel med radie r och centrum a i två variabler beskrivs som mängden av punkter vars avstånd till a är precis r, och motsvarande (öppna) disk består av de punkter vars avstånd till a är strikt mindre än r.

Definition 2.3.2. I $\mathbb{R}^n$ definieras ett öppet klot som en mängd som kan beskrivas som

$$\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{a}| < r\}$$

för något $r > 0$ och någon punkt $\mathbf{a} \in \mathbb{R}^n$. I så fall kallas r för klotets radie och $\mathbf{a}$ för klotets medelpunkt.

Vidare definieras en sfär som en mängd som kan beskrivas som

$$\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{a}| = r\}$$

för något $r > 0$ och någon punkt $\mathbf{a} \in \mathbb{R}^n$. På samma sätt som för sfären kallas i så fall r för sfärens radie och $\mathbf{a}$ för sfärens medelpunkt.

Vi kommer ibland använda notationen $B_r(\mathbf{a})$ och $S_r(\mathbf{a})$ för att beteckna det öppna klotet respektive sfären med radie r och centrum $\mathbf{a}$. △

Notera att en öppen sfär med radie $r > 0$ och centrum a i $\mathbb{R}^1$ bara blir det öppna intervallet $(a - r, a + r)$. Mer specifikt blir *det öppna enhetsklotet med centrum i origo* bara intervallet $(-1, 1)$.

Många viktiga definitioner i analys kräver att vi inte bara kan undersöka en funktion i en punkt, utan att vi kan uttala oss om funktionens beteende i ett litet område runt punkten. Till exempel kräver derivatans definition

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

att vi kan säga något om $f(x + h)$, i alla fall för *tillräckligt* små värden på h.

Om vi har en funktion $f(x)$ som bara är definierad på intervallet $[0, 1]$ så kan vi inte uttala oss om derivatan i punkten 1, för $f(1 + h)$ är inte definierat för *något* $h > 0$. I detta fall kan vi förvisso betrakta gränsvärdet från vänster ($h < 0$), men om vi har en funktion som är definierad på $[0, 1] \cup \{2\}$ finns det inget meningsfullt sätt att prata om en derivata i punkten 2.

Vi behöver alltså kunna prata om punkter i en mängd, som har egenskapen att ett litet område runt punkten också ligger i mängden.

Definition 2.3.3. Låt $M \subset \mathbb{R}^n$ och låt $\mathbf{a}$ vara i en punkt i $\mathbb{R}^n$. Vi säger att:

$\mathbf{a}$ är en *inre punkt* till M om det finns ett öppet klot med centrum i $\mathbf{a}$ som ligger helt i M.

$\mathbf{a}$ är en *yttre punkt* till M om det finns ett öppet klot med centrum i $\mathbf{a}$ som ligger helt i komplementet till M.

$\mathbf{a}$ är en *randpunkt* till M om *varje* öppet klot med centrum i $\mathbf{a}$ innehåller punkter från både M och komplementet till M. $\triangle$

Notera att varje punkt i $\mathbb{R}^n$ tillhör precis en av de ovanstående tre kategorierna med avseende på en given mängd M.

Det är huvudsakligen konceptet av en inre punkt vi är intresserade av i den här kursen, då det är i inre punkter av en definitionsmängd vi (eventuellt) kan definiera koncept i stil med derivata.

Det är viktigt att vara medveten om att huruvida en punkt $\mathbf{a} \in M$ är en inre punkt eller ej inte bara beror på M, utan även på den omkringliggande mängden $\mathbb{R}^n$.

Låt till exempel $I \subset \mathbb{R}$ vara intervallet

$$I = \{x \in \mathbb{R} : 0 < x < 2\}.$$

Då är punkten $x = 1$ en inre punkt till M, för till exempel är det öppna klotet med centrum 1 och radie $1/2$ en delmängd av M.

Men för motsvarande intervall som en delmängd till $\mathbb{R}^2$

$$I = \{(x, 0) \in \mathbb{R}^2 : 0 < x < 2\}$$

har vi nu att *inget* öppet klot med centrum i $(1, 0)$ är en delmängd till I. Till exempel har vi för varje $\delta > 0$ att

$$(1, \delta/2) \in B_\delta(1),$$

men $(1, \delta/2) \notin I$.

Problemet är i någon bemärkelse att den omkringliggande mängden är större relativt I, ocn att öppna klot därför börjar innehålla en massa extra punkter utanför I.

Definition 2.3.4. En mängd $M \subset \mathbb{R}^n$ kallas *öppen* om alla punkter i M är *inre* punkter till M. Den kallas sluten om alla randpunkter till M ligger i M. $\triangle$

Till exempel är det öppna intervallet $(0, 1) \subset \mathbb{R}$ en *öppen* mängd, och det slutna intervallet $[0, 1]$ är en *sluten* mängd då dess enda randpunkter är 0 och 1, och båda dessa ligger i $[0, 1]$. Att visa detta ordentligt lämnas som en övningsuppgift. Det är också värt att tänka igenom och verfiera att alla öppna klot faktiskt är öppna mängder enligt definitionen ovan. Att visa detta lämnas också som en övningsuppgift.

Notera dock att det finns mängder som varken är öppna eller slutna. Till exempel är $[0, 1)$ inte öppen då 0 inte är en inre punkt, men den är inte sluten heller då den har 1 som randpunkt, trots att 1 inte ligger i mängden.

Övningar

Övning 2.1 ($\star$)**.** Låt $\mathbf{x} = (1, 3, 5, 7)$ och $\mathbf{y} = (-1, 3, -5, 1)$.

(a) Beräkna $|\mathbf{x}|$ och $|\mathbf{y}|$

(b) Beräkna $\mathbf{x} \cdot \mathbf{y}$

(c) Är $\mathbf{x}$ och $\mathbf{y}$ parallella?

Övning 2.2 ($\star$)**.** Låt $\mathbf{x} = (1, 2, 3)$ och $\mathbf{y} = (-3, 2, -1)$.

(a) Verifiera Cauchy-Schwarz' olikhet för detta val av $\mathbf{x}$ och $\mathbf{y}$.

(b) Verifiera triangelolikheten för detta val av $\mathbf{x}$ och $\mathbf{y}$

Övning 2.3 ($\star\star$)**.** Använd induktion för att bevisa att

$$|x_1 + \ldots + x_n| \leqslant |x_1| + \ldots + |x_n|,$$

för alla positiva heltal $n \geqslant 2$.

Övning 2.4 ($\star\star$)**.** Bevisa att det slutna intervallet $[0, 1] \subset \mathbb{R}$ är en *sluten* mängd.

Övning 2.5 ($\star\star$)**.** Bevisa att det öppna intervallet $(0, 1) \subset \mathbb{R}$ är en *öppen* mängd.

Övning 2.6 ($\star\star\star$)**.** Låt $r > 0$ och $\mathbf{a} \in \mathbb{R}^n$ och betrakta det öppna klotet $B_r(\mathbf{a})$. Visa att $B_r(\mathbf{a})$ är en *öppen* mängd.

3 Optimering

I såväl maskininlärning som i många andra forskningsområden är man intresserad av att avgöra var en funktion antar sitt största respektive minsta värde. I den här kursen kommer vi vara intresserade av att konstruera modeller som ska förutspå utfall, och vi kommer att vilja hitta den modell som ger bäst förutsägelser. Vi är alltså intresserade av att veta vilken modell vi ska välja för att minimera skillnaden mellan modellens förutsägelse och det "förväntade utfallet". För att kunna göra detta behöver vi hitta ett villkor för när en funktion antar sitt minsta värde. Vi börjar med att göra detta för en funktion av en variabel, och därefter generaliserar vi detta till funktioner av flera variabler.

3.1 Optimering för funktioner av en variabel

För att undvika repetition definierar vi maximum och minimum för funktioner av n variabler redan här. För att få de relevanta definitionerna för en funktion av en variabel är det bara att välja att $n = 1$.

Definition 3.1.1. Låt $f(x)$ vara en funktion definierad på mängden $D_f \subset \mathbb{R}^n$ och låt x_0 vara en punkt i D_f. Vi säger att f har ett *lokalt maximum* i x_0 om det finns ett tal $\delta > 0$ så att

$$x \in \{x \in D_f : |x - x_0| < \delta\} \Rightarrow f(x) \leqslant f(x_0).$$

Vi kallar då x_0 en *lokal maximipunkt* för f och funktionsvärdet $f(x_0)$ för ett *lokalt maximivärde*.

Om vi dessutom har att $f(x) < f(x_0)$ då $x \neq x_0$ säger vi istället att vi har en *sträng* lokal maximipunkt och ett *strängt* lokalt maximivärde.

På motsvarande sätt (alltså med omvända olikhetstecken) definieras en (sträng) *lokal minimipunkt* och ett (strängt) *lokalt minimivärde*.

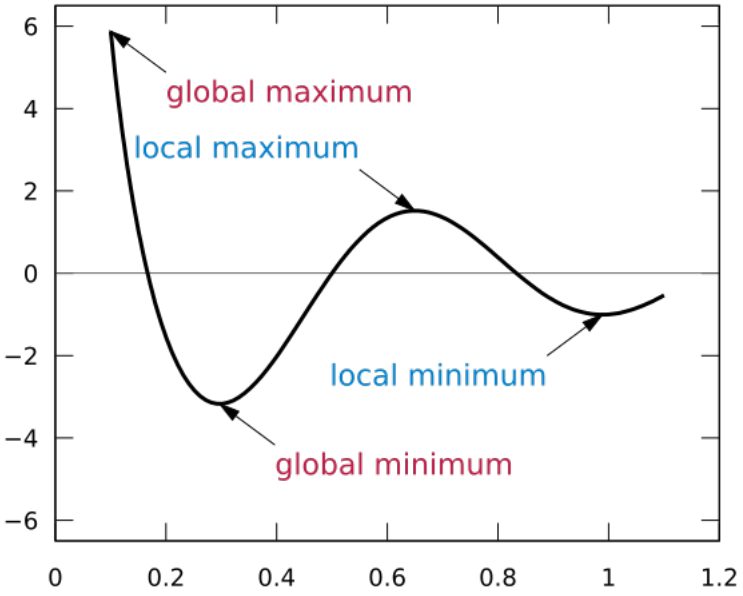
△

Lokala maximipunkter och lokala minimipunkter kallas med ett gemensamt namn för *lokala extremvärden*, och vi säger att f har ett *lokalt extremvärde* i dessa punkter.

Det är viktigt att komma ihåg att huruvida en punkt x_0 är en lokal extrempunkt eller ej beror på såväl funktionen som på definitionsmängden! Låt till exempel $f(x) = x$. Om $D_f = [0, 1]$ är $x_0 = 1$ både en lokal och en global maximipunkt. Om $D_f = [0, 1] \cup [2, 3]$ är $x_0 = 1$ en lokal maximipunkt, men inte en global maximipunkt. Om $D_f = [0, 2]$ är $x_0 = 1$ inte ens en lokal maximipunkt.

Sats 3.1.2. *Om funktionen f har ett lokalt extremvärde i en* inre *punkt $x_0 \in D_f$ och om f är deriverbar i x_0 så är $f'(x_0) = 0$.*

Bevis. Vi bevisar satsen i fallet då f har ett lokalt maximum i x_0. Beviset då f har ett lokalt minimum i x_0 är analogt.



Figur 3.1: Grafen för en funktion som har både lokala och globala extrempunkter. Bild från Wikipedia.

Eftersom f har ett lokalt maximum vet vi att för något tillräckligt litet $\delta > 0$ gäller det att

$$|h| < \delta \Rightarrow f(x_0 + h) < f(x_0).$$

och att $x_0 + h \in D_f$ för alla $h \in [-\delta, \delta]$.

Från detta följer det att

$$\frac{f(x_0 + h) - f(x_0)}{h} < 0$$

för alla $0 < h < \delta$ eftersom täljaren är negativ och nämnaren är positiv.

Vidare har vi att

$$\frac{f(x_0 + h) - f(x_0)}{h} > 0$$

för alla $-\delta < h < 0$ eftersom både täljaren och nämnaren är negativa i det fallet.

Vi har alltså att

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

och

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

Eftersom vi antog att f är deriverbar i x_0 måste dessa gränsvärden överensstämma, och den enda möjligheten är då att $f'(x_0) = 0$. □

Notera att det omvända påståendet *inte* gäller. Till exempel är $f'(0) = 0$ för $f(x) = x^3$ och $D_f = \mathbb{R}$, trots att $x_0 = 0$ inte är en extrempunkt för $f(x) = x^3$.

Satsen ger alltså ett *nödvändigt* men inte *tillräckligt* villkor för att en punkt x_0 ska vara en extrempunkt för en funktion f.

Notera även att vårt föregående exempel med den lokala extrempunkten $x_0 = 1$ för $f(x) = x$ och $D_f = [0, 1]$ visar att antagandet om att x_0 är en *inre* punkt i D_f är nödvändigt.

Vårt mål är att generalisera Sats (3.1.2) till funktioner som är definierade på $\mathbb{R}^n$, men för att kunna göra detta måste vi först definera vad vi menar med, kontinuitet, derivata och differentierbarhet för funktioner av flera variabler.

3.2 Partiella derivator

Vi börjar med att definiera begreppet kontinuitet för funktioner av flera variabler.

Definition 3.2.1. Låt $f(x_1, \ldots, x_n)$ vara en funktion av n variabler definierad på $D_f \subset \mathbb{R}^n$ och låt $\mathbf{a} \in D_f$ vara en inre punkt. Vi säger att f är *kontinuerlig i punkten* $\mathbf{a}$ om

$$\lim_{|\mathbf{h}|\to 0} f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}).$$

Om D_f är en öppen mängd och f är kontinuerlig i alla punkter i D_f så säger vi att f är en *kontinuerlig funktion.* △

Notera att kontinuitet i flera variabler är krångligare än kontinuitet i en variabel eftersom det helt enkelt finns fler sätt som $\mathbf{h}$ kan gå mot 0 på.

Betrakta till exempel funktion

$$f(x, y) = \begin{cases} \frac{x^2}{x^2+y^2} & \text{om } x, y \neq 0 \\ 0 & \text{om } x = y = 0. \end{cases}$$

Om vi nu betraktar $f(0 + \mathbf{h})$ där $\mathbf{h} = (h, 0)$ får vi

$$\lim_{|\mathbf{h}|\to 0} f(0 + \mathbf{h}) = \lim_{h\to 0} \frac{h^2}{h^2} = 1.$$

Om vi istället sätter $\mathbf{h} = (0, h)$ får vi

$$\lim_{|\mathbf{h}|\to 0} f(0 + \mathbf{h}) = \lim_{h\to 0} \frac{0^2}{h^2} = 0.$$

Och om vi sätter $\mathbf{h} = (h, h)$ får vi

$$\lim_{|\mathbf{h}|\to 0} f(0 + \mathbf{h}) = \lim_{h\to 0} \frac{h^2}{h^2 + h^2} = \frac{1}{2}.$$

Som vi ser kan vi få helt olika gränsvärden beroende på hur vi närmar oss 0, och som det sista exemplet visar räcker det inte att bara undersöka koordinataxlarna!

Låt nu $f(x_1, \ldots, x_n)$ vara en funktion av n variabler. Det är naturligt att fråga sig hur funktionens värde förändras när *en* variabel förändras, men de andra variablerna hålls konstanta.

Om till exempel $t(x, y, z)$ är en funktion som beskriver temperaturen på olika platser i ett rum, och x, y, z då alltså är koordinater som beskriver positionen i rummet, så är det naturligt att fråga sig hur temperaturen förändras om vi, till exempel, rör oss rakt uppåt från en punkt.

Detta leder oss till definitionen av en *partiell derivata*.

Definition 3.2.2. Låt $f(x_1, \ldots, x_n)$ vara en funktion definierad på $D_f \subset \mathbb{R}^n$ och låt $(a_1, \ldots, a_n)$ vara en *inre punkt* i D_f. Om gränsvärdet

$$\lim_{h \to 0} \frac{f(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{h}$$

existerar i punkten $(a_1, \ldots, a_n)$ säger vi att f är *partiellt deriverbar* med avseende på x_j i punkten $(a_1, \ldots, a_n)$. Själva gränsvärdet kallar vi för den *partiella derivatan* av f med avseende på x_j, och denna kommer vi beteckna med

$$\frac{\partial f}{\partial x_j}(a_1, \ldots, a_n), \quad \text{eller} \quad f'_j(a_1, \ldots, a_n).$$

Om alla de partiella derivatorna existerar i punkten $(a_1, \ldots, a_n)$ säger vi att f är partiellt deriverbar i punkten $(a_1, \ldots, a_n)$. Vidare säger vi att en funktion som är partiellt deriverbar i alla punkter i sin definitionsmängd helt enkelt är *partiellt deriverbar*. △

Det är värt att notera att om en funktion är partiellt deriverbar måste definitionsmängden vara öppen. Detta på grund av att definitionen av en partiell derivata förutsätter att det för varje punkt $(a_1, \ldots, a_n) \in D_f$ finns något tillräckligt litet $\delta > 0$ så att

$$|h| < \delta \Rightarrow (a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) \in D_f.$$

Om detta inte är fallet kommer ju differenskvoten

$$\frac{f(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{h}$$

från definitionen av den partiella derivatan inte vara väldefinierat!

Detta är skälet till att vi framöver nästan alltid kommer kräva att en punkt **a** i vilken vi ska undersöka någon slags derivata är en *inre* punkt.

Exempel 3.2.3. Låt $f(x_1, x_2) = x_1 x_2^2$. Från definitionen ser vi då att

$$\begin{aligned} f'_1(x_1, x_2) &= \lim_{h \to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2))}{h} \\ &= \lim_{h \to 0} \frac{(x_1 + h)x_2^2 - x_1 x_2^2}{h} \\ &= \lim_{h \to 0} \frac{h x_2^2}{h} \\ &= \lim_{h \to 0} x_2^2 = x_2^2, \end{aligned}$$

och

$$\begin{aligned} f_2'(x_1, x_2) &= \lim_{h \to 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2))}{h} \\ &= \lim_{h \to 0} \frac{x_1(x_2 + h)^2 - x_1 x_2^2}{h} \\ &= \lim_{h \to 0} \frac{x_1(x_2^2 + 2x_2 h + h^2) - x_1 x_2^2}{h} \\ &= \lim_{h \to 0} \frac{2x_1 x_2 h + x_1 h^2}{h} \\ &= \lim_{h \to 0} 2x_1 x_2 + x_1 h \\ &= 2x_1 x_2. \end{aligned}$$

▲

Att beräkna partiella derivator från definitionen är dock oftast onödigt krångligt. Eftersom partiella derivator definieras utifrån ”samma” differenskvot som används i definitionen av derivatan för en funktion av en variabel, kan vi istället tänka att alla variabler utom den vi deriverar med avseende på är konstanter. Därefter kan vi använda kända formler för derivator av elementära funktioner, samt derivationsregler för summor, produkter, kvoter och sammansättningar.

Exempel 3.2.4. Låt $f(x_1, x_2, x_3) = x_1 \cos(x_1 x_2 x_3)$ och låt $D_f = \mathbb{R}^3$. Vi vill beräkna $f_1'(x_1, x_2, x_3)$ och $f_2'(x_1, x_2, x_3)$.

I det första fallet tänker vi att

$$f(x_1, x_2, x_3) = x_1 \cos(c x_1),$$

där konstanten c (konstant med avseende på x_1 då alltså) är $x_2 x_3$. Genom att använda produktregeln och kedjeregeln ser vi att derivatan av $x_1 \cos(c x_1)$ är

$$\cos(c x_1) - x_1 c \sin(c x_1).$$

Eftersom $c = x_2 x_3$ blir alltså

$$f_1'(x_1, x_2, x_3) = \cos(x_1 x_2 x_3) - x_1 x_2 x_3 \sin(x_1 x_2 x_3).$$

När vi beräknar den partiella derivatan med avseende på x_2 är det istället alla uttryck i x_1 och x_3 som är konstanta istället. Vi tänker då istället att

$$f(x_1, x_2, x_3) = c_1 \cos(c_2 x_2),$$

där $c_1 = x_1$ och $c_2 = x_1 x_3$. Den partiella derivatan med avseende på x_2 blir då alltså

$$-c_1 c_2 \sin(c_2 x_2).$$

Genom att byta ut c_1 mot x_1 och c_2 mot $x_1 x_3$ ser vi att

$$f_2'(x_1, x_2, x_3) = -x_1^2 x_3 \sin(x_1 x_2 x_3).$$

3.3 Differentierbarhet

För många resultat är den naturliga generaliseringen av påståendet att en envariabelfunktion är *deriverbar* inte att en flervariabelfunktion har partiella derivator, utan ett aningen hårdare villkor. Nämligen att funktionen är *differentierbar*.

Låt $f(x)$ vara en funktion av en variabel. En ekvivalent formulering av påståendet att f är deriverbar i en punkt a är att säga att för något värde A gäller det att funktionen

$$g(h) = \frac{f(a+h) - f(a)}{h} - A$$

går mot 0 då $h \to 0$. En ekvivalent formulering är att säga att

$$f(a+h) - f(a) = Ah + hg(h)$$

för något värde A, och där $g(h)$ har egenskapen att

$$\lim_{h \to 0} g(h) = 0.$$

Detta är den formulering av deriverbarhet som av tekniska skäl bäst lämpar sig för att generalisera konceptet deriverbarhet till funktioner av flera variabler.

Definition 3.3.1. Låt f vara en funktion definierad på $D_f \subset \mathbb{R}^n$ och låt $\mathbf{a}$ vara en inre punkt i D_f. Vi säger att f är *differentierbar i punkten* $\mathbf{a}$ om det finns konstanter $A_1, \ldots, A_n$ och en funktion $g(\mathbf{h})$ så att

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = A_1 h_1 + \ldots + A_n h_n + |\mathbf{h}| g(\mathbf{h})$$

och

$$\lim_{\mathbf{h} \to 0} g(\mathbf{h}) = 0.$$

Om f är differentierbar i varje punkt $\mathbf{a} \in D_f$ säger vi att f är *differentierbar*. $\triangle$

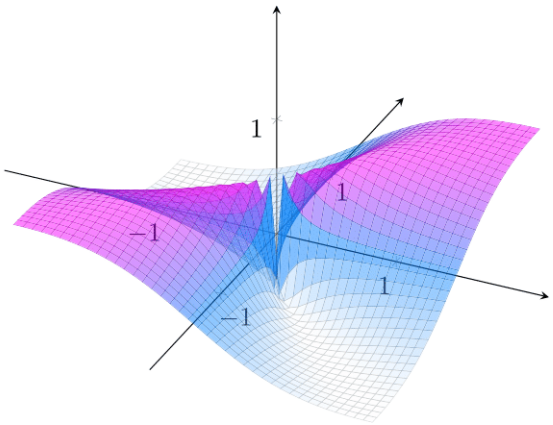
Notera att genom att låta $\mathbf{h} = h\mathbf{e}_j$, där $\mathbf{e}_j$ är enhetsvektorn som har en 1a på position j och 0 på alla andra positioner, i definitionen av differentierbarhet följer det direkt att konstanterna A_j måste överensstämma med de partiella derivatorna $f'_j(\mathbf{a})$. Att vara differentierbar i en punkt är alltså ett hårdare krav än att ha partiella derivator i den punkten.

Den intresserade läsaren kan verifiera att ovanstående definition faktiskt fungerar exakt likandant även om vi skulle göra (det märkliga) antagandet att $g(0)$ är något annat än 0. Men för enkelhets skull kommer vi alltid gör det naturliga antagandet att $g(0) = 0$, det vill säga att g är kontinuerlig i origo.

I praktiken visar man sällan att en funktion är differentierbar utifrån definitionen, utan man använder en sats som säger att om en funktion har partiella derivator *och alla partiella derivator är kontinuerliga* så är funktionen differentierbar. Vi kommer dock inte att bevisa denna implikation då det är för tekniskt avancerat och tidskrävande för denna kurs. Däremot kommer ett exempel ges i uppgifterna som visar att en funktion kan ha partiella derivator

i en punkt *utan* att vara differentierbar. Att anta något mer, till exempel att de partiella derivatorna är kontinuerliga, behövs alltså för att vi ska kunna garantera att funktionen är differentierbar.

Den subtila, men viktiga skillnaden mellan differentierbarhet och att vara partiellt deriverbar är att de partiella derivatorna bara ger information om beteendet längs koordinataxlarna, medan differentierbarhet ställer krav på alla möjliga sätt att närma sig en punkt. Till exempel måste en differentierbar funktion vara kontinuerlig, men detta är faktiskt inte sant för en funktion som bara är partiellt deriverbar! Ett exempel på detta kommer att ges i övningarna.



Figur 3.2: På bilden syns grafen för funktionen $xy/(x^2+y^2)$. Om värdet för funktionen sätts till 0 i origo är funktionen partiellt deriverbar, men den är inte differentierbar i origo.

3.4 Gradient

De partiella derivatorna ger var och en information om hur en funktion förändras längs en koordinatrikting, men det finns självklart mycket mer man kan vara intresserad av att undersöka. I vårt tidigare exempel där $t(x, y, z)$ beskriver temperaturen i ett rum, beskriver de partiella derivatorna hur temperaturen förändras om vi rör oss rakt framåt, rakt åt sidan, eller rakt uppåt. Men om man är intresserad av att veta hur temperaturen förändras när vi, till exempel, rör oss snett uppåt då? För att kunna svara på detta kommer vi sammaföra de partiella derivatorna till ett gemensamt begrepp.

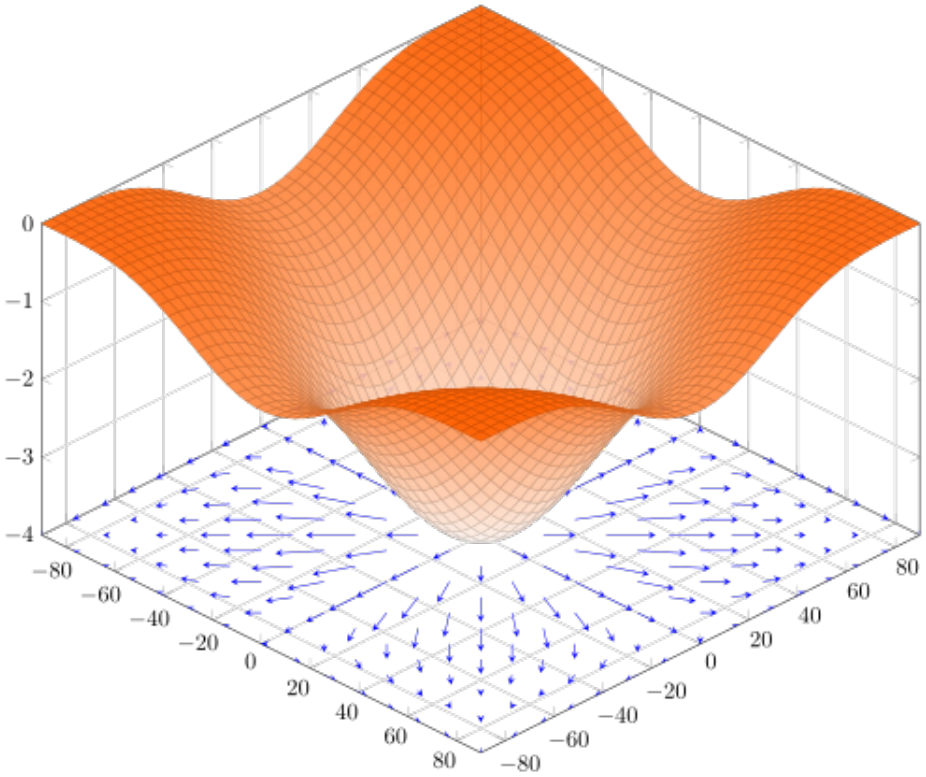
Definition 3.4.1. För en partiellt deriverbar funktion $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ definierar vi *gradienten* av f i punkten $\mathbf{x}$ som vektorn

$$\operatorname{grad} f(\mathbf{x}) = \left(f_1'(\mathbf{x}), \ldots, f_n'(\mathbf{x})\right).$$

△

Lägg märke till att $\operatorname{grad} f(\mathbf{x})$ är en vektor i $\mathbb{R}^n$, och att funktionen

$$\mathbf{x} \mapsto \operatorname{grad} f(\mathbf{x})$$



Figur 3.3: Gradienten för funktionen $f(x,y) = -(\cos(x)^2 + \cos(y)^2)^2$ utmarkerat som ett vektorfält i xy-planet under grafen för funktionen. Bild från Wikipedia.

är en avbildning från $D_f \subset \mathbb{R}^n$ till $\mathbb{R}^n$. Denna funktion, som vi på vanligt sätt betecknar med grad $f(\mathbf{x})$, är alltså ett n-dimensionellt *vektorfält*, som associerar en vektor i $\mathbb{R}^n$ till varje punkt i definitionsmängden till f.

Exempel 3.4.2. Låt $f(x,y,z) = xy^2z^3$. Derivering ger att

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

Gradienten ges alltså av

$$\begin{aligned}\operatorname{grad} f(x,y,z) &= \left(\frac{\partial f}{\partial x}(x,y,z), \frac{\partial f}{\partial y}(x,y,z), \frac{\partial f}{\partial z}(x,y,z)\right) \\ &= (y^2z^3, 2xyz^3, 3xy^2z^2).\end{aligned}$$

Det här är alltså en avbildning från $\mathbb{R}^3$ till $\mathbb{R}^3$. Till exempel är

$$\operatorname{grad} f(1,1,1) = (1,2,3).$$

3.5 Riktningsderivata

Vi är nu redo att undersöka tillväxten av en funktion f i en godtycklig inre punkt $\mathbf{a} \in D_f$ längs godtyckliga räta linjer $\mathbf{x} = \mathbf{a} + t\mathbf{v}$ som går genom punkten $\mathbf{a}$.

Här är $t \geqslant 0$ och $\mathbf{v} = (v_1, \ldots v_n)$ en *normerad* riktningsvektor, vilket innebär att

$$|\mathbf{v}| = \sqrt{v_1^2 + \ldots + v_n^2} = 1.$$

Detta innebär att t mäter avståndet från punkten $\mathbf{a}$ till punkten $\mathbf{a}+t\mathbf{v}$, eftersom

$$|(\mathbf{a} + t\mathbf{v}) - \mathbf{a}| = |t\mathbf{v}| = t|\mathbf{v}| = t.$$

Definition 3.5.1. Låt $\mathbf{v}$ vara en normerad riktningsvektor. Om gränsvärdet existerar kallar vi

$$f'_{\mathbf{v}}(\mathbf{a}) = \lim_{t\to 0} \frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t}$$

för *derivatan* av $f(x)$ i punkten $\mathbf{a}$ med avseende på riktiningen $\mathbf{v}$.

△

Om $\mathbf{v} = \mathbf{e}_j$ så får vi bara tillbaka den partiella derivatan $f'_j(\mathbf{a})$. Notera även att

$$f'_{-\mathbf{v}}(\mathbf{a}) = -f'_{\mathbf{v}}(\mathbf{a}). \tag{3.1}$$

I praktiken beräknar man sällan riktingsderivator från definitionen, utan istället använder man följande sats.

Sats 3.5.2. *Om f är en differentierbar funktion och $\mathbf{v}$ är en normerad riktningsvektor så är*

$$f'_{\mathbf{v}}(\mathbf{a}) = \operatorname{grad} f(\mathbf{a}) \cdot \mathbf{v}.$$

Bevis. Vi vill visa att

$$f'_{\mathbf{v}}(\mathbf{a}) = \lim_{t\to 0} \frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t} = \sum_{j=1}^{n} f'_j(\mathbf{a}) v_j.$$

Eftersom f är differentierbar gäller det enligt definitionen att

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = |\mathbf{h}| g(\mathbf{h}) + \sum_{j=1}^{n} f'_j(\mathbf{x}) h_j \tag{3.2}$$

där $g(\mathbf{h}) \to 0$ då $\mathbf{h} \to 0$ och $g(0) = 0$.

Genom att sätta $\mathbf{x} = \mathbf{a}$ och $\mathbf{h} = t\mathbf{v} = (tv_1, \ldots tv_n)$ i (3.2) och dividera med t får vi

$$\frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t} = \frac{|t\mathbf{v}| g(t\mathbf{v}) + \sum_{j=1}^{n} f'_j(\mathbf{x}) t v_j}{t} = \pm|\mathbf{v}| g(t\mathbf{v}) + \sum_{j=1}^{n} f'_j(\mathbf{x}) v_j,$$

där tecknet framför $|\mathbf{v}| g(t\mathbf{v})$ beror på om t är positivt eller negativt.

Eftersom $|\mathbf{v}| = 1$ och $\lim_{t\to 0} g(t\mathbf{v}) = 0$ är

$$\lim_{t\to 0} \frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t} = \lim_{t\to 0} \pm|\mathbf{v}| g(t\mathbf{v}) + \sum_{j=1}^{n} f'_j(\mathbf{x}) v_j = \sum_{j=1}^{n} f'_j(\mathbf{x}) v_j. \quad \square$$

Exempel 3.5.3. Vi vill bestäma derivatan av $f(x, y, z) = x^2y^3$ i punkten (1,1), längs riktingen $\mathbf{v} = (1/\sqrt{2}, 1/\sqrt{2})$.

Notera till att börja med att

$$|\mathbf{v}| = \sqrt{\frac{1}{2} + \frac{1}{2}} = \sqrt{1} = 1,$$

så v är en normerad riktningsvektor.

Vi har att

$$\operatorname{grad} f(x, y) = (2xy^3, 3x^2y^2),$$

så

$$\operatorname{grad} f(1, 1) = (2, 3).$$

Det följer att

$$f'_{\mathbf{v}}(1, 1) = (2, 3) \cdot (1/\sqrt{2}, 1/\sqrt{2}) = \frac{2}{\sqrt{2}} + \frac{3}{\sqrt{2}} = \frac{5}{\sqrt{2}},$$

vilket ger den önskade riktingsderivatan. ▲

Sats (3.5.2) är viktig dels för att den ger oss en enkel metod för att beräkna riktingsderivator — vi tar bara skalärprodukten mellan den normerade riktningsvektorn och gradienten — men även för att den ger oss ett verktyg för att härleda viktiga teoretiska resultat. Följande sats är en stor del av anledningen till varför vi intresserar oss för gradienten i den här kursen.

Sats 3.5.4. *Gradient* $\operatorname{grad} f(\mathbf{a})$ *pekar i den riktning i vilken funktionen* f *förändras snabbast i punkten* $\mathbf{a}$. *Vidare ges mätetalet på den maximala förändringshastigheten av* $|\operatorname{grad} f(\mathbf{a})|$.

Bevis. Påståendet är en direkt konsekvens av sats (3.5.2) och Cauchy–Schwarz' olikhet eftersom

$$f'_{\mathbf{v}}(\mathbf{a}) = \operatorname{grad} f(\mathbf{a}) \cdot \mathbf{v} \leqslant |\operatorname{grad} f(\mathbf{a})||\mathbf{v}| = |\operatorname{grad} f(\mathbf{a})|.$$

Vidare säger Cauchy–Schwarz' olikhet också att likhet i olikheten ovan inträffar om och endast om de två vektorerna är parallella, det vill säga då

$$\mathbf{v} = \frac{1}{|\operatorname{grad} f(\mathbf{a})|} \operatorname{grad} f(\mathbf{a}).$$

Det här innebär precis att den maximala tillväxten erhålls i gradientens riktning, och att den maximala tillväxten ges av $|\operatorname{grad} f(\mathbf{a})|$. □

Senare i kursen kommer vi använda detta resultat för att ta fram en algoritm för att finna maximum och minimum för en funktion; vi följer helt enkelt funktionen uppåt (eller neråt) i gradientens riktning tills vi har kommit så högt upp (eller långt ner) vi kan komma. Men hur ska algoritmen veta när den är klar? För detta behöver vi en motsvarighet till sats (3.1.2) för funktioner av flera variabler.

3.6 Optimering för funktioner av flera variabler

Kom ihåg definitionen för lokala extrempunkter som gavs i definition (3.1.1). I flera variabler gäller följande generalisering av sats (3.1.2).

Sats 3.6.1. *Om funktionen f har ett lokalt extremvärde i en* inre *punkt* $\mathbf{a}$ *i definitionsmängden D_f till f och om f är partiellt deriverbar i* $\mathbf{a}$ *så är*

$$f'_j(\mathbf{a}) = 0, \quad j = 1, \ldots, n.$$

Detta kan ekvivalent skrivas som att

$$\operatorname{grad} f(\mathbf{a}) = 0.$$

Bevis. Detta följer av motsvarande resultat i en variabel — det vill säga sats (3.1.2) — genom att betrakta restriktionerna av f till de räta linjer genom $\mathbf{a}$ som är parallella med koordinataxlarna.

Eftersom f har ett lokalt extremvärde i $\mathbf{a}$ måste även envariabelfunktionen

$$x_j \mapsto f(a_1, \ldots, x_j, \ldots, a_n)$$

ha ett lokalt extremvärde för $x_j = a_j$, och därmed är dess derivata 0 för $x_j = a_j$. Men detta betyder exakt att

$$f'_j(\mathbf{a}) = 0$$

för $j = 1, \ldots, n$. □

Övningar

Övning 3.1 ($\star$)**.** Beräkna de partiella derivatorna av $f(x, y) = \sin(\cos(xy))$.

Övning 3.2 ($\star$)**.** Beräkna de partiella derivatorna av $f(x, y) = x^5y^3z + xyz$.

Övning 3.3 ($\star$)**.** Beräkna gradienten av $f(x, y) = e^{x^2y}$.

Övning 3.4 ($\star$)**.** Beräkna gradienten av $f(x, y, z) = x^3 + xy + y^3$.

Övning 3.5 ($\star$)**.** Använd definitionen av globalt minimum för att visa att

$$f(x, y, z) = x^4 + y^2 + z^2$$

har ett lokalt minimum i origo.

Övning 3.6 ($\star$)**.** Beräkna riktningsderivatan i riktningen $(1, 1, 1)/\sqrt{3}$ av funktionen $f(x, y, z) = x^2 + y^3 + z^2x$.

Övning 3.7 ($\star$)**.** Beräkna riktningsderivatan i riktningen $v = (1, 2)$ av funktionen $f(x, y) = x^2y$.

Övning 3.8 ($\star\star$)**.** Låt $f(x, y) = x^4 - y^4$. Visa att $\operatorname{grad}(f)(x, y) = 0$ för $(x, y) = 0$, men visa att f varken har ett lokalt maximum eller minimum i origo.

Detta visar att villkoret att gradienten är 0 i en punkt är ett *nödvändigt* villkor för att punkten ska vara en extrempunkt, men *inte* ett *tillräckligt* villkor.

Övning 3.9 ($\star$)**.** Beräkna gradienten av $f(x, y) = e^{x\cos(y)}$ och visa att f inte har ett lokalt maximum i origo.

Övning 3.10 ($\star\star$)**.** Använd definitionen av differentierbarhet för att visa att om en funktion f är differentierbar i en inre punkt $\mathbf{a} \in D_f$ så är f kontinuerlig i $\mathbf{a}$.

Övning 3.11 ($\star$)**.** Betrakta funktionen $\sin(xyz^2)$. I vilken riktning förändras funktionen snabbast i punkten $(0, 1, 2)$? Vad är mätetalet på den maximala förändringshastigheten?

Övning 3.12 ($\star\star\star$)**.** I den här uppgiften ska vi visa att det finns funktioner som är partiellt deriverbara, men som inte är differentierbara.

Betrakta funktionen

$$f(x, y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{om } x, y \neq 0 \\ 0 & \text{om } x = y = 0. \end{cases}$$

(a) Visa att f är partiellt deriverbar i origo.

(b) Visa att f *inte* är kontinuerlig i origo.

(c) Använd (b) och resultatet av en tidigare övningsuppgift för att visa att f *inte* är differentierbar i origo.

Övning 3.13 ($\star$)**.** Betrakta funktionen x^2y^2z. I vilken riktning förändras funktionen snabbast i punkten $(-1, -1, 3)$? Vad är mätetalet på den maximala förändringshastigheten?

Övning 3.14 ($\star$)**.** Låt $t(x, y, z) = z/(x^2 + y^2)$ beskriva temperaturen i ett rum som bland anant innehåller punkten $(1, 1, 1)$. I vilken riktning ska man röra sig från punkten $(1, 1, 1)$ för att temperaturen ska förändras så snabbt som möjligt? Hur snabbat förändras temperaturen i den riktningen?

4 Sannolikhetsteorins grunder

Senare i kursen kommer det, vagt formulerat, ges information till en modell, vilken därefter ska gissa vad det är för något som har givits till den. Den nya informationen är från modellens perspektiv i någon bemärkelse slumpmässig, och dess uppgift är att använda tidigare information som den har fått samt den nya informationen för att avgöra vad som *mest troligt* är den bästa klassificeringen av den nya informationen. Till exempel kan den få en bild i form av massor med pixlar, och ska därefter försöka avgöra om det är mest sannolikt att pixlarna föreställer en hund, en människa, eller en katt.

Vi kommer även i ett tekniskt skede att behöva lägga på brus i en algoritm för att den ska undvika falska extrempunkter, och på grund av detta kommer vi även behöva gå igenom den teori som är nödvändig för att matematiskt beskriva ”att lägga på brus”.

4.1 Händelser och utfallsrum

I grunden för all sannolikhetsteori ligger hela tiden ett *slumpexperiment*. Med detta menas en situation där något kommer att inträffa, men där vi inte med säkerhet kan säga vad. Ett slumpexperiement kan till exempel vara situationen ”att vi slår en tärning”.

Definition 4.1.1. Resultatet av ett slumpexperiment kallas ett *utfall*. Mängden av möjliga utfall för ett slumpexperiment kallas *utfallsrum*. Vidare kallar vi en viss specificerad delmängd av möjliga utfall för en *händelse*. △

Vi kommer att beteckna enskilda utfall med $u_1, u_2, \ldots$, händelser med versaler, $A, B, \ldots$, och utfallsrummet med Ω. Ett utfallsrum med ändligt eller uppräkneligt många utfall kallas *diskreta* utfallsrum, medans utfallsrum med oupräkneligt många utfall kallas *kontinuerliga* utfallsrum.

Det är värt att påpeka att ett utfall *inte* är ett tal, och man kan därför inte i allmänhet prata om att addera, subtrahera och utföra andra klassiska operationer på utfall. Ett utfall skulle till exempel kunna vara en persons födelsedag, eller färgen på nästa bil som åker förbi. Utfall och händelser är bara *element* och *mängder* av element, så vi kan dock utföra vanliga mängdoperationer på dem.

Exempel 4.1.2. Betrakta återigen experimentet av att slå en tärning. Ett utfall är då en av siffrorna $1, 2, 3, 4, 5, 6$, och utfallsrummet är helt enkelt mängden $\{1, 2, 3, 4, 5, 6\}$. En händelse är alltså en delmängd av dessa 6 tal. En händelse skulle till exempel kunna vara ”att vi slår ett udda tal”, eller ”att vi slår ett tal mindre än 4”. Dessa händelser beskrivs då av mängderna $A = \{1, 3, 5\}$ respektive $B = \{1, 2, 3\}$.

Händelsen ”att vi slår ett tal som både är ett udda tal *och* ett tal mindre än 4” beskrivs av *snittet* av dessa två mängder, det vill säga $A \cap B = \{1, 3\}$.

Att vi slår ett tal som är udda eller mindre än 4 (eller båda) beskrivs som *unionen* av dessa mängder, det vill säga av $A \cup B = \{1, 2, 3, 5\}$.

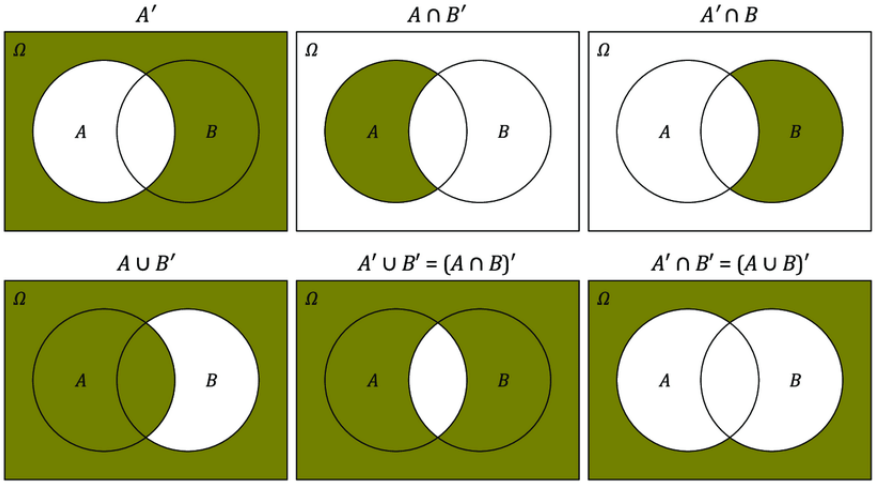
Att vi slår ett tal som *inte* är mindre än 4 beskrivs av komplementet till mängden ”att vi slår ett tal mindre än 4”, det vill säga av $B^c = \{4, 5, 6\}$.

Att talet är mindre än 4 men *inte* udda beskrivs som snittet av händelsen ”talet är mindre än 4” och komplementet till händelsen ”talet är udda”, vilket är samma sak som att ta bort alla udda tal från händelsen ”talet är mindre än 4”. Händelsen beskrivs alltså av $B \cap A^c = B \setminus A = \{2\}$. ▲

Som föregående exempel illustrerar beskriver resultaten mängdoperationerna på två givna händelser A och B rätt naturliga händelser, nämligen

- att *minst en* av A och B inträffar skrivs som $A \cup B$,
- att *både* A och B inträffar skrivs som $A \cap B$,
- att A inte inträffar skrivs som A^c,
- att A inträffar men inte B skrivs som $A \cap B^c = A \setminus B$.

En speciell händelse är den *tomma mängden*, ”inget utfall”, vilket vi kommer beteckna med $\emptyset$. Två händelser sägs vara oförenliga eller disjunkta om $A \cap B = \emptyset$. Till exempel är alltid en mängd oförenlig med sitt komplement. Resultatet av ett tärningsslag kan inte vara både jämnt och udda!



Figur 4.1: Venndiagram som illustrerar hur komplement fungerar tillsammans med snitt och unioner. I bilden används $S^{'}$ istället för S^c för att beteckna komplmentet till en mängd S. Bild av J. Scott Cardinal, hämtad från ResearchGate.

4.2 Sannolikheter på utfallsrum

När vi nu har preciserat vad vi menar med utfall och utfallsrum är vi redo att definiera slumpförsök och sannolikheter på sådana.

Ett *slumpförsök* på ett utfallsrum består av ett försök som resulterar i något av utfallen i utfallsrummet, men där man på förhand inte kan veta exakt vilket av utfallen som kommer att inträffa. För att beskriva slumpförsöket preciserar

man därför sannolikheterna för alla möjliga händelser i utfallsrummet. Sannolikheten för en händelse A brukar betecknas med $P(A)$, där P alltså är en reell funktion som definierad på mängden av händelser i utfallsrummet. Dock är inte vilken funktion som helst som är definierad på mängden av händelser rimlig som en sannolikhetsfunktion. Det är till exempel orimligt att sannolikheten för en händelse är negativ, vi väntar oss att sannolikheten för alla möjliga händelser ska vara 1 (det vill säga 100 procent) och så vidare. Följande krav på vad vi förväntar oss av något som beskriver sannolikheter på ett utfallsrum kallas för *Kolmogorovs axiomsystem.*

Definition 4.2.1. En reell funktion P definierad på händelser i ett utfallsrum Ω är en *sannolikhetsfunktion* om den uppfyller följande tre axiom.

- $0 \leqslant P(A) \leqslant 1$ för alla händelser $A \subset \Omega$
- $P(\Omega) = 1$
- om $A \cap B = \emptyset$ så gäller $P(A \cup B) = P(A) + P(B)$.

Om utfallsrummet är oändligt ersätts det sista villkoret med:

Om $A_1, A_2, \ldots$ är en oändlig följd av parvis oförenliga händelser, det vill säga $A_i \cap A_j = \emptyset$ för alla $i \neq j$, så gäller

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

△

Det första villkoret säger helt enkelt att sannolikheten för en händelse är mellan 0 och 100 procent. Det andra axiomet säger att sannolikheten att *någonting* händer är 100 procent. Det sista axiomet säger att sannolikheten för två oförenliga händelser är summan av var och en av händelserna.

Den vanligaste tolkningen av en sannolikhet är att om man utför ett experiment ett stort antal gånger bör andelen gånger då händelsen inträffar, det vill säga antalet gånger händelsen inträffar delat på det totala antalet utförda experiment, gå mot sannolikheten för händelsen.

Om vi återigen tänker på slumpförsöket att slå en tärning, så tänker vi oss att det bör vara lika stor sannolikhet att få vilket som helst av de 6 möjliga utfallen, det vill säga $1/6$ eftersom axiom 2 säger att $P(\{1,2,3,4,5,6\}) = 1$. Sannolikheten att slå ett tal mindre än 4 bör således vara $1/6+1/6+1/6 = 1/2$ enligt axiom 3, och tolkningen av $P(\{1,2,3\}) = 1/2$ är alltså att om vi slår en tärning 1000 gånger väntar vi oss att vi ska slå ett tal mindre än 4 väldigt nära hälften av gångerna.

Detta är helt i linje med vår intuition, vilket är ett bra tecken för vår definition eftersom den är en matematisk konkretisering av ett intuitivt begrepp.

Från axiomen kan vi dra några ytterligare, väldigt intuitiva, slutsatser om hur sannolikhetsfunktioner och mängdoperationer interagerar med varandra.

Sats 4.2.2. *Låt A och B vara godtyckliga händelser i utfallsrummet Ω. Då gäller*

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Bevis. Eftersom $A \cap A^c = \emptyset$ och $A \cup A^c = \Omega$ gäller det att

$$1 = P(A \cup A^c) = P(A) + P(A^c) \Rightarrow P(A^c) = 1 - P(A),$$

vilket bevisar det första påståendet. Eftersom $\emptyset^c = \Omega$ och $P(\Omega) = 1$ följer det andra påståendet från det första.

För att visa det sista påståendet behöver vi först notera att $A \cup B$ är samma sak som A, och att vi därefter lägger till de element i B som *inte* också ligger i A. Det vill säga

$$A \cup B = A \cup (B \cap A^c).$$

Eftersom $B \cap A^c \subset A^c$ är $A \cap (B \cap A^c) = \emptyset$.

Det tredje axiomet ger oss därför att

$$P(A \cup B) = P(A) + P(B \cap A^c).$$

Nu kan vi dela upp B i två disjunkta delar; de element som finns i både B och A, och de element som finns i B men *inte* i A. Vi har alltså att $B = (B \cap A) \cup (B \cap A^c)$, och $(B \cap A) \cap (B \cap A^c) = \emptyset$. Det tredje axiomet ger nu att

$$P(B) = P(B \cap A) + P(B \cap A^c) \Rightarrow P(B \cap A^c) = P(B) - P(A \cap B).$$

Genom att sätta in detta i föregående ekvation får vi att

$$P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B),$$

vilket bevisar den sista ekvationen. □

4.3 Betingning och oberoende

Vi kommer nu introducera koncepten *betingning* och *oberoende* som berör huruvida en händelse inträffar påverkar huruvida en annan händelse inträffar.

Om vi till exempel drar två kort ur en vanlig kortlek med 52 kort påverkar resultatet av den första dragningen sannolikheterna för resultatet av den andra. Om till exempel det första kortet vi drar är ett ess är sannolikheten att kort nummer 2 är ett ess 3/51 eftersom det finns 3 ess i leken och 51 kort totalt, och alla kort har samma sannolikhet att dras. Om vi istället drar en kung är sannolikheten att dra ett ess 4/51, eftersom det i detta fall finns 4 ess kvar i leken.

Vad vi är intresserade av här är så kallad *betingad sannolikhet*. Vi är intresserade av att veta sannolikheten för en händelse B betingat av att en annan händelse A redan har inträffat. Detta skrivs som $P(B|A)$ och läses som sannolikheten för B betingat av att A har inträffat.

I exemplet ovan är händelsen B att vi drar ett ess som vårt andra kort. Om händelsen A är händelsen att vi drar ett ess som vårt första kort är alltså $P(B|A) = 3/51$. Om A är händelsen att vi drar en kung som vårt första kort är $P(B|A) = 4/51$.

Förutsatt att det finns någon underförstådd tidsordning som gör att A måste inträffa innan eller samtidigt som B så bör sannolikheten att B *och* A inträffar vara samma som sannolikheten att först A inträffar, multiplicerat med sannolikheten att B inträffar betingat på att A redan har inträffat. Alltså att $P(B|A)P(A) = P(A \cap B)$. Med detta som motivering ger vi följande *definition* av betingad sannolikhet i termer av redan kända begrepp.

Definition 4.3.1. Antag att A är en händelse som upppfyller att $P(A) > 0$. Den *betingade sannolikheten* för händelsen B givet att A har inträffat skrivs $P(B|A)$ och definieras som

$$P(B|A) := \frac{P(B \cap A)}{P(A)}.$$

△

Exempel 4.3.2. Vid en hastighetskontroll utanför en skola visar det sig att 30 procent av fordonen kör för fort, och att 6 procent av fordonen kör mer än 20 km/h för fort, vilket resutlerar i indraget körkort. Vad är då sannolikheten att en person som kör för fort får indraget körkort?

Vi låter A vara händelsen att ett fordon kör för fort, och B händelsen att ett fordon kör mer än 20 km/h för fort. Notera att $B \subset A$, för kör man mer än 20 km/h för fort så kör man för fort, så $P(B \cap A) = P(B) = 0.06$. Sannolikheten att en fortkörning resulterar i indraget körkort är alltså sannolikheten att B inträffar (att personen kör mer än 20 km/h för fort) betingat av att A redan har inträffat (personen körde för fort), det vill säga

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = 0.06/0.3 = 0.2.$$

När man är intresserad av sannolikheter för flera olika händelser i ett slumpexperiment är man ofta intresserad av huruvida händelserna beror av varandra eller inte.

Till exempel noterade vi tidigare att resultaten av att dra kort ur en kortlek beror på de tidigare resultaten. Om det första kortet var ett ess gick ju sannolikheten ner för att det andra kortet ska vara ett ess. Däremot påverkar inte resultatet av ett tärningsslag resultatet av ytterligare ett tärningsslag.

I vardagligt tal tolkar man påståendet ”att två händelser är oberoende av varandra” som att de inte har något med varandra att göra. Vetskap om huruvida den ena händelsen har inträffat eller ej påverkar inte sannolikheterna för den andra händelsen.

Detta motiverar följande definition.

Definition 4.3.3. Två händelser A och B är oberoende om $P(A|B) = P(A)$ förutsatt att $P(B) > 0$, och $P(B|A) = P(B)$ förutsatt att $P(A) > 0$. △

Notera att om både $P(A)$ och $P(B)$ är större än 0 så är de två relationerna ekvivalenta. Att visa detta lämnas som en övning.

Ovanstående definition är inutivt tydlig då den på ett matematiskt sätt säger att sannolikheten att A händer är samma som sannolikheten att A händer betingat av att B redan har hänt och vice versa, vilket helt enkelt betyder att huruvida den ena händelsen har inträffat eller ej inte påverkar sannolikheten för den andra att inträffa.

Definitionen har dock nackdelen att den förutsätter att $P(A)$ och $P(B)$ är större än noll. En alternativ definition, som dock inte är lika intuitivt tydlig, har inte detta problem.

Definition 4.3.4. Två händelser A och B sägs vara *oberoende* om

$$P(A \cap B) = P(A)P(B).$$

△

Notera att de två definitionerna är ekvivalenta om vi förutsätter att $P(A)$ och $P(B)$ är större än noll.

Vidare säger vi att en mängd av händelser $\{A_1, A_2, \ldots\}$ är *parvis oberoende* om det för alla par (i, j) med $i \neq j$ gäller att $P(A_i \cap A_j) = P(A_i)P(A_j)$.

Som tidigare har nämnts är de vanligaste situationerna då man stöter på oberoende händelser situationer då man utför ett experiment flera gånger, och senare experiment inte påverkas av resultaten av tidigare, eller situationer då man utför helt olika experiment. Till exempel påverkas inte sannolikheten att man får triss i poker av att man fick triss förra gången, och sannolikheten för att få triss i poker påverkas inte av huruvida den senaste bilen som såldes i Sverige var svart eller inte.

Det finns dock situationer där olika händelser relaterade till *samma* experiment är oberoende. Tänk till exempel att vi kastar en pil på en piltavla och att sannolikheten är lika stor att träffa vilken punkt som helst, det vill säga sannolikheten att pilen landar i ett område av tavlan ges av

$$\frac{\text{area(område)}}{\text{area(hela tavlan)}}$$

(såhär ser troligtvis inte den riktiga sannolikhetsfunktionen ut för någon person, men det är tillräckligt för att illustrera poängen).

Är händelserna A som betecknar att ”pilen träffar vänsrta halvan av tavlan”, och B som betecknar ”att pilen träffar högra halvan av tavlan” oberoende?

Absolut inte! Träffar pilen högra halvan av tavlan *vet* vi att den inte träffade vänstra halvan av tavlan, och vice versa.

Matematiskt kan detta visas genom att visa att $P(A \cap B) \neq P(A)P(B)$.

Vi har att

$$P(A) = \frac{\text{area(vänstra halvan av tavlan)}}{\text{area(hela tavlan)}} = 1/2,$$

och

$$P(B) = \frac{\text{area(högra halvan av tavlan)}}{\text{area(hela tavlan)}} = 1/2,$$

så $P(A)P(B) = 1/4$.

Men om vi låter C beteckna mängden av punkter som ligger på *både* vänstra och högra halvan av tavlan är

$$P(A \cap B) = \frac{\text{area}(C)}{\text{area(hela tavlan)}} = 0 \neq 1/4.$$

Händelserna är alltså inte oberoende.

Däremot är händelserna ”att träffa övre halvan av tavlan” och ”att träffa högra halvan av tavlan” oberoende! Detta på grund av att det är $50-50$ om vi träffar högra halvan eller ej, och det förändras inte av att vi vet att pilen träffade den övre halvan, för den övre vänstra halvan och den övre högra halvan av tavlan är lika stora, så det är fortfarande $50 - 50$ att pilen hamnade på den högra halvan. Att visa detta rigoröst lämnas som en övningsuppgift.

4.4 Slumpvariabler

Vi har nu bekantat oss med slumpförsök och sannolikheter. Som tidigare har nämnts kan man i allmänhet inte utföra klassiska matematiska operationer på resultaten av slumpförsök; vad är till exempel differensen av resultaten av två slumpförsök om färgen på nyligen sålda bilar? Men vissa slumpförsök ger upphov till numeriska resultat, och sådana slumpförsök är av extra intresse. Dels för att de är väldigt vanligt förekommande, men även för att vi har mer möjligheter att jämföra olika sådana resultat. Till exempel kan vi då prata om differenser mellan utfall.

När slumpexperiment ger upphov till numeriska värden kallas de för *slumpvariabler* eller *stokastiska variabler*.

Exempel 4.4.1. Ett kafé väljer att räkna antalet personer som går in i affären varje dag. Antalet personer som går in i affären varje dag kan betraktas som en slumpvariabel. Ett annat exempel på en slumpvariabel är antalet arbetsdagar som påbörjas innan kaféet är uppe i sammanlagt 1000 besökare (från att de började räkna då såklart). Båda dessa antal kommer med nödvändighet vara heltal (och icke-negativa). Men säger då att man har en *diskret* slumpvariabel. ▲

Exempel 4.4.2. Ett barns vikt vid födsel kan betraktas som en slumpvariabel. Eftersom vikten i princip kan vara vilket postivt reellt tal som helst (eller man kan väl egentligen sätta någon form av övre begränsning), och inte bara heltal, sägs slumpvariabeln vara *kontinuerlig*. Den uppmätta vikten är däremot *inte* kontinuerlig, då den i allmänhet avrundas, oftast till hela gram. ▲

Definition 4.4.3. En *slumpvariabel*, eller *stokastisk variabel* $X(u)$ är en reellvärd funktion definierad på utfallsrummet för något slumpförsök

$$X : \Omega \to \mathbb{R}.$$

När slumpförsöket har genomförts och ett utfall har erhållits sägs funktionens värde för utfallet vara en *observation* av slumpvariabeln. △

När det inte är explicit nödvändigt brukar inte funktionsargumentet skrivas ut, och vi skriver då bara X istället för $X(u)$. Slumpvariabler kommer att becknas med stora bokstäver X, Y, Z, och motsvarande observationer kommer att betecknas med små bokstäver x, y, z. Vi använder alltså stora bokstäver för att beteckna själva funktionerna på utfallsrummet, och små bokstäver för att beteckna de konkreta tal som kommer ut ur dessa funktioner efter ett genomfört slumpförsök.

Vi kommer ofta att vara intresserade av att veta sannolikheten att en slumpvariabel hamnar i en viss delmängd. Låt alltså $A \subset \mathbb{R}$, vi är då ofta intresserade av att veta

$$P(u \in \Omega : X(u) \in A).$$

Alltså, vad är sannolikheten att vi får ett utfall u som avbildas in i A? Detta kommer dock av bekvämlighetsskäl ofta att förkortas till $P(X \in A)$.

Kom ihåg att P är en reell funktion som tar en delmängd av Ω som sitt argument, medan för en given slumpvariabel X är $P(X \in A)$ en reell funktion som tar en delmängd av $\mathbb{R}$ som sitt argument.

Vi har tidigare stött på vissa slumpförsök vars utfallsrum är en delmängd av $\mathbb{R}$, och således kan vi direkt tänka på dessa slumpförsök som slumpvariabler. Ett exempel på detta är resultatet av ett tärningskast. Så behöver det dock inte alltid vara, vi kan utföra andra operationer på resultaten av slumpförsök för ett erhålla nya slumpvariabler.

Exempel 4.4.4. Betrakta slumpförsöket av att kasta två tärningar. Varje tärning kommer ge ett tal mellan 1 och 6, och utfallsrummet Ω är då alltså mängden av par av heltal mellan 1 och 6, dvs

$$\Omega = \{(x, y) : x, y \in \{1, 2, 3, 4, 5, 6\}\}.$$

Från detta slumpförsök kan vi skapa massor med nya slumpvariabler. Ett av de vanligaste är att betrakta summan av de två träningskasten, det vill säga

$$X : \Omega \to \mathbb{R}, \quad X : (x, y) \to x + y.$$

Men det finns såklart mycket mer vi kan göra. För två heltal $n, m \in \mathbb{Z}$ kan vi till exempel betrakta

$$X : \Omega \to \mathbb{R}, \quad X : (x, y) \to x^n y^m.$$

Om $n = m = 1$ får vi alltså produkten av våra tärningskast, medan om $n = 1, m = 0$ får vi bara resultatet av den första tärningen.

Det finns dock ingenting som säger att vi måste begränsa oss till "vanliga" funktioner. Om vi spelar ett spel är vi kanske intresserade av sannolikheten att vi får par, det vill säga att vi får samma resultat på båda tärningarna. Då kan till exempel slumpvariabeln

$$X : \Omega \to \mathbb{R}, \quad X : (x, y) \to \begin{cases} 1 & \text{om } x = y \\ 0 & \text{annars} \end{cases}$$

vara av intresse.

Med denna slumpvariabel X ges då sannolikheten av att få par av

$$P(u \in \Omega : X(u) \in \{1\}).$$

▲

Övningar

Övning 4.1 ($\star$)**.** Bevisa att för två godtyckliga händelser A och B gäller det att

$$P(A \cup B) \leqslant P(A) + P(B).$$

Övning 4.2 ($\star$)**.** Finns det någon sannolikhetsfunktion P för vilken det existerar händelser A och B så att

$$P(A) = 0.3, \quad P(B) = 0.2, \quad P(A \cap B) = 0.6?$$

Övning 4.3 ($\star$)**.** Betrakta ett slumpexperiment med två händelser A och B som uppfyller att $P(A) = 0.3$, $P(B) = 0.5$ och $P(A \cup B) = 0.6$. Beräkna $P(A \cap B)$.

Övning 4.4 ($\star$)**.** En vanlig tärning kastas. Låt A vara händelsen att tärningen visar ett udda tal, B händelsen att talet blir minst 4, och C, händelsen att talet är delbart med 3.

(a) Är A och B oberoende?

(b) Är A och C oberoende?

(c) Är B och C oberoende?

Övning 4.5 ($\star\star$)**.** Betrakta slumpexperimentet att vi kastar en pil på en piltavla. Låt Ω beteckna piltavlan och anta att sannolikheten att pilen landar i en delmängd $A \subset \Omega$ ges av

$$P(A) := \frac{\text{area}(A)}{\text{area}(\Omega)}.$$

(a) Visa att $P(A)$ uppfyller alla axiomen i Kolomogorovs axiomsystem, och således beskriver en sannolikhetsfunktion.

(b) Visa att händelserna $A =$ ”pilen landar på den övre halvan av piltavlan” och $B =$ ”pilen landar på den högra halvan av piltavla” är oberoende.

Övning 4.6 ($\star\star\star$)**.** Visa att om A och B är oberoende så är även A^c och B oberoende, A och B^c oberoende, och A^c och B^c oberoende.

Övning 4.7 ($\star\star$)**.** För två händelser A och B gäller det att $P(A) = 0.4$, $P(A|B) = 0.6$, och $P(B|A) = 0.75$. Vad är sannolikheten att minst en av A och B inträffar?

Övning 4.8 ($\star$)**.** Betrakta slumpexperimentet att person X och person Y kastar en vanlig tärning var. Vad är sannolikheten att Ys tärning visar ett större tal än Xs tärning?

Övning 4.9 ($\star\star$)**.** Betrakta slumpexperimentet att person X kastar en vanlig tärning och att person Y kastar 2 vanliga tärningar.

Vad är sannolikheten att någon av person Ys tärningar visar ett större tal än talet på person Xs tärning?

Övning 4.10 ($\star\star\star$)**.** Låt $k \geqslant 1$ vara ett heltal och betrakta slumpexperimentet att person X kastar en vanlig tärning och att person Y kastar k vanliga tärningar.

Vad är sannolikheten att någon av person Ys tärningar visar ett högre tal än talet på person Xs tärning?

Övning 4.11 ($\star\star$)**.** Bevisa att om $A \subset B$ så är $P(A) \leqslant P(B)$.

Övning 4.12 ($\star$)**.** Betrakta slumpexperimentet att vi två gånger väljer något heltal mellan 1 och 5, och att det är lika sannolikt att varje tal väljs. Utfallet kommer vara ett par (x, y), där $x, y \in \{1, 2, 3, 4, 5\}$. Betrakta slumpvariabeln

$$X : \Omega \to \mathbb{N}, \quad X : (x, y) \to x + y,$$

det vill säga X är slumpvariabeln som ger summan av de två talen.

Vad är sannolikheten att $X = 10$? Vad är sannolikheten att $X = 3$?

Övning 4.13 ($\star$)**.** Betrakta slumpexperimentet att vi slår tre vanliga tärningar. Utfallet kommer vara en trippel (x, y, z), där $x, y, z \in \{1, 2, 3, 4, 5, 6\}$. Betrakta slumpvariabeln

$$X : \Omega \to \mathbb{N}, \quad X : (x, y, z) \to x + y + z,$$

det vill säga X är slumpvariabeln som ger summan av de tre tärningskasten.

Vad är sannolikheten att $X = 18$? Vad är sannolikheten att $X = 17$?

5 Mått på slumpvariabler och normalfördelningar

Vårt nästa mål är att introducera de funktioner som beskriver slumpstrukturen hos diskreta och kontinuerliga slumpvariabler, nämligen sannolikhetsfunktionen respektive fördelningsfunktionen. Dessa ligger till grund för nästan all analys av slumpvariabler, och kommer konkret behövas i den här kursen för att kunna definiera väntevärde och varians. Slutligen kommer vi att introducera den kanske viktigaste fördelningen av alla, nämligen normalfördelningen.

5.1 Diskreta slumpvariabler och sannolikhetsfunktioner

Definition 5.1.1. En slumpvariabel X är *diskret* om den endast kan anta ändligt eller uppräkneligt oändligt antal värden $x_1, x_2, \ldots$. △

Definition 5.1.2. *Sannolikhetsfunktionen*, p_X, för en diskret slumpvariabel X definieras av

$$p_X(x) := P(X \in \{x\}) = P(X \text{ antar värdet } x), \quad x = x_1, x_2, \ldots.$$

△

Exempel 5.1.3. Betrakta exemplet från föregående kapitel där vi slår två tärningar och X är sannolikhetsfunktionen som är 1 om vi får par, det vill säga samma tal på båda tärningarna, och 0 annars.

Denna slumpvariabel är diskret då den endast kan anta värdena 0 och 1.

Sannolikheten att den ska bli 1 är

$$\begin{aligned} p_X(1) &= P(u \in \Omega : X(u) = 1) \\ &= P(u \in \{(1,1),(1,2),\ldots(6,6)\} : X(u) = 1) \\ &= P(u \in \{(1,1),(2,2),\ldots,(5,5),(6,6)\}) \\ &= 1/6. \end{aligned}$$

Att den sista sannolikheten är 1/6 följer eftersom vi har 6 möjliga utfall som ger värdet 1 av totalt 36 stycken, och alla utfall är lika troliga.

Vidare är $X(u) = 0 \iff X(u) \neq 1$, så

$$p_X(0) = P(X = 0) = 1 - P(X = 1) = 1 - 1/6 = 5/6.$$

▲

Några viktiga egenskaper som gäller för sannolikhetsfunktioner formuleras i följande sats.

Sats 5.1.4. *För en diskret slumpvariabel X gäller*

- $0 \leqslant p_X(k) \leqslant 1$ *för alla* k
- $\sum_k p_X(k) = 1$

- $P(a \leqslant X \leqslant b) = \sum_{k:a\leqslant k\leqslant b} p_X(k)$
- $P(X \leqslant a) = \sum_{k:k\leqslant a} p_X(k)$
- $P(x > a) = \sum_{k:k>a} p_X(k) = 1 - \sum_{k:k\leqslant a} p_X(k) = 1 - P(X \leqslant a)$.

Dessa egenskaper följer direkt från Kolmogorovs axiom och vi uppmuntrar läsaren att verifera dessa själv.

Exempel 5.1.5. Tre personer korrekturläser en text i vilken det finns exakt ett stavfel. För var och en av dessa personer är det 10 procents chans att de missar detta stavfel. Låt X vara det antal bland de tre personerna som missar stavfelet. Vilka observationsvärden k kan X ha? Vad är $p_X(k)$ för dessa observationsvärden?

Eftersom 3 personer läser texten kan anting inga, $1, 2$ eller 3 personer missa stavfelet.

Vi börjar med att härleda $p_X(k)$ för de enklaste värdena, nämligen 0 och 3. Eftersom det är 10 procents sannolikhet att en person missar stavfelet är sannolikheten att alla missar stavfelet $p_X(3) = 0.1^3$. Eftersom det är 90 procents sannolikhet att en enskild person hittar stavfelet är sannolikheten att ingen missar stavfelet $p_X(0) = 0.9^3$.

Att exakt en person missar stavfelet är lite svårare. Sannolikheten att den *första* personen missar stavfelet är 0.1, och därefter är sannolikheten att den andre hittar det 0.9, och att den tredje hittar det också 0.9. Sammanlagt är sannolikheten för att den första missar stavfelet och de andra hittar det $0.1 \cdot 0.9^2$. Motsvarande beräkningar ger samma sannolikhet för att exakt den andre personen missar och de andra hittar, och att exakt den tredje personen missar och de andra hittar, så sammanlagt får vi att sannolikheten för att *precis* en person missar är

$$p_X(1) = 3 \cdot 0.1 \cdot 0.9^2.$$

Slutligen kan vi använda egenskaperna för sannolikhetsfunktioner för att härleda att

$$p_X(2) = 1 - p_X(0) - p_X(1) - p_X(3) = 1 - 0.9^3 - 3 \cdot 0.1 \cdot 0.9^2 - 0.1^3.$$

▲

Denna typ av sannolikhetsfördelning är väldigt vanligt förekommande och kallas en *binomialfördelning*. De uppkommer i situationer som den ovan, alltså då vi utför ett experiment ett visst antal gånger, n, och experimentet lyckas med någon sannolikhet p, och misslyckas med någon sannolikhet $1 - p$ och vi därefter undrar vad sannolikheten är att exakt k av de n utförda experimenten lyckas.

5.2 Fördelningsfunktioner

Ofta är man intresserad av att beräkna uttryck som

$$P(a < X < b), P(X \leqslant a), \text{ och } P(X > a).$$

Om man till exempel vill beräkna sannolikheten att summan av fem tärningsslag är större än 25 är man intresserad av att beräkna $P(X > 25)$ för någon lämplig slumpvariabel X.

För en diskret slumpvariabel såg vi tidigare att vi kan använda oss av sannolikhetsfunktionen för att få ut information av den typen och för att få ut information om slumpvariabeln i allmänhet. För en kontinuerlig sannolikhetsfunktion är ofta sannolikheten att få *exakt* ett specifikt värde k helt obefintlig, och vi kan därför inte prata om en sannolikhetsfunktion, $P(X = k)$, på ett meningsfullt sätt. Vi introducerar istället följande funktion som ger relevant information för både diskreta och kontinuerliga slumpvariabler.

Definition 5.2.1. *Fördelningsfunktionen* $F_X(t)$ för en slumpvariabel X definieras av

$$F_X(t) := P(X \leqslant t), -\infty < t < \infty.$$

△

Notera att för en diskret slumpvariabel gäller $F_X(t) = \sum_{x \leqslant t} p_X(x)$.

Summor för diskreta objekt motsvaras av integraler för kontinuerliga objekt. Med detta som motivation introducerar vi följande begrepp som är kontinuerliga motsvarigheter till diskreta slumpvariabler och deras sannolikhetsfunktioner.

Definition 5.2.2. En slumpvariabel X sägs vara *kontinuerlig* om det finns en funktion $f_X(t)$ så att det för ”alla” mängder A gäller att

$$P(X \in A) = \int_A f_X(t)dt.$$

Funktionen f_X kallas för slumpvariabelns *täthetsfunktion.* △

Notera att definitionen ovan inte är helt entydig, vi kan till exempel alltid ändra funktionens värde i någon enstaka punkt utan att det påverkar integralen till höger. Det är underförstått att vi alltid väljer ”den mest kontinuerliga” täthetsfunktionen.

Vidare är det så att till och med för väldigt snälla täthetsfunktioner är inte integralen till höger nödvändigtvis väldefinierad. Till exempel kan man betrakta täthetsfunktionen $f_X(t)$ som är 1 om $0 \leqslant t \leqslant 1$, och 0 annars. Den här funktionen fördelar all sannolikhet på intervallet $[0, 1]$, och $P(x \in A)$ är direkt proportionerlig till storleken på A. Det finns dock delmängder av $S \subset [0, 1]$ för vilka

$$\int_S f_X(t)dt = \int_S 1dt$$

inte är väldefinierat. Det vanligaste exemplet är mängden av rationella eller irrationella tal mellan 0 och 1. Pröva att beräkna integralen ovan med hjälp av definitionen för en integral och se vad som går fel!

Definitionen fungerar dock bra för mängder A som kan beskrivas som unioner av öppna intervall, vilket är tillräckligt för våra ändamål.

För en kontinuerlig slumpvariabel är uppenbarligen täthetsfunktionen och fördelningsfunktionen relaterade. Sambandet mellan dem beskrivs i följande sats.

Sats 5.2.3. *För en kontinuerlig slumpvariabel X med täthetsfunktion f_X och fördelningsfunktion F_X gäller*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

och omvänt gäller det att

$$f_X(x) = F'_X(x) = \lim_{h\to 0} \frac{F_X(x+h) - F_X(x)}{h}$$

för de punkter där $f_X(x)$ är kontinuerlig.

Bevis. Det första påståendet följer direkt från definitionerna av fördelningsfunktionen respektive täthetsfunktionen.

Det andra påståendet är en direkt konsekvens av det första påståendet och analysens fundamentalsats. □

Notera även att

$$\lim_{t\to\infty} \int_{-\infty}^{t} f_X(t)dt = \lim_{t\to\infty} F_X(t) = \lim_{t\to\infty} P\left(u \in \Omega : X(u) \in (-\infty, t]\right) = 1.$$

Man kan faktiskt visa att varje funktion $f(x)$ som uppfyller att $f(x) \geqslant 0$ och att $\int_{-\infty}^{\infty} f(x)dx = 1$ motsvarar en slumpvariabel X genom att man *definierar* X utifrån likheten

$$P(X \in A) := \int_A f(x)dx.$$

Exempel 5.2.4. Låt $a > 0$ vara ett reellt tal och låt X vara en slumpvariabel med täthetsfunktion $f_X(t) = ae^{-at}$ för $t > 0$ och $f_X(t) = 0$ för $t \leqslant 0$.

En sådan slumpvariabel sägs vara *exponentialfördelad*. Exponentialfördelningar uppkommer i situationer då man mäter tiden fram tills att något inträffar, och denna sak har samma sannolikhet att inträffa i varje givet ögonblick. Till exempel är tiden till att nästa person kommer in genom dörren på ett kafé (under öppettider och någon relativt fix tid på dagen, säg 10:00 - 10:30) exponentialfördelad för någon parameter a, och samma sak gäller för hur lång tid det tar innan en radioaktiv partikel sönderfaller.

För en exponentialfördelad slumpvariabel kan vi till exempel räkna ut att fördelningsfunktionen ges av

$$F_X(x) = \int_0^x ae^{-at}dt = [-e^{-at}]_0^x = 1 - e^{-ax}.$$

Härifrån är det tydligt att $\lim_{x\to\infty} F_X(x) = 1$, vilket är tur eftersom det var ett villkor för att F_X ska vara en riktig fördelningsfunktion till att börja med.

Om vi nu till exempel tänker oss att X är en slumpvariabel som beskriver tiden det tar innan nästa person går in genom dörren på ett kafé, och täthetsfunktionen för X ges av $f_X(t) = 2e^{-2t}$ där t mäter tiden i timmar, så säger formeln ovan att sannolikheten att nästa person har kommit in på kaféet inom en halvtimme är $F_X(1/2) = 1 - e^{-2\cdot(1/2)} \approx 0.63$, och sannolikheten att nästa person har kommit in inom en timme är $F_X(1) = 1 - e^{-2} \approx 0.86$.

▲

5.3 Väntevärde

Man vill ofta uttala sig om storleken av en slumpvariabel. Det finns givetvis massor med sätt att göra detta på, man kan till exempel kolla på det största värdet som möjligtvis kan antas, eller det minsta värdet som kan antas, men det vanligaste måttet för en slumpvariabels storlek är *väntevärdet*.

Väntevärdet är på många sätt den stokastiska versionen av medelvärde; man beräknar summan respektive integralen av alla möjliga utfall viktat med hur sannolikt det är att slumpvariabeln antar varje utfall. När man mäter medelvärdet av någonting så sker själva viktningen implicit genom att vissa utfall är mer vanligt förekommande i datamängden än andra. Om man till exempel mäter medellängden av alla män i Sverige, och det är dubbelt så många som är mellan 175 och 185 än som är mellan 185 och 195, så får den första gruppen ”dubbelt så mycket vikt” i beräkningen av medelvärdet, helt enkelt genom att den gruppen är dubbelt så vanligt förekommande.

Definition 5.3.1. *Väntevärdet* för en slumpvariabel X betecknas med $E(X)$, μ_X, eller bara μ om det inte kan förväxlas med andra väntevärden. För en diskret slumpvariabel definieras väntevärdet genom summan

$$E(X) := \sum_k k \cdot p_X(k),$$

och för en kontinuerlig slumpvariabel genom integralen

$$E(X) := \int_{-\infty}^{\infty} x \cdot f_X(x)dx.$$

△

Notera att definitionen bara gäller om summan eller integralen ovan är väldefinierad. Om integralen eller summan ovan är oändlig säger vi att X saknar väntevärde.

Notera även att väntevärdet för en slumpvariabel är ett fixt reellt tal, till skillnad från själva slumpvariabeln som är en funktion på utfallsrummet och kan anta olika värden.

Exempel 5.3.2. Betrakta återigen exemplet med att vi slår en tärning och låt X vara slumpvariabeln som ger oss talet som kommer upp när vi slagit tärningen. Eftersom alla utfall är lika sannolika, det vill säga $p_X(j) = P(X = j) = 1/6$ för $j = 1, 2, \ldots, 6$, ges väntevärdet av

$$E(X) = \sum_{k=1}^{6} kp_X(k) = \sum_{k=1}^{6} \frac{k}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$$

▲

Exempel 5.3.3. Betrakta återigen exemplet där X är en exponentialfördelad slumpvariabel, alltså att x har täthetsfunktion $f_X(t) = ae^{-at}$ för något $a > 0$ om $t > 0$ och $f_X(t) = 0$ om $t \leqslant 0$. Väntevärdet för X ges då av

$$E(X) = \int_0^{\infty} tae^{-at}dt = \left[-te^{-at} - \frac{e^{-at}}{a}\right]_0^{\infty} = \frac{1}{a}.$$

▲

5.4 Varians och standardavvikelse

Väntevärdet beskriver, som namnet antyder, det förväntade utfallet av en slumpvariabel. Detta är dock bara *en* siffra, och ger givetvis inte all relevant information om en slumpvariabel. En annan kvantitet som ofta är av intresse är *variansen*, vilket är ett mått på hur mycket spridning slumpvariabeln har. Betrakta till exempel två grupper med sex personer i varje, i den ena gruppen är personerna $100, 200, 120, 180, 190$ och 110, cm långa, och i den andra gruppen är alla personerna 150 cm långa. De båda grupperna har samma medellängd, men i den ena gruppen är fördelningen av längder uppenbarligen mycket mer (helt) koncentrerad kring medelvärdet. Variansen kvantifierar den förväntade avvikelsen från medelvärdet för en slumpvariabel.

Definition 5.4.1. *Variansen* $V(X)$ för en slumpvariabel X med väntevärde μ definieras som $V(X) := E((X-\mu)^2)$ om detta uttryck är ändligt. Variansen betecknas ofta med σ^2 eller σ_μ^2. △

Om X med 100 procents sannolikhet antar sitt medelvärde blir variansen 0, medan om X med stor sannolikhet hamnar väldigt långt ifrån sitt medelvärde blir variansen stor, varav detta ger ett mått på spridningen.

Notera även att det inte kan ske någon kancellering av termer i beräkningen av väntevärdet i formeln då alla termer är positiva eftersom vi betraktar kvadraten av avvikelsen från väntevärdet, $(X-\mu)^2$.

Exempel 5.4.2. Som vi tidigare har beräknat är väntevärdet för ett tärningsslag 3.5. Eftersom $p_X(k) = 1/6$ för $k = 1, \ldots, 6$ ges variansen av

$$E((X-3.5)^2) = \sum_{k=1}^{6}(k-3.5)^2 p_X(k) = \frac{1}{6}\sum_{k=1}^{6}(k-3.5)^2 = \frac{1}{6}\frac{35}{2} = \frac{35}{12}.$$

▲

Att beräkna varians utifrån formeln är dock ofta onödigt krångligt. Istället använder man ofta följande formel.

Sats 5.4.3. *För en slumpvariabel X med väntevärde μ gäller*

$$V(X) = E(X^2) - \mu^2 = E(X^2) - E(X)^2.$$

Bevis. Vi visar formlen i fallet då X är en kontinuerlig slumpvariabel, och väntevärdet alltså ges av en integral. Fallet då X är en diskret slumpvariabel visas på samma sätt.

Från definitionen av varians har vi att

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty}(x-\mu)^2 f_X(x)dx \\ &= \int_{-\infty}^{\infty}(x^2 - 2x\mu + \mu^2) f_X(x)dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x)dx - 2\mu\int_{-\infty}^{\infty} x f_X(x)dx + \mu^2\int_{-\infty}^{\infty} f_X(x)dx. \end{aligned}$$

Den första termen är definitionsmässigt $E(X^2)$, och integralen i den andra termen är definitionsmässigt $E(X) = \mu$. Den andra termen blir alltså $-2\mu^2$, och eftersom

$$\int_{-\infty}^{\infty} f_X(x)dx = 1,$$

blir den tredje termen μ^2. Tillsammans får vi alltså att

$$V(X) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2.$$

□

Ett betydigt bättre och mer använt mått på spridningen än variansen är *standardavvikelsen*. Man behöver dock räkna ut variansen för att kunna räkna ut standardavvikelsen, varav vi definerade variansen först.

Definition 5.4.4. *Standardavvikelsen* $D(X)$ för en slumpvariabel X definieras som

$$D(X) := \sqrt{V(X)}.$$

Standardavvikelsen betecknas ofta med σ eller σ_X. △

Notera att standardavvikelsen, precis som variansen och väntevärdet är ett reellt tal, återigen till skillnad från slumpvariabeln som är en funktion och kan anta olika värden vid skilda observationer.

Huvudanledningen till varför standardavvikelsen är ett bättre mått än variansen är att standardavvikelsen, precis som väntevärdet men till skillnad från variansen, har samma enhet som slumpvariabeln själv. Om slumpvariabeln till exempel ger längden på fotbollsspelare i meter kommer väntevärdet och standardavvikelsen också vara uttryck i meter, medan variansen ger ett uttryck i kvadratmeter eftersom variansen är väntevärdet av avvikelsen från väntevärdet *i kvadrat*.

5.5 Normalfördelning

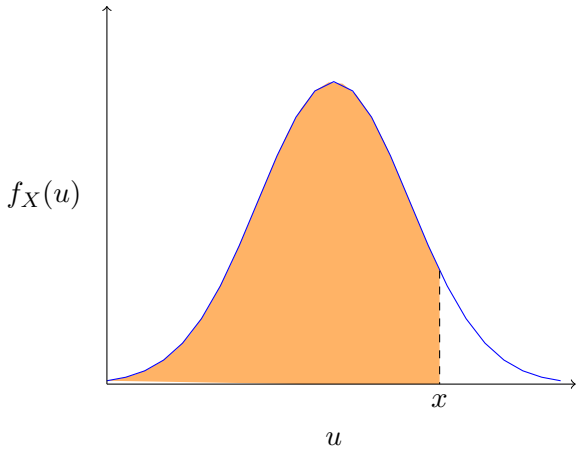
Vi avslutar det här kapitlet och vår korta genomgång av sannolikhetsteori med att introducera den kanske viktigaste fördelningen av alla, nämligen *normalfördelningen*.

Anledningen till detta är en oerhört viktig sats som kallas *centrala gränsvärdessatsen*. Den säger att om man adderar ett stort antal oberoende slumpvariabler, eventuellt med olika fördelningsfunktioner men med ändliga varianser, så kommer resultatet att vara normalfördelat. Detta gör att normalfördelningar uppkommer hela tiden, i alla möjliga olika sammanhang. Det enda som krävs är att man betraktar en kumulativ effekt av något som sker tillräckligt ofta.

Definition 5.5.1. En kontinuerlig slumpvariabel X sägs vara *normalfördelad* med parametrar μ och σ^2 om täthetsfunktionen ges av

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

För sådana X används ofta beteckningen $X \sim N(\mu, \sigma^2)$. △



Figur 5.1: Bilden visar en normalfördelning, och arean av det orangefärgade området ger värdet av fördelningsfunktionen evaluerad i punkten x.

Som valet av symboler för parametrarna antyder har $X \sim N(\mu, \sigma^2)$ väntevärde μ, varians σ^2, och standardavvikelse σ. Att visa detta kräver dock att man kan beräkna

$$\int_{-\infty}^{\infty} e^{-x^2} dx,$$

vilket görs genom ett smart trick och formeln för variabelbyte för funktioner av flera variabler, vilket vi tyvärr inte hinner gå igenom i den här kursen.

Övningar

Övning 5.1 ($\star$)**.** Betrakta slumpexperimentet att vi slår två vanliga tärningar och betrakta slumpvariabeln

$$X : \Omega \to \mathbb{N}, \quad X : (x, y) \to x^2 + y.$$

Beräkna $p_X(34)$ och $F_X(34)$.

Övning 5.2 ($\star$)**.** Betrakta slumpexperimentet att vi slår två vanliga tärningar och betrakta slumpvariabeln

$$X : \Omega \to \mathbb{N}, \quad X : (x, y) \to xy.$$

Beräkna $p_X(33)$ och $F_X(33)$.

Övning 5.3 ($\star\star\star$)**.** Låt X och Y vara två slumpvariabler. Bevisa att väntevärdet av $X + Y$ ges av $E(X) + E(Y)$.

Övning 5.4 ($\star$)**.** Betrakta slumpexperimentet att vi slår två vanliga tärningar och betrakta slumpvariabeln

$$X : \Omega \to \mathbb{N}, \quad X : (x, y) \to x + y.$$

Beräkna sannolikhetsfunktionen av X.

Övning 5.5 ($\star$)**.** Beräkna väntevärdet för summan av två tärningsslag med vanliga tärningar.

Övning 5.6 ($\star\star\star$)**.** Låt X och Y vara två *oberoende* slumpvariabler. Visa att $E(XY) = E(X)E(Y)$.

Övning 5.7 ($\star$)**.** Betrakta ett (slump)experiment som har sannolikhet p att lyckas, och därmed sannolikhet $1-p$ att misslyckas. Låt X vara slumpvariabeln som ger 1 om experimentet lyckas, och 0 annars.

Vad är väntevärdet för X? Vad är variansen för X?

Övning 5.8 ($\star\star$)**.** Betrakta ett (slump)experiment som har sannolikhet p att lyckas, och därmed sannolikhet $1 - p$ att misslyckas. Antag nu att vi utför experimentet n gånger, och att resultatet av varje utförande är oberoende av de andra resultaten. Låt X_j vara slumpvariabeln som ger 1 om experiment j lyckas, och 0 annars.

Vad är väntevärdet av $\sum_{j=1}^{n} X_j$? Vad är väntevärdet av $\prod_{j=1}^{n} X_j$?

Övning 5.9 ($\star\star$)**.** Låt X vara en slumpvariabel vars utfall ligger i intervallet (a, b) där $a, b \in \mathbb{R}$ och $a < b$. Vi säger att X är *likformigt* fördelad om alla intervall av samma längd har samma sannolikhet. Man kan visa att detta innebär att

$$P(X \in A) = \frac{\text{längd}(A)}{b - a},$$

för någorlunda naturliga delmängder $A \subset (a, b)$ (det måste till exempel gå att mäta längden av delmängden).

(a) Hitta fördelningsfunktionen och täthetsfunktionen för X.

(b) Beräkna väntevärdet för X.

(c) Beräkna variansen för X.

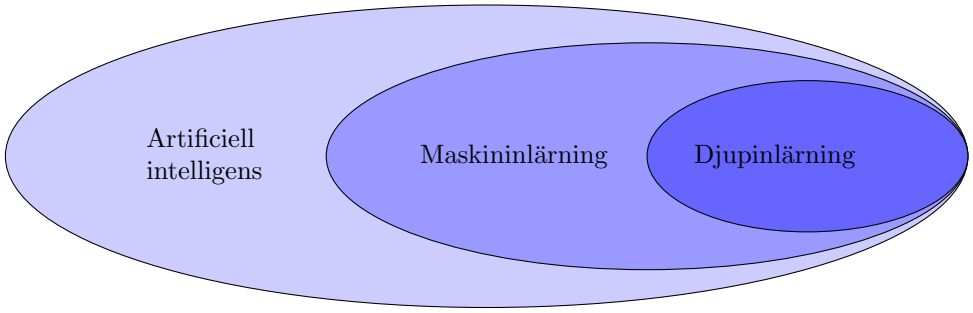
6 Maskininlärning

6.1 Vad är maskininlärning?

Med data kan man lära datorer att själva upptäcka och lära sig regler för att lösa en uppgift, utan att programmera datorerna med regler för just den specifika uppgiften i fråga. Detta kallas för maskininlärning.

Maskininlärning är ett område inom artificiell intelligens och används till exempel inom robotik, programvaruutveckling, självkörande bilar, medicinsk diagnostik och strömningstjänster på internet.

Eftersom samhället blir mer och mer digitaliserat har vi samlat mer och mer data. Det har varit svårt att hantera all data men maskininlärning kan hjälpa oss att bättre analysera och förstå vår data. I denna kurs ska vi diskutera djupinlärning som är en del av området maskininlärning (se Figur 6.1).



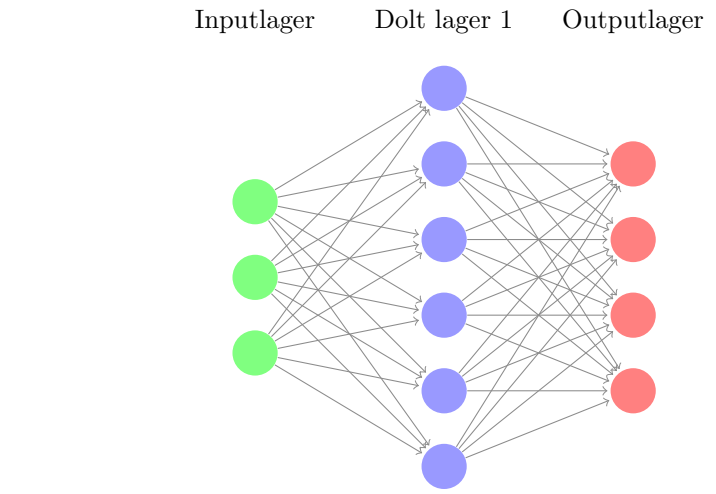
Figur 6.1: Artificiell intelligens, maskininlärning och djupinlärning

6.2 Vad är djupinlärning?

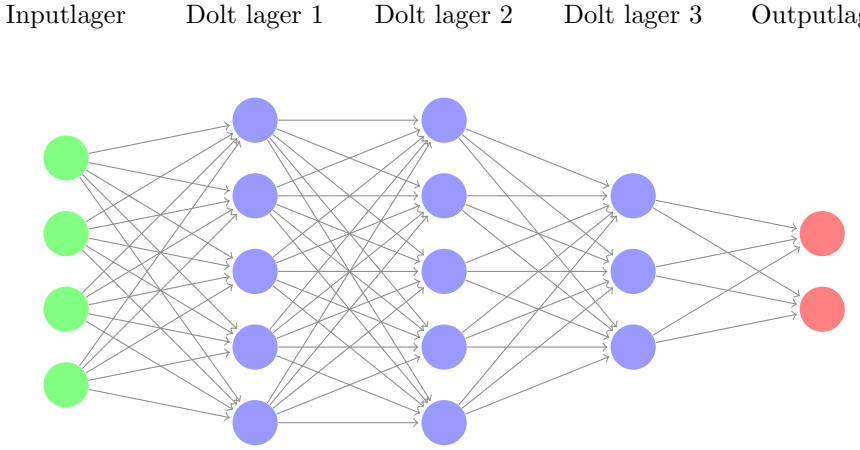
Djupinlärning är inspirerad av hjärnans struktur och skapar representationer i flera steg som kallas **lager**. I Figur 6.2 ser man **ett djupt neuralt nätverk** med ett **inputlager**, ett **dolt lager** och ett **outputlager** och i Figur 6.3 ser man ett djupt neuralt nätverk med ett inputlager, tre dolda lager och ett outputlager. Var och en av cirklarna kallas en **nod**. Noder representerar informationen som flyter in i nätverket.

Poängen med allt det här är att skapa bättre representationer för varje lager. Djupa neurala nätverket förvandlar data som matas in (**input**) till andra representationer som blir mer informativa (**output**). Generellt kan man säga att nätverket filtrerar informationen så att endast de användbara och karaktäristiska dragen är kvar i det sista lagret. Man kan använda flera hundra lager med hundratals noder i varje lager om man har en dator som kan klara av det.

I slutet av kursen ska ni förstå grunden till djupinlärning och använda det för att implementera maskininlärning. Mer specifikt kommer ni att kunna approximera en funktion från data med hjälp av artificiell intelligens.



Figur 6.2: Ett djupt neuralt nätverk med ett dolt lager



Figur 6.3: Ett djupt neuralt nätverk med tre dolda lager

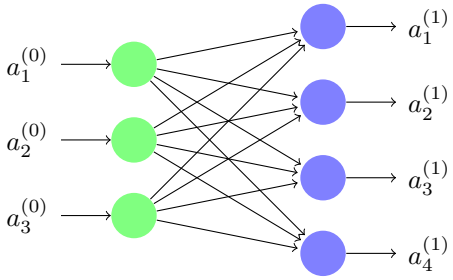
6.3 Ett djupt neuralt nätverk

Varje lager tar emot information från föregående lager i form av flera tal och därefter gör beräkningar med talen. Talen representerar någon typ av information. De nya talen skickas vidare till nästa lager. Vi säger att **ingångsnoderna** tar **ingångsvärden**. Ingångsnoderna är de gröna noderna i Figur 6.2. De kan t.ex. vara en binär 1 eller 0, en del av ett RGB-färgvärde m.m. Talen som kommer ut från modellen heter **utgångsnoderna**. Utgångsnoderna är de röda noderna i Figur 6.2.

6.3.1 Att beräkna det andra lagret

Vi illustrerar hur man beräknar noderna i det andra lagret i Figur 6.4 med det följande exemplet. Vi noterar att det finns andra sätt att arrangera modellen än det som följer. Poängen är att parametrarna är kopplade till varandra, som kan ses i en mängd olika möjliga modeller. Vi har valt ett enkelt sätt att börja

med i detta kapitel. Kapitel 7 visar en mer komplicerad modell.



Figur 6.4: Beräkna andra lagret

Exempel 6.3.1. Varje nod håller ett tal. Vi vill först beräkna nod $a_1^{(1)}$ i Figur 6.4. Noden

$$a_1^{(1)}$$

är första noden (subscript) i det första dolda lagret (superscript). Noder i inputlagret har superscript noll, d.v.s.

$$a_j^{(0)},\ j = 1, \ldots, n_0$$

där $n_0 = 3$ är det totala antalet noder i inputlagret. Vi beräknar $a_1^{(1)}$ som

$$a_1^{(1)} = \sigma(w_{1,1}^{(0)} a_1^{(0)} + w_{1,2}^{(0)} a_2^{(0)} + w_{1,3}^{(0)} a_3^{(0)} + b_1^{(0)}) \tag{6.1}$$

med **vikten** $w_{1,j}^{(0)}$, $j = 1, 2, 3$, **bias** $b_1^{(0)}$ och

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{6.2}$$

I allmänhet har vi $w_{i,j}^{(k)}$, $a_j^{(k)}$ och $b_i^{(k)}$, där

$$i = 1, \ldots, n_{k+1},$$

$$j = 1, \ldots, n_k$$

och

$$k = 0, 1, \ldots, \ell, \ell + 1.$$

Variabeln k representerar vilket lager vi började på, ℓ är det totala antalet dolda lager i det djupa neurala nätverket och n_k representerar antalet noder i det k:te lagret av nätverket. ▲

I ekvation (6.1) har vi en **aktiveringsfunktion** σ. Syftet med aktiveringsfunktionen är att omvandla slutresultatet till ett tal som är lättare att hantera. I det här exemplet har vi en sigmoidal aktiveringsfunktion som definieras i (6.2). I Figur 6.5 ser vi även en graf av σ.

En sigmoidal aktiveringsfunktionen omvandlar alla inkommande värden till ett tal mellan 0 och 1. Denna aktiveringsfunktion tar stora positiva tal till ett tal nära 1 och stora negativa tal till ett tal nära 0. En sigmoidal aktiveringsfunktion är inte det enda möjliga valet för en aktiveringsfunktion.

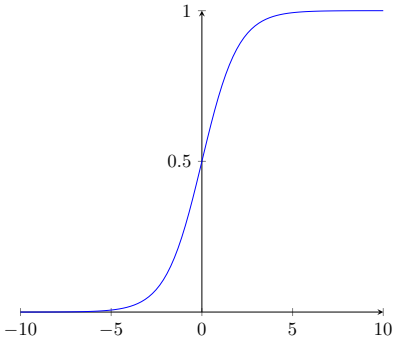
Vi noterar att n_k kan varierar beroende på vilket lager vi är i, t.ex. i Figur 6.2 har vi $n_0 = 3$ i inputlagret, $n_1 = 6$ i det dolda lagret och $n_2 = 4$ i outputlagret. Här har vi $\ell = 1$ eftersom vi bara har ett dolt lager.

6.3.2 Mer om noder, vikter och bias

Mer allmänt är vikter tal som är koefficienter och bias är tal. Med andra ord multipliceras noderna med vikter och därefter summeras alla input och ett bias adderas (se Figur 6.4 där var och en av linjerna representerar en vikt). Bias är precis som vikterna ytterligare en parameter som vi beräknar. Genom att välja bias kan man öka eller minska en nods värde.

Vikter och bias är de parametrar vi justerar när vi tränar modellen. Det är viktigt att välja bra parametrar. Att ha tränat modellen väl betyder att vi har valt vikter och bias så att det neurala nätverket klarar av uppgiften den utvecklats för. I Avsnitt 6.3.4 kommer vi att ge en mer rigorös definition av en väl fungerande modell och i Avsnitt 6.3.5 diskuterar vi hur träningen går till.

Det krävs mycket jobb för att hitta de rätta vikterna och bias eftersom neurala nätverk kan innehålla miljoner vikter och bias. Dessutom är det ännu svårare att hitta dem eftersom en justering i en enda vikt kan påverka hela neuronnätet. Den stora utmaningen är alltså att justera alla dessa vikter så att resultatet blir optimalt. Vi illustrerar hur många parametrar vi har och hur de påverkar varandra med de två exemplen som följer.



Figur 6.5: Aktiveringsfunktion σ, s.k. sigmoidal

Exempel 6.3.2. Nedanför beräknar vi alla noder i det första dolda lagret i Figur 6.4. Vi skriver beräkningar med en matris och vektorer som gör det lättare att organiserar alla vikter och bias. Det blir också lättare att analysera hur många parametrar vi måste träna i ett visst problem om vi skriver så här.

$$\begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \\ a_4^{(1)} \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} \end{bmatrix} \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \end{bmatrix} + \begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \\ b_4^{(0)} \end{bmatrix} \right) \tag{6.3}$$

$$\iff \begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \\ a_4^{(1)} \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{1,1}^{(0)} a_1^{(0)} + w_{1,2}^{(0)} a_2^{(0)} + w_{1,3}^{(0)} a_3^{(0)} + b_1^{(0)} \\ w_{2,1}^{(0)} a_1^{(0)} + w_{2,2}^{(0)} a_2^{(0)} + w_{2,3}^{(0)} a_3^{(0)} + b_2^{(0)} \\ w_{3,1}^{(0)} a_1^{(0)} + w_{3,2}^{(0)} a_2^{(0)} + w_{3,3}^{(0)} a_3^{(0)} + b_3^{(0)} \\ w_{4,1}^{(0)} a_1^{(0)} + w_{4,2}^{(0)} a_2^{(0)} + w_{4,3}^{(0)} a_3^{(0)} + b_4^{(0)} \end{bmatrix} \right). \tag{6.4}$$

Vi noterar att första raden i (6.4) motsvarar Exempel 6.3.1 och att ekvationerna (6.3) och (6.4) är ekvivalenta.[5] Dessutom evaluerar vi aktiveringsfunktionen elementvis, d.v.s.

$$\sigma\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} \sigma(x_1) \\ \sigma(x_2) \\ \vdots \\ \sigma(x_n) \end{bmatrix}.$$

För att beräkna alla noder i det första dola lagret i Figur 6.4, d.v.s.

$$a_i^{(1)},\ i = 1, \ldots, n_1,$$

där $n_1 = 4$, behöver vi

$$n_0 \times n_1 = 3 \times 4 = 12$$

vikter och $n_1 = 4$ bias. Här använder vi den sigmoidala aktiveringsfunktionen σ som i (6.2). ▲

Ännu mer generellt kan vi beräkna $a_1^{(1)}, \ldots, a_{n_1}^{(1)}$ som

$$\begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_{n_1}^{(1)} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & \cdots & w_{1,n_0}^{(0)} \\ w_{2,1}^{(0)} & w_{2,1}^{(0)} & \cdots & w_{2,n_0}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_1,1}^{(0)} & w_{n_1,2}^{(0)} & \cdots & w_{n_1,n_0}^{(0)} \end{bmatrix} \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_{n_0}^{(0)} \end{bmatrix} + \begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_{n_1}^{(0)} \end{bmatrix}\right),$$

där n_0 är det totala antalet noder i inputlagret och n_1 är det totala antalet noder i det första dolda lagret.

Exempel 6.3.3. Vi kan beräkna alla vikter och bias i det djupa neurala nätverket med tre dolda lager i Figur 6.3 på samma sätt. Vi delar upp det här i 4 olika steg som följer. Varje steg representerar anslutningen mellan två lager och vi beräknar från vänster till höger.

Vi noterar att vi börjar med ingångsnoderna och varje steg efter det första börjar med resultatet från det föregående lagret.

Steg 1: Inputlager till det första dolda lagret

$$\begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \\ a_4^{(1)} \\ a_5^{(1)} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} & w_{1,4}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} & w_{2,4}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} & w_{3,4}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} & w_{4,4}^{(0)} \\ w_{5,1}^{(0)} & w_{5,2}^{(0)} & w_{5,3}^{(0)} & w_{5,4}^{(0)} \end{bmatrix} \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \\ a_4^{(0)} \end{bmatrix} + \begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \\ b_4^{(0)} \\ b_5^{(0)} \end{bmatrix}\right)$$

[5] Multiplikation av två matriser A och B är möjlig då A är av typ $m \times n$ och B är av typ $n \times p$. Då är

$$A \cdot B = C = (c_{ij})_{m \times p}$$

där $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$, $i = 1, 2, \ldots, m$ och $j = 1, 2, \ldots, p$. Addition av två matriser A och B är möjlig då A och B är av samma dimension, och beräknas genom att addera elementen parvis.

Totalt antal parametrar att välja i Steg 1		
antal vikter	antal bias	totalt
$4 \times 5 = 20$	5	$20 + 5 = 25$

Steg 2: Det första dolda lagret till det andra dolda lagret

$$\begin{bmatrix} a_1^{(2)} \\ a_2^{(2)} \\ a_3^{(2)} \\ a_4^{(2)} \\ a_5^{(2)} \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} & w_{1,3}^{(1)} & w_{1,4}^{(1)} & w_{1,5}^{(1)} \\ w_{2,1}^{(1)} & w_{2,2}^{(1)} & w_{2,3}^{(1)} & w_{2,4}^{(1)} & w_{2,5}^{(1)} \\ w_{3,1}^{(1)} & w_{3,2}^{(1)} & w_{3,3}^{(1)} & w_{3,4}^{(1)} & w_{3,5}^{(1)} \\ w_{4,1}^{(1)} & w_{4,2}^{(1)} & w_{4,3}^{(1)} & w_{4,4}^{(1)} & w_{4,5}^{(1)} \\ w_{5,1}^{(1)} & w_{5,2}^{(1)} & w_{5,3}^{(1)} & w_{5,4}^{(1)} & w_{5,5}^{(1)} \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \\ a_4^{(1)} \\ a_5^{(1)} \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \\ b_4^{(1)} \\ b_5^{(1)} \end{bmatrix} \right)$$

Totalt antal parametrar att välja i Steg 2		
antal vikter	antal bias	totalt
$5 \times 5 = 25$	5	$25 + 5 = 30$

Steg 3: Det andra dolda lagret till det tredje dolda lagret

$$\begin{bmatrix} a_1^{(3)} \\ a_2^{(3)} \\ a_3^{(3)} \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{1,1}^{(2)} & w_{1,2}^{(2)} & w_{1,3}^{(2)} & w_{1,4}^{(2)} & w_{1,5}^{(2)} \\ w_{2,1}^{(2)} & w_{2,2}^{(2)} & w_{2,3}^{(2)} & w_{2,4}^{(2)} & w_{2,5}^{(2)} \\ w_{3,1}^{(2)} & w_{3,2}^{(2)} & w_{3,3}^{(2)} & w_{3,4}^{(2)} & w_{3,5}^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(2)} \\ a_2^{(2)} \\ a_3^{(2)} \\ a_4^{(2)} \\ a_5^{(2)} \end{bmatrix} + \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \\ b_3^{(2)} \end{bmatrix} \right)$$

Totalt antal parametrar att välja i Steg 3		
antal vikter	antal bias	totalt
$5 \times 3 = 15$	3	$15 + 3 = 18$

Steg 4: Det tredje dolda lagret till outputlagret

$$\begin{bmatrix} a_1^{(4)} \\ a_2^{(4)} \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{1,1}^{(3)} & w_{1,2}^{(3)} & w_{1,3}^{(3)} \\ w_{2,1}^{(3)} & w_{2,2}^{(3)} & w_{2,3}^{(3)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \\ a_2^{(3)} \\ a_3^{(3)} \end{bmatrix} + \begin{bmatrix} b_1^{(3)} \\ b_2^{(3)} \end{bmatrix} \right)$$

Totalt antal parametrar att välja i Steg 4		
antal vikter	antal bias	totalt
$3 \times 2 = 6$	2	$6 + 2 = 8$

Då har vi i totalt:

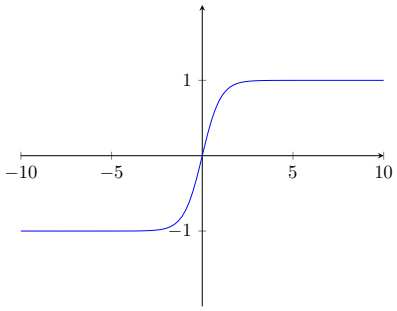
$$25 + 30 + 18 + 8 = 81$$

parametrar att välja. Vi ska senare diskutera vad det betyder att välja parametrar men, kort sagt, att välja alla parametrar är samma sak som att träna ett djupt neuralt nätverk. ▲

6.3.3 Andra aktiveringsfunktioner

I Exempel 6.3.1 och 6.3.2 tar vi en sigmoidal aktiveringsfunktion som i (6.2) men det är inte det enda valet. En annan möjlighet är tangens hyperbolicus (Figur 6.6) som ges av följande ekvation.

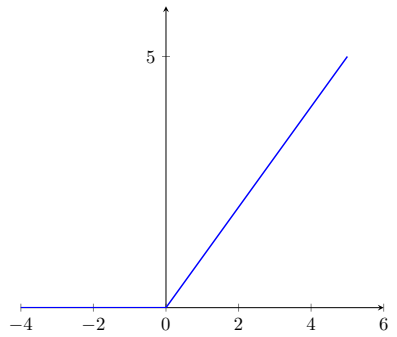
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{6.5}$$



Figur 6.6: Aktiveringsfunktion tanh som i (6.5), tangens hyperbolicus

Den mest använda aktiveringsfunktionen idag heter ReLU (eng: *Rectified Linear Unit*) och definieras som

$$R(x) = \max(0, x). \tag{6.6}$$



Figur 6.7: Aktiveringsfunktion R, s.k. ReLU

I Figur 6.7 ser vi en graf av ReLU som i (6.6). I många tillämpningar har ReLU visat sig vara bättre att träna modellen med än en sigmoidal.

Alla aktiveringsfunktioner har fortfarande samma syfte: att omvandla slutresultatet till ett tal som är lättare att hantera. Mer konkret ser vi t.ex. i Figur 6.6 att funktionen tanh tar in alla tal i $(-\infty, \infty)$ men ger tillbaka ett tal i det lilla intervallet $(-1, 1)$.

6.3.4 Förlustfunktionen

Vikterna justeras med hjälp av **förlustfunktionen** (eng: *loss function*). Förlustfunktionen beräknar hur stor skillnaden är mellan nätverkets output och sanningen genom att testa modellen på exempel där vi redan vet sanningen. Mer speficikt testar vi nätverket på ett exempel och definierar förlustfunktionen L som

$$L(X) = |Y(X) - \hat{Y}(X)|^2 \tag{6.7}$$

där

- X är ingångsnoderna på ett exampel
- $Y(X)$ är nätverkets gissning på X (utgångsnoderna)
- $\hat{Y}(X)$ är etiketten på exemplet (sanningen) och
- $|z|^2 = z_1^2 + \ldots + z_n^2$ är längd i kvadrat av vektorn $z \in \mathbb{R}^n$.

Inom maskininlärning tittar vi på många exempel när vi tränar modellen. Därför tittar vi på medelvärdet av (6.7) över N exempel, d.v.s.,

$$\frac{1}{N}\sum_{l=1}^{N}|Y(X_l) - \hat{Y}(X_l)|^2, \; X = \{X_l\}_{l=1}^{N} \tag{6.8}$$

där varje X_l i mängden $\{X_l\}_{l=1}^{N}$ representerar ett exempel.

Just den här förlustfunktionen i (6.8) heter **medelkvadratfel** (eng: *mean square error*). Den är inte den enda möjliga förlustfunktionen. Vi ska använda en förlustfunktion som heter **korsentropifunktionen** (eng: *cross entropy function*) när vi progammerar i Kapitel 7.

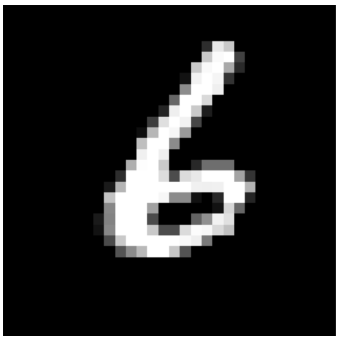
Om värdet på förlustfunktionen i (6.7) är ett stort värde på ett exempel X så är skillnaden mellan nätverkets gissning och sanningen stor. Om värdet på förlustfunktionen i (6.8) är stort på en mängd $X = \{X_l\}_{l=1}^{N}$ betyder det att det neurala nätverket fungerar dåligt. Om värdet är litet så har det neurala nätverket gjort många bra gissningar och modellen fungerar bra. Vi tar ett exempel för att gör detta konkret.

Exempel 6.3.4. Tänk att vår uppgift är att klassifiera handskriva siffror med ett djupt neuralt nätverk. (Se Övningar i Kapitel 7 där vi experimenterar med just det här.) Vi kan använda MNIST-databasen (eng: *Modified National Institute of Standards and Technology*) för att träna modellen. MNIST-databasen är en samling av sextio tusen små, kvadratiska gråskalebilder. Varje bild har 28×28 pixlar, med en enda handskriven siffra mellan 0 och 9 (se Figur 6.8 och Figur 6.9). Bilderna kommer med en etikett med den rätta siffran som en **enhetsvektor**.

Enhetsvektorn är sådan att ett element är 1 och alla andra element är 0. Elementet som är 1 motsvarar siffran som är på bilden. Mer specifikt låt X representerar en bild av siffran 6. Motsvarande etiketten är då

$$\hat{Y}(X) = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0\end{bmatrix}^T \in \mathbb{R}^{10}, \tag{6.9}$$

Figur 6.8: Handskrivna siffror från MNIST-databasen, bild från Wikipedia



Figur 6.9: En handskriven siffra från MNIST-databasen, bild från Stack Overflow

där T betecknar transponat.[6] Vi noterar att siffran 0 har etiketten

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \in \mathbb{R}^{10}.$$

MNIST-databasen är utmärkt för att träna neurala nätverk eftersom det finns så många märkta exempel. Först tar vi de

$$28 \times 28 = 784$$

pixlarna i en bild (se Figur 6.9). Varje pixel är ett värde som bestämmer pixelns färg. Här har vi att 0 representerar svart och 1 är vit och alla tal i intervallet $(0, 1)$ är nyanser av grått.

Vi skapar en lång vektor med 784 element från varje bild. Vektorn går in till det neurala nätverket som

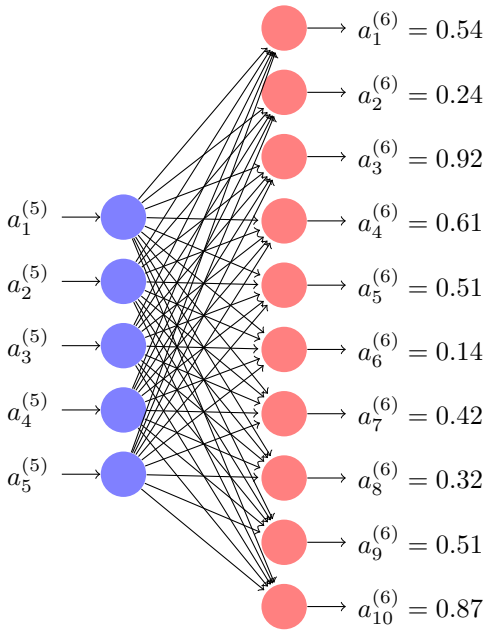
$$X = \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_{784}^{(0)} \end{bmatrix} \in \mathbb{R}^{784},$$

[6] Transponatet av en vektor är sådan att $\begin{bmatrix} a & b & c \end{bmatrix}^T = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$.

där bilden ges av matrisen

$$\begin{bmatrix} a_1^{(0)} & a_{29}^{(0)} & \cdots & a_{757}^{(0)} \\ a_2^{(0)} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{28}^{(0)} & \cdots & \cdots & a_{784}^{(0)} \end{bmatrix} \in \mathbb{R}^{28\times 28}$$

där $a_i^{(0)} \in [0,1]$, $i = 1, \ldots, 784$. Med andra ord består vektorn X av ingångsnoderna. Vi noterar att vektorn har all information vi behöver från bilden. Från



Figur 6.10: Det sista steget i klassificering av handskrivna siffror i ett djupt neuralt nätverk med 5 dolda lager

ingångsnoderna beräknar vi utgångsnoderna genom samma process som beskrevs innan. I detta fall har vi 10 utgångsnoder. Varje utgångsnod motsvarar en siffra mellan 0 och 9. Vi säger att siffran som motsvarar utgångsnoden med det största värdet är modellens klassificering.

I Figur 6.10 ser vi ett exempel av det sista steget i klassificering av handskrivna siffror i ett djupt neuralt nätverk med 5 dolda lager. Modellen gissar att exemplet är en tvåa eftersom utgångsnoden

$$a_3^{(6)} = 0.92$$

har det största värdet. Vi kollar hur bra nätverket har fungerat med följande analys.

Anta först att nätverket klassificerar exempelet X i Figur 6.10 rätt, d.v.s. att etiketten $\hat{Y}(X)$ ges av enhetsvektorn

$$\hat{Y}(X) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \in \mathbb{R}^{10}.$$

Då beräknar förlustfunktionens värdet på exemplet i Figur 6.10 som längden i kvadrat av vektorn

$$\begin{bmatrix}(0.54-0.00)\\(0.24-0.00)\\(0.92-1.00)\\(0.61-0.00)\\(0.51-0.00)\\(0.14-0.00)\\(0.42-0.00)\\(0.32-0.00)\\(0.51-0.00)\\(0.87-0.00)\end{bmatrix} \in \mathbb{R}^{10}, \tag{6.10}$$

precis som i (6.7).

Anta nu att nätverket klassificerar exempelet X i Figur 6.10 fel. I detta fall ges etiketten $\hat{Y}(X)$ av enhetsvektorn

$$\hat{Y}(X) = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}^T \in \mathbb{R}^{10},$$

d.v.s. siffra är en 9a i verkligheten men nätverket gissar att det är en 2a. Då beräknar vi förlustfunktionens värdet som längden i kvadrat av

$$\begin{bmatrix}(0.54-0.00)\\(0.24-0.00)\\(0.92-0.00)\\(0.61-0.00)\\(0.51-0.00)\\(0.14-0.00)\\(0.42-0.00)\\(0.32-0.00)\\(0.51-0.00)\\(0.87-1.00)\end{bmatrix} \in \mathbb{R}^{10}. \tag{6.11}$$

Förlustfunktionens värde av vektorn i (6.10) beräknas som följer:

$$\begin{aligned} L(X) = \quad & (0.54-0.00)^2 + (0.24-0.00)^2 + (0.92-1.00)^2 \\ + \quad & (0.61-0.00)^2 + (0.51-0.00)^2 + (0.14-0.00)^2 \\ + \quad & (0.42-0.00)^2 + (0.32-0.00)^2 + (0.51-0.00)^2 \\ + \quad & (0.87-0.00)^2 \end{aligned}$$

och

$$L(X) = 2.3032. \tag{6.12}$$

Förlustfunktionens värde av vektorn i (6.11) beräknas som

$$\begin{aligned} L(X) = \quad & (0.54-0.00)^2 + (0.24-0.00)^2 + (0.92-0.00)^2 \\ + \quad & (0.61-0.00)^2 + (0.51-0.00)^2 + (0.14-0.00)^2 \\ + \quad & (0.42-0.00)^2 + (0.32-0.00)^2 + (0.51-0.00)^2 \\ + \quad & (0.87-1.00)^2 \end{aligned}$$

och

$$L(X) = 2.4032. \tag{6.13}$$

Värdet i (6.12) är mindre än värdet i (6.13) eftersom nätverket klassificerade exemplet rätt. ▲

6.3.5 Att träna modellen

Eftersom förlustfunktionen visar hur mycket modellens beräkningar skiljer sig från verklighetens söker vi det **globala minimumet** i förlustfunktionen, d.v.s. vikter och bias sådan att förlustfunktion minimeras. Ett djupt neuralt nätverk kan ha tusentals vikter och bias. På grund av detta hittar vi i praktiken en approximation till vikterna och bias som minimerar förlustfunktionen. Mer specifikt ska vi använda **iterativa metoder** för att approximera vikterna och bias. Iterativa metoder börjar med en startgissning och genom successiva förändringar av densamma åstadkommer vi en successivt förbättrad approximation av lösningen till problemet.

Först, för enkelhetens skull, minimerar vi en funktion

$$f : \mathbb{R}^2 \to \mathbb{R}.$$

I Kapitel 7 minimerar vi en förlustfunktion för att träna ett neuralt nätverk.

Gradientnedstigning

Metoden **gradientnedstigning** är en iterativ metod för att lösa ett minimeringsproblem. Så med varje iteration av metoden får vi en bättre approximation av **minimipunkten**. Gradientnedstigning tar *steg* i motsatt riktning till gradienten av funktionen[7] eftersom funktionen minskar snabbast i den riktningen. Algoritmen för Gradientnedstigning ges i Algoritm 1.

Parametern γ i Algoritm 1 kallas för **inlärningshastighet**. I allmänhet tar vi $\gamma \in \mathbb{R}$ som en liten konstant. Gradientnedstigning kommer att stanna om den når fram (eller kommer väldigt nära) till ett minimum eftersom gradienten är noll (eller nästan noll) där (se Kapitel 3). Vi ser att algoritmen därmed inte gör så mycket i varje steg, eftersom det j:te steg ges av

$$\gamma \nabla f(x^{(j)}).$$

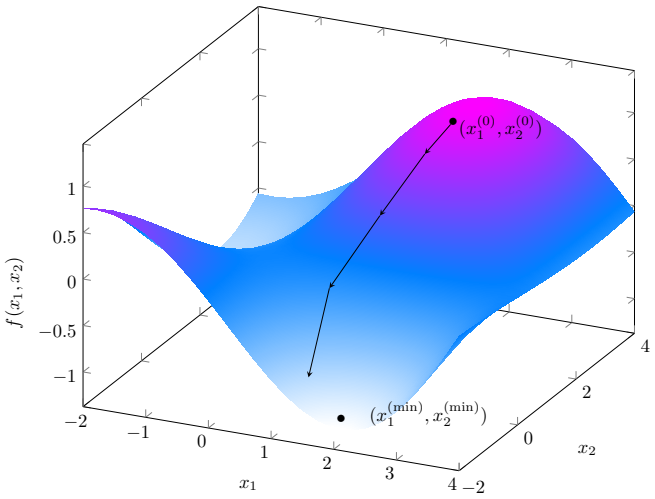
Ett exempel av gradientnedstigning för att hitta en minimipunkt på en funktion med 2 variabler finns i Figur 6.11. Här har vi

$$(x_1^{(0)}, x_2^{(0)}) = (2, 2)$$

som startgissning och funktionen f ges av

$$f(x_1, x_2) = \sin(0.8x_1)\sin(0.6x_2)e^{0.1x_1}$$

[7]Notera: ∇f betyder samma sak som grad$f$ i Kapitel 3



Figur 6.11: Fyra steg av metoden gradientnedstigning för att approximera en minimipunkt på en funktion $f : \mathbb{R}^2 \to \mathbb{R}$

Algoritm 1: Gradientnedstigning

---

input : Startgissning $(x_1^{(0)}, \ldots, x_k^{(0)}) \in \mathbb{R}^k$
Gradient av funktionen $f : \mathbb{R}^k \to \mathbb{R}$ sådan att $\nabla f \in \mathbb{R}^k$
$\gamma \in \mathbb{R}$
Parameter $m \in \mathbb{Z}$ (antal steg)

output: $(x_1^{(m+1)}, \ldots, x_k^{(m+1)}) \in \mathbb{R}^k$ sådant att
$f(x_1^{(m+1)}, \ldots, x_k^{(m+1)}) \leqslant f(x_1^{(0)}, \ldots, x_k^{(0)})$

1 **for** $j = 0, 1, \ldots, m$ **do**

2 $$\begin{bmatrix} x_1^{(j+1)} \\ \vdots \\ x_k^{(j+1)} \end{bmatrix} = \begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_k^{(j)} \end{bmatrix} - \gamma \nabla f \left(\begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_k^{(j)} \end{bmatrix} \right)$$

3 **end**

---

där

$$D_f = \{(x_1, x_2) : -2 \leqslant x_1 \leqslant 3, -2 \leqslant x_2 \leqslant 2\},$$

minimipunkten finns i

$$(x_1^{(\min)}, x_2^{(\min)}) \approx (2.11894, -2)$$

och

$$f(x_1^{(\min)}, x_2^{(\min)}) \approx -1.14312.$$

Vi ser i Figur 6.11 att med varje steg av gradientnedstigning kommer vi närmare minimipunkten och efter 4 steg är vi ganska nära vårt mål.

Det är omöjligt att visualisera minimeringsproblemet om dimensionen är större än 3, t.ex. problemet att minimera förlustfunktionen (6.8) där X är en mängd med N bilder som i Exempel 6.3.4. Dock är processen densamma och gradientnedstigning fungerar fortfarande.

Stokastisk gradientnedstigning

Metoden **stokastisk gradientnedstigning** är en iterativ metod som ofta används för att minimera förlustfunktionen i ett neuralt nätverk under träningens gång. Metoden är ganska lik gradientnedstigning. Skillnaden mellan de två metoderna är att metoden stokastisk gradientnedstigning approximerar gradienten istället för att beräkna den exakt. Det som vi förlorar är att vissa steg inte är så effektiva när det gäller att minimera förlustsfunktionen. I djupinlärning **konvergerar** metoden stokastisk gradientnedstigning ofta bättre än gradientnedstigning. Det betyder att metoden hittar en approximation på ett bättre sätt (t.ex. snabbare) eller hittar en approximation som är närmare sanningen. Vi kommer att studera denna metod närmare i de följande kapitlen.

Övningar

Övning 6.1 ($\star$)**.** Rita en bild av ett djupt neuralt nätverk med ett input lager med 2 noder, ett dolt lager med 4 noder och ett output lager med 3 noder.

Övning 6.2 ($\star$)**.** Rita en bild av följande neurala nätverk. Nätverket har

- ett input lager
- 2 dolda lager
- ett output lager.

Nätverket tar in 3 ingångsnoder och ger 3 utgångsnoder. Det finns 5 noder i det första dolda lagret och 4 noder i det andra dolda lagret.

Övning 6.3 ($\star\star$)**.** Titta igen på Exempel 6.3.2. Vi tar

$$\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} \end{bmatrix} = \begin{bmatrix} 10 & 20 & 30 \\ -10 & -1 & -18 \\ 40 & 20 & 50 \\ -5 & -10 & -3 \end{bmatrix}, \quad \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}$$

$$\begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \\ b_4^{(0)} \end{bmatrix} = \begin{bmatrix} -3 \\ -10 \\ 1 \\ -4 \end{bmatrix}$$

och aktiveringsfunktionen σ som i (6.2). Beräkna $\begin{bmatrix} a_1^{(1)} a_2^{(1)} a_3^{(1)} a_4^{(1)} \end{bmatrix}^T$.

Övning 6.4 ($\star\star$)**.** Titta igen på Exempel 6.3.2. Vi tar

$$\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} \end{bmatrix} = \begin{bmatrix} -10 & -20 & -30 \\ 10 & 1 & 18 \\ 40 & 20 & 50 \\ 5 & 10 & 3 \end{bmatrix}, \quad \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ -1 \end{bmatrix}$$

$$\begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \\ b_4^{(0)} \end{bmatrix} = \begin{bmatrix} -3 \\ 10 \\ 1 \\ -4 \end{bmatrix}$$

och aktiveringsfunktionen σ som i (6.2). Beräkna $\begin{bmatrix} a_1^{(1)} a_2^{(1)} a_3^{(1)} a_4^{(1)} \end{bmatrix}^T$.

Övning 6.5 ($\star$)**.** Titta igen på Exempel 6.3.3. Hur många parametrar måste vi beräkna för att träna modellen i Övning 6.1? Förklara ditt svar.

Övning 6.6 ($\star$)**.** Titta igen på Exempel 6.3.3. Hur många parametrar måste vi beräkna för att träna modellen i Övning 6.2? Förklara ditt svar.

Övning 6.7 ($\star\star\star$)**.** Vi har en funktion f sådant att

$$f(x) = 2x^2 - x + 2.$$

Använd Algoritm 1 (gradientnedstigning) för att approximerar minimipunkten. Ta $\gamma = 0.2$, $m = 5$ och $x^{(0)} = 2$.

Övning 6.8 ($\star\star\star$)**.** Vi har en funktion f sådan att

$$f(x) = 4x^2 + 2x + 2.$$

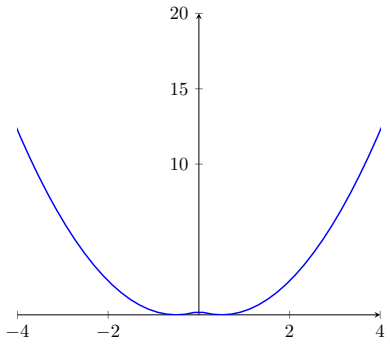
Använd Algoritm 1 (gradientnedstigning) för att approximerar minimipunkten. Ta $\gamma = 0.1$, $m = 6$ och $x^{(0)} = 0$.

7 Att approximera en funktion från datapunkter

Betrakta nu uppgiften att approximera en kvadratisk funktion f som ges av följande ekvation.

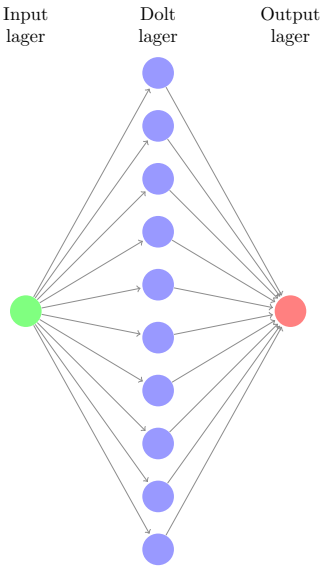
$$f(x) = \left| x - \frac{1}{2} \right|^2 \tag{7.1}$$

Denna funktion är mycket enkel. Tanken här är att vi kan använda de verktyg vi har studerat för att approximera denna funktion. Idag används maskininlärning för att lösa mycket mer komplicerade problem, men dessa problem kräver mycket mer avancerade verktyg än vi har tid för i denna kurs.



Figur 7.1: f som i (7.1)

En graf av funktionen i (7.1) finns i Figur 7.1. För att approximera (7.1) ska vi använda ett neuralt nätverk med ett dolt lager. Vi har en ingångsnod, 10 noder i det dolda lagret och en utgångsnod. Figur 7.2 visar det här neurala nätverket.



Figur 7.2: Ett djupt neuralt nätverk med ett dolt lager

7.1 Modellen och träningen

Vi approximerar (7.1) med en funktion $\alpha_{\boldsymbol{\theta}}$ som ges av

$$\alpha_{\boldsymbol{\theta}}(x) = \sum_{k=1}^{K} \tilde{w}_{1,k}^{(0)} \sigma\big(w_{1,k}^{(0)} x + b_k^{(0)}\big) \tag{7.2a}$$

$$\boldsymbol{\theta} = [\tilde{w}_{1,k}^{(0)}, w_{1,k}^{(0)}, b_k^{(0)}], k = 1, \ldots, K \tag{7.2b}$$

där

$$\boldsymbol{\theta} \in \mathbb{R}^{3\times K}.$$

Här har vi att x är ingångsnoden, σ är aktiveringsfunktion som i (6.2), och $\boldsymbol{\theta}$ representerar vikter och bias. Vi har $K = 10$ eftersom vi har 10 noder i det dolda lagret.

Vi noterar att (7.2a) är lite annorlunda än modellen som beskrevs i Kapitel 6 där

$$\tilde{w}_{1,k}^{(0)},\ k = 1, \ldots, K$$

saknades. Idén här är densamma även om modellen har fler parametrar.[8] Målet är att välja $\boldsymbol{\theta}$ för att minimera värdet av förlustfunktionen (medelkvadratfel) över N exempel, d.v.s., minimera

$$\frac{1}{N}\sum_{l=1}^{N} |\alpha_{\boldsymbol{\theta}}(X_l) - f(X_l)|^2,\ X = \{X_l\}_{l=1}^{N} \tag{7.3}$$

där $X_l, l = 1, \ldots, N$ i detta exempel är ett stickprov från

$$\mathcal{U}(-4, 4) \tag{7.4}$$

med $\alpha_{\boldsymbol{\theta}}$ och $\boldsymbol{\theta}$ som i (7.2). Ekvation (7.4) betyder att varje X_l väljs enligt en likformig fördelning på intervallet $(-4, 4)$, d.v.s. inget utfall är mer eller mindre sannolikt än något annat och alla utfall är större än -4 och mindre än 4 (se Övning 5.9).

Vi formulerar vårt problem som

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{3\times K}} \frac{1}{N}\sum_{l=1}^{N} |\alpha_{\boldsymbol{\theta}}(X_l) - f(X_l)|^2,\ X = \{X_l\}_{l=1}^{N}. \tag{7.5}$$

Med andra ord börjar vi med att slumpmässigt generera N datapunkter på intervallet $(-4, 4)$. Dessa datapunkter kallas **träningsdata** eftersom vi kommer att träna modellen på dem. Vi beräknar

$$f(X_l),\ l = 1, \ldots, N,$$

d.v.s. funktionsvärdena. Detta motsvarar märkning av träningsdata (etiketter). Därefter approximerar vi $\boldsymbol{\theta}$ som minimerar (7.5) med en iterativ metod. Detta är vad som kallas **träning**.

[8] Att implementera just den här modellen i Keras är faktiskt enklare. Keras tar hand om de extra parametrarna, d.v.s. det finns inte något mer vi behöver göra för att hantera dem. Vi ska diskutera det här mer i Kapitel 8 där vi diskuterar koden.

Algoritm 2: Approximera (7.5) med stokastisk gradientnedstigning

input : Startgissning $\boldsymbol{\theta}_0 \in \mathbb{R}^{3\times K}$
$X_1, \ldots, X_N$ stickprov från $\mathcal{U}(-4, 4)$
$\gamma \in \mathbb{R}$
Parameter $m \in \mathbb{Z}$ (antal steg)
output: $\boldsymbol{\theta}_{m+1}$ som approximerar (7.5)

Beräkna $f_l = f(X_l)$, $l = 1, \ldots, N$ där f är som i (7.1)
for $j = 0, 1, \ldots, m$ **do**
  Välj slumpmässigt $i \in \{1, \ldots, N\}$
  $\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \gamma \nabla_{\boldsymbol{\theta}} \left(|\alpha_{\boldsymbol{\theta}_j}(X_i) - f(X_i)| \right)^2$
end

Här använder vi metoden stokastisk gradientnedstigning som diskuterades i Kapitel 6 för att approximera $\boldsymbol{\theta}$. Approximationen betecknas $\boldsymbol{\theta}_{m+1}$ eftersom vi tar $(m+1)$ steg av stokasisk gradientnedstigning. Vi noterar att det som skiljer stokastisk gradientnedstigning från gradientnedstigning är att vi approximerar gradienten i varje iteration istället för att beräkna den exakt. Mer specifikt approximerar vi gradienten med information från träningsdatan. Vi sammanfattar hur man använder stokastisk gradientnedstigning för att approximera (7.5) i Algoritm 2.

7.2 Att evaluera ett neuralt nätverk

För att evaluera modellen som vi skapade i Algoritm 2 måste vi diskutera två olika sorters fel. Först har vi **träningsfel**. Träningsfelet är vad vi minimerar med stokastisk gradientnedstigning. Metoden tar in träningsdata

$$\{X_l, f(X_l)\}, l = 1, \ldots, N$$

och hittar $\boldsymbol{\theta}_{m+1}$ för att approximera (7.5). Då är träningsfelet

$$\frac{1}{N} \sum_{l=1}^{N} |\alpha_{\boldsymbol{\theta}_{m+1}}(X_l) - f(X_l)|^2.$$

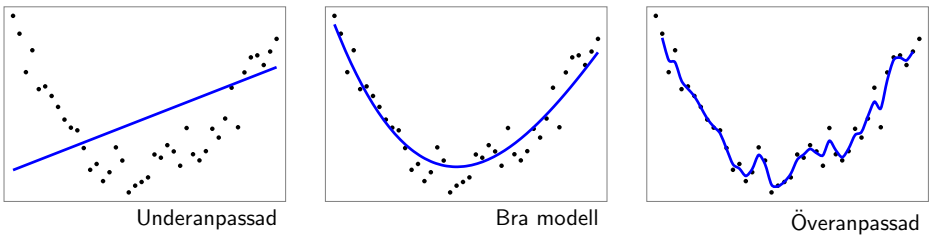
Vi beräknar också **generaliseringsfelet** som mäter hur stor förlustfunktionen är för data som modellen inte har tränat på. Mer specifikt genererar vi ett stickprov av $\tilde{N}$ punkter

$$\tilde{X}_1, \ldots, \tilde{X}_{\tilde{N}}$$

från $\mathcal{U}(-4, 4)$. Vi kallar det här stickprovet **testmängden** eftersom vi bara använder det för att testa modellen. Vi beräknar generaliseringsfelet på testmängden som

$$\frac{1}{\tilde{N}} \sum_{l=1}^{\tilde{N}} |\alpha_{\boldsymbol{\theta}_{m+1}}(\tilde{X}_l) - f(\tilde{X}_l)|^2.$$

Underanpassning innebär att en modell är för enkel och inte kan lära sig från träningsdata. Här kan man se ett stort värde på träningsfelet och ett



Figur 7.3: En datamängd, tre olika modeller

ännu större värde på generaliseringsfelet. **Överanpassning** innebär att en modell funkar bra på träningsdata men fungerar dåligt på osedd data, d.v.s. på testmängden. I Figur 7.3 ser vi tre olika modeller för en datamängd. Den på vänstra sidan visar underanpassning, den i mitten är en bra modell och den på högra sidan visar en modell som är överanpassad.

7.3 Andra viktiga begrepp inom djupinlärning

Att utbildas i djupinlärning innebär att man måste lära sig många nya begrepp. Här definierar vi några andra viktiga begrepp i området.

- **Övervakad inlärning**: modellen tränas genom att behandla en datamängd med etiketter.
- **Oövervakad inlärning**: modellens uppgift är att hitta mönster och avvikelser i en datamängd utan att ha fått instruktioner om vad som är rätt och vad som är fel.
- **Backpropagation**: en algoritm som beräknar gradienten vid träning av det neurala nätverket.
- **Framåtkopplade nätverk**: ett nätverk där informationen bara rör sig i en riktning (framåt) d.v.s. från ingångsnoderna, genom dolda noder (om sådana finns) och till utgångsnoderna.

Övningar

Övning 7.1 ($\star\star$)**.** Antag att vi vill approximera funktionen f som ges av

$$f(x) = \left| x - \frac{1}{2} \right|^2$$

med en funktion $\alpha_{\boldsymbol{\theta}}$ som ges av

$$\alpha_{\boldsymbol{\theta}}(x) = \sum_{k=1}^{K} \sigma\big(w_{1,k}^{(0)} x + b_k^{(0)}\big).$$

Vi har $K = 3$, d.v.s.,

$$\boldsymbol{\theta} = [w_{1,k}^{(0)}, b_k^{(0)}], k = 1, \ldots, 3$$

med

$$\begin{bmatrix} w_{1,1}^{(0)} \\ w_{1,2}^{(0)} \\ w_{1,3}^{(0)} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0.5 \end{bmatrix},$$

och

$$\begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix}.$$

Använd aktiveringsfunktionen σ som i (6.2).

Beräkna värdet av förlustfunktionen som i (6.8) där

$$X = \{X_l\}_{l=1}^{4} = \{0.5, 1, 1.25, 2\}.$$

Övning 7.2 ($\star\star$)**.** Nu ska vi jobba med att klassificera handskrivna siffror.

Läs instruktionerna för nedladdning av TensorFlow på kursens hemsida. Ladda också ner filen `ovning7.py` där vi har skrivit kod som ni kan använda.

Koden använder en annan förlustfunktionen än den som vi diskuterade i föreläsningen. Den heter *cross entropy* och ges av

$$H(y, z) = -\sum_{i=1}^{10} y_i \log(z_i), \ y, z \in \mathbb{R}^{10}$$

där y är etiketten och z är värdena på utgångsnoderna. I koden är det också en funktion som heter *softmax* för att förvandla utgångsnoderna till sannolikheter. Funktionen ges av

$$S(z) = \left(\frac{e^{z_1}}{\sum_{i=1}^{10} e^{z_i}}, \ldots, \frac{e^{z_{10}}}{\sum_{i=1}^{10} e^{z_i}} \right) \in \mathbb{R}^{10}$$

och summan av alla element i *softmax*-funktionen är 1. Vi vill minimera

$$E\left[H(y, S(z))\right]$$

med hjälp av vikten och bias.

I stället för den stokastiska gradientmetoden använder denna kod en optimerare som heter *Adam*. *Adam* är en populär algoritm inom djupinlärning eftersom den snabbt uppnår bra resultat. Vi kommer inte att ge en fullständig förklaring av *Adam* optimizer i den här kursen.

En epok (eng. *epoch*) är när en hel träningsdatamängd skickats igenom ett neuralt nätverk en gång.

Allt detta är redan skrivet i filen så det är inget du behöver ändra. Några frågor följer här.

- Titta på det första exemplet på träningsdata. Beskriv vad du ser.
- Hur många dolda lager finns i den här modellen?

- Hur många noder finns det i varje lager?
- Vad är värdet på förlustfunktionen efter 5 epoker?
- Vad får du för *accuracy* (noggrannhet) efter 5 epoker?
- Vad tror du kommer att hända om vi lägger till fler epoker?
- Vad händer om vi lägger till ett dolt lager?
- Gör några små ändringar i koden och kontrollera om de första tio försöken klassificerades korrekt eller inte.
- Sammanfatta det du har sett här.

8 Att använda Keras och TensorFlow

Vi jobbar med **Keras** i **TensorFlow** för att approximera funktionen f som i (7.1) med en funktion $\alpha_{\boldsymbol{\theta}}$ som i (7.2). Tanken här är att vi inte gör beräkningarna av alla vikterna, bias och noderna själva utan snarare använder befintlig programvara.

TensorFlow är ett programvarubibliotek av öppen källkod för maskininlärning. Det har utvecklats och används av Google. Keras är ett djupt lärande **API** (eng: *application programming interface*) skrivet i Python. API:er gör det möjligt att återanvända redan utvecklad och kvalitetssäkrad mjukvara som är i någon form av kodbibliotek. Några fördelar med keras är att det är enkelt, flexibelt och kraftfullt.

8.1 En guidad implementering

Vi börjar[9] med följande för att importera nödvändiga bibliotek.

```
import tensorflow as tf
from tensorflow import keras
import numpy as np
import matplotlib.pyplot as plt
```

Variabeln `N` motsvarar det totala antalet datapunkter (träningsdata och testmängden tillsammans) och `K` definierar hur många datapunkter vi har i det första dolda lagret.

```
N=1000
K=10
```

När vi tränar modellen måste vi ta tillräckligt många steg av stokastisk gradientnedstigning så att algoritmen konvergerar. Då tar vi ett stort `M`.

```
M=40000
```

Parametern `dt` är inlärningshastighet (parametern γ i Algoritm 2). Idén här är att om vi tar ett litet värde här kommer steglängderna att vara små. Då kan stokastisk gradientnedstigning försiktigt ta små steg mot minimipunkten.

```
dt=0.008
```

När stokastisk gradientnedstigning approximerar gradienten i Algoritm 2 använder det bara en punkt x_i där i väljs slumpmässigt. Ett alternativ är att approximera gradienten på mer än en punkt och därefter beräknar genomsnittet. Det motsvarar att ersätta rad (3) och (4) i Algoritm 2 med

$$\text{Välj slumpmässigt } k \text{ parametrar i mängden } \{1,\dots,N\}$$
$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \gamma \left(\frac{1}{k} \sum_{i=1}^{k} \nabla_{\boldsymbol{\theta}} \big(|\alpha_{\boldsymbol{\theta}_j}(x_i) - f_i| \big)^2 \right)$$

Vi tar bara en punkt som beskrivs av variabeln `Nbatch`.

[9] Det som följer var inspirerad av ett projekt i kursen FSF3581 på KTH, vårtermin 2021. Se referens till *Stochastic Differential Equations: Models and Numerics*.

```
Nbatch=1
```

De allra flesta datapunkter används som träningdata. Följande parameter anger kvoten mellan träningsdatan och testmängden.

```
validation_split_ratio=0.2
```

En epok (eng. *epoch*) är en hyperparameter som refererar till hur många gånger algoritmen kommer att arbeta genom hela träningsdatasetet. Vi har

```
M*Nbatch=(N*(1-validation_split_ratio))*EPOCHS
```

där den vänstra sidan motsvarar det totala antalet datapunkter stokastisk gradientnedstigning kommer att hantera. Då löser vi ekvationen för `EPOCHS` som följande. Notera att `np.int` förvandlar ett decimaltal till ett heltal.

```
EPOCHS=np.int64(M/(N*(1-validation_split_ratio)/Nbatch))
```

Vi ställer in slumptalsgeneratorn så att vi får samma datapunkter varje gång vi kör koden. Detta gör det möjligt att återge våra resultat.

```
np.random.seed(123)
tf.random.set_seed(124)
```

Vi tar ett stickprov av N punkter från en likformig fördelning på intervallet $(-4, 4)$

```
xs = np.random.uniform(-4,4,size=(N,1))
```

och definierar funktionen f som i (7.1).

```
def f(x):
    return np.square(np.abs(x-0.5))
```

Sedan evaluerar vi funktionen f på träningsdatan och testmängden.

```
ys=f(xs)
```

Vi definierar inputlagret med en ingångsnod

```
Input_layer=tf.keras.Input(shape=(1,))
```

och det dolda lagret med `K` noder och sigmoid aktiveringsfunktionen som i (6.2). Vi initialiserar vikterna enligt en normalfördelning och med bias satt till noll.

```
Hidden_layer= keras.layers.Dense(units=K,
                  activation="sigmoid",
                  use_bias=True,
                  kernel_initializer='random_normal',
                  bias_initializer='zeros')
```

Outputlagret med en utgångsnod och inga bias definieras som följer.

```
Output_layer=keras.layers.Dense(units=1,use_bias=False)
```

Vi anger ordningen på lagren och väljer stokastisk gradientnedstigning som den iterativa metoden med inlärningshastighet `dt`. Förlustfunktionen är inställd på medelkvadratfel som i (6.8).

```
model=keras.Sequential([Input_layer,Hidden_layer,Output_layer])
optimizer=tf.keras.optimizers.SGD(learning_rate=dt)
model.compile(optimizer=optimizer,loss='mean_squared_error')
```

Vi tränar modellen på de angivna parametrarna och därefter skapar vi en lista med ekvidistanta (jämnt utspridda) punkter mellan -4 och 4. Vi evaluerar funktionen f på dem, där f är som i (7.1).

```
history=model.fit(x=xs,
                  y=ys,
                  batch_size=Nbatch,
                  epochs=EPOCHS,
                  validation_split=validation_split_ratio,
                  verbose=1)
pts = np.linspace(-4,4,300).reshape(-1,1)
target_fcn_vals=f(pts)
```

Att ställa in `verbose=1` visar oss en utskrift av förlustfunktionen på vår skärm när algoritmen körs. Vi noterar att utskriften visar `loss` och `val_loss` där `loss` avser förlustfunktionen evaluerad på träningsdata och `val_loss` avser värdet av förlustfunktionen evaluerad på testmängden.

Vi evaluerar också modellen som algoritmen har skapat på samma punkter. Här motsvarar `model` (7.2a).

```
alpha_vals=model(pts)
```

Slutligen skapar vi en graf av funktionen f och modellens approximation $\alpha_{\boldsymbol{\theta}}$

```
plt.figure('Alfa',figsize=(15,10))
plt.plot(pts,target_fcn_vals,label='f')
plt.plot(pts,alpha_vals,label='alfa')
plt.legend(['f','alfa'])
plt.show()
```

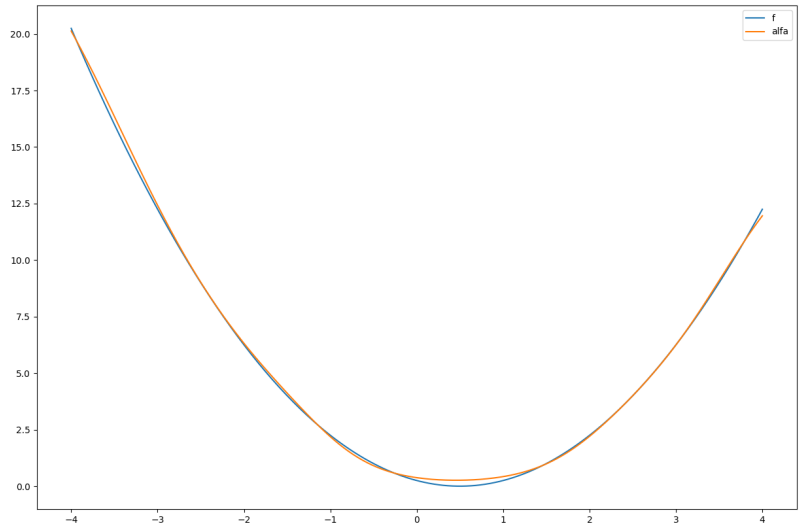
och en graf av träningsfelet och generaliseringsfelet.

```
plt.figure('Fel',figsize=(15,10))
plt.semilogy(history.history['loss'],label='Traningsfelet')
plt.semilogy(history.history['val_loss'],
    label='Generaliseringsfelet')
plt.xlabel('Epok')
plt.ylabel('Medelkvadratfel')
plt.legend()
plt.grid(True)
plt.show()
```

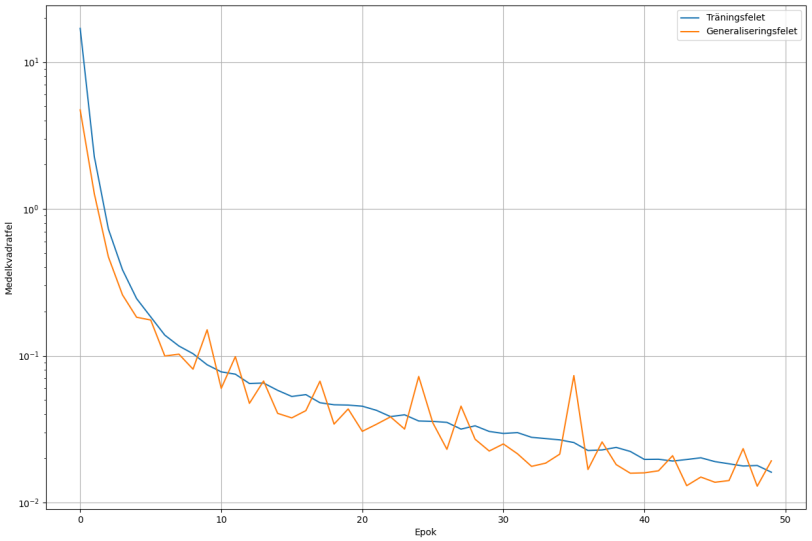
8.2 Resultat

I Figur 8.1a ser vi en graf av funktionen f som i (7.1) och algoritmens approximation som i (7.2a). Figur 8.1b visar träningsfelet och generaliseringsfelet av samma approximation. Vi ser att båda felen minskar med varje epok men träningsfelet minskar på ett smidigare sätt.

Vi ser i Figur 8.2a och Figur 8.2b resultatet och felet från modellen när vi tar fler punkter, fler steg med stokastisk gradientnedstigning och en långsammare inlärningshastighet. Vi ser att båda felen minskar på ett smidigt sätt då.



(a) Resultat

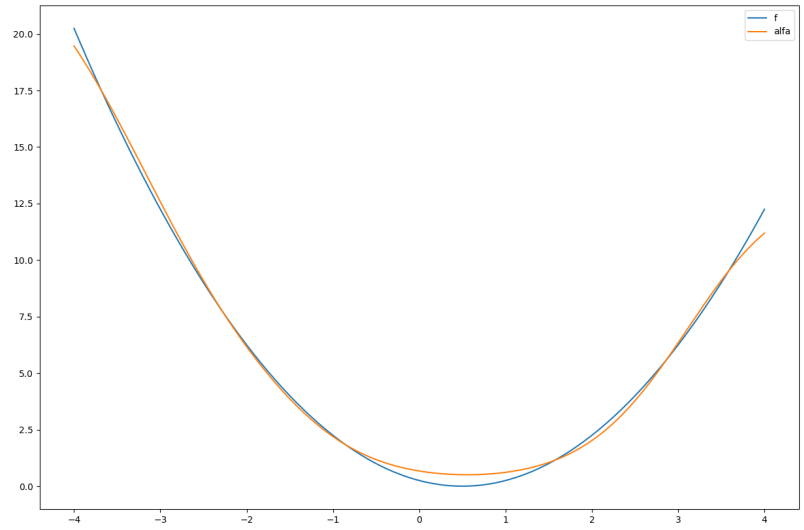


(b) Fel

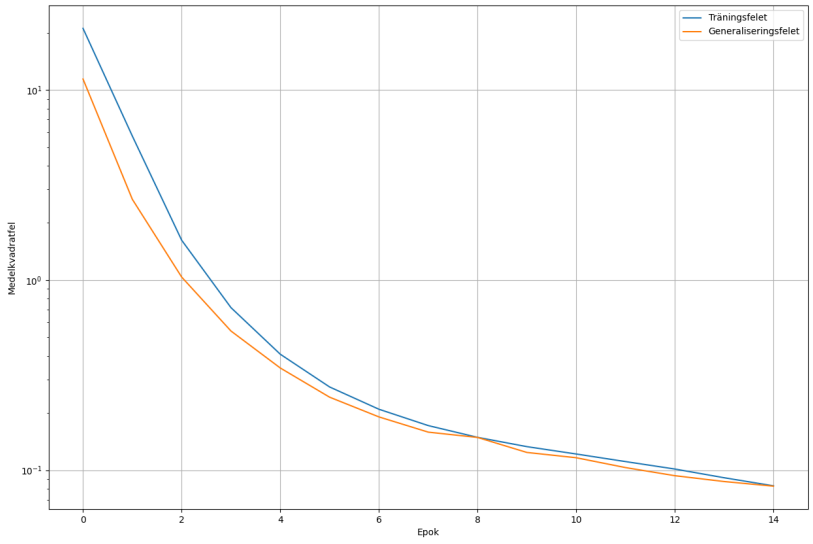
Figur 8.1: `N=1000, K=10, M=40000, dt=0.008`

8.3 Sammanfattning

Dessa är dock bara två exempel utförda med en mycket enkel modell. I verkligheten finns det väldigt många mer komplicerade modeller som kan användas för att approximera mycket mer komplicerade fenomen. Den här kursen fungerar endast som en introduktion men vi uppmuntrar er att fortsätta experimentera



(a) Resultat



(b) Fel

Figur 8.2: `N = 5000, K = 10, M = 60000, dt = 0.001`

och fortsätta att söka efter ny information om detta ämne.

Övningar

Övning 8.1 ($\star\star\star$)**.** Kör koden som är skriven ovan. Några frågor följer här.

- Beskriv resultaten du får med koden som skriven ovan. Kommentera de olika typerna av fel.
- Vad händer om du ökar antalet noder i det dolda lagret?
- Beskriv vad det innebär att öka värdet på M. Är dina resultat annorlunda om du ökar M? Om du minskar M?
- Beskriv vad det innebär att öka värdet på N. Är dina resultat annorlunda om du ökar N? Om du minskar N?
- Ändra funktionen f till en valfri funktion. Hur fungerar modellen? Notera vad du har valt till funktionen f.
- Ändra parametrarna så att du har en modell med lågt träningsfel men högt generaliseringsfel. Vad är detta ett exempel på? (Se Kapitel 7.)
- Sammanfatta det du har sett här.

Lösningar till udda övningsuppgifter

Kapitel 1

Övning 1.1. (i) 0, 1, 2, 3, 4.

(ii) 1, 2, $\{2, 3\}$.

(iii) -3, -2, -1, 0, 1, 2, 3.

(iv) 1, 2.

(v) -3, -2, -1, 0, 1, 2, $\{2, 3\}$.

Övning 1.3. (i) B är en äkta delmängd av A.

(ii) A och B är lika.

(iii) Mängderna är disjunkta.

(iv) Mängderna är varken disjunkta, lika eller äkta delmängder av varandra.

(v) Mängderna är disjunkta.

Övning 1.5. Observera att dessa svar är förslag. Uppgifterna har flera korrekta svar.

(i) $\{n \in \mathbb{Z} \mid n = 2k \text{ för något heltal } k \geqslant 1\}$.

(ii) $\{p/2 \mid p \text{ är ett heltal }\}$.

(iii) $\{r \in \mathbb{R} \mid r \notin \mathbb{Q} \text{ och } |r| < 1\}$.

Övning 1.7. (i) Definitionsmängd: $\{1, 2, 3, \ldots\}$. Målmängd: $\mathbb{N}$.

(ii) Definitionsmängd: $\{T \mid T \text{ är en triangel }\}$. Målmängd: $\mathbb{R}$.

(iii) Definitionsmängd: $\{p(x) \mid p(x) \text{ är ett andragradspolynom}\}$. Målmängd: $\{p(x) \mid p(x) \text{ är ett förstagradspolynom}\}$.

Övning 1.9. (i) Detta är en funktion. Den är definierad för alla värden i definitionsmängden, är determinerad och alla funktionsvärden ligger i målmängden.

(ii) Detta är inte en funktion, då funktionens värden inte ligger i målmängden ($f(2) = \sqrt{2} \notin \mathbb{Q}$).

(iii) Detta är inte en funktion, eftersom dess värden är slumpmässiga.

(iv) Detta är en funktion. Eftersom ordet Balkong börjar på B, så har funktionen ett definierat värde som ligger i målmängden.

Övning 1.11. (i) Funktionerna är lika. De har samma definitionsmängd, målmängd och $\sqrt{x^2} = |x|$ för alla x.

(ii) Funktionerna är inte lika, då deras definitionsmängder inte är samma.

(iii) Funktionerna är inte lika, då deras målmängder är olika.

Övning 1.13. Om n är jämnt så finns det per definition ett heltal k så att $n = 2k$. Då gäller att $n + 1 = 2k + 1$, det vill säga $n + 1$ är udda.

Övning 1.15. Antag motsatsen, det vill säga att det finns ett rationellt tal p/q och ett irrationellt tal r vars summa är rationell. Skriv summan som kvoten s/t. Då gäller

$$\frac{p}{q} + r = \frac{s}{t} \implies r = \frac{s}{t} - \frac{p}{q} = \frac{sq - pt}{tq}$$

genom att skriva vänsterledet på gemensamt bråkstreck. Men detta bevisar att r är rationellt, vilket är en motsägelse.

Övning 1.17. Basfallet är $n = 1$, och då gäller att $1 = 1^2$. För induktionssteget, antag att

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2$$

för något n. Då gäller att

$$1 + 3 + 5 + \cdots + (2n - 1) + (2n + 1) = n^2 + 2n + 1 = (n + 1)^2,$$

vilket avslutar induktionssteget och därmed beviset.

Övning 1.19. Antag motsatsen till satsen. Den kan delas in i två fall. I det ena fallet gäller att $ab = c$ och $a < \sqrt{c}$ och $b < \sqrt{c}$. Då gäller att

$$ab < a\sqrt{c} < \sqrt{c}\sqrt{c} = \sqrt{c}^2 = c.$$

Alltså gäller $ab < c$, vilket är en motsägelse.

I det andra fallet gäller att $ab = c$ och $a > \sqrt{c}$ och $b > \sqrt{c}$. Då gäller att

$$ab > a\sqrt{c} > \sqrt{c}\sqrt{c} = \sqrt{c}^2 = c.$$

Alltså gäller $ab > c$, vilket är en motsägelse.

Övning 1.21. (i) Per definition gäller $n = 0 + n = n + 0$, det vill säga $n \geqslant 0$ och $n \geqslant n$.

(ii) Enligt antagandet $n \geqslant m$ så finns det a så att $n = m + a$. Om dessutom $m \geqslant k$ så finns det b så att $m = k + b$. Då gäller

$$n = m + a = k + (b + a),$$

det vill säga $n \geqslant k$.

(iii) Om $n \geqslant m$ och $m \geqslant n$ så gäller att $n = m + k$ och $m = n + l$. Så

$$n = m + k = n + l + k \implies l + k = 0.$$

Det ger $l = k = 0$, det vill säga $n = m$.

(iv) Om $n \geqslant m$ så finns det ett a så att $n = m + a$. Då gäller

$$n + k = m + a + k = m + k + a,$$

alltså att $n + k \geqslant m + k$.

(v) Låt $n \geqslant m$, så att det finns ett tal a så att $n = m + a$. Vi ska använda induktion över k för att bevisa att $nk \geqslant mk$. Basfallet är $k = 0$, och då gäller att

$$nk = 0 \geqslant 0 = mk.$$

Antag att $nk \geqslant mk$ för något k. Det finns då b så att $nk = mk + b$. Då gäller att

$$n(k+1) = nk + n = mk + a + m + b = m(k+1) + a + b,$$

vilket bevisar att $n(k+1) \geqslant m(k+1)$. Detta avslutar induktionssteget och hela beviset.

Övning 1.23. Basfallet är $n = 2$, och då gäller att $1^2 + 2^2 = 5 < 2^3 = 8$.

Antag nu att $1^2 + 2^2 + \cdots + m^2 < m^3$ gäller för något m. Då har vi att

$$1^2 + 2^2 + \cdots + m^2 + (m+1)^2 < m^3 + m^2 + 2m + 1.$$

Om man multiplicerar ihop $(m+1)^3$ så ser vi att

$$(m+1)^3 = m^3 + 3m^2 + 3m + 1.$$

Differensen mellan högerleden blir

$$m^3 + 3m^2 + 3m + 1 - (m^3 + m^2 + 2m + 1) = 2m^2 + m$$

vilket är större än 0 när $m \geqslant 2$. Alltså gäller

$$1^2+2^2+\cdots+m^2+(m+1)^2 < m^3+m^2+2m+1 < m^3+3m^2+3m+1 = (m+1)^3.$$

Detta avslutar induktionssteget och därmed hela beviset.

Övning 1.25. Satsen att alla icke-tomma delmängder av $\mathbb{N}$ har ett minsta element kan omformulera som att alla mängder $S \subset \mathbb{N}$ som inte har ett minsta element är tom. Så om S inte har ett minsta element så gäller att $n \notin S$ för alla $n \in \mathbb{N}$. Detta är ekvivalent med att

$$0, 1, 2, \ldots, n \notin S$$

för alla n. Detta kan vi bevisa med induktion.

Låt S vara en mängd utan minsta element. Basfallet är att $0 \notin S$, vilket måste stämma eftersom $0 \leqslant n$ för alla $n \in \mathbb{N}$, så om $0 \in S$ har S ett minsta element, vilket är en motsägelse.

Anta nu att inget av talet 0, 1, …, m ligger i S. Antag att $m + 1 \in S$. Då kommer $m + 1$ vara ett minsta element i S, eftersom S inte innehåller något mindre tal. Detta motsäger att S inte har ett minsta element. Alltså gäller att 0, 1, …, $m + 1$ inte ligger i S Detta avslutar induktionen, och bevisar att S är tom.

Kommentar: *Detta resultat kallas för välordningsprincipen, och säger att* $\mathbb{N}$ *utgör en så kallad välordning. Dessa är användbara, då de tillåter oss att använda induktionsliknande resonemang.*

Vissa hävdar att induktion och välordningsprincipen är ekvivalenta. Detta stämmer inte, då induktion förutsätter att alla element i mängden kan nås genom ändligt många steg, vilket inte välordningsprincipen gör.

Övning 1.27. Mening (iv) är den vanliga definitionen av en rätvinklig triangel.

Mening (i) fungerar inte för att alla trianglar, även de som inte har en rät vinkel, har vinkelsumma 180 grader.

Mening (iii) är för snäv, då det finns rätvinkla trianglar som inte kan bildas genom att dra en diagonal i en kvadrat.

Mening (ii) är lurig. Eftersom en triangel har vinkelsumma 180 grader, kommer en triangel där två vinklar har summan 90 att vara rätvinklig, och vice versa. För trianglar är alltså påståendena ekvivalenta.

Men det finns många geometriska figurer med fler hörn som har två vinklar vars summa är 90 grader. Om (ii) var definitionen av en rätvinklig triangel, skulle vissa fyrhörningar rent formellt kunna vara rätvinkliga trianglar, vilket vore absurt.

Kapitel 2

Övning 2.1. (a)

$$|\mathbf{x}| = \sqrt{1^2 + 3^2 + 5^2 + 7^2} = \sqrt{84}$$

och

$$|\mathbf{y}| = \sqrt{(-1)^2 + 3^2 + (-5)^2 + 1^2} = \sqrt{36} = 6.$$

(b) $\mathbf{x} \cdot \mathbf{y} = -1 + 3^2 - 5^2 + 7 = -10.$

(c) Nej. Ett λ som uppfyller att $\lambda\mathbf{x} = \mathbf{y}$ Måste både uppfylla att $\lambda = -1$ och $3\lambda = 3$, vilket är omöjligt.

Övning 2.3. Basfallet $n = 2$ är den vanliga triangelolikheten.

För att bevisa induktionssteget gör vi induktionsantagandet att olikheten vi vill bevisa håller för n termer, och vill nu visa att detta implicerar att olikheten också håller för $n + 1$ termer.

Genom att använda triangelolikheten för 2 termer ser vi att

$$|x_1 + \ldots + x_{n+1}| = |(x_1 + \ldots + x_n) + x_{n+1}| \leqslant |x_1 + \ldots + x_n| + |x_{n+1}|.$$

Genom att nu använda induktionsantagandet på den första termen ser vi att

$$|x_1 + \ldots + x_n| + |x_{n+1}| \leqslant |x_1| + \ldots + |x_n| + |x_{n+1}|.$$

Tillsammans ger dessa två olikheter det önskade resultatet.

Övning 2.5. Låt $0 < a < 1$ vara en godtycklig punkt i $(0, 1)$. Vi vill visa att a är en inre punkt, det kommer vi göra genom att visa att det finns något tillräckligt litet δ så att $B_\delta(a) \subset (0, 1)$.

Låt nu

$$\delta = \frac{\min((a - 0), (1 - a))}{2},$$

det vill säga halva avståndet från a till den närmsta av punkterna 0 och 1 (samma bevis kommer såklart att fungera om vi väljer något ännu mindre δ, och i själva verket för vissa δ som är lite större).

Antag först att $\delta = a/2$, vilket då även betyder att $a \leqslant (1-a)$. Då är

$$B_\delta(a) = (a - a/2, a + a/2) \subset (a - a/2, a + (1-a)/2)) = (a/2, 1/2 + a/2) \\ \subset (a/2, 1) \subset (0, 1).$$

Notera att $(1+a)/2 < 1$ eftersom $a < 1$.

Antag nu att $\delta = (1-a)/2$, vilket då även betyder att $(1-a) \leqslant a$. Då är

$$B_\delta(a) = (a - (1-a)/2, a + (1-a)/2) \subset (a - a/2, a + (1-a)/2)) \\ = (a/2, 1/2 + a/2) \subset (a/2, 1) \subset (0, 1).$$

Notera återigen att $(1+a)/2 < 1$ eftersom $a < 1$.

Eftersom a var en godtycklig punkt i mängden är beviset nu klart.

Kapitel 3

Övning 3.1. För ett fixt värde på x, y är

$$f'_x(x, y) = \cos(\cos(xy))(-\sin(xy))y$$

och

$$f'_y(x, y) = \cos(\cos(xy))(-\sin(xy))x$$

Övning 3.3. De partiella derivatorna ges av

$$f'_x(x, y) = 2xye^{x^2y} \text{ och } f'_y(x, y) = x^2e^{x^2y},$$

och gradienten i en punkt (x, y) är således vektorn

$$\left(2xye^{x^2y}, x^2e^{x^2y}\right).$$

Övning 3.5. Till att börja med är definitionsmängden D_f hela $\mathbb{R}^3$.

Eftersom var och en av termerna är icke-negativ följer det omedelbart att

$$f(x, y, z) \geqslant 0 = f(0, 0, 0).$$

Vidare är $x^4 > 0$ om $x \neq 0$, $y^2 > 0$ om $y \neq 0$, och $z^2 > 0$ om $z \neq 0$, vilket innebär att

$$f(x, y, z) > 0 = f(0, 0, 0) \text{ för alla } (x, y, z) \neq 0,$$

och f har således ett globalt maximum i orgio.

Övning 3.7. Till att börja med ges gradienten av

$$\text{grad}(f)(x, y) = (2xy, x^2).$$

Dock är den givna riktningen inte normerad, så vi måste först normera den till

$$\frac{v}{\|v\|} = \frac{(1, 2)}{\sqrt{1^2 + 2^2}} = \frac{(1, 2)}{\sqrt{5}}.$$

Riktningsderivatan ges nu av

$$\text{grad}(f)(x, y) \cdot \frac{v}{\|v\|} = \frac{2xy + 2x^2}{\sqrt{5}}.$$

Övning 3.9. Gradienten ges av

$$\operatorname{grad}(f)(x,y) = (\cos(y)e^{x\cos(y)}, -x\sin(y)e^{x\cos(y)}),$$

så

$$\operatorname{grad}(f)(0,0) = (1,0) \neq (0,0),$$

så origo kan omöjligt vara en lokal extrempunkt.

Övning 3.11. Till att börja med ges gradienten till f av

$$\operatorname{grad}(f)(x,y,z) = (yz^2\cos(xyz^2), xz^2\cos(xyz^2), 2xyz\cos(xyz^2)),$$

så

$$\operatorname{grad}(f)(0,1,2) = (4,0,0).$$

Funktionen förändras snabbast i gradientens riktning, det vill säga i riktiningen $(1,0,0)$, och mätetalet på förändringen i den riktningen ges av

$$|\operatorname{grad}(f)(0,1,2)| = |(4,0,0)| = 4.$$

Övning 3.13. Till att börja med ges gradienten till f av

$$\operatorname{grad}(f)(x,y,z) = (2xy^2z, 2x^2yz, x^2y^2),$$

så

$$\operatorname{grad}(f)(-1,-1,3) = (-6,-6,1).$$

Funktionen förändras snabbast i gradientens riktning, det vill säga i riktiningen $(-6,-6,1)/\sqrt{73}$, och det maximala mätetalet på förändringen i den riktningen ges av

$$|\operatorname{grad}(f)(-1,-1,3)| = |(-6,-6,1)| = \sqrt{73}.$$

Kapitel 4

Övning 4.1. Eftersom $0 \leqslant P(A \cap B)$ har vi att

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leqslant P(A) + P(B).$$

Övning 4.3.

$$0.6 = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.5 - P(A \cap B)$$
$$\Rightarrow P(A \cap B) = 0.8 - 0.6 = 0.2.$$

Övning 4.5. (a) Till att börja med är $0 \leqslant \operatorname{area}(A) \leqslant \operatorname{area}(\Omega)$, så

$$0 \leqslant P(A) = \frac{\operatorname{area}(A)}{\operatorname{area}(\Omega)} \leqslant 1.$$

Vidare har vi att

$$P(\Omega) = \frac{\operatorname{area}(\Omega)}{\operatorname{area}(\Omega)} = 1.$$

Slutligen har vi att om $A \cap B = \emptyset$ så är $\operatorname{area}(A \cup B) = \operatorname{area}(A) + \operatorname{area}(B)$, så

$$P(A \cup B) = \frac{\operatorname{area}(A \cup B)}{\operatorname{area}(\Omega)} = \frac{\operatorname{area}(A)}{\operatorname{area}(\Omega)} + \frac{\operatorname{area}(B)}{\operatorname{area}(\Omega)} = P(A) + P(B),$$

och således är alla axiomen i Kolmogorovs axiomsystem uppfyllda.

(b) Vi vill visa att $P(A \cap B) = P(A)P(B)$.

Händelsen $A \cap B$ är alltså händelsen att pilen landar i den övre högra kvadranten av piltavlan, och alltså är

$$P(A \cap B) = \frac{\text{area(”övre högra kvadranten”)}}{\text{area}(\Omega)} = \frac{1}{4}.$$

Vidare har vi att

$$P(A) = \frac{\text{area(”övre halvan av tavlan”)}}{\text{area}(\Omega)} = \frac{1}{2}$$

och

$$P(B) = \frac{\text{area(”högra halvan av tavlan”)}}{\text{area}(\Omega)} = \frac{1}{2},$$

så

$$P(A)P(B) = \frac{1}{4} = P(A \cap B),$$

och händelserna är således oberoende.

Övning 4.7. Vi vill beräkna $P(A \cup B)$. Vi vet att

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

och att

$$0.6 = P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad 0.75 = P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Från detta får vi att

$$P(A \cap B) = 0.75P(A) = 0.3,$$

och

$$P(B) = \frac{P(A \cap B)}{P(A|B)} = \frac{0.3}{0.6} = 0.5.$$

Tillsammans ger detta att

$$P(A \cup B) = 0.4 + 0.5 - 0.3 = 0.6.$$

Övning 4.9. I den här uppgiften är det troligtvis lättare att beräkna sannolikheten för komplementhändelsen, det vill säga sannolikheten att person Xs tärning är större än eller lika med person Ys högsta tärning, och sen använda att

$$P(A) = 1 - P(A^c).$$

Det finns totalt 36 möjliga utfall för person Ys tärningsslag, och alla har samma sannolikhet, och det finns totalt 6 möjliga utfall för Xs tärningsslag, och alla har samma sannolikhet.

Sannolikheten att X slår en 6a är $1/6$, och i så fall vinner X alltid. Sannolikheten att X vinner genom att ha slagit en 6a är således $1/6$. Vad detta betyder är alltså att

$$P(\text{vinst för X}|X \text{ slog en sexa})P(X \text{ slog en sexa}) = 1/6.$$

Sannolikheten att person X slår en 5a är $1/6$, och i så fall vinner X om Y inte slår en sexa. Det finns totalt 25 möjliga utfall där person Y inte har någon 6a, så sannolikheten att person X vinner efter att ha slagit en 5a är $25/36$ (sannolikheten att den första tärningen är 5 eller mindre är $5/6$, och samma för den andra, så sannolikheten att båda är 5 eller mindre är $(5/6)^2$ eftersom de båda tärningsslagen är oberoende).

Sannolikheten att person X slår en 4a är $1/6$, och i så fall vinner X om Y inte slår en 5a eller 6a. Det finns totalt 16 sådana utfall, så sannolikheten att X vinner genom att ha slagit en 4a är $16/36$.

På samma sätt ser man att sannolikheten att vinna genom att ha slagit en 3a är $9/36$, att vinna med en 2a är $4/36$, och att vinna med en 1a är $1/36$.

Den totala sannolikheten för vinst för X är således

$$\sum_{n=1}^{6} \frac{1}{6}\frac{n^2}{36} = \frac{91}{6^3}$$

(vilket ungefär är 42 procent, men det är behövs inte för svaret). Notera att faktorn $1/6$ framför $n^2/36$ behövs eftersom varje givet värde på tärningsslaget för X har sannolikhet $1/6$ att inträffa.

Sannolikheten att någon av Ys tärningar visar en större siffra än Xs tärning är således

$$1 - \frac{91}{6^3}.$$

Övning 4.11. Eftersom $A \subset B$ är $B = A \cup (B \setminus A) = A \cup (B \cap A^c)$, där $A \cap (B \cap A^c) = \emptyset$ Det följer att

$$P(B) = P(A) + P(B \cap A^c),$$

och eftersom $P(B \cap A^c) \geqslant 0$ innebär detta att

$$P(A) \leqslant P(B).$$

Övning 4.13. Det finns totalt 6^3 möjliga utfall, och alla är lika sannolika.

Att $X = 18$ kan endast inträffa på ett sätt, nämligen genom att alla tärningar visar en sexa. Sannolikheten att $X = 18$ är alltså $1/6^3$.

Att $X = 17$ kan inträffa på 3 sätt, nämligen genom att var och en av de tre tärningarna visar en femma, medan de övriga två tärningarna visar sexor. Alltså utfallen

$$(5, 6, 6), (6, 5, 6), (6, 6, 5).$$

Sannolikheten att $X = 17$ är således $3/6^3 = 1/72$.

Kapitel 5

Övning 5.1. Till att börja med är $p_X(34)$ sannolikheten att resultaten av tärningsslagen x och y uppfyller att $x^2 + y = 34$. Vi vet att y kommer vara $1, 2, 3, 4, 5$ eller 6, så för att likheten ska uppfyllas måste x^2 vara $33, 32, 31, 30, 29$

eller 28. Men inget av dessa tal är ett kvadrattal, och det finns alltså *inga* möjliga utfall (x, y) så att $X = x^2 + y = 34$. Det följer att $p_X(34) = 0$.

Vidare är $F_X(34)$ sannolikheten att $X \leqslant 34$. Om $x \leqslant 5$ är $x^2 \leqslant 25$, och $x^2 + y \leqslant 31 < 34$. Och således är $X \leqslant 34$ alltid uppfyllt om $x \leqslant 5$. Om däremot $x = 6$ är $x^2 + y > x^2 = 36 > 34$ och således är $X \leqslant 34$ *aldrig* uppfyllt.

Det följer att

$$F_X(34) = P(X \leqslant 34) = P(x \leqslant 5) = 5/6.$$

Notera att värdet på den andra tärningen inte spelar någon roll för resultatet. Det vill säga att händelserna $X \leqslant 34$ och $y =$ ”vad som helst”, är oberoende händelser. Den intresserade läsaren kan verifiera detta med hjälp av definitionen av oberoende händelser.

Övning 5.3. Vi vill alltså bevisa att $E(X + Y) = E(X) + E(Y)$.

Vi bevisar påståendet för diskreta slumpvariabler X och Y. Det kontinuerliga fallet bevisas på samma sätt, men med integraler istället för summor.

Det är underförstått att summorna går igenom alla möjliga värden för X respektive Y.

Detta ger nu att

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y) P(X = x, Y = y) \\ &= \sum_x \sum_y x P(X = x, Y = y) + \sum_x \sum_y y P(X = x, Y = y) \\ &= \sum_x x \sum_y P(X = x, Y = y) + \sum_y y \sum_x P(X = x, Y = y) \\ &= \sum_x x P(X = x) + \sum_y y P(Y = y) = E(X) + E(Y). \end{aligned}$$

Notera att vi har använt att $\sum_y P(X = x, Y = y) = P(X = x)$ och att $\sum_x P(X = x, Y = y) = P(Y = y)$ för varje fixt värde på x respektive y. Detta är sant eftersom vi helt enkelt går igenom *alla* möjliga utfall för X respektive Y i summorna.

Övning 5.5. Låt X beteckna slumpvaraibeln som ger resultatet av det första tärningsslaget och låt Y beteckna slumpvariabeln som ger resultatet av det andra tärningsslaget. Vi vill alltså beräkna $E(X + Y)$. En tidigare uppgift ger att $E(X + Y) = E(X) + E(Y)$ eftersom resultaten av de två tärningsslagen är oberoende, och vi har tidigare sett att $E(X) = E(Y) = 3.5$. Det följer att

$$E(X + Y) = 3.5 + 3.5 = 7.$$

Övning 5.7. $p_X(0) = (1 - p)$ och $p_X(1) = p$, så

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Vidare är

$$E(X^2) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p,$$

så enligt formeln för variansen är

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p).$$

Övning 5.9. (a) Till att börja med är

$$F_X(x) = P(X \leqslant x) = \begin{cases} 0 & \text{om } x \leqslant a \\ \frac{x-a}{b-a} & \text{om } a < x < b \\ 1 & \text{om} b \leqslant x. \end{cases}$$

Eftersom täthetsfunktionen är derivatan av fördelningsfunktionen (där derivatan är definierad) ges den av

$$f_X(x) = \frac{\partial}{\partial x} P(X \leqslant x) = \begin{cases} 0 & \text{om } x < a \\ \frac{1}{b-a} & \text{om } a \leqslant x \leqslant b \\ 0 & \text{om } b < x. \end{cases}$$

(b) Väntevärdet ges av

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2},$$

det vill säga punkten mittemellan a och b, vilket är vad vi intuitivt hade väntat oss.

(c) Variansen ges av

$$V(X) = E(X^2) - E(X)^2,$$

och eftersom vi vet att $E(X) = (a+b)/2$ återstår det att beräkna $E(X^2)$. Vi har att

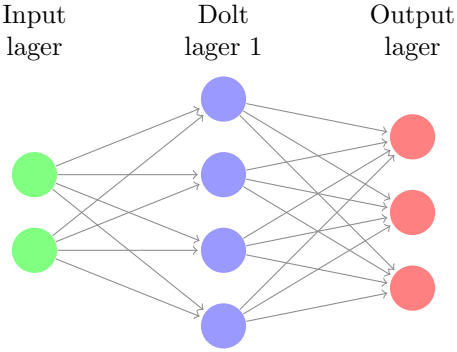
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

Tillsammans ger detta att

$$V(X) = E(X^2) - E(X)^2 = \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{2^2} = \frac{4(b^2 + ab + a^2) - 3(a^2 + 2ab + b^2))}{12} = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}.$$

Kapitel 6

Övning 6.1. Det här neurala nätverket ser ut som på följande bild.



Övning 6.3. Vi gör så här. Först, beräknar vi

$$\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} \end{bmatrix} \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \end{bmatrix}$$

som

$$\begin{bmatrix} 10 & 20 & 30 \\ -10 & -1 & -18 \\ 40 & 20 & 50 \\ -5 & -10 & -3 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 20 \\ -13 \\ 90 \\ -22 \end{bmatrix}.$$

Därefter beräknar vi

$$\begin{bmatrix} w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)} \\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)} \\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)} \\ w_{4,1}^{(0)} & w_{4,2}^{(0)} & w_{4,3}^{(0)} \end{bmatrix} \begin{bmatrix} a_1^{(0)} \\ a_2^{(0)} \\ a_3^{(0)} \end{bmatrix} + \begin{bmatrix} b_1^{(0)} \\ b_2^{(0)} \\ b_3^{(0)} \\ b_4^{(0)} \end{bmatrix}$$

som

$$\begin{bmatrix} 20 \\ -13 \\ 90 \\ -22 \end{bmatrix} + \begin{bmatrix} -3 \\ -10 \\ 1 \\ -4 \end{bmatrix} = \begin{bmatrix} 17 \\ -23 \\ 91 \\ -26 \end{bmatrix}.$$

Sedan evaluerar vi aktiveringsfunktionen σ på alla 4 element ovanför som

$$\sigma\left(\begin{bmatrix} 17 \\ -23 \\ 91 \\ -26 \end{bmatrix}\right) \approx \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

Övning 6.5. Vi måste beräkna

$$2 \times 4 + 4 = 12,$$

parametrar mellan inputlagret och det dolda lagret,

$$4 \times 3 + 3 = 15$$

mellan det dolda lagret och outputlagret. Totalt blir det då

$$12 + 15 = 27$$

parametrar för att träna nätverket.

Övning 6.7. Vi har att

$$\nabla f = 4x - 1.$$

Vi beräknar $x^{(1)}$ som

$$\begin{aligned} x^{(1)} &= x^{(0)} - \gamma \nabla f(x^{(0)}) \\ &= 2 - .2 * 7 \\ &= .6 \end{aligned}$$

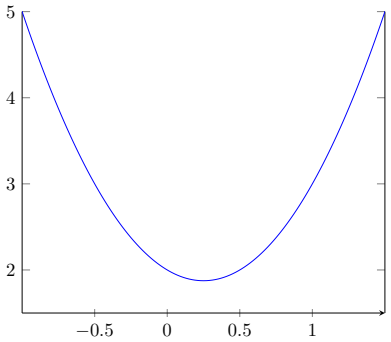
Vi fortsätter på samma sätt och får

$$\begin{aligned} x^{(2)} &= x^{(1)} - \gamma \nabla f(x^{(1)}) \\ x^{(3)} &= x^{(2)} - \gamma \nabla f(x^{(2)}) \\ x^{(4)} &= x^{(3)} - \gamma \nabla f(x^{(3)}) \\ x^{(5)} &= x^{(4)} - \gamma \nabla f(x^{(4)}) \end{aligned}$$

och att $x^{(5)} = 0.25056$. Då har vi att

$$\left(x^{(\min)}, f(x^{(\min)})\right) \approx (0.25056, 1.87500).$$

Vi ser en graf av f nedanför.



Kapitel 7

Övning 7.1. Vi har

$$\begin{aligned} L(X) = \quad & \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(X_1) - f(X_1)\right)^2 + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(X_2) - f(X_2)\right)^2 \\ & + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(X_3) - f(X_3)\right)^2 + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(X_4) - f(X_4)\right)^2 \end{aligned}$$

där

$$\alpha_{\boldsymbol{\theta}}(x) = \sum_{k=1}^{3} \sigma\big(w_{1,k}^{(0)} x + b_k^{(0)}\big).$$

Vi får

$$\alpha_{\boldsymbol{\theta}}(x) = \sigma\big(2x + 3\big) + \sigma\big(x - 1\big) + \sigma\big(.5x + 5\big)$$

och

$$\begin{aligned} L(X) = \quad & \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(.5) - f(.5)\right)^2 + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(1) - f(1)\right)^2 \\ & + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(1.25) - f(1.25)\right)^2 + \frac{1}{4}\left(\alpha_{\boldsymbol{\theta}}(2) - f(2)\right)^2. \end{aligned}$$

Då har vi

$$\begin{aligned} L(X) \approx \quad & \frac{1}{4}\left(2.3543 - 0\right)^2 + \frac{1}{4}\left(2.4892 - 0.2500\right)^2 \\ & + \frac{1}{4}\left(2.5545 - 0.5625\right)^2 + \frac{1}{4}\left(2.7277 - 2.2500\right)^2 \end{aligned}$$

som ger

$$\begin{aligned} L(X) &\approx \frac{1}{4}\left(5.5429 + 5.0142 + 3.9681 + 0.2282\right) \\ &\approx 3.6883. \end{aligned}$$

Kapitel 8

Övning 8.1. Vi kommer att gå igenom detta tillsammans under lektionen.

Referenser

Jesper Carlsson, Kyoung-Sook Moon, Anders Szepessy, Raúl Tempone and Georgios Zouraris *Stochastic Differential Equations: Models and Numerics*
2021

Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*
MIT Press, 2016

Michael Nielsen *Neural Networks and Deep Learning*
2019

3Blue1Brown,
`https://www.youtube.com/c/3blue1brown`

Introduktion till Artificiell Intelligens (AI),
`https://science.nu/lektion/introduktion-till-artificiell-intelligens-ai/`

Förslag till vidare läsning

Aku Kammonen, Jonas Kiessling, Petr Plecháč , Mattias Sandberg and Anders Szepessy *Adaptive random Fourier features with Metropolis sampling*
American Institute of Mathematical Sciences, 2020

Timothy Sauer, *Numerical Analysis*
Pearson, 2014, Andra upplagan

Arne Persson, Lars-Christer Böiers *Analys i en variabel*
Studentlitteratur, 2010, Tredje upplagan

Arne Persson, Lars-Christer Böiers *Analys i flera variabel*
Studentlitteratur, 2005, Tredje upplagan

Sven Erick Alm, Tom Britton *Stokastik*
Liber, 2008, Första upplagan