Konvexitet och optimering

Lars-Åke Lindahl

2016

Innehåll

Förord **vii**
Symbollista **ix**

I Konvexitet **1**

1 Notation och rekvisita **3**

2 Konvexa mängder **21**
2.1 Affina mängder och avbildningar 21
2.2 Konvexa mängder 26
2.3 Konvexitetsbevarande operationer 27
2.4 Konvext hölje 32
2.5 Topologiska egenskaper 33
2.6 Koner 37
2.7 Recessionskonen 42
Övningar 49

3 Separation **51**
3.1 Separerande hyperplan 51
3.2 Dualkonen 58
3.3 Lösbarhet för system av linjära olikheter 60
Övningar 65

4 Mer om konvexa mängder **67**
4.1 Extremalpunkter och fasader 67
4.2 Struktursatser för konvexa mängder 72
Övningar 76

5 Polyedrar **79**
5.1 Extremalpunkter och extremalstrålar 79
5.2 Polyedriska koner 83
5.3 Polyederns inre struktur 84

5.4 Polyederbevarande operationer . . . 86
5.5 Separation . . . 87
Övningar . . . 89

6 Konvexa funktioner 91
6.1 Grundläggande definitioner . . . 91
6.2 Konvexitetsbevarande operationer . . . 98
6.3 Maximum och minimum . . . 104
6.4 Några viktiga olikheter . . . 106
6.5 Lösbarhet för system av konvexa olikheter . . . 109
6.6 Kontinuitet . . . 111
6.7 Konvexa funktioners recessiva delrum . . . 113
6.8 Slutna konvexa funktioner . . . 116
6.9 Stödfunktionen . . . 118
6.10 Minkowskifunktionalen . . . 121
Övningar . . . 123

7 Släta konvexa funktioner 125
7.1 Konvexa funktioner på $\mathbf{R}$ . . . 125
7.2 Differentierbara konvexa funktioner . . . 131
7.3 Stark konvexitet . . . 133
7.4 Konvexa funktioner med Lipschitzkontinuerlig derivata . . . 135
Övningar . . . 139

8 Subdifferentialen 141
8.1 Subdifferentialen . . . 141
8.2 Slutna konvexa funktioner . . . 146
8.3 Konjugatfunktionen . . . 150
8.4 Riktningsderivatan . . . 155
8.5 Subdifferentieringsregler . . . 158
Övningar . . . 161

II Optimering – grundläggande teori 163

9 Optimering 165
9.1 Optimeringsproblem . . . 165
9.2 Klassificering av optimeringsproblem . . . 169
9.3 Ekvivalenta problemformuleringar . . . 172
9.4 Några modellexempel . . . 176
Övningar . . . 189

10 Lagrangefunktionen **191**
10.1 Lagrangefunktionen och det duala problemet 191
10.2 Johns sats 199
Övningar 203

11 Konvex optimering **205**
11.1 Stark dualitet 205
11.2 Karush–Kuhn–Tuckers sats 207
11.3 Tolkning av Lagrangemultiplikatorerna 209
Övningar 212

12 Linjär programmering **217**
12.1 Optimala lösningar 217
12.2 Dualitet 222
Övningar 232

III Simplexalgoritmen **235**

13 Simplexalgoritmen **237**
13.1 Standarform 237
13.2 Informell beskrivning av simplexalgoritmen 239
13.3 Baslösningar 245
13.4 Simplexalgoritmen 253
13.5 Blands anticyklingsregel 266
13.6 Simplexalgoritmen, fas 1 270
13.7 Känslighetsanalys 276
13.8 Duala simplexalgoritmen 279
13.9 Komplexitet 282
Övningar 284

IV Inrepunktsmetoder **289**

14 Descentmetoder **291**
14.1 Allmänna principer 291
14.2 Brantaste lutningsmetoden 296
Övningar 300

15 Newtons metod **301**
15.1 Newtonriktning och Newtondekrement 301
15.2 Newtons metod 309

15.3 Bivillkor i form av likheter 318
Övningar 322

16 Självkonkordanta funktioner 325
16.1 Självkonkordanta funktioner 326
16.2 Slutna självkonkordanta funktioner 330
16.3 Grundläggande olikheter för den lokala seminormen 333
16.4 Minimering 338
16.5 Newtons metod för självkonkordanta målfunktioner 342
Övningar 346
Appendix 348

17 Den vägföljande metoden 353
17.1 Barriärer och den centrala vägen 354
17.2 Vägföljande metoder 357

18 Vägföljande metoden med självkonkordant barriär 361
18.1 Självkonkordanta barriärer 361
18.2 Vägföljande metoden 370
18.3 LP-problem 382
18.4 Komplexitet 386
Övningar 395

Bibliografiska och historiska notiser 397
Referenser 401
Svar och lösningar till övningarna 407
Sakregister 424

Förord

Som utlovas av titeln har den här boken två teman, konvexitet och optimering, och konvex optimering är den gemensamma nämnaren. Konvexitet spelar en mycket viktig roll inom många delar av matematiken, och bokens del I, som behandlar ändligdimensionell konvexitetsteori, innehåller därför väsentligt mer om konvexitet än vad som sedan används i de efterkommande tre delarna om optimering, där del II ger den grundläggande klassiska teorin för linjär och konvex optimering, del III ägnas åt simplexalgoritmen, och del IV beskriver Newtons algoritm och en inrepunktsmetod med självkonkordant barriär.

I boken presenteras ett flertal algoritmer, men tyngdpunkten ligger hela tiden på den matematiska teorin, så vi går inte in på hur algoritmerna bör implementeras rent numeriskt, utan den som är intresserad av denna viktiga aspekt får söka sig till speciallitteraturen.

Matematiska optimeringsmetoder används numera rutinmässigt som redskap i samband med ekonomisk och industriell planering, vid produktionsstyrning och ingenjörsmässig produktdesign, i civil och militär logistik, i medicinsk bildanalys, etc., och utvecklingen inom optimeringsområdet har varit enorm sedan andra världskriget − år 1945 studerade George Stigler ett dietproblem med 77 födoämnen och 9 bivillkor utan att kunna bestämma den optimala dieten, idag är det möjligt att lösa optimeringsproblem som innehåller hundratusentals variabler och bivillkor. Det är två faktorer som möjliggjort detta − datorer och effektiva algoritmer. Naturligtvis är det den explosiva utvecklingen inom datorområdet som varit mest synbar för gemene man, men på teori- och algoritmsidan har det också skett en fantastisk utveckling, och utan effektiva algoritmer skulle datorerna stå sig slätt.

Maximerings- och minimeringsproblem har man naturligtvis löst sedan den matematiska analysens begynnelse, men optimeringsteori i modern mening kan sägas starta kring 1948 med George Dantzig, som introducerade och populariserade begreppet linjär programmering (LP) och anvisade en effektiv lösningsalgoritm, simplexalgoritmen, för sådana problem. Simplexalgoritmen är en iterativ algoritm, där för normala och i verkligheten förekommande LP-problem antalet iterationer erfarenhetsmässigt är ungefärligen proportionellt mot antalet variabler. Dess värstafalluppförande är emellertid dåligt; ett exempel av Victor Klee och George Minty 1972 visar att det finns LP-problem

i n variabler som för sin lösning kräver 2^n iterationer. En naturlig följdfråga är därför hur svårt det är att lösa generella LP-problem.

En algoritm för att lösa en klass $\mathcal{K}$ av problem kallas *polynomiell* om det finns ett polynom P så att algoritmen löser varje problem av storlek s i $\mathcal{K}$ med högst $P(s)$ aritmetiska operationer; ett problems storlek mäts då i antalet binära bitar som behövs för att representera det. Klassen $\mathcal{K}$ kallas vidare *lättlöst* om det finns en polynomiell algoritm som löser samtliga problem i klassen, och *svårlöst* om det inte finns någon sådan algoritm.

Klee–Mintys exempel visar att (deras variant av) simplexalgoritmen inte är polynomiell. Huruvida LP-problem är lättlösta eller svårlösta förblev dock ett öppen fråga fram till år 1979 då Leonid Khachiyan visade att LP-problem kan lösas med en polynomiell algoritm, ellipsoidmetoden. LP-problem är således i teknisk mening lättlösta.

Ellipsoidmetoden kom emellertid inte att få någon praktisk betydelse beroende på att den för normala LP-problem uppför sig sämre än simplexalgoritmen. Simplexalgoritmen var därför ohotad som praktiskt lösningsverktyg för LP-problem fram till år 1984, då Narendra Karmarkar presenterade en polynomiell inrepunktsalgoritm med lika goda prestanda som simplexalgoritmen då den tillämpas på i praktiken förekommande LP-problem.

Karmarkars upptäckt blev startpunkten för ett intensivt utvecklingsarbete av olika inrepunktsmetoder, och ett nytt genombrott skedde i slutet av 1980-talet, då Yurii Nesterov och Arkadi Nemirovski introducerade en speciell typ av konvexa barriärfunktioner (s. k. självkonkordanta funktioner), som gör att en klassisk inrepunktsmetod får polynomiella konvergensegenskaper, inte bara för LP-problem utan också för en stor klass av konvexa optimeringsproblem. Detta gör det möjligt att idag lösa optimeringsproblem som tidigare låg utom räckhåll.

Embryot till den här boken är ett kompendium som Christer Borell och undertecknad skrev 1978–79, men olika tillägg, uteslutningar och omarbetningar under årens lopp har medfört att framställningen nu är helt annorlunda jämfört med ursprunget. Det viktigaste tillägget är del IV med en beskrivning av självkonkordanta funktioner som i allt väsentligt bygger på Nesterovs och Nemirovskis arbeten.

Framställningen i boken är fullständig i så mening att alla satser bevisas. Några av bevisen är ganska tekniska, men i princip behövs det ingenstans andra förkunskaper än goda kunskaper i linjär algebra och flervariabelanalys.

Uppsala, april 2016
Lars-Åke Lindahl

Symbollista

$\operatorname{aff} X$	affina höljet till X, sid. 22
$\operatorname{bdry} X$	randen till X, sid. 11
$\operatorname{cl} f$	tillslutningen av funktionen f, sid. 148
$\operatorname{cl} X$	slutna höljet till X, sid. 12
$\operatorname{con} X$	koniska höljet till X, sid. 40
$\operatorname{cvx} X$	konvexa höljet till X, sid. 32
$\dim X$	dimensionen hos X, sid. 23
$\operatorname{dom} f$	f:s effektiva domän $\{x \mid -\infty < f(x) < \infty\}$, sid. 5
$\operatorname{epi} f$	epigrafen till f, sid. 91
$\operatorname{exr} X$	mängden av extremalstrålar till X, sid. 68
$\operatorname{ext} X$	mängden av extremalpunkter till X, sid. 67
$\operatorname{int} X$	det inre av X, sid. 11
$\operatorname{lin} X$	recessiva delrummet till X, sid. 46
$\operatorname{rbdry} X$	relativa randen till X, sid. 34
$\operatorname{recc} X$	recessionskonen till X, sid. 42
$\operatorname{rint} X$	relativa inre av X, sid. 34
$\operatorname{sublev}_\alpha f$	α-subnivåmängden till f, sid. 91
$\mathbf{e}_i$	i:te standardbasvektorn $(0, \ldots, 1, \ldots, 0)$, sid. 5
f'	derivatan eller gradienten till f, sid. 16
$f'(x; v)$	riktad derivata till f i punkten x med riktning v, sid. 155
f''	andraderivatan eller hessianen till f, sid. 18
f^*	konjugatfunktionen till f, sid. 150
$v_{\max}$, $v_{\min}$	max- resp. minproblems optimala värde, sid. 166
$B(a; r)$	öppna bollen med centrum i a och radie r, sid. 10
$\overline{B}(a; r)$	slutna bollen med centrum i a och radie r, sid. 11
$Df(a)[v]$	differentialen av f i punkten a, sid. 16
$D^2 f(a)[u, v]$	$\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i v_j$, sid. 18
$D^3 f(a)[u, v, w]$	$\sum_{i,j,k=1}^{n} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(a) u_i v_j w_k$, sid. 19
$\mathcal{E}(x; r)$	ellipsoiden $\{y \mid \|y - x\|_x \leq r\}$, sid. 365
$I(x)$	mängden av aktiva bivillkor i punkten x, sid. 199
L	inputlängd, sid. 387
$L(x, \lambda)$	Lagrangefunktionen, sid. 191
$M_{\hat{r}}[x]$	det objekt som fås genom att i M ersätta elementet på plats r med x, sid. 246

$\mathbf{R}_+$, $\mathbf{R}_{++}$	$\{x \in \mathbf{R} \mid x \geq 0\}$ resp. $\{x \in \mathbf{R} \mid x > 0\}$, sid. 3
$\mathbf{R}_-$	$\{x \in \mathbf{R} \mid x \leq 0\}$, sid. 3
$\overline{\mathbf{R}}$, $\underline{\mathbf{R}}$, $\overline{\underline{\mathbf{R}}}$	$\mathbf{R} \cup \{\infty\}$, $\mathbf{R} \cup \{-\infty\}$, resp. $\mathbf{R} \cup \{\infty, -\infty\}$, sid. 3
S_X	stödfunktionen till X, sid. 118
$S_{\mu,L}(X)$	klassen av μ-starkt konvexa funktioner på X med L-Lipschitzkontinuerlig derivata, sid. 136
$\mathrm{Var}_X(v)$	$\sup_{x\in X}\langle v, x\rangle - \inf_{x\in X}\langle v, x\rangle$, sid. 369
X^+	dualkonen till X, sid. 58
$\mathbf{1}$	vektorn $(1, 1, \ldots, 1)$, sid. 5
$\partial f(a)$	subdifferentialen till f i punkten a, sid. 141
$\lambda(f, x)$	Newtondekrementet till f i punkten x, sid. 304, 319
π_y	translaterad Minkowskifunktional, sid. 366
$\rho(t)$	$-t - \ln(1-t)$, sid. 333
ϕ_X	Minkowskifunktionalen till X, sid. 121
$\phi(\lambda)$	duala funktionen $\inf_x L(x, \lambda)$, sid. 192
Δx_{nt}	Newtonriktning i punkten x, sid. 303, 319
∇f	gradienten till f, sid. 16
$\overrightarrow{x}$	strålen från 0 genom x, sid. 37
$[x, y]$	sträckan mellan x och y, sid. 7
$]x, y[$	öppna sträckan mellan x och y, sid. 7
$\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$	ℓ^1-norm, euklidisk norm resp. maxnorm, sid 10
$\|\cdot\|_x$	lokala seminormen $\sqrt{\langle \cdot\,, f''(x)\cdot\rangle}$, sid. 305
$\|v\|_x^*$	duala lokala seminormen $\sup_{\|w\|_x\leq 1}\langle v, w\rangle$, sid. 368

Del I

Konvexitet

Kapitel 1

Notation och rekvisita

I det här inledande kapitlet skall vi etablera den notation som vi kommer att använda oss av samt repetera några grundläggande begrepp och resultat från analys och linjär algebra.

Reella tal

Vi använder standardbeteckningen $\mathbf{R}$ för mängden av alla reella tal. Vi sätter

$$\begin{aligned}\mathbf{R}_+ &= \{x \in \mathbf{R} \mid x \geq 0\},\\ \mathbf{R}_- &= \{x \in \mathbf{R} \mid x \leq 0\},\\ \mathbf{R}_{++} &= \{x \in \mathbf{R} \mid x > 0\}.\end{aligned}$$

$\mathbf{R}_+$ består med andra ord av alla icke-negativa reella tal, och $\mathbf{R}_{++}$ betecknar mängden av alla positiva reella tal.

Utvidgade reella tallinjen

Varje uppåt begränsad icke-tom delmängd A av de reella talen har som bekant en minsta övre begränsning som betecknas $\sup A$, och varje nedåt begränsad icke-tom mängd B har på motsvarande sätt en största nedre begränsning, betecknad $\inf B$. För att dessa båda begrepp skall bli väldefinierade för godtyckliga delmängder av $\mathbf{R}$ (och också av andra skäl) utvidgar vi de reella talen med de båda symbolerna $-\infty$ och ∞ samt inför beteckningarna

$$\overline{\mathbf{R}} = \mathbf{R} \cup \{\infty\}, \quad \underline{\mathbf{R}} = \mathbf{R} \cup \{-\infty\} \quad \text{och} \quad \overline{\underline{\mathbf{R}}} = \mathbf{R} \cup \{-\infty, \infty\}.$$

Vi utvidgar ordningsrelationen på $\mathbf{R}$ till den utvidgade tallinjen $\overline{\underline{\mathbf{R}}}$ genom att för alla reella tal x definiera

$$-\infty < x < \infty.$$

De aritmetiska operationerna på $\mathbf{R}$ utvidgas partiellt med hjälp av följande "naturliga" definitioner, där x betecknar ett godtyckligt reellt tal:

$$x + \infty = \infty + x = \infty + \infty = \infty$$
$$x + (-\infty) = -\infty + x = -\infty + (-\infty) = -\infty$$
$$x \cdot \infty = \infty \cdot x = \begin{cases} \infty & \text{om } x > 0 \\ 0 & \text{om } x = 0 \\ -\infty & \text{om } x < 0 \end{cases}$$
$$x \cdot (-\infty) = -\infty \cdot x = \begin{cases} -\infty & \text{om } x > 0 \\ 0 & \text{om } x = 0 \\ \infty & \text{om } x < 0 \end{cases}$$
$$\infty \cdot \infty = (-\infty) \cdot (-\infty) = \infty$$
$$\infty \cdot (-\infty) = (-\infty) \cdot \infty = -\infty.$$

Nu kan vi på ett konsistent sätt definiera supremum och infimum för godtyckliga icke-tomma delmängder av den utvidgade reella tallinjen; för icke uppåt begränsade mängder A definieras $\sup A = \infty$, och för icke nedåt begränsade mängder A definieras $\inf A = -\infty$. Slutligen definierar vi infimum och supremum också för den tomma mängden $\emptyset$ genom att sätta

$$\inf \emptyset = \infty \quad \text{och} \quad \sup \emptyset = -\infty.$$

Mängder och funktioner

Vi kommer att använda oss av mängdlärans standardbeteckningar, och dessa är förhoppningsvis välbekanta för läsaren, men måhända är snitt och union av godtyckligt många mängder nya begrepp.

Låt $\{X_i \mid i \in I\}$ vara en familj av mängder X_i; med deras *snitt*, betecknat

$$\bigcap \{X_i \mid i \in I\} \quad \text{eller} \quad \bigcap_{i \in I} X_i,$$

menas mängden av alla element som tillhör alla mängderna X_i. *Unionen*

$$\bigcup \{X_i \mid i \in I\} \quad \text{eller} \quad \bigcup_{i \in I} X_i$$

består av alla element som tillhör X_i för åtminstone något $i \in I$.

Vi skriver $f\colon X \to Y$ för att ange att funktionen f är definierad på mängden X och antar sina värden i Y. I allmänhet kommer X att vara $\mathbf{R}^n$ eller någon delmängd av $\mathbf{R}^n$, medan Y oftast kommer att vara $\mathbf{R}$ eller $\mathbf{R}^m$ för ett allmänt $m \geq 1$ men ibland också $\overline{\mathbf{R}}$, $\underline{\mathbf{R}}$ eller $\overline{\underline{\mathbf{R}}}$.

Om A är en godtycklig delmängd av definitionsmängden X kallas mängden

$$f(A) = \{f(x) \mid x \in A\}$$

för *bilden av* A under funktionen f, och om B är en delmängd av målmängden Y kallas mängden

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}$$

för *inversa bilden av* B under f. Observera att $f^{-1}(B)$ existerar oavsett om funktionen f har en invers eller ej.

För funktioner $f\colon X \to \overline{\mathbf{R}}$ använder vi $\operatorname{dom} f$ som beteckning för den inversa bilden av $\mathbf{R}$, dvs.

$$\operatorname{dom} f = \{x \in X \mid -\infty < f(x) < \infty\}.$$

Mängden $\operatorname{dom} f$ består med andra ord av alla $x \in X$ med ändliga funktionsvärden $f(x)$ och kallas f:s *(effektiva) domän.*

Vektorrummet $\mathbf{R}^n$

Vi utgår ifrån att läsaren är väl bekant med grundläggande vektorrumsbegrepp såsom linjärt delrum, linjärt oberoende, bas och dimension. $\mathbf{R}^n$ betecknar som vanligt vektorrummet av alla n-tipler $(x_1, x_2, \ldots, x_n)$ av reella tal. Elementen i $\mathbf{R}^n$, som vi omväxlande kallar punkter och vektorer, kommer att betecknas med små bokstäver från alfabetets början eller slut, och om bokstäverna inte räcker till förser vi dem med sub- eller superindex. Subindex används även för att ange koordinaterna till en vektor, men risken för förväxling är obefintlig, ty av sammanhanget kommer alltid att framgå om exempelvis x_1 är en vektor i $\mathbf{R}^n$ eller första koordinaten i vektorn x.

Vi kommer att identifiera vektorerna i $\mathbf{R}^n$ med *kolonnmatriser*. För oss är därför

$$(x_1, x_2, \ldots, x_n) \qquad \text{och} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

samma objekt.

Vi låter $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ beteckna de naturliga basvektorerna i $\mathbf{R}^n$, dvs.

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \quad \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \quad \ldots, \quad \mathbf{e}_n = (0, 0, \ldots, 0, 1).$$

Vi låter vidare $\mathbf{1}$ beteckna vektorn vars alla koordinater är lika med ett så att

$$\mathbf{1} = (1, 1, \ldots, 1).$$

Standardskalärprodukten $\langle \cdot\,, \cdot \rangle$ på $\mathbf{R}^n$ definieras av att

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Om vi använder oss av matrismultiplikation är tydligen

$$\langle x, y \rangle = x^T y = y^T x,$$

där ^{T} står för transponering; allmänt betecknar A^T transponatet av matrisen A.

Lösningsmängden till ett homogent linjärt ekvationssystem med n obekanta är ett linjärt delrum till $\mathbf{R}^n$, och omvänt är varje linjärt delrum till $\mathbf{R}^n$ lika med lösningsmängden till något homogent linjärt ekvationssystem

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = 0 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = 0. \end{cases}$$

På matrisform får systemet ovanför utseendet

$$Ax = 0,$$

där A är systemets koefficientmatris. Dimensionen hos systemets lösningsrum är $n - r$, där r är lika med matrisen A:s rang.

Speciellt finns det för varje linjärt delrum X till $\mathbf{R}^n$ av dimension $n - 1$ en nollskild vektor $c = (c_1, c_2, \ldots, c_n)$ så att

$$X = \{x \in \mathbf{R}^n \mid c_1 x_1 + c_2 x_2 + \cdots + c_n x_n = 0\}.$$

Mängdsummor

Låt X och Y vara två icke-tomma delmängder av $\mathbf{R}^n$ och låt α vara ett reellt tal. Med *(vektor-)summan* $X + Y$*, (vektor-)differensen* $X - Y$ och *produkten* αX menas mängderna

$$\begin{aligned} X + Y &= \{x + y \mid x \in X,\, y \in Y\}, \\ X - Y &= \{x - y \mid x \in X,\, y \in Y\}, \\ \alpha X &= \{\alpha x \mid x \in X\}. \end{aligned}$$

För att summor, differenser och produkter också ska vara definierade för den tomma mängden utvidgar vi ovanstående definitioner genom att sätta $\alpha\emptyset = \emptyset$ och $X \pm \emptyset = \emptyset \pm X = \emptyset$ för godtyckliga mängder X.

Om $\{a\}$ är en enpunktsmängd skriver man $a+X$ istället för $\{a\}+X$ och kallar mängden $a+X$ för ett *translat* av X.

För godtyckliga mängder X, Y och Z och godtyckliga reella tal α och β gäller, som man lätt verifierar, följande räkneregler

$$\begin{aligned} X+Y &= Y+X \\ (X+Y)+Z &= X+(Y+Z) \\ \alpha X+\alpha Y &= \alpha(X+Y) \\ (\alpha+\beta)X &\subseteq \alpha X+\beta X\,. \end{aligned}$$

Man bör i anslutning till den sistnämnda av ovanstående räkneregler notera att den omvända inklusionen $\alpha X+\beta X \subseteq (\alpha+\beta)X$ **inte** gäller för godtyckliga mängder X.

Olikheter i $\mathbf{R}^n$

Låt $x=(x_1,x_2,\ldots,x_n)$ och $y=(y_1,y_2,\ldots,y_n)$ vara vektorer i $\mathbf{R}^n$. Vi skriver $x\geq y$ om $x_j\geq y_j$ för alla index j, och $x>y$ om $x_j>y_j$ för alla j. Speciellt betyder alltså $x\geq 0$ att alla koordinaterna i x är icke-negativa.

Mängden

$$\mathbf{R}^n_+ = \mathbf{R}_+\times\mathbf{R}_+\times\cdots\times\mathbf{R}_+ = \{x\in\mathbf{R}^n \mid x\geq 0\}$$

kallas *icke-negativa ortanten* i $\mathbf{R}^n$.

Ordningsrelationen $\geq$ är en s. k. partiell ordning på $\mathbf{R}^n$, ty den är reflexiv ($x\geq x$ för alla x), transitiv ($x\geq y$ & $y\geq z \Rightarrow x\geq z$) och antisymmetrisk ($x\geq y$ & $y\geq x \Rightarrow x=y$). Däremot är den förstås inte fullständig om $n>1$; två vektorer x, y kan vara orelaterade.

En viktig egenskap, som vi kommer att utnyttja då och då, är de triviala implikationerna

$$\begin{aligned} x\geq 0 \;\&\; y\geq 0 \;&\Rightarrow\; \langle x,y\rangle \geq 0 \\ x\geq 0 \;\&\; y\geq 0 \;\&\; \langle x,y\rangle = 0 \;&\Rightarrow\; x=y=0. \end{aligned}$$

Sträckor

Låt x och y vara två punkter i $\mathbf{R}^n$. Om punkterna är skilda kallas mängden

$$[x,y] = \{(1-\lambda)x+\lambda y \mid 0\leq\lambda\leq 1\}$$

för *sträckan* mellan x och y, och mängden

$$]x,y[= \{(1-\lambda)x+\lambda y \mid 0<\lambda<1\}$$

kallas den *öppna sträckan* mellan x och y. Om punkterna sammanfaller, dvs. om $x=y$, så är förstås $[x,x]=]x,x[=\{x\}$.

Linjära avbildningar och linjära former

Vi påminner om att en avbildning $S\colon \mathbf{R}^n \to \mathbf{R}^m$ kallas *linjär* om identiteten

$$S(\alpha x + \beta y) = \alpha Sx + \beta Sy$$

gäller för alla vektorer x, $y \in \mathbf{R}^n$ och alla skalärer (dvs. reella tal) α, β.

En linjär avbildning $S\colon \mathbf{R}^n \to \mathbf{R}^n$ kallas också för en *linjär operator* på $\mathbf{R}^n$.

Till varje linjär avbildning $S\colon \mathbf{R}^n \to \mathbf{R}^m$ hör en unik $m \times n$-matris $\tilde{S}$ så att

$$Sx = \tilde{S}x,$$

dvs. så att avbildningsvärdet Sx beräknas som matrisprodukten $\tilde{S}x$. Av det skälet kommer vi att använda samma bokstav för avbildningen och avbildningens matris. Vi uppfattar således urskiljningslöst Sx som ett avbildningsvärde och som en matrisprodukt.

Genom att beräkna skalärprodukten $\langle x, Sy\rangle$ som en matrisprodukt får vi sambandet

$$\langle x, Sy\rangle = x^T Sy = (S^T x)^T y = \langle S^T x, y\rangle$$

mellan en linjär avbildning $S\colon \mathbf{R}^n \to \mathbf{R}^m$ (dvs. $m \times n$-matris S) och den *transponerade avbildningen* $S^T\colon \mathbf{R}^m \to \mathbf{R}^n$ (dvs. den transponerade matrisen S^T).

En $n \times n$-matris $A = [a_{ij}]$, och motsvarande linjära avbildning, kallas *symmetrisk* om $A^T = A$, dvs. om $a_{ij} = a_{ji}$ för alla index i, j.

En linjär avbildning $f\colon \mathbf{R}^n \to \mathbf{R}$ kallas en *linjär form*. De linjära formerna har utseendet

$$f(x) = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n,$$

där $c = (c_1, c_2, \ldots, c_n)$ är en vektor i $\mathbf{R}^n$. Med hjälp av standardskalärprodukten på $\mathbf{R}^n$ kan linjärformen f enklare skrivas

$$f(x) = \langle c, x\rangle,$$

och på matrisform har vi

$$f(x) = c^T x.$$

Om $f(x) = \langle c, y\rangle$ är en linjär form på $\mathbf{R}^m$ och avbildningen $S\colon \mathbf{R}^n \to \mathbf{R}^m$ är linjär, så är den sammansatta avbildningen $f \circ S$ en linjär form på $\mathbf{R}^n$, och det finns därför en unik vektor $d \in \mathbf{R}^n$ så att $(f \circ S)(x) = \langle d, x\rangle$ för alla $x \in \mathbf{R}^n$. Eftersom $f(Sx) = \langle c, Sx\rangle = \langle S^T c, x\rangle$, är tydligen $d = S^T c$.

Kvadratiska former

En funktion $q\colon \mathbf{R}^n \to \mathbf{R}$ kallas en *kvadratisk form* om det finns en symmetrisk $n \times n$-matris $Q = [q_{ij}]$ så att

$$q(x) = \sum_{i,j=1}^{n} q_{ij} x_i x_j.$$

Detta innebär att

$$q(x) = \langle x, Qx \rangle = x^T Q x.$$

Den kvadratiska formen q bestämmer den symmetriska matrisen Q entydigt, så vi kommer därför i fortsättningen att identifiera formen q med matrisen (operatorn) Q.

Med hjälp av linjära och kvadratiska former kan vi nu skriva godtyckliga andragradspolynom $p(x)$ i n variabler på formen

$$p(x) = \langle x, Ax \rangle + \langle b, x \rangle + c,$$

där $x \mapsto \langle x, Ax \rangle$ är en kvadratisk form bestämd av en symmetrisk operator (eller matris) A, $x \mapsto \langle b, x \rangle$ är en linjär form bestämd av en vektor b, och c är ett reellt tal.

Exempel. För att skriva andragradspolynomet

$$p(x_1, x_2, x_3) = x_1^2 + 4x_1x_2 - 2x_1x_3 + 5x_2^2 + 6x_2x_3 + 3x_1 + 2x_3 + 2$$

på denna form ersätter vi först termerna dx_ix_j för $i < j$ med $\frac{1}{2}dx_ix_j + \frac{1}{2}dx_jx_i$. Detta ger

$$\begin{aligned} p(x_1, x_2, x_3) &= (x_1^2 + 2x_1x_2 - x_1x_3 + 2x_2x_1 + 5x_2^2 + 3x_2x_3 - x_3x_1 + 3x_3x_2) \\ &\quad + (3x_1 + 2x_3) + 2 = \langle x, Ax \rangle + \langle b, x \rangle + c \end{aligned}$$

med $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 5 & 3 \\ -1 & 3 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$ och $c = 2$. □

En kvadratisk form q på $\mathbf{R}^n$ (och motsvarande symmetriska operator och matris) kallas *positivt semidefinit* om $q(x) \geq 0$ och *positivt definit* om $q(x) > 0$ för alla vektorer $x \neq 0$ i $\mathbf{R}^n$.

Normer och bollar

Med en *norm* $\|\cdot\|$ på $\mathbf{R}^n$ menas en funktion $\mathbf{R}^n \to \mathbf{R}_+$ med följande egenskaper:

(i)	$\|x+y\| \le \|x\| + \|y\|$	för alla x, y
(ii)	$\|\lambda x\| = \|\lambda\| \, \|x\|$	för alla $x \in \mathbf{R}^n$, $\lambda \in \mathbf{R}$
(iii)	$\|x\| = 0 \Leftrightarrow x = 0.$	

Den för oss viktigaste normen är den *euklidiska normen*, som definieras via standardskalärprodukten som

$$\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

Det är den normen som vi använder oss av, om inte annat sägs explicit. Om vi speciellt behöver markera att en norm är den euklidiska normen, använder vi beteckningen $\|\cdot\|_2$ för densamma.

Andra normer, som kommer att förekomma då och då, är *maxnormen*

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|,$$

och ℓ^1*-normen*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

Man verifierar omedelbart att dessa verkligen är normer, dvs. att villkoren (i)–(iii) är uppfyllda.

Alla normer på $\mathbf{R}^n$ är ekvivalenta i den meningen att om $\|\cdot\|$ och $\|\cdot\|'$ är två godtyckliga normer så finns det positiva konstanter c och C så att

$$c\|x\|' \le \|x\| \le C\|x\|'$$

för alla $x \in \mathbf{R}^n$. Exempelvis är

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\, \|x\|_\infty.$$

Givet en norm $\|\cdot\|$ definieras *avståndet* mellan två punkter x och a i $\mathbf{R}^n$ som $\|x-a\|$. Mängden

$$B(a;r) = \{x \in \mathbf{R}^n \mid \|x-a\| < r\},$$

som alltså består av alla punkter x vars avstånd till a är mindre än r, kallas en *öppen boll* med centrum i punkten a och radie r. För att denna boll skall

vara icke-tom krävs förstås att $r > 0$. Med motsvarande *slutna boll* menas mängden

$$\overline{B}(a;r) = \{x \in \mathbf{R}^n \mid \|x - a\| \leq r\}.$$

Hur bollarna ser ut beror naturligtvis på den underliggande normen. I $\mathbf{R}^2$ och med maxnormen är bollen $\overline{B}(0;1)$ en kvadrat med hörn i punkterna $(\pm 1, \pm 1)$. Med avseende på ℓ^1-normen är bollen istället en kvadrat med hörn i punkterna $(\pm 1, 0)$ och $(0, \pm 1)$, och med avseende på den euklidiska normen är bollen enhetscirkelskivan.

Av ovan nämnda ekvivalensegenskap för normer följer emellertid att om B betecknar bollar som definieras med hjälp av en norm och B' betecknar bollar som definieras med hjälp av en annan norm, så finns det positiva konstanter c och C så att inklusionerna

$$B'(a;cr) \subseteq B(a;r) \subseteq B'(a;Cr) \tag{1.1}$$

gäller för alla punkter $a \in \mathbf{R}^n$ och alla $r > 0$.

När inget annat sägs förutsätts bollarna i fortsättningen vara definierade relativt den euklidiska normen.

Topologiska begrepp

Med hjälp av våra bollar skall vi nu definiera ett antal s.k. topologiska begrepp. Som den uppmärksamme läsaren lätt kan konstatera blir resultaten på grund av inklusionerna (1.1) oberoende av vilken underliggande norm som används, men låt oss för enkelhets skull hela tiden anta att våra bollar är euklidiska.

Låt X vara en godtycklig delmängd av $\mathbf{R}^n$. En punkt $a \in \mathbf{R}^n$ kallas en

- *inre punkt* till X om det finns ett $r > 0$ så att $B(a;r) \subseteq X$;
- *randpunkt* till X om $X \cap B(a;r) \neq \emptyset$ och $\complement X \cap B(a;r) \neq \emptyset$ för alla $r > 0$;
- *yttre punkt* till X om det finns ett $r > 0$ så att $X \cap B(a;r) = \emptyset$.

En punkt är tydligen antingen en inre punkt, en randpunkt eller en yttre punkt till X. En inre punkt till X tillhör nödvändigtvis X, en yttre punkt ligger alltid i komplementet till X, medan en randpunkt kan tillhöra X men inte behöver göra det. En yttre punkt till X är tydligen en inre punkt i komplementet $\complement X$ och vice versa, och de båda mängderna X och $\complement X$ har samma randpunkter.

Mängden av alla inre punkter till X kallas *det inre* av X och betecknas $\operatorname{int} X$. Mängden av alla randpunkter kallas *randen* till X och betecknas $\operatorname{bdry} X$.

En mängd X kallas *öppen* om alla punkter i X är inre punkter, dvs. om $\operatorname{int} X = X$.

Det är lätt att se att unionen av ett godtyckligt antal öppna mängder är öppen och att snittet av ett ändligt antal öppna mängder är öppet. Hela $\mathbf{R}^n$ och den tomma mängden $\emptyset$ är per definition öppna mängder.

För varje mängd X är $\operatorname{int} X$ en öppen mängd (som kan vara tom), och $\operatorname{int} X$ är den största öppna mängden som är inkluderad i X.

En mängd X kallas *sluten* om dess komplement $\complement X$ är en öppen mängd. Detta är ekvivalent med att alla randpunkter till X tillhör X. En mängd X är därför sluten om och endast om $\operatorname{bdry} X \subseteq X$.

Snittet av godtyckligt många slutna mängder är slutet, unionen av ändligt många slutna mängder är slutet, och $\mathbf{R}^n$ och $\emptyset$ är slutna mängder.

För varje mängd X är mängden

$$\operatorname{cl} X = X \cup \operatorname{bdry} X$$

en sluten mängd som innehåller X. Denna mängd kallas *slutna höljet* (eller *tillslutningen*) av X. Slutna höljet $\operatorname{cl} X$ är den minsta slutna mängden som omfattar X.

Exempelvis är för $r > 0$

$$\operatorname{cl} B(a;r) = \{x \in \mathbf{R}^n \mid \|x - a\| \leq r\} = \overline{B}(a;r),$$

så det är alltså konsistent att kalla $\overline{B}(a;r)$ för en sluten boll.

För godtyckliga icke-tomma delmängder X av $\mathbf{R}^n$ och tal $r > 0$ sätter vi

$$X(r) = \{y \in \mathbf{R}^n \mid \exists x \in X \colon \|y - x\| < r\}.$$

Mängden $X(r)$ består av alla punkter vars avstånd till X är mindre än r.

En punkt x är per definition en yttre punkt till X om och endast x har ett positivt avstånd till X, dvs. om och endast om det finns ett $r > 0$ så att $x \notin X(r)$. Detta innebär att en punkt x tillhör slutna höljet $\operatorname{cl} X$, dvs. är en inre punkt eller randpunkt, om och endast om x tillhör mängderna $X(r)$ för alla $r > 0$. Med andra ord är

$$\operatorname{cl} X = \bigcap_{r>0} X(r).$$

En mängd X säges vara *begränsad* om den är innehållen i någon boll med centrum i 0, dvs. om det finns något $R > 0$ så att $X \subseteq B(0;R)$.

En mängd X som är både sluten och begränsad kallas *kompakt*.

En viktig egenskap hos kompakta delmängder X av $\mathbf{R}^n$ är att varje oändlig följd $(x_n)_{n=1}^{\infty}$ av punkter $x_n \in X$ innehåller en delföljd $(x_{n_k})_{k=1}^{\infty}$ som konvergerar mot en punkt i X (Bolzano–Weierstrass sats).

Om X är en kompakt delmängd av $\mathbf{R}^m$ och Y är en kompakt delmängd av $\mathbf{R}^n$, så är produktmängden $X \times Y$ en kompakt delmängd av $\mathbf{R}^m \times \mathbf{R}^n$ $(= \mathbf{R}^{m+n})$.

Kontinuitet

En funktion $f\colon X \to \mathbf{R}^m$, som är definierad på en delmängd X av $\mathbf{R}^n$, säges vara *kontinuerlig i punkten* $a \in X$ om det för varje $\epsilon > 0$ finns ett $r > 0$ så att

$$f(X \cap B(a;r)) \subseteq B(f(a);\epsilon).$$

(Här är förstås bollen i högerledet en boll i $\mathbf{R}^m$ och bollen i vänsterledet en boll i $\mathbf{R}^n$.) Om funktionen är kontinuerlig i varje punkt $a \in X$ säges funktionen rätt och slätt vara *kontinuerlig* (eller *kontinuerlig på* X).

Om funktionen $f\colon \mathbf{R}^n \to \mathbf{R}$ är kontinuerlig och I är ett öppet delintervall av $\mathbf{R}$, så är inversa bilden $f^{-1}(I)$ en öppen mängd i $\mathbf{R}^n$. Speciellt är alltså mängderna $\{x \mid f(x) < a\}$ och $\{x \mid f(x) > a\}$, dvs. mängderna $f^{-1}(]-\infty, a[)$ och $f^{-1}(]a, \infty[)$, öppna för alla $a \in \mathbf{R}$. Deras komplementmängder, mängderna $\{x \mid f(x) \geq a\}$ och $\{x \mid f(x) \leq a\}$, är förstås slutna.

Summor och (skalär)produkter av kontinuerliga funktioner är kontinuerliga, och kvoter av reellvärda kontinuerliga funktioner är kontinuerliga överallt där kvoterna är definierade. Sammansättningar av kontinuerliga funktioner är kontinuerliga.

Om mängden X är kompakt och funktionen $f\colon X \to \mathbf{R}^m$ är kontinuerlig, så är bilden $f(X)$ kompakt. Detta gäller förstås speciellt om $m = 1$ och innebär i detta fall att funktionen är begränsad och att maximum och minimum existerar, dvs. att det finns två punkter x_1, $x_2 \in X$ så att $f(x_1) \leq f(x) \leq f(x_2)$ för alla $x \in X$.

Lipschitzkontinuitet

En funktion $f\colon X \to \mathbf{R}^m$, som är definierad på en delmängd X av $\mathbf{R}^n$, kallas *Lipschitzkontinuerlig* med Lipschitzkonstant L om

$$\|f(y) - f(x)\| \leq L\|y - x\| \qquad \text{för alla } x, y \in X.$$

Eftersom alla normer på ett ändligdimensionellt rum är ekvivalenta, beror begreppet Lipschitzkontinuitet inte på vilka normer som används. Däremot beror förstås konstanten L på valet av normer.

Lipschitzkontinuerliga funktioner är uppenbarligen (likformigt) kontinuerliga.

Operatornormen

Låt $\|\cdot\|$ vara en given norm på $\mathbf{R}^n$. Eftersom slutna enhetsbollen är kompakt och linjära operatorer på $\mathbf{R}^n$ är kontinuerliga, är

$$\|S\| = \sup_{\|x\|\leq 1} \|Sx\|$$

ett ändligt tal för varje linjär operator S på $\mathbf{R}^n$. Talet $\|S\|$ kallas *normen* av operatorn S.

Att operatornormen verkligen är en norm på rummet av linjära operatorer på $\mathbf{R}^n$, dvs. har egenskaperna (i)–(iii) i normdefinitionen, följer omedelbart av motsvarande egenskaper hos den underliggande normen på $\mathbf{R}^n$.

För varje $x \neq 0$ är vidare per definition $S(x/\|x\|) \leq \|S\|$, så det följer att

$$\|Sx\| \leq \|S\|\|x\|$$

för alla $x \in \mathbf{R}^n$.

Av denna olikhet följer i sin tur att $\|STx\| \leq \|S\|\|Tx\| \leq \|S\|\|T\|\|x\|$, vilket ger oss den viktiga olikheten

$$\|ST\| \leq \|S\|\|T\|$$

för normen av en produkt av två operatorer.

Identitetsoperatorn I på $\mathbf{R}^n$ har uppenbarligen norm 1. Om operatorn S är inverterbar, så får vi därför genom att välja $T = S^{-1}$ i olikheten ovan att

$$\|S^{-1}\| \geq 1/\|S\|.$$

Operatornormen beror uppenbarligen av den underliggande normen på $\mathbf{R}^n$, men återigen ger olika normer på $\mathbf{R}^n$ upphov till ekvivalenta normer på operatorrummet. I den här boken kommer vi emellertid, när vi använder oss av operatornormen alltid att förutsätta att den underliggande normen på $\mathbf{R}^n$ är den euklidiska normen, även om inte detta utsägs explicit.

Symmetriska operatorer, egenvärden och normer

Varje symmetrisk operator S på $\mathbf{R}^n$ kan enligt spektralsatsen diagonaliseras. Detta betyder att det finns en ON-bas $e_1, e_2, \ldots, e_n$ av egenvektorer och att motsvarande egenvärden $\lambda_1, \lambda_2, \ldots, \lambda_n$ är reella.

Operatorns största och minsta egenvärden $\lambda_{\max}$ och $\lambda_{\min}$ erhålls som maximi- resp. minimivärden till den kvadratiska formen $\langle x, Sx\rangle$ över enhetssfären $\|x\| = 1$:

$$\lambda_{\max} = \max_{\|x\|=1} \langle x, Sx\rangle \quad \text{och} \quad \lambda_{\min} = \min_{\|x\|=1} \langle x, Sx\rangle.$$

För $x = \sum_{i=1}^{n} \xi_i e_i$ är nämligen

$$\langle x, Sx\rangle = \sum_{i=1}^{n} \lambda_i \xi_i^2 \leq \lambda_{\max} \sum_{i=1}^{n} \xi_i^2 = \lambda_{\max} \|x\|^2$$

med likhet då x är den till egenvärdet $\lambda_{\max}$ hörande egenvektorn e_i, och motsvarande olikhet åt andra hållet gäller för $\lambda_{\min}$.

För operatornormen (med avseende på den euklidiska normen) gäller vidare att

$$\|S\| = \max_{1\leq i\leq n} |\lambda_i| = \max\{|\lambda_{\max}|, |\lambda_{\min}|\}.$$

Med x som ovan är nämligen $Sx = \sum_{i=1}^{n} \lambda_i \xi_i e_i$, och följaktligen

$$\|Sx\|^2 = \sum_{i=1}^{n} \lambda_i^2 \xi_i^2 \leq \max_{1\leq i\leq n} |\lambda_i|^2 \sum_{i=1}^{n} \xi_i^2 = (\max_{1\leq i\leq n} |\lambda_i|)^2 \, \|x\|^2,$$

och likhet råder i denna olikhet då x är den mot $\max_i |\lambda_i|$ svarande egenvektorn.

Operatorn S är inverterbar om alla egenvärden är nollskilda, och då är förstås också inversen S^{-1} symmetrisk med $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}$ som egenvärden. Inversens norm fås därför som

$$\|S^{-1}\| = 1/\min_{1\leq i\leq n} |\lambda_i|.$$

En symmetrisk operator S är positivt semidefinit om alla egenvärden är icke-negativa och positivt definit om alla egenvärden är positiva. För positivt definita operatorer är tydligen

$$\|S\| = \lambda_{\max} \qquad \text{och} \qquad \|S^{-1}\| = 1/\lambda_{\min}.$$

Av spektralsatsen följer det vidare enkelt att varje positivt semidefinit symmetrisk operator S på $\mathbf{R}^n$ har en unik positivt semidefinit symmetrisk kvadratrot $S^{1/2}$, och av identiteten

$$\langle x, Sx\rangle = \langle x, S^{1/2}(S^{1/2}x)\rangle = \langle S^{1/2}x, S^{1/2}x\rangle = \|S^{1/2}x\|$$

följer att operatorerna S och $S^{1/2}$ har samma nollrum samt att nollrummet

$$\mathcal{N}(S) = \{x \in \mathbf{R}^n \mid Sx = 0\} = \{x \in \mathbf{R}^n \mid \langle x, Sx\rangle = 0\}.$$

Differentierbarhet

En funktion $f\colon U \to \mathbf{R}$, som är definierad på en öppen delmängd U av $\mathbf{R}^n$, kallas *differentierbar i punkten* $a \in U$ om de partiella derivatorna $\frac{\partial f}{\partial x_i}$ existerar i punkten x och likheten

$$f(a+v) = f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)\, v_i + r(v) \tag{1.2}$$

gäller för alla v i någon omgivning av origo med en restterm $r(v)$ som uppfyller villkoret

$$\lim_{v\to 0} \frac{r(v)}{\|v\|} = 0.$$

Vi sätter

$$Df(a)[v] = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)\, v_i,$$

och kallar den linjära formen $Df(a)[v]$ för *differentialen* av funktionen f i punkten a.

Differentialens koefficientvektor

$$\Big(\frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), \ldots, \frac{\partial f}{\partial x_n}(a)\Big)$$

kallas för *derivatan* eller *gradienten* av f i punkten a och betecknas $f'(a)$ eller $\nabla f(a)$. Vi kommer mestadels att använda den förstnämnda beteckningen.

Ekvation (1.2) kan nu på kompakt form skrivas

$$f(a+v) = f(a) + Df(a)[v] + r(v),$$

och i termer av derivatan är

$$Df(a)[v] = \langle f'(a), v\rangle.$$

En funktion $f\colon U \to \mathbf{R}$ kallas *differentierbar (på U)* om den är differentierbar i varje punkt i U. Detta förutsätter alltså speciellt att U är en öppen mängd.

För funktioner av en variabel är differentierbarhet och deriverbarhet samma sak, men så är inte fallet för funktioner av flera variabler. Ett tillräckligt villkor för att en funktion, som är definierad på en öppen delmängd U av $\mathbf{R}^n$, skall vara differentierbar är att de partiella derivatorna existerar och är kontinuerliga på U.

Medelvärdessatsen

Antag att funktionen $f\colon U \to \mathbf{R}$ är differentierbar och att sträckan $[x, x+v]$ ligger i U. Sätt $\phi(t) = f(x+tv)$; funktionen ϕ är då definierad och deriverbar på intervallet $[0,1]$ med derivata

$$\phi'(t) = Df(x+tv)[v] = \langle f'(x+tv), v\rangle.$$

Detta är förstås ett specialfall av kedjeregeln men följer i föreliggande fall mycket enkelt ur derivatans definition. Medelvärdessatsen för envariabelfunktioner ger nu att det finns ett tal $s \in]0,1[$ så att $\phi(1) - \phi(0) = \phi'(s)(1-0)$. Eftersom $\phi(1) = f(x+v)$, $\phi(0) = f(x)$ och $x + sv$ är en punkt på den öppna sträckan $]x, x+v[$, har vi därmed härlett följande *medelvärdessats* för flervariabelfunktioner.

Sats 1.1.1. *Antag att funktionen $f\colon U \to \mathbf{R}$ är differentierbar och att sträckan $[x, x+v]$ ligger i U. Då finns det en punkt $c \in]x, x+v[$ så att*

$$f(x+v) = f(x) + Df(c)[v].$$

Funktioner med Lipschitzkontinuerlig derivata

I många fall kommer vi att behöva bättre information om resttermen $r(v)$ i likheten (1.2) än den som följer av definitionen för differentierbara funktioner. För funktioner med Lipschitzkontinuerlig derivata har vi följande resultat.

Sats 1.1.2. *Antag att funktionen $f\colon U \to \mathbf{R}$ är differentierbar med Lipschitzkontinuerlig derivata, dvs. att $\|f'(y) - f'(x)\| \le L\|y - x\|$ för alla $x, y \in U$. Antag vidare att sträckan $[x, x+v]$ ligger i U. Då är*

$$|f(x+v) - f(x) - Df(x)[v]| \le \frac{L}{2}\,\|v\|^2.$$

Bevis. Sätt

$$\Phi(t) = f(x+tv) - t\,Df(x)[v].$$

Funktionen Φ är definierad på intervallet $[0,1]$ med derivata

$$\Phi'(t) = Df(x+tv)[v] - Df(x)[v] = \langle f'(x+tv) - f'(x), v\rangle.$$

Det följer av Cauchy–Schwarz olikhet och Lipschitzkontinuiteten att

$$|\Phi'(t)| \le \|f'(x+tv) - f'(x)\| \cdot \|v\| \le Lt\,\|v\|^2.$$

Eftersom $f(x+v) - f(x) - Df(x)[v] = \Phi(1) - \Phi(0) = \int_0^1 \Phi'(t)\,dt$, följer det nu att

$$|f(x+v) - f(x) - Df(x)[v]| \le \int_0^1 |\Phi'(t)|\,dt \le L\|v\|^2 \int_0^1 t\,dt = \frac{L}{2}\,\|v\|^2. \quad \square$$

Två gånger differentierbara funktioner

Om f och samtliga partiella derivator $\frac{\partial f}{\partial x_i}$ är differentierbara i U, säges funktionen f vara två gånger differentierbar. De blandade partiella andraderivatorna är i så fall automatiskt lika, dvs.

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

för alla i, j och alla $a \in U$.

Ett tillräckligt villkor för att funktionen f skall vara två gånger differentierbar i U är att de partiella derivatorna upp till och med ordning 2 existerar och är kontinuerliga i U.

För två gånger differentierbara funktioner $f\colon U \to \mathbf{R}$, punkter $a \in U$, och godtyckliga vektorer u, v i $\mathbf{R}^n$ sätter vi nu

$$D^2 f(a)[u,v] = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i v_j.$$

Funktionen $(u,v) \mapsto D^2 f(a)[u,v]$ är en symmetrisk bilinjär form på $\mathbf{R}^n$, och motsvarande symmetriska linjära operator kallas *andraderivatan* av f i punkten a och betecknas $f''(a)$. Andraderivatans matris, dvs. matrisen

$$\Big[\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\Big]_{i,j=1}^{n},$$

kallas *hessianen* (eller *Hessematrisen*) till f (i punkten a), och eftersom vi inte skiljer på matriser och operatorer, använder vi $f''(a)$ också som beteckning på hessianen.

Uttryckt med hjälp av $f''(a)$, uppfattad som operator resp. matris, är tydligen

$$D^2 f(a)[u,v] = \langle u, f''(a)v\rangle = u^T f''(a) v.$$

Vi erinrar om *Taylors formel*, som för två gånger differentierbara funktioner får följande utseende.

Sats 1.1.3. *Antag att funktionen f är två gånger differentierbar i en omgivning av punkten a. Då är*

$$f(a+v) = f(a) + Df(a)[v] + \tfrac{1}{2} D^2 f(a)[v,v] + r(v)$$

med en restterm som uppfyller $\lim_{v\to 0} r(v)/\|v\|^2 = 0$.

Tre gånger differentierbara funktioner

Vi kommer också att få anledning att betrakta tre gånger differentierbara funktioner f som är definierade på någon öppen delmängd U av $\mathbf{R}^n$. För $a \in U$ och godtyckliga vektorer $u, v, w \in \mathbf{R}^n$ sätter vi då

$$D^3 f(a)[u,v,w] = \sum_{i,j,k=1}^{n} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(a) u_i v_j w_k,$$

och får på så sätt för varje a en trilinjär symmetrisk form.

Vi överlåter åt läsaren att formulera Taylors formel för tre gånger differentierbara funktioner och noterar istället följande deriveringsregler som följer av kedjeregeln:

$$\begin{aligned}
\frac{d}{dt} f(x+tv) &= Df(x+tv)[v] \\
\frac{d}{dt}\Big(Df(x+tv)[u]\Big) &= D^2 f(x+tv)[u,v], \\
\frac{d}{dt}\Big(D^2 f(x+tw)[u,v]\Big) &= D^3 f(x+tw)[u,v,w].
\end{aligned}$$

Om ϕ betecknar restriktionen av funktionen f till linjen genom punkten x med riktningen v, dvs.

$$\phi(t) = f(x+tv),$$

så är alltså speciellt

$$\begin{aligned}
\phi'(t) &= Df(x+tv)[v], \\
\phi''(t) &= D^2 f(x+tv)[v,v], \\
\phi'''(t) &= D^3 f(x+tv)[v,v,v].
\end{aligned}$$

Kapitel 2

Konvexa mängder

2.1 Affina mängder och avbildningar

Affina mängder

Definition. En delmängd av $\mathbf{R}^n$ kallas *affin* om den för varje par av skilda punkter i mängden också innehåller hela linjen genom dessa punkter.

En mängd X är med andra ord affin om och endast om

$$x, y \in X,\ \lambda \in \mathbf{R} \Rightarrow\ \lambda x + (1-\lambda)y \in X.$$

Den tomma mängden $\emptyset$, hela rummet $\mathbf{R}^n$, linjära delrum av $\mathbf{R}^n$, enpunktsmängder $\{x\}$ och linjer är exempel på affina mängder.

Definition. En linjärkombination $y = \sum_{j=1}^{m} \alpha_j x_j$ av vektorer $x_1, x_2, \ldots, x_m$ kallas en *affin kombination* om $\sum_{j=1}^{m} \alpha_j = 1$.

Sats 2.1.1. *En affin mängd innehåller alla affina kombinationer av sina element.*

Bevis. Låt X vara en godtycklig affin mängd. En affin kombination av ett element är elementet självt, så X innehåller alla affina kombinationer som kan bildas av ett element i mängden.

Antag induktivt att X innehåller alla affina kombinationer som kan bildas av $m-1$ stycken element ur X, där $m \geq 2$, och betrakta en godtycklig affin kombination $x = \sum_{j=1}^{m} \alpha_j x_j$ av m element $x_1, x_2, \ldots, x_m$ i X. Eftersom $\sum_{j=1}^{m} \alpha_j = 1$, måste någon koefficient α_j vara skild från 1; antag utan inskränkning att $\alpha_m \neq 1$, och sätt $s = 1 - \alpha_m = \sum_{j=1}^{m-1} \alpha_j$. Då är $s \neq 0$ och

$\sum_{j=1}^{m-1} \alpha_j/s = 1$, vilket innebär att elementet

$$y = \sum_{j=1}^{m-1} \frac{\alpha_j}{s} x_j$$

är en affin kombination av $m-1$ stycken element i X. Enligt induktionsantagandet ligger därför y i X. Men $x = sy + (1-s)x_m$, så det följer av affinitetsdefinitionen att x ligger i X, och därmed är induktionssteget genomfört och satsen bevisad. □

Definition. Låt A vara en godtycklig icke-tom mängd i $\mathbf{R}^n$. Mängden av alla affina kombinationer $\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_m a_m$ som kan bildas av ett godtyckligt antal element $a_1, a_2, \ldots, a_m$ från A, kallas A:s *affina hölje* och betecknas aff A .

För att det affina höljet även skall vara definierat för den tomma mängden sätter vi aff $\emptyset = \emptyset$.

Sats 2.1.2. *Affina höljet* aff A *är en affin mängd som innehåller* A *som delmängd, och det är den minsta affina delmängden med denna egenskap, dvs. om mängden* X *är affin och* $A \subseteq X$*, så gäller* aff $A \subseteq X$.

Bevis. Att en affin kombination av två element i aff A är en ny affin kombination av element från A, dvs. tillhör aff A, är uppenbart, så aff A är en affin mängd. Att A är en delmängd av aff A är också uppenbart, ty varje element är en affin kombination av sig självt.

En affin mängd X innehåller enligt sats 2.1.1 varje affin kombination av sina element; om $A \subseteq X$ så innehåller därför speciellt X alla affina kombinationer av element hämtade från A, vilket innebär att aff A är en delmängd av X. □

Karakterisering av affina mängder

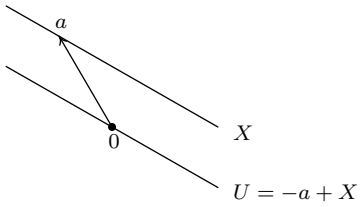
Icke-tomma affina mängder är translat till linjära delrum. Mer precist gäller:

Sats 2.1.3. *Antag att* X *är affin mängd i* $\mathbf{R}^n$ *och att* $a \in X$*. Då är translatet* $-a+X$ *ett linjärt delrum till* $\mathbf{R}^n$*. För varje* $b \in X$ *är vidare* $-b+X = -a+X$.

Till varje affin icke-tom mängd X *hör med andra ord ett entydigt bestämt linjärt delrum* U *så att* $X = a + U$.

Bevis. Sätt $U = -a + X$. Om $u_1 = -a + x_1$ och $u_2 = -a + x_2$ är två element i U och α_1, α_2 är godtyckliga reella tal, så är linjärkombinationen

$$\alpha_1 u_1 + \alpha_2 u_2 = -a + (1 - \alpha_1 - \alpha_2)a + \alpha_1 x_1 + \alpha_2 x_2$$



Figur 2.1. Illustration till sats 2.1.3: En affin mängd X och motsvarande linjära delrum U.

också ett element i U beroende på att $(1-\alpha_1-\alpha_2)a+\alpha_1x_1+\alpha_2x_2$ är en affin kombination av element i X och därför tillhör X enligt sats 2.1.1. Detta visar att U är ett linjärt delrum.

Antag vidare att $b \in X$ och att $v = -b + x$ är ett element i $-b + X$. Genom att skriva v på formen $v = -a + (a - b + x)$ ser vi att v också ligger i $-a + X$, ty $a - b + x$ är en affin kombination av element i X. Detta visar inklusionen $-b + X \subseteq -a + X$, och den omvända inklusionen följer förstås av symmetriskäl. Således är $-a + X = -b + X$. □

Dimension

Sats 2.1.3 möjliggör följande definition.

Definition. Med *dimensionen* $\dim X$ hos en icke-tom affin mängd X menas dimensionen hos det linjära delrummet $-a + X$, där a är ett godtyckligt element i X.

Eftersom varje icke-tom affin delmängd har en väldefinierad dimension, kan vi utvidga dimensionsbegreppet till godtyckliga icke-tomma mängder på följande vis.

Definition. Låt A vara en godtycklig icke-tom delmängd av $\mathbf{R}^n$. Med mängdens *(affina) dimension* $\dim A$ menas dimensionen hos mängdens affina hölje $\operatorname{aff} A$.

I $\mathbf{R}^n$ har varje sträcka $[x, y]$ dimension 1, och varje öppen boll $B(a; r)$ har dimension n.

Dimensionen är uppenbarligen *invariant under translation* och *växande*, dvs. för alla vektorer a och icke-tomma mängder A, B gäller:

$$\dim(a + A) = \dim A \qquad \text{och} \qquad A \subseteq B \Rightarrow \dim A \leq \dim B.$$

Lösningsmängder till linjära ekvationssystem

Följande sats ger en fullständig beskrivning av de affina mängderna i $\mathbf{R}^n$.

Sats 2.1.4. *Varje affin delmängd av $\mathbf{R}^n$ är lösningsmängd till ett linjärt ekvationssystem*

$$\begin{cases} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n = b_1 \\ c_{21}x_1 + c_{22}x_2 + \cdots + c_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ c_{m1}x_1 + c_{m2}x_2 + \cdots + c_{mn}x_n = b_m \end{cases}$$

och omvänt. Icke-tomma affina mängders dimension är lika med $n - r$, där r är rangen hos koefficientmatrisen C.

Bevis. Den tomma affina mängden fås som lösningsmängd till ett inkonsistent system, så vi behöver bara betrakta icke-tomma affina mängder X, och dessa har formen $X = x_0 + U$, där x_0 ligger i X och U är ett linjärt delrum av $\mathbf{R}^n$. Varje linjärt delrum är lösningsmängd till något homogent ekvationssystem, så det finns alltså en matris C så att $U = \{x \mid Cx = 0\}$, och $\dim U = n - \operatorname{rang} C$. Med $b = Cx_0$ gäller därför att $x \in X$ om och endast om $Cx - Cx_0 = C(x - x_0) = 0$, dvs. om och endast om x är en lösning till ekvationssystemet $Cx = b$.

Omvänt, om $Cx_0 = b$ så är x en lösning till ekvationssystemet $Cx = b$ om och endast om vektorn $x - x_0$ ligger i lösningsrummet U till det homogena ekvationssystemet $Cx = 0$. Det följer att lösningsmängden till ekvationssystemet $Cx = b$ har formen $x_0 + U$, dvs. är en affin mängd. □

Hyperplan

Definition. Affina delmängder till $\mathbf{R}^n$ av dimension $n - 1$ kallas *hyperplan.*

Sats 2.1.4 har följande korollarium:

Korollarium 2.1.5. *En delmängd X av $\mathbf{R}^n$ är ett hyperplan om och endast om det finns en nollskild vektor $c = (c_1, c_2, \ldots, c_n)$ och ett reellt tal b så att $X = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = b\}$.*

Varje affin äkta delmängd av $\mathbf{R}^n$ kan följaktligen enligt sats 2.1.4 framställas som ett snitt av hyperplan.

Affina avbildningar

Definition. Låt X vara en affin delmängd av $\mathbf{R}^n$. En avbildning $T\colon X \to \mathbf{R}^m$ kallas *affin* om

$$T(\lambda x + (1-\lambda)y) = \lambda Tx + (1-\lambda)Ty$$

för alla $x, y \in X$ och alla $\lambda \in \mathbf{R}$.

Med induktion visar man lätt att om $T\colon X \to \mathbf{R}^m$ är en affin avbildning och $x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$ är en affin kombination av element i X, så är

$$Tx = \alpha_1 Tx_1 + \alpha_2 Tx_2 + \cdots + \alpha_m Tx_m.$$

Om Y är en affin delmängd av X, så är vidare bildmängden $T(Y)$ en affin delmängd av $\mathbf{R}^m$, och om Z är en affin delmängd av $\mathbf{R}^m$ så är inversa bilden $T^{-1}(Z)$ en affin delmängd av X.

Sammansättningen av två affina avbildningar är uppenbarligen affin. Speciellt är en linjär avbildning följd av en translation affin, och nästa sats visar att varje affin avbildning kan skrivas som en sådan sammansättning.

Sats 2.1.6. *Antag att $T\colon X \to \mathbf{R}^m$ är en affin avbildning och att $X \subseteq \mathbf{R}^n$. Då finns det en linjär avbildning $C\colon \mathbf{R}^n \to \mathbf{R}^m$ och en vektor v i $\mathbf{R}^m$ så att*

$$Tx = Cx + v$$

för alla $x \in X$.

Bevis. Skriv definitionsmängden på formen $X = x_0 + U$ med $x_0 \in X$ och U som ett linjärt delrum av $\mathbf{R}^n$, och definiera avbildningen C på delrummet U genom att sätta

$$Cu = T(x_0 + u) - Tx_0.$$

För $u_1, u_2 \in U$ och $\alpha_1, \alpha_2 \in \mathbf{R}$ blir då

$$\begin{aligned} C(\alpha_1 u_1 + \alpha_2 u_2) &= T(x_0 + \alpha_1 u_1 + \alpha_2 u_2) - Tx_0 \\ &= T\big(\alpha_1(x_0 + u_1) + \alpha_2(x_0 + u_2) + (1 - \alpha_1 - \alpha_2)x_0\big) - Tx_0 \\ &= \alpha_1 T(x_0 + u_1) + \alpha_2 T(x_0 + u_2) + (1 - \alpha_1 - \alpha_2)Tx_0 - Tx_0 \\ &= \alpha_1\big(T(x_0 + u_1) - Tx_0\big) + \alpha_2\big(T(x_0 + u_2) - Tx_0\big) \\ &= \alpha_1 Cu_1 + \alpha_2 Cu_2. \end{aligned}$$

Avbildningen C är med andra ord linjär på U och kan förstås utvidgas till en linjär avbildning på hela $\mathbf{R}^n$.

För $x \in X$ blir nu, eftersom $x - x_0$ ligger i U,

$$Tx = T(x_0 + (x - x_0)) = C(x - x_0) + Tx_0 = Cx - Cx_0 + Tx_0,$$

vilket visar att satsen gäller med $v = Tx_0 - Cx_0$. □

2.2 Konvexa mängder

Grundläggande definitioner och egenskaper

Definition. En delmängd X av $\mathbf{R}^n$ kallas *konvex* om $[x, y] \subseteq X$ för alla x och y i X.

En mängd X är med andra ord konvex om och endast om den innehåller sträckan mellan varje par av sina punkter.

Figur 2.2. Konvex och icke-konvex mängd

Exempel 2.2.1. Affina mängder är uppenbarligen konvexa. Speciellt är den tomma mängden $\emptyset$, hela rummet $\mathbf{R}^n$ och linjära delrum konvexa mängder. Öppna och slutna sträckor är konvexa mängder. □

Exempel 2.2.2. För godtyckliga normer $\|\cdot\|$ är motsvarande öppna bollar $B(a; r)$ konvexa mängder. Detta följer av triangelolikheten och homogenitet; för $x, y \in B(a; r)$ och $0 \leq \lambda \leq 1$ är nämligen

$$\begin{aligned}\|\lambda x + (1-\lambda)y - a\| &= \|\lambda(x-a) + (1-\lambda)(y-a)\| \\ &\leq \lambda\|x-a\| + (1-\lambda)\|y-a\| < \lambda r + (1-\lambda)r = r,\end{aligned}$$

vilket innebär att varje punkt $\lambda x + (1-\lambda)y$ på sträckan $[x, y]$ ligger i $B(a; r)$.

Motsvarande slutna bollar $\overline{B}(a; r) = \{x \in \mathbf{R}^n \mid \|x - a\| \leq r\}$ är förstås också konvexa. □

Definition. En linjärkombination $y = \sum_{j=1}^{m} \alpha_j x_j$ av vektorer $x_1, x_2, \ldots, x_m$ kallas en *konvex kombination* om $\sum_{j=1}^{m} \alpha_j = 1$ och $\alpha_j \geq 0$ för alla j.

Sats 2.2.1. *En konvex mängd innehåller alla konvexa kombinationer av sina element.*

Bevis. Låt X vara en godtycklig konvex mängd. En konvex kombination av ett element är elementet självt, så X innehåller alla konvexa kombinationer som kan bildas av ett element i mängden. Antag induktivt att X innehåller alla konvexa kombinationer som kan bildas av $m-1$ stycken element ur X,

och betrakta en godtycklig konvex kombination $x = \sum_{j=1}^{m} \alpha_j x_j$ av $m \geq 2$ stycken element $x_1, x_2, \ldots, x_m$ i X. Eftersom $\sum_{j=1}^{m} \alpha_j = 1$, måste någon koefficient α_j vara strikt mindre än 1; antag utan inskränkning att $\alpha_m < 1$, och sätt $s = 1 - \alpha_m = \sum_{j=1}^{m-1} \alpha_j$. Då är förstås $s > 0$ och $\sum_{j=1}^{m-1} \alpha_j/s = 1$, vilket innebär att

$$y = \sum_{j=1}^{m-1} \frac{\alpha_j}{s} x_j$$

är en konvex kombination av $m-1$ stycken element i X. Enligt induktionsantagandet ligger därför y i X. Men $x = sy + (1-s)x_m$, så det följer av konvexitetsdefinitionen att x ligger i X, och därmed är induktionssteget genomfört, vilket bevisar satsen. □

2.3 Konvexitetsbevarande operationer

Vi skall nu beskriva ett antal sätt att konstruera konvexa mängder utifrån givna konvexa mängder.

Bilder och inversa bilder till affina avbildningar

Sats 2.3.1. *Antag att avbildningen $T\colon V \to \mathbf{R}^m$ är affin, att X är en konvex delmängd av V och att Y är en konvex delmängd av $\mathbf{R}^m$. Då är bilden $T(X)$ och inversa bilden $T^{-1}(Y)$ konvexa mängder.*

Bevis. Antag att y_1, y_2 är två punkter i $T(X)$. Då finns det $x_1, x_2 \in X$ så att $y_i = T(x_i)$, och för $0 \leq \lambda \leq 1$ är

$$\lambda y_1 + (1-\lambda)y_2 = \lambda T x_1 + (1-\lambda) T x_2 = T(\lambda x_1 + (1-\lambda)x_2)$$

ett element i $T(X)$ eftersom $\lambda x_1 + (1-\lambda)x_2$ ligger i X. Detta visar att bildmängden $T(X)$ är konvex.

För att visa att inversa bilden $T^{-1}(Y)$ är konvex antar vi istället att $x_1, x_2 \in T^{-1}(Y)$, dvs. att $Tx_1, Tx_2 \in Y$, och att $0 \leq \lambda \leq 1$. Eftersom Y är konvex, är

$$T(\lambda x_1 + (1-\lambda)x_2) = \lambda T x_1 + (1-\lambda) T x_2$$

ett element i Y, och detta innebär att $\lambda x_1 + (1-\lambda)x_2$ ligger i $T^{-1}(Y)$. □

Som specialfall av föregående sats följer att translaten $a+X$ till en konvex mängd X är konvexa.

EXEMPEL 2.3.1. Mängder av typen

$$\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq b\} \quad \text{och} \quad \{x \in \mathbf{R}^n \mid \langle c, x\rangle \leq b\},$$

där b är ett godtyckligt reellt tal och $c = (c_1, c_2, \ldots, c_n)$ är en godtycklig nollskild vektor, kallas motsatta *slutna halvrum*, medan deras komplement, dvs.

$$\{x \in \mathbf{R}^n \mid \langle c, x\rangle < b\} \quad \text{och} \quad \{x \in \mathbf{R}^n \mid \langle c, x\rangle > b\},$$

kallas *öppna halvrum*.

Halvrummen $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq b\}$ och $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > b\}$ är inversa bilder av intervallen $[b, \infty[$ och $]b, \infty[$ till den linjära avbildningen $x \mapsto \langle c, x\rangle$. Det följer därför av sats 2.3.1 att *halvrum är konvexa mängder.* □

Snitt och union

Sats 2.3.2. *Låt $\{X_i \mid i \in I\}$ vara en familj av konvexa mängder i $\mathbf{R}^n$. Då är snittet $\bigcap\{X_i \mid i \in I\}$ konvext.*

Bevis. Antag att x, y är punkter i snittet Y. Definitionen av snitt innebär att x och y ligger i X_i för alla $i \in I$, och konvexiteten medför att $[x, y] \subseteq X_i$ för alla $i \in I$. Enligt definitionen av snitt gäller därför $[x, y] \subseteq Y$. Detta visar att snittet är konvext. □

En union av konvexa mängder är förstås i allmänhet inte konvex. I ett trivialt fall blir dock unionen konvex, nämligen om mängderna kan ordnas så att de bildar en ”växande kedja”.

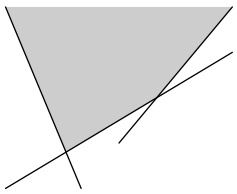
Sats 2.3.3. *Antag att $\{X_i \mid i \in I\}$ är en familj av konvexa mängder X_i och att för varje par $i, j \in I$ antingen $X_i \subseteq X_j$ eller $X_j \subseteq X_i$. Då är unionen $\bigcup\{X_i \mid i \in I\}$ konvex.*

Bevis. Förutsättningarna innebär att det för varje par av punkter x, y i unionen finns ett index $i \in I$ så att båda punkterna ligger i X_i, och då ligger hela sträckan $[x, y]$ i X_i och därmed också i unionen. □

EXEMPEL 2.3.2. Konvexiteten hos slutna bollar följer av konvexiteten hos öppna bollar eftersom $\overline{B}(a; r_0) = \bigcap\{B(a; r) \mid r > r_0\}$.

Omvänt följer konvexiteten hos öppna bollar av konvexiteten hos slutna bollar, eftersom $B(a; r_0) = \bigcup\{\overline{B}(a; r) \mid r < r_0\}$ och mängderna $\overline{B}(a; r)$ bildar en växande kedja. □

Definition. En delmängd X av $\mathbf{R}^n$ kallas en *polyeder* om X kan skrivas som ett snitt av ändligt många slutna halvrum eller om $X = \mathbf{R}^n$.[†]



Figur 2.3. Polyeder i $\mathbf{R}^2$

Polyedrar är konvexa mängder enligt sats 2.3.2, och de kan representeras som lösningsmängder till system av linjära olikheter. Vi kan, genom att eventuellt multiplicera några av olikheterna med -1, anta att alla olikheterna är av formen $c_1x_1 + c_2x_2 + \cdots + c_nx_n \geq d$. Detta innebär att varje polyeder är lösningsmängd till ett system av följande slag

$$\begin{cases} c_{11}x_1 + c_{12}x_2 + \cdots + c_{1n}x_n \geq b_1 \\ c_{21}x_1 + c_{22}x_2 + \cdots + c_{2n}x_n \geq b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ c_{m1}x_1 + c_{m2}x_2 + \cdots + c_{mn}x_n \geq b_m. \end{cases}$$

På matrisform får förstås systemet formen

$$Cx \geq b.$$

Ett hyperplan $\{x \mid \langle c, x\rangle = b\}$ är en polyeder eftersom det är lika med snittet av de två slutna halvrummen $\{x \mid \langle c, x\rangle \geq b\}$ och $\{x \mid \langle c, x\rangle \leq b\}$. Varje affin mängd (utom hela rummet) är ett snitt av hyperplan och är därför en polyeder. Speciellt är den tomma mängden en polyeder.

Cartesiansk produkt

Sats 2.3.4. *Den cartesianska produkten $X \times Y$ av två konvexa mängder X och Y är en konvex mängd.*

Bevis. Antag att X ligger i $\mathbf{R}^n$ och Y ligger i $\mathbf{R}^m$. Projektionerna

$$P_1 \colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n \quad \text{och} \quad P_2 \colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^m,$$

[†]Snittet av en tom familj av mängder brukar definieras som hela rummet, och med den konventionen kan även polyedern $\mathbf{R}^n$ uppfattas som ett snitt av halvrum.

definierade som $P_1(x, y) = x$ och $P_2(x, y) = y$, är linjära avbildningar, och

$$X \times Y = (X \times \mathbf{R}^m) \cap (\mathbf{R}^n \times Y) = P_1^{-1}(X) \cap P_2^{-1}(Y).$$

Påståendet följer därför av satserna 2.3.1 och 2.3.2. □

Summa

Sats 2.3.5. *Summan $X + Y$ av två konvexa delmängder X och Y av $\mathbf{R}^n$ är konvex, och produkten αX av ett tal och en konvex mängd i $\mathbf{R}^n$ är konvex.*

Bevis. Avbildningarna $S\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ och $T\colon \mathbf{R}^n \to \mathbf{R}^n$, definierade som $S(x, y) = x + y$ och $Tx = \alpha x$, är linjära. Eftersom $X + Y = S(X \times Y)$ och $\alpha X = T(X)$, följer därför påståendena av satserna 2.3.1 och 2.3.4. □

Exempel 2.3.3. Mängden $X(r)$ av alla punkter vars avstånd till en given mängd X är mindre än det positiva talet r kan skrivas som en summa, nämligen

$$X(r) = X + B(0; r).$$

Det följer att mängden $X(r)$ är konvex om X är konvex, ty bollar $B(0; r)$ är konvexa mängder. □

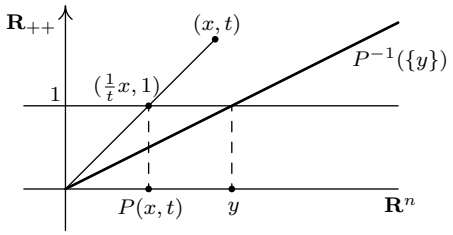
Bilder och inversa bilder till perspektivavbildningen

Definition. *Perspektivavbildningen $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$* definieras av att

$$P(x, t) = t^{-1}x$$

för alla $x \in \mathbf{R}^n$ och $t > 0$.

Perspektivavbildningen skalar alltså om punkter i $\mathbf{R}^n \times \mathbf{R}_{++}$ så att sista koordinaten blir 1 och kastar sedan bort den sista koordinaten. Figur 2.4 illustrerar det hela.



Figur 2.4. Perspektivavbildningen. Inversa bilden av $y \in \mathbf{R}^n$ är en halvlinje.

Sats 2.3.6. *Antag att X är en konvex delmängd till $\mathbf{R}^n \times \mathbf{R}_{++}$ och att Y är en konvex delmängd av $\mathbf{R}^n$. Då är bilden $P(X)$ av X och inversa bilden $P^{-1}(Y)$ av Y under perspektivavbildningen $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$ konvexa mängder.*

Bevis. För att visa att bilden $P(X)$ är konvex antar vi att $y, y' \in P(X)$. Då finns $t, t' > 0$ så att punkterna (ty, t) och $(t'y', t')$ ligger i X. Vi har att visa att punkten $\lambda y + (1-\lambda)y'$ ligger i $P(X)$ om $0 < \lambda < 1$. Sätt för den skull

$$\alpha = \frac{\lambda t'}{\lambda t' + (1-\lambda)t};$$

då är $0 < \alpha < 1$ och punkten

$$z = \alpha(ty, t) + (1-\alpha)(t'y', t') = \Big(\frac{tt'(\lambda y + (1-\lambda)y')}{\lambda t' + (1-\lambda)t}, \frac{tt'}{\lambda t' + (1-\lambda)t}\Big)$$

ligger därför i X. Punkten $P(z)$ ligger således i $P(X)$, och eftersom $P(z) = \lambda y + (1-\lambda)y'$ är beviset klart.

För att visa att inversa bilden $P^{-1}(Y)$ är konvex antar vi istället att (x, t) och (x', t') är punkter i $P^{-1}(Y)$ och att $0 < \lambda < 1$. Då ligger punkterna $\frac{1}{t}x$ och $\frac{1}{t'}x'$ i Y, och

$$\alpha = \frac{\lambda t}{\lambda t + (1-\lambda)t'}$$

är ett tal mellan 0 och 1. Eftersom Y är en konvex mängd, ligger punkten

$$z = \alpha\frac{1}{t}x + (1-\alpha)\frac{1}{t'}x' = \frac{\lambda x + (1-\lambda)x'}{\lambda t + (1-\lambda)t'}$$

också i Y, och följaktligen ligger punkten

$$\big((\lambda t + (1-\lambda)t')z, \lambda t + (1-\lambda)t'\big) = \lambda(x, t) + (1-\lambda)(x', t')$$

i $P^{-1}(Y)$, vilket visar att $P^{-1}(Y)$ är en konvex mängd. □

Exempel 2.3.4. Inversa bilden av enhetsbollen $B(0;1)$ under perspektivavbildningen är lika med mängden

$$\{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\| < x_{n+1}\},$$

som alltså är en konvex mängd i $\mathbf{R}^{n+1}$ för varje val av norm $\|\cdot\|$. Speciellt är således mängderna

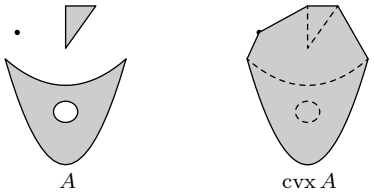
$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > |x_1| + |x_2| + \cdots + |x_n|\},$$
$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}\} \quad \text{och}$$
$$\{x \in \mathbf{R}^{n+1} \mid x_{n+1} > \max_{1\le i\le n} |x_i|\},$$

som fås genom att som normer välja ℓ^1-normen, den euklidiska normen och maxnormen, konvexa. □

2.4 Konvext hölje

Definition. Låt A vara en icke-tom mängd i $\mathbf{R}^n$. Mängden av alla konvexa kombinationer $\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_m a_m$ av ett godtyckligt antal element $a_1, a_2, \ldots, a_m$ i A kallas A:s *konvexa hölje* och betecknas $\operatorname{cvx} A$.

Vi sätter vidare $\operatorname{cvx} \emptyset = \emptyset$ så att det konvexa höljet även blir definierat för den tomma mängden.



Figur 2.5. En mängd A och dess konvexa hölje

Sats 2.4.1. *Konvexa höljet* $\operatorname{cvx} A$ *är en konvex mängd som innehåller* A *och är den minsta mängden med den egenskapen, dvs. om* X *är en konvex mängd och* $A \subseteq X$*, så gäller att* $\operatorname{cvx} A \subseteq X$*.*

Bevis. $\operatorname{cvx} A$ är en konvex mängd, ty konvexa kombinationer av två element av typen $\sum_{j=1}^{m} \lambda_j a_j$, där $m \geq 1$, $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$, $\sum_{j=1}^{m} \lambda_j = 1$ och $a_1, a_2, \ldots, a_m \in A$, är uppenbarligen ett element av samma slag. Vidare gäller att $A \subseteq \operatorname{cvx} A$, ty varje element i A är en konvex kombination av sig självt $(a = 1a)$.

En konvex mängd X innehåller enligt sats 2.2.1 alla konvexa kombinationer av sina element. Om $A \subseteq X$, så innehåller därför speciellt X alla konvexa kombinationer av elementen i A, vilket betyder att $\operatorname{cvx} A \subseteq X$. □

Konvexa höljet av en mängd i $\mathbf{R}^n$ består av alla konvexa kombinationer av ett godtyckligt antal element i mängden, men varje element i höljet är i själva verket en konvex kombination av högst $n + 1$ stycken element.

Sats 2.4.2. *Låt* $A \subseteq \mathbf{R}^n$ *och antag att* $x \in \operatorname{cvx} A$*. Då innehåller* A *en delmängd* B *med högst* $n + 1$ *element sådan att* $x \in \operatorname{cvx} B$*.*

Bevis. Enligt definitionen av konvext hölje finns det en ändlig delmängd B av A sådan att $x \in \operatorname{cvx} B$. Välj en sådan mängd $B = \{b_1, b_2, \ldots, b_m\}$ med så få element som möjligt. Minimalitetsantagandet innebär att $x = \sum_{j=1}^{m} \lambda_j b_j$, där $\sum_{j=1}^{m} \lambda_j = 1$ och $\lambda_j > 0$ för alla j.

Sätt $c_j = b_j - b_m$ för $j = 1, 2, \ldots, m-1$. Vi skall visa att mängden $C = \{c_1, c_2, \ldots, c_{m-1}\}$ är en linjärt oberoende delmängd av $\mathbf{R}^n$, något som förstås implicerar att $m \leq n+1$.

Antag motsatsen, dvs. att mängden C är linjärt beroende. Då finns det reella tal μ_j, som inte alla är 0, så att $\sum_{j=1}^{m-1} \mu_j c_j = 0$. Om vi definierar $\mu_m = -\sum_{j=1}^{m-1} \mu_j$, så är alltså $\sum_{j=1}^{m} \mu_j = 0$ och $\sum_{j=1}^{m} \mu_j b_j = 0$. Vidare är säkert något av de m talen $\mu_1, \mu_2, \ldots, \mu_m$ positivt.

Betrakta nu för $t > 0$ talen $\nu_j = \lambda_j - t\mu_j$. Vi observerar att

$$\sum_{j=1}^{m} \nu_j = \sum_{j=1}^{m} \lambda_j - t \sum_{j=1}^{m} \mu_j = 1 \quad \text{och} \quad \sum_{j=1}^{m} \nu_j b_j = \sum_{j=1}^{m} \lambda_j b_j - t \sum_{j=1}^{m} \mu_j b_j = x.$$

Vidare är $\nu_j \geq \lambda_j > 0$ om $\mu_j \leq 0$, och $\nu_j \geq 0$ om $\mu_j > 0$ och $t \leq \lambda_j/\mu_j$. Genom att välja t som det minsta av talen λ_j/μ_j med positiv nämnare μ_j, uppnår vi därför att $\nu_j \geq 0$ för alla j och att $\nu_{j_0} = 0$ för åtminstone ett index j_0. Detta innebär att x är en konvex kombination av de $m-1$ stycken elementen i mängden $B \setminus \{b_{j_0}\}$, vilket strider mot minimalitetsantagandet beträffande mängden B. □

2.5 Topologiska egenskaper

Slutna höljet

Sats 2.5.1. *Slutna höljet* $\operatorname{cl} X$ *av en konvex mängd* X *är konvext.*

Bevis. Vi erinrar om att $\operatorname{cl} X = \bigcap_{r>0} X(r)$, där $X(r)$ är mängden av alla punkter vars avstånd till X är mindre än r. För konvexa mängder X är mängderna $X(r)$ konvexa (se exempel 2.3.3), och snittet av konvexa mängder är konvext. □

Inre och relativt inre

För att en konvex mängd skall ha inre punkter krävs att mängdens dimension är lika med det omgivande rummets − exempelvis är det inre av en sträcka i $\mathbf{R}^n$ tomt om $n \geq 2$.

Sats 2.5.2. *En konvex mängd* X *i* $\mathbf{R}^n$ *har inre punkter om och endast om* $\dim X = n$.

Bevis. Om X har en inre punkt a, så finns det en öppen boll $B = B(a; r)$ med centrum i a så att $B \subseteq X$, och det följer därför att $\dim X \geq \dim B = n$, dvs. $\dim X = n$.

Antag omvänt att $\dim X = n$; vi skall visa att $\operatorname{int} X \neq \emptyset$. Eftersom dimensionen är invariant under translation och $\operatorname{int}(a + X) = a + \operatorname{int} X$, kan vi utan inskränkning anta att $0 \in X$.

Låt nu $a_1, a_2, \ldots, a_m$ vara ett maximalt antal linjärt oberoende vektorer i X; då är X en delmängd av det linjära delrum av dimension m som spänns upp av vektorerna, så det följer av dimensionalitetsantagandet att $m = n$. Mängden X innehåller som delmängd det konvexa höljet av vektorerna 0, $a_1, a_2, \ldots, a_n$, och därmed speciellt den icke-tomma öppna mängden

$$\{\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n \mid 0 < \lambda_1 + \cdots + \lambda_n < 1,\ \lambda_1 > 0, \ldots, \lambda_n > 0\}.$$

Detta visar att $\operatorname{int} X \neq \emptyset$. □

En sträcka $[a, b]$ på x-axeln i det tvådimensionella talplanet saknar inre punkter om vi betraktar sträckan som en delmängd av $\mathbf{R}^2$, men som delmängd av $\mathbf{R}$ är sträckans inre lika med den öppna sträckan $]a, b[$. Motsvarande gäller förstås för det konvexa höljet $T = \operatorname{cvx}\{a, b, c\}$ av tre punkter som inte ligger i linje; om triangeln T uppfattas som delmängd av $\mathbf{R}^2$ har den inre punkter, men om den uppfattas som delmängd av $\mathbf{R}^3$ saknar den sådana. Detta är naturligtvis otillfredsställande om man vill ha begrepp som är oberoende av det omgivande rummets dimension. Lösningen på dilemmat ligger i att använda sig av den relativa topologi som mängdens affina hölje "ärver" från det omgivande rummet $\mathbf{R}^n$.

Definition. Låt X vara en mängd av dimension m i $\mathbf{R}^n$, och låt V beteckna X:s affina hölje, dvs. V är den affina mängd av dimension m som innehåller X.

En punkt $x \in X$ säges vara en *relativt inre punkt* i X om det finns ett $r > 0$ så att $B(x; r) \cap V \subseteq X$, och mängden av alla relativt inre punkter i X kallas mängdens *relativa inre* och betecknas $\operatorname{rint} X$.

En punkt $x \in \mathbf{R}^n$ kallas en *relativ randpunkt* till X om det för varje $r > 0$ gäller att snittet $B(x; r) \cap V$ innehåller minst en punkt från X och minst en punkt från $V \setminus X$. Mängden av alla relativa randpunkter kallas den *relativa randen* till till X och betecknas $\operatorname{rbdry} X$.

En relativt inre punkt i X ligger naturligtvis i X, och en relativ randpunkt till X är också en randpunkt till X i vanlig mening och ligger följaktligen i tillslutningen $\operatorname{cl} X$. Detta betyder att $\operatorname{rint} X \cup \operatorname{rbdry} X \subseteq \operatorname{cl} X$.

Omvänt, om x är en punkt i tillslutningen $\operatorname{cl} X$, så är

$$B(x,r) \cap V \cap X = B(x,r) \cap X \neq \emptyset$$

för varje $r > 0$, och följaktligen är x antingen en relativ randpunkt eller en relativt inre punkt till X. Detta visar den omvända inklusionen, och vi drar slutsatsen att

$$\operatorname{rint} X \cup \operatorname{rbdry} X = \operatorname{cl} X,$$

eller med andra ord att

$$\operatorname{rbdry} X = \operatorname{cl} X \setminus \operatorname{rint} X.$$

Det följer av sats 2.5.2 att alla icke-tomma konvexa mängder har icke-tomt relativt inre. Notera att för enpunktsmängder $\{a\}$ är $\operatorname{rint}\{a\} = \{a\}$. För sträckor är $\operatorname{rint}[a,b] =]a,b[$, vilket är konsistent med vårt språkbruk att kalla $]a,b[$ en öppen sträcka.

Sats 2.5.3. *Det relativa inre* $\operatorname{rint} X$ *av en konvex mängd* X *är konvext.*

Bevis. Eftersom $\operatorname{rint} X \subseteq \operatorname{cl} X$, är satsen ett korollarium till följande sats. □

Sats 2.5.4. *Om* X *är en konvex mängd,* $a \in \operatorname{rint} X$ *och* $b \in \operatorname{cl} X$*, så ligger hela den öppna sträckan* $]a,b[$ *i* $\operatorname{rint} X$*.*

Bevis. Låt $V = \operatorname{aff} X$ beteckna den affina mängd av minimal dimension som innehåller X, och låt $c = \lambda a + (1-\lambda)b$, där $0 < \lambda < 1$, vara en godtycklig punkt på den öppna sträckan $]a,b[$. Vi skall visa att c är en relativt inre punkt i X genom att konstruera en öppen boll B som innehåller c och vars snitt med V ligger helt i X.

Välj för den skull $r > 0$ så att $B(a;r) \cap V \subseteq X$ och en punkt $b' \in X$ så att $\|b' - b\| < \lambda r/(1-\lambda)$; detta går eftersom a är en relativt inre punkt i X och b är en punkt i slutna höljet till X. Sätt

$$B = \lambda B(a;r) + (1-\lambda)b',$$

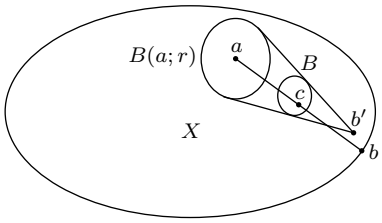
och notera att $B = B(\lambda a + (1-\lambda)b'; \lambda r)$. Den öppna bollen B innehåller punkten c eftersom

$$\|c - (\lambda a + (1-\lambda)b')\| = \|(1-\lambda)(b-b')\| = (1-\lambda)\|b-b'\| < \lambda r.$$

Vidare gäller

$$B \cap V = \lambda(B(a;r) \cap V) + (1-\lambda)b' \subseteq \lambda X + (1-\lambda)X \subseteq X$$

på grund av konvexitet, och därmed är beviset klart. □



Figur 2.6. Illustration till beviset för sats 2.5.4. Konvexa höljet till bollen $B(a;r)$ och punkten b' bildar en "strut" med icke-tomt inre som innehåller punkten c.

Sats 2.5.5. *För konvexa mängder X är*

(i) $\operatorname{cl}(\operatorname{rint} X) = \operatorname{cl} X$;

(ii) $\operatorname{rint}(\operatorname{cl} X) = \operatorname{rint} X$;

(iii) $\operatorname{rbdry}(\operatorname{cl} X) = \operatorname{rbdry}(\operatorname{rint} X) = \operatorname{rbdry} X$.

Bevis. Likheterna i (iii) för relativa ränder följer av de övriga två likheterna samt definitionen av relativ rand.

De båda inklusionerna $\operatorname{cl}(\operatorname{rint} X) \subseteq \operatorname{cl} X$ och $\operatorname{rint} X \subseteq \operatorname{rint}(\operatorname{cl} X)$ är triviala, eftersom $A \subseteq B$ medför $\operatorname{cl} A \subseteq \operatorname{cl} B$ och $\operatorname{rint} A \subseteq \operatorname{rint} B$ för godtyckliga mängder A och B.

Det återstår således enbart att bevisa de båda inklusionerna

$$\operatorname{cl} X \subseteq \operatorname{cl}(\operatorname{rint} X) \quad \text{och} \quad \operatorname{rint}(\operatorname{cl} X) \subseteq \operatorname{rint} X.$$

Fixera för den skull en punkt $x_0 \in \operatorname{rint} X$.

Om $x \in \operatorname{cl} X$, så ligger enligt sats 2.5.4 varje punkt på den öppna sträckan $]x_0, x[$ i $\operatorname{rint} X$, och det följer förstås av detta att punkten x antingen är en inre punkt eller en randpunkt till $\operatorname{rint} X$, dvs. en punkt i tillslutningen $\operatorname{cl}(\operatorname{rint} X)$. Detta visar inklusionen $\operatorname{cl} X \subseteq \operatorname{cl}(\operatorname{rint} X)$.

För att visa den återstående inklusionen $\operatorname{rint}(\operatorname{cl} X) \subseteq \operatorname{rint} X$ antar vi istället att $x \in \operatorname{rint}(\operatorname{cl} X)$ och sätter $y_t = (1-t)x_0 + tx$ för $t > 1$. Eftersom $y_t \to x$ då $t \to 1$, ligger punkterna y_t i $\operatorname{cl} X$ förutsatt att t ligger tillräckligt nära 1. Välj ett tal $t_0 > 1$ så att y_{t_0} tillhör $\operatorname{cl} X$. Enligt sats 2.5.4 är alla punkter på den öppna sträckan $]x_0, y_{t_0}[$ relativt inre punkter i X, och x är en sådan punkt eftersom $x = \frac{1}{t_0} y_{t_0} + (1 - \frac{1}{t_0})x_0$. Detta bevisar implikationen $x \in \operatorname{rint}(\operatorname{cl} X) \Rightarrow x \in \operatorname{rint} X$, och därmed är beviset klart. □

Kompakthet

Sats 2.5.6. *Konvexa höljet* $\operatorname{cvx} A$ *av en kompakt mängd A i* $\mathbf{R}^n$ *är kompakt.*

Bevis. Sätt $S = \{\lambda \in \mathbf{R}^{n+1} \mid \lambda_1, \lambda_2, \ldots, \lambda_{n+1} \geq 0,\ \sum_{j=1}^{n+1} \lambda_j = 1\}$, och låt $f\colon S \times \mathbf{R}^n \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n \to \mathbf{R}^n$ vara funktionen

$$f(\lambda, x_1, x_2, \ldots, x_{n+1}) = \sum_{j=1}^{n+1} \lambda_j x_j.$$

Funktionen f är givetvis kontinuerlig, och mängden S är kompakt eftersom den är sluten och begränsad. Enligt sats 2.4.2 kan varje element $x \in \operatorname{cvx} A$ skrivas som en konvex kombination $x = \sum_{j=1}^{n+1} \lambda_j a_j$ av högst $n+1$ stycken element $a_1, a_2, \ldots, a_{n+1}$ ur mängden A. Detta betyder att konvexa höljet $\operatorname{cvx} A$ är lika med bilden $f(S \times A \times A \times \cdots \times A)$ under f av den kompakta produktmängden $S \times A \times A \times \cdots \times A$. Eftersom kontinuerliga funktioner avbildar kompakta mängder på kompakta mängder, är således konvexa höljet $\operatorname{cvx} A$ kompakt. □

2.6 Koner

Definition. Om x är en punkt i $\mathbf{R}^n$ och $x \neq 0$, så kallar vi mängden

$$\overrightarrow{x} = \{\lambda x \mid \lambda \geq 0\}$$

för *strålen* genom x eller *halvlinjen* från origo genom x.

En *kon* X i $\mathbf{R}^n$ är en icke-tom mängd som för varje punkt $x \in X$ innehåller hela strålen genom x.

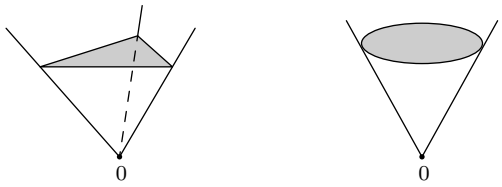
En kon X är med andra ord en icke-tom mängd som är sluten under multiplikation med icke-negativa tal, dvs.

$$x \in X,\ \lambda \geq 0 \Rightarrow \lambda x \in X.$$

Speciellt innehåller alltså alla koner 0.

Vi ska studera konvexa koner. Strålar och linjära delrum till $\mathbf{R}^n$ är förstås konvexa koner; speciellt är alltså hela rummet $\mathbf{R}^n$ och det triviala delrummet $\{0\}$ konvexa koner. Ytterligare exempel på enkla konvexa koner ges av följande exempel.

Exempel 2.6.1. Ett slutet halvrum $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$, som begränsas av ett hyperplan genom origo, är en konvex kon och kallas ett *koniskt halvrum.* Även unionen $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\} \cup \{0\}$ av motsvarande öppna halvrum och origo är en konvex kon. □



Figur 2.7. Plant snitt genom två äkta konvexa koner i $\mathbf{R}^3$

EXEMPEL 2.6.2. Icke-negativa ortanten

$$\mathbf{R}^n_+ = \{x = (x_1, \ldots, x_n) \in \mathbf{R}^n \mid x_1 \geq 0, \ldots, x_n \geq 0\},$$

i $\mathbf{R}^n$ är en konvex kon. □

Definition. En kon, som inte innehåller någon linje genom 0, kallas en *äkta* kon.‡

En kon innehåller en linje genom origo om och endast om det finns en vektor x sådan att både x och $-x$ ligger i konen. En kon X är med andra ord äkta om och endast om $X \cap (-X) = \{0\}$.

Slutna koniska halvrum i $\mathbf{R}^n$ är oäkta koner om $n \geq 2$, medan icke-negativa ortanten $\mathbf{R}^n_+$ är en äkta kon. Konerna $\{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\} \cup \{0\}$ är också äkta koner.

Här följer nu två alternativa sätt att uttrycka att en mängd är en konvex kon.

Sats 2.6.1. *Följande tre villkor är ekvivalenta för en icke-tom delmängd X av $\mathbf{R}^n$:*

(i) *X är en konvex kon.*

(ii) *X är en kon och $x + y \in X$ för alla $x, y \in X$.*

(iii) *$\lambda x + \mu y \in X$ för alla $x, y \in X$ och alla $\lambda, \mu \in \mathbf{R}_+$.*

Bevis. (i) ⇒ (ii): Om X är en konvex kon och $x, y \in X$, så tillhör $z = \frac{1}{2}x + \frac{1}{2}y$ mängden X på grund av konvexitet, och $x + y$ $(= 2z)$ tillhör X på grund av att X är en kon.

(ii) ⇒ (iii): Om (ii) gäller, $x, y \in X$ och $\lambda, \mu \in \mathbf{R}_+$, så följer först att $\lambda x \in X$ och $\mu y \in X$ på grund av konvillkoret och sedan att $\lambda x + \mu y \in X$ på grund av additivitetsvillkoret.

‡Terminologin skiftar från författare till författare. På engelska kallas en äkta kon vanligtvis för *salient cone*, medan begreppet *proper cone* kan betyda att konen ifråga är sluten, har ett icke-tomt inre och inte innehåller någon linje genom origo.

(iii) $\Rightarrow$ (i): Om (iii) gäller så följer att X är en kon genom att välja $y = x$ och $\mu = 0$, samt att konen är konvex genom att speciellt välja $\lambda + \mu = 1$. □

Definition. En linjärkombination $\sum_{j=1}^{m} \lambda_j x_j$ av vektorer $x_1, x_2, \ldots, x_m$ i $\mathbf{R}^n$ kallas en *konisk kombination* om samtliga koefficienter $\lambda_1, \lambda_2, \ldots, \lambda_m$ är icke-negativa.

Sats 2.6.2. *En konvex kon innehåller alla koniska kombinationer av sina element.*

Bevis. Följer omedelbart med induktion av karakteriseringen (iii) av konvexa koner i föregående sats. □

Konbevarande operationer

Bevisen för nedanstående fyra satser är analoga med bevisen för motsvarande satser om konvexa mängder och lämnas därför som övning.

Sats 2.6.3. *Antag att avbildningen $T\colon \mathbf{R}^n \to \mathbf{R}^m$ är linjär, att X är en konvex kon i $\mathbf{R}^n$ och att Y är en konvex kon i $\mathbf{R}^m$. Då är bildmängden $T(X)$ en konvex kon i $\mathbf{R}^m$ och inversa bildmängden $T^{-1}(Y)$ en konvex kon i $\mathbf{R}^n$.*

Sats 2.6.4. *Snittet $\bigcap_{i\in I} X_i$ av en godtycklig familj av konvexa koner X_i är en konvex kon.*

Sats 2.6.5. *Antag att X är en konvex kon i $\mathbf{R}^n$ och Y är en konvex kon i $\mathbf{R}^m$. Då är cartesianska produkten $X \times Y$ en konvex kon i $\mathbf{R}^{n+m}$.*

Sats 2.6.6. *Antag att X och Y är konvexa koner i $\mathbf{R}^n$. Då är $-X$ och $X+Y$ konvexa koner i $\mathbf{R}^n$.*

Exempel 2.6.3. Ett snitt

$$X = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \geq 0\}$$

av ändligt många slutna koniska halvrum kallas en *polyedrisk kon* eller en *konisk polyeder*.

En polyedrisk kon är med andra ord lösningsmängd till ett system av homogena linjära olikheter. Om vi låter C vara den $m \times n$-matris vars *rader* består av radmatriserna c_i^T, kan vi skriva konen X på formen

$$X = \{x \in \mathbf{R}^n \mid Cx \geq 0\}.$$

□

Koniskt hölje

Definition. Låt A vara en godtycklig icke-tom delmängd av $\mathbf{R}^n$. Mängden av alla koniska kombinationer av element ur A kallas A:s *koniska hölje* och betecknas $\operatorname{con} A$. Elementen i A kallas *generatorer* till $\operatorname{con} A$.

För att begreppet även skall vara definierat för den tomma mängden sätter vi $\operatorname{con} \emptyset = \{0\}$.

Sats 2.6.7. *För varje mängd A är $\operatorname{con} A$ en konvex kon som innehåller A som delmängd, och $\operatorname{con} A$ är den minsta konvexa konen med denna egenskap, dvs. om X är en godtycklig konvex kon och $A \subseteq X$, så gäller att $\operatorname{con} A \subseteq X$.*

Bevis. En konisk kombination av två koniska kombinationer av element från A är uppenbarligen en ny konisk kombination av element från A, så $\operatorname{con} A$ är en konvex kon. Att $A \subseteq \operatorname{con} A$ är uppenbart, och eftersom en godtycklig konvex kon X innehåller alla koniska kombinationer av sina element, måste en konvex kon X som innehåller A som delmängd speciellt innehålla alla koniska kombinationer av element från A, dvs. konen $\operatorname{con} A$. □

Sats 2.6.8. *Antag att $X = \operatorname{con} A$ är en kon i $\mathbf{R}^n$, att $Y = \operatorname{con} B$ är en kon i $\mathbf{R}^m$ och att $T\colon \mathbf{R}^n \to \mathbf{R}^m$ är en linjär avbildning. Då är*

(i) $T(X) = \operatorname{con} T(A)$;

(ii) $X \times Y = \operatorname{con}\big((A \times \{0\}) \cup (\{0\} \times B)\big)$;

(iii) $X + Y = \operatorname{con}(A \cup B)$, *förutsatt att $m = n$ så att summan $X + Y$ är väldefinierad.*

Bevis. (i) Konen X består av alla koniska kombinationer $x = \sum_{j=1}^{p} \lambda_j a_j$ av element a_j i A. För en sådan konisk kombination är $Tx = \sum_{j=1}^{p} \lambda_j T a_j$, så det följer att bildkonen $T(X)$ består av alla koniska kombinationer av elementen $Ta_j \in T(A)$. Detta innebär att $T(X) = \operatorname{con} T(A)$.

(ii) Konen $X \times Y$ består av alla par $(x, y) = \big(\sum_{j=1}^{p} \lambda_j a_j, \sum_{k=1}^{q} \mu_k b_k\big)$ av koniska kombinationer av element i A resp. B. Men

$$(x, y) = \sum_{j=1}^{p} \lambda_j (a_j, 0) + \sum_{k=1}^{q} \mu_k (0, b_k),$$

så (x, y) är en konisk kombination av element i $(A \times \{0\}) \cup (\{0\} \times B)$, dvs. ett element i konen $Z = \operatorname{con}\big((A \times \{0\}) \cup (\{0\} \times B)\big)$. Detta bevisar inklusionen $X \times Y \subseteq Z$.

Den omvända inklusionen $Z \subseteq X \times Y$ följer direkt av den triviala inklusionen $(A \times \{0\}) \cup (\{0\} \times B) \subseteq X \times Y$, och det faktum att $X \times Y$ är en kon.

(iii) Ett typiskt element i $X+Y$ har formen $\sum_{j=1}^{p} \lambda_j a_j + \sum_{k=1}^{q} \mu_k b_k$, vilket är en konisk kombination av element i $A \cup B$. Detta bevisar påståendet. □

Ändligt genererade koner

Definition. En konvex kon X säges vara *ändligt genererad* om $X = \operatorname{con} A$ för någon ändlig mängd A.

EXEMPEL 2.6.4. Icke-negativa ortanten $\mathbf{R}^n_+$ är ändligt genererad av standardbasen $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ för $\mathbf{R}^n$. □

Sats 2.6.8 har följande omedelbara korollarium.

Korollarium 2.6.9. *Cartesianska produkter $X \times Y$, summor $X+Y$ och bilder $T(X)$ under linjära avbildningar T av ändligt genererade koner X och Y, är ändligt genererade koner.*

Även snitt $X \cap Y$ och inversa bilder $T^{-1}(Y)$ av ändligt genererade koner är ändligt genererade, men beviset för denna utsaga får anstå till kapitel 5, där vi visar att varje ändligt genererad kon är en polyedrisk kon och vice versa.

Sats 2.6.10. *Antag att A är en delmängd av $\mathbf{R}^n$ och att $x \in \operatorname{con} A$. Då finns det en linjärt oberoende delmängd B av A sådan att $x \in \operatorname{con} B$.*

Anmärkning. Notera att mängden B innehåller högst n element.

Bevis. Eftersom x är en konisk kombination av element från A, är x per definition en konisk kombination av ändligt många element valda ur A. Välj nu en ändlig delmängd B av A med så få element som möjligt sådan att $x \in \operatorname{con} B$. Vi skall visa att mängden B är linjärt oberoende.

Om $B = \emptyset$ (dvs. om $x = 0$), är saken klar, ty den tomma mängden är linjärt oberoende. Antag därför att $B \neq \emptyset$ och sätt $B = \{b_1, b_2, \ldots, b_m\}$. Enligt förutsättningarna är $x = \sum_{j=1}^{m} \lambda_j b_j$, där varje $\lambda_j > 0$.

Vi gör ett motsägelsebevis och antar därför att mängden B är linjärt beroende; då finns det skalärer $\mu_1, \mu_2, \ldots, \mu_m$, där minst en av dem är positiv, så att $\sum_{j=1}^{m} \mu_j b_j = 0$. Det följer att $x = \sum_{j=1}^{m} (\lambda_j - t\mu_j) b_j$ för alla $t \in \mathbf{R}$.

Sätt nu $t_0 = \min \lambda_j/\mu_j$, där minimum tas över alla index j för vilka $\mu_j > 0$, och låt j_0 vara ett index som ger minimum. Då är $\lambda_j - t_0\mu_j \geq 0$ för alla index j samtidigt som $\lambda_{j_0} - t_0\mu_{j_0} = 0$. Detta innebär emellertid att x ligger i konen som genereras av $B \setminus \{b_{j_0}\}$, vilket motsäger minimalitetsantagandet. Följaktligen är B linjärt oberoende. □

Sats 2.6.11. *Varje ändligt genererad kon är sluten.*

Bevis. Antag att X är en ändligt genererad kon i $\mathbf{R}^n$, dvs. $X = \operatorname{con} A$ för någon ändlig mängd A.

Betrakta först fallet att $A = \{a_1, a_2, \ldots, a_m\}$ är en linjärt oberoende mängd. Då är $m \leq n$, och vi kan utvidga A med vektorer a_{m+1}, …, a_n till en bas för $\mathbf{R}^n$.

Låt nu $(c_1(x), c_2(x), \ldots, c_n(x))$ beteckna koordinaterna för vektorn x med avseende på basen $a_1, a_2, \ldots, a_n$, dvs. $x = \sum_{j=1}^{n} c_j(x)\, a_j$. Koordinatfunktionerna $c_j(x)$ är linjära former på $\mathbf{R}^n$.

En vektor x tillhör X om och endast om x är en konisk kombination av de m första nya basvektorerna. Detta innebär att

$$X = \{x \in \mathbf{R}^n \mid c_1(x) \geq 0, \ldots, c_m(x) \geq 0,\ c_{m+1}(x) = \cdots = c_n(x) = 0\},$$

dvs. X är lika med snittet av de m slutna halvrummen $\{x \in \mathbf{R}^n \mid c_j(x) \geq 0\}$, $1 \leq j \leq m$, och de $n - m$ hyperplanen $\{x \in \mathbf{R}^n \mid c_j(x) = 0\},\ m + 1 \leq j \leq n$. Följaktligen är X en sluten polyedrisk kon.

Vi övergår nu till det allmänna fallet, dvs. vi låter A vara en godtycklig ändlig mängd. Eftersom det för varje $x \in \operatorname{con} A$ finns en linjärt oberoende delmängd B så att $x \in \operatorname{con} B$, är $\operatorname{con} A = \bigcup \operatorname{con} B$, där unionen skall tas över alla linjärt oberoende delmängder B av A. Det finns förstås bara ändligt många sådana delmängder B, så vi har en union av ändligt många koner $\operatorname{con} B$, som samtliga är slutna enligt första delen av beviset. Konen $\operatorname{con} A$ är därför sluten. □

2.7 Recessionskonen

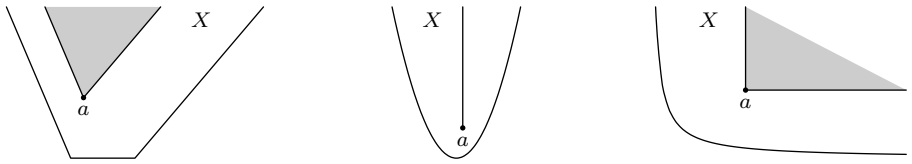
En mängds recessionskon beskriver mängdens uppförande i oändligheten genom att ge information om i vilka riktningar den är obegränsad. Här följer den formella definitionen av begreppet.

Definition. Låt X vara en delmängd av $\mathbf{R}^n$ och v vara en vektor i $\mathbf{R}^n$ skild från nollvektorn. Vi säger att mängden X *recederar* i riktningen v och att v är en *recessionsvektor* till X om alla halvlinjer med riktningsvektor v och start i en godtycklig punkt i X ligger helt i X.

Mängden som består av alla recessionsvektorer till X och nollvektorn kallas *recessionskonen* till X och betecknas $\operatorname{recc} X$. För icke-tomma mängder X är alltså

$$\operatorname{recc} X = \{v \in \mathbf{R}^n \mid x + tv \in X \text{ för alla } x \in X \text{ och alla } t > 0\},$$

medan $\operatorname{recc} \emptyset = \{0\}$.



Figur 2.8. Figuren visar tre konvexa mängder X och motsvarande translaterade recessionskoner $a + \operatorname{recc} X$.

Sats 2.7.1. *Recessionskonen till en godtycklig mängd X är en konvex kon och*

$$X = X + \operatorname{recc} X.$$

Bevis. Att $\operatorname{recc} X$ verkligen är en kon följer omedelbart av definitionen och detsamma gäller inklusionen $X + \operatorname{recc} X \subseteq X$. Den omvända inklusionen $X \subseteq X + Y$ är trivialt sann för alla mängder Y som innehåller nollvektorn, och då speciellt för konen $\operatorname{recc} X$.

Om v och w är två recessionsvektorer till X, x är en godtycklig punkt i X och t är ett godtyckligt positivt tal, så ligger per definition först punkten $x + tv$ i X och sedan också punkten $x + t(v + w) = (x + tv) + tw$ i X, vilket betyder att summan $v + w$ är en recessionsvektor. Recessionskonen har således additivitetsegenskapen $v, w \in \operatorname{recc} X \Rightarrow v + w \in \operatorname{recc} X$, vilket enligt sats 2.6.1 betyder att den är konvex. □

Exempel 2.7.1. Här följer några enkla exempel på recessionskoner:

$$\begin{aligned}
&\operatorname{recc}(\mathbf{R}_+ \times [0,1]) = \operatorname{recc}(\mathbf{R}_+ \times]0,1[) = \mathbf{R}_+ \times \{0\},\\
&\operatorname{recc}(\mathbf{R}_+ \times]0,1[\cup \{(0,0)\}) = \{(0,0)\},\\
&\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_1^2 + x_2^2 \leq 1\} = \{(0,0)\},\\
&\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_2 \geq x_1^2\} = \{0\} \times \mathbf{R}_+,\\
&\operatorname{recc}\{x \in \mathbf{R}^2 \mid x_2 \geq 1/x_1,\ x_1 > 0\} = \mathbf{R}_+^2.
\end{aligned}$$

□

För konvexa mängder förenklas beräkningen av recessionskonen av följande sats.

Sats 2.7.2. *En vektor v är en recessionsvektor till en icke-tom konvex mängd X om och endast om $x + v \in X$ för alla $x \in X$.*

Bevis. Om v är en recessionsvektor till X, så ligger förstås speciellt $x + v$ i X för alla $x \in X$.

Antag för att visa omvändningen att $x + v \in X$ för alla $x \in X$. Då följer genom induktion att $x+nv \in X$ för alla naturliga tal n och varje givet $x \in X$. Eftersom mängden X är konvex ligger därför den slutna sträckan $[x, x+nv]$ i X för alla naturliga tal n, och detta medför att $x + tv \in X$ för varje positivt tal t och varje $x \in X$. Så v är en recessionsvektor till X. □

För konvexa koner gäller speciellt:

Korollarium 2.7.3. *Om X är en konvex kon, så är* $\operatorname{recc} X = X$.

Bevis. Inklusionen $\operatorname{recc} X \subseteq X$ gäller för alla mängder X som innehåller 0 och då speciellt för koner. Den omvända inklusionen $X \subseteq \operatorname{recc} X$ följer av föregående sats och additivitetsegenskapen $x, v \in X \Rightarrow x + v \in X$ för konvexa koner X. □

EXEMPEL 2.7.2. $\operatorname{recc} \mathbf{R}^2_+ = \mathbf{R}^2_+$, $\operatorname{recc}(\mathbf{R}^2_{++} \cup \{(0,0)\}) = \mathbf{R}^2_{++} \cup \{(0,0)\}$. □

En sluten konvex mängds recessionsvektorer karakteriseras av följande sats.

Sats 2.7.4. *Antag att X är en icke-tom sluten konvex mängd. Då är följande tre villkor ekvivalenta för en vektor v.*

(i) v är en recessionsvektor till X.

(ii) Det finns en punkt $x \in X$ sådan att $x + nv \in X$ för alla $n \in \mathbf{Z}_+$.

(iii) Det finns en följd $(x_n)_1^\infty$ av punkter x_n i X och en följd $(\lambda_n)_1^\infty$ av positiva tal med egenskapen att $\lambda_n \to 0$ och $\lambda_n x_n \to v$ då $n \to \infty$.

Bevis. (i) ⇒ (ii): Trivialt eftersom $x + tv \in X$ for alla punkter $x \in X$ och alla positiva tal t, om v är en recessionsvektor till X.

(ii) ⇒ (iii): Om (ii) gäller, så uppfylls villkoret (iii) av punkterna $x_n = x+nv$ och talen $\lambda_n = 1/n$.

(iii) ⇒ (i): Antag att $(x_n)_1^\infty$ och $(\lambda_n)_1^\infty$ är följder av punkter i X och positiva tal sådana att $\lambda_n \to 0$ och $\lambda_n x_n \to v$ då $n \to \infty$, och låt x vara en godtycklig punkt i X. Punkterna $z_n = (1 - \lambda_n)x + \lambda_n x_n$ ligger då också i X för alla tillräckligt stora n, och eftersom $z_n \to x + v$ då $n \to \infty$ medför slutenheten hos X att $x + v \in X$. Det följer nu av sats 2.7.2 att v är en recessionsvektor till X. □

Sats 2.7.5. *Om X är en sluten konvex mängd, så är recessionskonen* $\operatorname{recc} X$ *en sluten konvex kon.*

Bevis. Fallet $X = \emptyset$ är trivialt, så antag att X är en icke-tom sluten konvex mängd. För att visa att recessionskonen $\operatorname{recc} X$ är sluten antar vi att v är

en randpunkt till konen och väljer en följd $(v_n)_1^\infty$ av punkter i konen som konvergerar mot v då $n \to \infty$. Om x är en godtycklig punkt i X, så ligger punkterna $x + v_n$ i X för varje naturligt tal n, och detta medför att deras gränsvärde $x + v$ också ligger i X, eftersom X är en sluten mängd. Detta visar att v är en recessionsvektor, dvs. tillhör recessionskonen $\operatorname{recc} X$, som således innehåller alla sina randpunkter och därför är sluten. □

Sats 2.7.6. *Låt $\{X_i \mid i \in I\}$ vara en familj av slutna konvexa mängder med icke-tomt snitt. Då är* $\operatorname{recc}(\bigcap_{i\in I} X_i) = \bigcap_{i\in I} \operatorname{recc} X_i$.

Bevis. Låt x_0 vara en punkt i snittet $\bigcap_i X_i$. Det följer då av sats 2.7.4 att $v \in \operatorname{recc}(\bigcap_i X_i)$ om och endast om $x_0 + nv$ ligger i X_i för alla positiva heltal n och alla $i \in I$, vilket gäller om och endast om $v \in \operatorname{recc} X_i$ för alla $i \in I$. □

För polyedrar ges recessionskonen av följande sats.

Sats 2.7.7. *För icke-tomma polyedrar $X = \{x \in \mathbf{R}^n \mid Cx \geq b\}$ är*

$$\operatorname{recc} X = \{x \in \mathbf{R}^n \mid Cx \geq 0\}.$$

Bevis. Uppenbarligen är recessionskonen till ett slutet halvrum lika med motsvarande koniska halvrum, så satsen följer därför av föregående sats eftersom varje polyeder är ett snitt av slutna halvrum. □

Observera att recessionskonen till en delmängd Y av X kan vara större än X:s recessionskon. Exempelvis är

$$\operatorname{recc} \mathbf{R}^2_{++} = \mathbf{R}^2_{+} \supsetneq \mathbf{R}^2_{++} \cup \{(0,0)\} = \operatorname{recc}(\mathbf{R}^2_{++} \cup \{(0,0)\}).$$

Detta kan emellertid inte inträffa om den större mängden är sluten.

Sats 2.7.8. *(i) Antag att X är en sluten konvex mängd och att $Y \subseteq X$. Då är* $\operatorname{recc} Y \subseteq \operatorname{recc} X$.
(ii) Antag att X är en konvex mängd. Då är $\operatorname{recc}(\operatorname{rint} X) = \operatorname{recc}(\operatorname{cl} X)$.

Bevis. (i) Fallet $Y = \emptyset$ är trivialt, så antag att Y är en icke-tom delmängd av X och låt y vara en godtycklig punkt i Y. Om v är ett element i $\operatorname{recc} Y$, så ligger vektorerna $y + nv$ i Y och därmed också i X för varje naturligt tal n, och det följer därför av sats 2.7.4 att v också tillhör $\operatorname{recc} X$.
(ii) Inklusionen $\operatorname{recc}(\operatorname{rint} X) \subseteq \operatorname{recc}(\operatorname{cl} X)$ följer av del (i) eftersom $\operatorname{cl} X$ är en sluten konvex mängd. För att visa omvändningen antar vi att $v \in \operatorname{recc}(\operatorname{cl} X)$. Givet en godtycklig punkt $x \in \operatorname{rint} X$ ligger då speciellt punkten $x+2v$ i $\operatorname{cl} X$, och det följer därför av sats 2.5.4 att den öppna sträckan $]x, x + 2v[$ ligger i $\operatorname{rint} X$. Speciellt ligger alltså punkten $x + v$ i $\operatorname{rint} X$, och därmed har vi visat

implikationen $x \in \operatorname{rint} X \Rightarrow x + v \in \operatorname{rint} X$, som enligt sats 2.7.2 innebär att $v \in \operatorname{recc}(\operatorname{rint} X)$. □

Sats 2.7.9. *En sluten konvex mängd X är begränsad om och endast om* $\operatorname{recc} X = \{0\}$.

Bevis. Att $\operatorname{recc} X = \{0\}$ om mängden X är begränsad är uppenbart. För att visa omvändningen antar vi att X är en obegränsad sluten mängd. Då finns det en följd $(x_n)_1^\infty$ av punkter i X sådan att $\|x_n\| \to \infty$ då $n \to \infty$. Den begränsade följden $(x_n/\|x_n\|)_1^\infty$ har enligt Bolzano–Weierstrass sats en konvergent delföljd, och vi kan genom att vid behov stryka element i den ursprungliga följden anta att följden är konvergent. Gränsvärdet v är en vektor med norm 1, vilket garanterar att $v \neq 0$. Med $\lambda_n = 1/\|x_n\|$ har vi nu en följd av punkter x_n i X och en följd av positiva tal λ_n sådana att $\lambda_n \to 0$ och $\lambda_n x_n \to v$ då $n \to \infty$, och detta betyder enligt sats 2.7.4 att v är en recessionsvektor, dvs. ett nollskilt element i $\operatorname{recc} X$. Därmed har vi visat att $\operatorname{recc} X \neq \{0\}$ om X är en obegränsad sluten konvex mängd. □

Definition. Om X är en godtycklig mängd så är snittet $\operatorname{recc} X \cap (-\operatorname{recc} X)$ ett linjärt delrum, som kallas mängden X:s *recessiva delrum* och betecknas $\operatorname{lin} X$.

En sluten konvex mängd kallas *linjefri* om $\operatorname{lin} X = \{0\}$. Mängden X är med andra ord linjefri om och endast om $\operatorname{recc} X$ är en äkta kon.

Om X är en icke-tom sluten konvex delmängd av $\mathbf{R}^n$ och x är en godtycklig punkt i X, så är tydligen

$$\operatorname{lin} X = \{v \in \mathbf{R}^n \mid x + tv \in X \text{ för alla } t \in \mathbf{R}\}.$$

Bilden $T(X)$ av en sluten konvex mängd X under en linjär avbildning T behöver inte vara sluten. Ett motexempel ges av $X = \{x \in \mathbf{R}_+^2 \mid x_1 x_2 \geq 1\}$ och projektionen $T(x_1, x_2) = x_1$ av $\mathbf{R}^2$ på första faktorn med bildmängd $T(X) =]0\,, \infty[$. Anledningen till att bildmängden i detta fall inte är sluten är att X har en recessionsvektor v som avbildas på 0 av T, nämligen $v = (0, 1)$.

Vi har följande generella resultat, där $\mathcal{N}(T)$ betecknar nollrummet till avbildningen T, dvs. $\mathcal{N}(T) = \{x \mid Tx = 0\}$.

Sats 2.7.10. *Låt $T\colon \mathbf{R}^n \to \mathbf{R}^m$ vara en linjär avbildning, låt X vara en sluten konvex delmängd av $\mathbf{R}^n$ och antag att*

$$\mathcal{N}(T) \cap \operatorname{recc} X \subseteq \operatorname{lin} X.$$

Då är bildmängden $T(X)$ sluten, och

$$\operatorname{recc} T(X) = T(\operatorname{recc} X).$$

Speciellt är alltså bildmängden $T(X)$ sluten om X är en sluten konvex delmängd och $x = 0$ är den enda vektorn i $\operatorname{recc} X$ *med* $Tx = 0$.

Bevis. Snittet

$$L = \mathcal{N}(T) \cap \operatorname{lin} X = \mathcal{N}(T) \cap \operatorname{recc} X$$

är ett linjärt delrum till $\mathbf{R}^n$. Låt $L^\perp$ beteckna dess ortogonala komplement. Då är $X = X \cap L + X \cap L^\perp$, och eftersom $Tx = 0$ för alla $x \in L$ är

$$T(X) = T(X \cap L^\perp).$$

Låt nu y vara en godtycklig randpunkt till bildmängden $T(X)$. Då finns det på grund av likheten ovan en följd $(x_n)_1^\infty$ av punkter $x_n \in X \cap L^\perp$ sådan att $\lim_{n\to\infty} Tx_n = y$. Vi påstår att följden $(x_n)_1^\infty$ måste vara begränsad.

Antag nämligen motsatsen. Då har följden $(x_n)_1^\infty$ en delföljd $(x_{n_k})_1^\infty$ sådan att $\|x_{n_k}\| \to \infty$ då $k \to \infty$ och den begränsade följden $(x_{n_k}/\|x_{n_k}\|)_1^\infty$ konvergerar. Gränsvärdet v är förstås en vektor i det linjära delrummet $L^\perp$ med norm 1. Eftersom $x_{n_k} \in X$ och $1/\|x_{n_k}\| \to 0$ följer det vidare av sats 2.7.4 att $v \in \operatorname{recc} X$. Slutligen är

$$Tv = \lim_{k\to\infty} T(x_{n_k}/\|x_{n_k}\|) = \lim_{k\to\infty} \|x_{n_k}\|^{-1} Tx_{n_k} = 0 \cdot y = 0,$$

så vektorn v ligger i nollrummet $\mathcal{N}(T)$, och därmed också i L. Detta betyder att $v \in L \cap L^\perp$, vilket är en motsägelse eftersom $L \cap L^\perp = \{0\}$ och $v \neq 0$.

Följden $(x_n)_1^\infty$ är således begränsad. Låt $(x_{n_k})_1^\infty$ vara en konvergent delföljd, och sätt $x = \lim_{k\to\infty} x_{n_k}$. Då ligger x i X eftersom mängden X är sluten, och $y = \lim_{k\to\infty} Tx_{n_k} = Tx$, vilket innebär att $y \in T(X)$. Bildmängden $T(X)$ innehåller således alla sina randpunkter, och den är därför sluten.

Inklusionen $T(\operatorname{recc} X) \subseteq \operatorname{recc} T(X)$ gäller för alla mängder X. Antag nämligen att $v \in \operatorname{recc} X$ och låt y vara en godtycklig punkt i $T(X)$. Då är $y = Tx$ för någon punkt $x \in X$, och eftersom $x + tv \in X$ för alla $t > 0$ och $y + tTv = T(x + tv)$, ligger punkterna $y + tTv$ i $T(X)$ för alla $t > 0$. Detta innebär att $Tv \in \operatorname{recc} T(X)$.

För att visa den omvända inklusionen $\operatorname{recc} T(X) \subseteq T(\operatorname{recc} X)$ för slutna konvexa mängder X och linjära avbildningar T som uppfyller förutsättningarna i satsen, antar vi att $w \in \operatorname{recc} T(X)$ och ska visa att det finns en vektor $v \in \operatorname{recc} X$ sådan att $w = Tv$. Vi noterar då först att det finns en följd $(y_n)_1^\infty$ av punkter $y_n \in T(X)$ och en följd $(\lambda_n)_1^\infty$ av positiva tal så att $\lambda_n \to 0$ och $\lambda_n y_n \to w$ då $n \to \infty$. Välj för varje n en punkt $x_n \in X \cap L^\perp$ sådan att $y_n = T(x_n)$.

Följden $(\lambda_n x_n)_1^\infty$ måste vara begränsad. Ty antag motsatsen; då finns det en delföljd sådan att $\|\lambda_{n_k} x_{n_k}\| \to \infty$ och följden $(x_{n_k}/\|x_{n_k}\|)_1^\infty$ konvergerar

mot en vektor z då $k \to \infty$. Det följer av sats 2.7.4 att $z \in \operatorname{recc} X$, ty x_{n_k} är punkter i X och $\|x_{n_k}\| \to \infty$ då $k \to \infty$. Givetvis ligger z också i delrummet $L^\perp$, och vidare är

$$\begin{aligned} Tz &= \lim_{k\to\infty} T(x_{n_k}/\|x_{n_k}\|) = \lim_{k\to\infty} T(\lambda_{n_k}x_{n_k}/\|\lambda_{n_k}x_{n_k}\|) \\ &= \lim_{k\to\infty} \lambda_{n_k}y_{n_k}/\|\lambda_{n_k}x_{n_k}\| = 0 \cdot w = 0. \end{aligned}$$

Detta innebär att $z \in \mathcal{N}(T) \cap \operatorname{recc} X = L$, så z ligger i $L \cap L^\perp$, vilket är en motsägelse till att $\|z\| = 1$.

Eftersom följden $(\lambda_n x_n)_1^\infty$ är begränsad, har den en delföljd som konvergerar mot en vektor v, som enligt sats 2.7.4 ligger i $\operatorname{recc} X$, och eftersom $T(\lambda_n x_n) = \lambda_n y_n \to w$, drar vi slutsatsen att $Tv = w$, och detta betyder att $w \in T(\operatorname{recc} X)$. □

Sats 2.7.11. *Låt X och Y vara icke-tomma slutna konvexa delmängder av $\mathbf{R}^n$, och antag att*

$$x \in \operatorname{recc} X \;\&\; y \in \operatorname{recc} Y \;\&\; x + y = 0 \quad\Rightarrow\quad x \in \operatorname{lin} X \;\&\; y \in \operatorname{lin} Y.$$

Då är summan $X + Y$ en sluten konvex mängd och

$$\operatorname{recc}(X + Y) = \operatorname{recc} X + \operatorname{recc} Y.$$

Anmärkning. Förutsättningen i satsen är uppfyllt om $\operatorname{recc} X$ och $-\operatorname{recc} Y$ inte har någon annan gemensam vektor än nollvektorn 0.

Bevis. Låt $T\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ vara den linjära avbildningen $T(x, y) = x + y$. Vi lämnar som enkel övning att visa att $\operatorname{recc}(X \times Y) = \operatorname{recc} X \times \operatorname{recc} Y$ och att $\operatorname{lin}(X \times Y) = \operatorname{lin} X \times \operatorname{lin} Y$. Eftersom $\mathcal{N}(T) = \{(x, y) \mid x + y = 0\}$, innebär förutsättningen i satsen att

$$\mathcal{N}(T) \cap \operatorname{recc}(X \times Y) \subseteq \operatorname{lin}(X \times Y),$$

och det följer därför av sats 2.7.10 att $T(X \times Y)$, dvs. summan $X + Y$, är sluten och att

$$\begin{aligned} \operatorname{recc}(X + Y) &= \operatorname{recc} T(X \times Y) = T(\operatorname{recc}(X \times Y)) = T(\operatorname{recc} X \times \operatorname{recc} Y) \\ &= \operatorname{recc} X + \operatorname{recc} Y. \end{aligned}$$

□

Korollarium 2.7.12. *Summan $X + Y$ av en icke-tom sluten konvex mängd X och en icke-tom kompakt konvex mängd Y är en sluten konvex mängd, och* $\operatorname{recc}(X + Y) = \operatorname{recc} X$.

Bevis. Eftersom $\operatorname{recc} Y = \{0\}$ är förutsättningarna i sats 2.7.11 trivialt uppfyllda. □

Korollarium 2.7.13. *Antag att C är en sluten konvex kon och Y är en icketom kompakt konvex mängd. Då är* $\operatorname{recc}(C + Y) = C$.

Bevis. Korollariet följer som specialfall av föregående korollarium eftersom $\operatorname{recc} C = C$. □

Övningar

2.1 Visa att mängden $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq a\}$ är konvex, och mer generellt att mängden $\{x \in \mathbf{R}^n_+ \mid x_1 x_2 \cdots x_n \geq a\}$ är konvex.
[Ledning: Använd olikheten $x_i^\lambda y_i^{1-\lambda} \leq \lambda x_i + (1-\lambda) y_i$ mellan vägt geometriskt och vägt aritmetiskt medelvärde; se sats 6.4.1.]

2.2 Bestäm konvexa höljet $\operatorname{cvx} A$ till följande delmängder A av $\mathbf{R}^2$:
a) $A = \{(0,0), (1,0), (0,1)\}$ b) $A = \{x \in \mathbf{R}^2 \mid \|x\| = 1\}$
c) $A = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 = 1\} \cup \{(0,0)\}$.

2.3 Ge exempel på en sluten mängd med icke-slutet konvext hölje.

2.4 Bestäm den inversa bilden $P^{-1}(X)$ till mängden $X = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ under perspektivavbildningen $P\colon \mathbf{R}^2 \times \mathbf{R}_{++} \to \mathbf{R}^2$.

2.5 Visa att mängden $\{x \in \mathbf{R}^{n+1} \mid \left(\sum_{j=1}^n x_j^2\right)^{1/2} \leq x_{n+1}\}$ är en kon.

2.6 Låt $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ vara den kanoniska basen i $\mathbf{R}^n$ och sätt $\mathbf{e}_0 = -\sum_1^n \mathbf{e}_j$. Visa att hela rummet $\mathbf{R}^n$ genereras som kon av de $n+1$ stycken vektorerna $\mathbf{e}_0$, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

2.7 Visa att ett koniskt halvrum i $\mathbf{R}^n$ genereras som kon av $n+1$ stycken vektorer.

2.8 Visa att om X är en sluten kon i $\mathbf{R}^2$, så är $X = \operatorname{con} A$ för någon mängd A med högst tre element.

2.9 Visa att summan av två slutna koner i $\mathbf{R}^2$ är en sluten kon.

2.10 Bestäm $\operatorname{recc} X$ och $\operatorname{lin} X$ för följande konvexa mängder:
a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4\}$
d) $X = \{x \in \mathbf{R}^3 \mid x_1^2 - x_2^2 \geq 1,\ x_1 \geq 0\}$.

2.11 Låt X vara en godtycklig icke-tom delmängd av $\mathbf{R}^m$ och Y vara en godtycklig icke-tom delmängd av $\mathbf{R}^n$. Visa att $\operatorname{recc}(X \times Y) = \operatorname{recc} X \times \operatorname{recc} Y$ och att $\operatorname{lin}(X \times Y) = \operatorname{lin} X \times \operatorname{lin} Y$.

2.12 Låt $P\colon \mathbf{R}^n \times \mathbf{R}_{++} \to \mathbf{R}^n$ vara perspektivavbildningen. För konvexa delmängder X av $\mathbf{R}^n$ sätter vi $c(X) = P^{-1}(X) \cup \{(0,0)\}$.

a) Bevisa att $c(X)$ är en kon, närmare bestämt att $c(X) = \operatorname{con}(X \times \{1\})$.

b) Bestäm ett explicit uttryck för konerna $c(X)$ och $\operatorname{cl}(c(X))$ om

(i) $n = 1$ och $X = [2,3]$; (ii) $n = 1$ och $X = [2,\infty[$;

(iii) $n = 2$ och $X = \{x \in \mathbf{R}^2 \mid x_1 \geq x_2^2\}$.

c) Bestäm $c(X)$ om $X = \{x \in \mathbf{R}^n \mid \|x\| \leq 1\}$ och $\|\cdot\|$ är en godtycklig norm på $\mathbf{R}^n$.

d) Visa att $\operatorname{cl}(c(X)) = c(\operatorname{cl} X) \cup (\operatorname{recc}(\operatorname{cl} X) \times \{0\})$.

e) Visa att $\operatorname{cl}(c(X)) = c(\operatorname{cl} X)$ om och endast om mängden X är begränsad.

f) Visa att konen $c(X)$ är sluten om och endast om X är kompakt.

2.13 Konen $Y = \{x \in \mathbf{R}^3 \mid x_1 x_3 \geq x_2^2,\ x_3 > 0\} \cup \{x \in \mathbf{R}^3 \mid x_1 \geq 0,\ x_2 = x_3 = 0\}$ är sluten. (Jmf b) (iii) i föregående övning). Sätt

$$Z = \{x \in \mathbf{R}^3 \mid x_1 \leq 0,\ x_2 = x_3 = 0\}$$

och visa att

$$Y + Z = \{x \in \mathbf{R}^3 \mid x_3 > 0\} \cup \{x \in \mathbf{R}^3 \mid x_2 = x_3 = 0\}.$$

Summan av två slutna koner i $\mathbf{R}^3$ behöver med andra ord inte vara en sluten kon.

2.14 Visa att summan $X + Y$ av en godtycklig sluten mängd X och en godtycklig kompakt mängd Y är sluten.

Kapitel 3

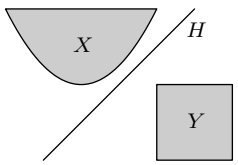
Separation

3.1 Separerande hyperplan

Definition. Låt X och Y vara två mängder i $\mathbf{R}^n$. Man säger att hyperplanet H *separerar* mängderna om de ligger i var sitt av de båda slutna halvrum som definieras av hyperplanet H och hyperplanet inte innehåller både X och Y som delmängder.[†]

Separationen kallas *strikt* om det finns två med H parallella hyperplan, ett på vardera sidan om H, som separerar X och Y.

Hyperplanet $H = \{x \mid \langle c, x\rangle = b\}$ separerar med andra ord mängderna X och Y om $\langle c, x\rangle \leq b$ för alla $x \in X$, $\langle c, y\rangle \geq b$ för alla $y \in Y$ och $\langle c, x\rangle \neq b$ för åtminstone något element $x \in X \cup Y$. Separationen är strikt om det finns tal b_1 och b_2 sådana att $\langle c, x\rangle \leq b_1 < b < b_2 \leq \langle c, y\rangle$ för alla $x \in X$ och alla $y \in Y$.



Figur 3.1. Strikt separerande hyperplan H

I många fall är vi enbart intresserade av själva existensen av separerande hyperplan, och den hänger på ett naturligt sätt ihop med extremvärden hos linjära funktioner.

[†]Ett hyperplan som separerar två delmängder till ett hyperplan H måste alltså med vår terminologi vara skilt från H.

Sats 3.1.1. *Låt X och Y vara två icke-tomma mängder i $\mathbf{R}^n$.*

(i) Det finns ett hyperplan som separerar X och Y om och endast om det finns en vektor c sådan att

$$\sup_{x\in X}\langle c,x\rangle \le \inf_{y\in Y}\langle c,y\rangle \quad \textit{och} \quad \inf_{x\in X}\langle c,x\rangle < \sup_{y\in Y}\langle c,y\rangle.$$

(ii) Det finns ett hyperplan som separerar X och Y strikt om och endast om det finns en vektor c sådan att

$$\sup_{x\in X}\langle c,x\rangle < \inf_{y\in Y}\langle c,y\rangle.$$

Bevis. En vektor c som uppfyller villkoren i (i) eller (ii) är förstås nollskild.

Antag att c uppfyller villkoren i (i) och välj talet b så att

$$\sup_{x\in X}\langle c,x\rangle \le b \le \inf_{y\in Y}\langle c,y\rangle.$$

Då är $\langle c,x\rangle \le b$ för alla $x \in X$ och $\langle c,y\rangle \ge b$ för alla $y \in Y$. Vidare är $\langle c,x\rangle \ne b$ för något $x \in X \cup Y$ på grund av den andra olikheten i (i). Hyperplanet $H = \{x \mid \langle c,x\rangle = b\}$ separerar därför mängderna X och Y.

Om c uppfyller villkoret i (ii) väljer vi istället b så att

$$\sup_{x\in X}\langle c,x\rangle < b < \inf_{y\in Y}\langle c,y\rangle,$$

och drar nu slutsatsen att hyperplanet H separerar X och Y strikt.

Omvänt, om hyperplanet H separerar X och Y, så kan vi genom att vid behov byta tecken på c och b anta att $\langle c,x\rangle \le b$ för alla $x \in X$ och $\langle c,y\rangle \ge b$ för alla $y \in Y$, och detta medför att $\sup_{x\in X}\langle c,x\rangle \le b \le \inf_{y\in Y}\langle c,y\rangle$. Eftersom H vidare inte innehåller både X och Y, finns det punkter $x_1 \in X$ och $y_1 \in Y$ med $\langle c,x_1\rangle < \langle c,y_1\rangle$, och detta ger oss den andra olikheten i (i).

Om separationen dessutom är strikt, så finns det två med H parallella hyperplan $H_i = \{x \mid \langle c,x\rangle = b_i\}$, där $b_1 < b < b_2$, som separerar X och Y, och det följer att $\sup_{x\in X}\langle c,x\rangle \le b_1 < b < b_2 \le \inf_{y\in Y}\langle c,y\rangle$, dvs. c uppfyller villkoret i (ii). □

Om ett hyperplan separerar en enpunktsmängd $\{x_0\}$ och en godtycklig mängd Y, så säger vi också att hyperplanet separerar punkten x_0 från mängden Y. Följande enkla lemma reducerar problemet att separera två mängder till problemet att separera punkten 0 från den algebraiska differensmängden.

Lemma 3.1.2. *Låt X och Y vara två icke-tomma mängder.*

(i) Om det finns ett hyperplan som separerar 0 från mängden $X - Y$, så finns det ett hyperplan som separerar X och Y.

(ii) Om det finns ett hyperplan som separerar punkten 0 strikt från mängden $X - Y$, så finns det ett hyperplan som separerar X och Y strikt.

Bevis. (i) Om det finns ett hyperplan som separerar 0 från $X - Y$, så finns det på grund av sats 3.1.1 en vektor c så att

$$\begin{cases} 0 = \langle c, 0\rangle \leq \inf\limits_{x\in X,\, y\in Y} \langle c, x - y\rangle = \inf\limits_{x\in X} \langle c, x\rangle - \sup\limits_{y\in Y} \langle c, y\rangle \\ 0 = \langle c, 0\rangle < \sup\limits_{x\in X,\, y\in Y} \langle c, x - y\rangle = \sup\limits_{x\in X} \langle c, x\rangle - \inf\limits_{y\in Y} \langle c, y\rangle \end{cases}$$

dvs. $\sup_{y\in Y}\langle c, y\rangle \leq \inf_{x\in X}\langle c, x\rangle$ och $\inf_{y\in Y}\langle c, y\rangle < \sup_{x\in X}\langle c, x\rangle$, och vi drar slutsatsen att det finns ett hyperplan som separerar X och Y.

(ii) Om det finns ett hyperplan som separerar 0 strikt från $X - Y$, så finns det istället en vektor c så att

$$0 = \langle c, 0\rangle < \inf_{x\in X,\, y\in Y} \langle c, x - y\rangle = \inf_{x\in X} \langle c, x\rangle - \sup_{y\in Y} \langle c, y\rangle$$

och det följer nu att $\sup_{y\in Y}\langle c, y\rangle < \inf_{x\in X}\langle c, x\rangle$, vilket visar att Y och X kan separeras strikt av ett hyperplan. □

Följande sats ligger till grund för det här avsnittets resultat om separation av konvexa mängder.

Sats 3.1.3. *Antag att mängden X är konvex och att $a \notin \operatorname{cl} X$. Då finns det ett hyperplan som separerar a och X strikt.*

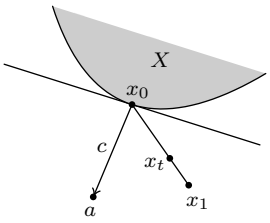
Bevis. Mängden $\operatorname{cl} X$ är konvex och sluten, och ett hyperplan som separerar a och $\operatorname{cl} X$ strikt separerar naturligtvis också a och X strikt eftersom X är en delmängd till $\operatorname{cl} X$. Det räcker således att visa att vi kan separera punkten a strikt från varje sluten konvex mängd som inte innehåller a.

Vi kan därför utan inskränkning anta att den konvexa mängden X är sluten och icke-tom. Sätt $d(x) = \|x - a\|^2$, dvs. kvadraten på det euklidiska avståndet mellan x och a. Vi börjar med att visa att funktionen $d\colon X \to \mathbf{R}$ har en minimipunkt i X.

Välj därför ett positivt reellt tal r som är så stort att den slutna bollen $\overline{B}(a; r)$ skär mängden X. För $x \in X \setminus \overline{B}(a; r)$ är uppenbarligen $d(x) > r^2$, och för $x \in X \cap \overline{B}(a; r)$ är $d(x) \leq r^2$. Eftersom mängden $X \cap \overline{B}(a; r)$ är kompakt, finns det en punkt $x_0 \in X \cap \overline{B}(a; r)$ där den kontinuerliga funktionen d antar sitt minsta värde på $X \cap \overline{B}(a; r)$, och denna punkt blir då automatiskt en minimipunkt i hela X, dvs. olikheten $d(x_0) \leq d(x)$ gäller för alla $x \in X$.

Sätt nu $c = a - x_0$. Vi påstår att $\langle c, x - x_0\rangle \leq 0$ för alla $x \in X$. Antag därför motsatsen, dvs. att det finns en punkt $x_1 \in X$ med $\langle c, x_1 - x_0\rangle > 0$. Vi ska visa att detta antagande leder till en motsägelse.

Betrakta punkterna $x_t = tx_1 + (1 - t)x_0$; de tillhör X för $0 \leq t \leq 1$ på grund av att X är konvex. Sätt $f(t) = d(x_t) = \|x_t - a\|^2$. Sambandet mellan



Figur 3.2. Illustration till beviset för sats 3.1.3.

norm och skalärprodukt ger

$$\begin{aligned} f(t) &= \|x_t - a\|^2 = \|t(x_1 - x_0) + (x_0 - a)\|^2 = \|t(x_1 - x_0) - c\|^2 \\ &= t^2\|x_1 - x_0\|^2 - 2t\,\langle c, x_1 - x_0\rangle + \|c\|^2. \end{aligned}$$

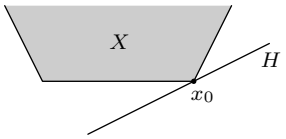
Funktionen $f(t)$ är ett andragradspolynom i t, och för derivatan i origo gäller $f'(0) = -2\,\langle c, x_1 - x_0\rangle < 0$. Funktionen $f(t)$ är således strängt avtagande i en omgivning av $t = 0$, vilket innebär att $d(x_t) < d(x_0)$ för tillräckligt små positiva tal t. Detta strider mot att x_0 är funktionens minimipunkt och bevisar påståendet att $\langle c, x - x_0\rangle \leq 0$ för alla $x \in X$. Följaktligen är $\langle c, x\rangle \leq \langle c, x_0\rangle = \langle c, a - c\rangle = \langle c, a\rangle - \|c\|^2$ för alla $x \in X$, vilket innebär att $\sup_{x\in X}\langle c, x\rangle < \langle c, a\rangle$. Enligt sats 3.1.1 finns det därför ett hyperplan som separerar a strikt från X. □

Definition. Låt X vara en delmängd av $\mathbf{R}^n$ och låt x_0 vara en punkt i X. Ett hyperplan H genom x_0 kallas ett *stödhyperplan* till X om det separerar x_0 och X. Man säger i så fall att hyperplanet H *stöder* X *i punkten* x_0.

Existensen av ett stödhyperplan till mängden X i punkten $x_0 \in X$ är tydligen ekvivalent med villkoret att det finns en vektor c sådan att

$$\langle c, x_0\rangle = \inf_{x\in X}\langle c, x\rangle \quad \text{och} \quad \langle c, x_0\rangle < \sup_{x\in X}\langle c, x\rangle.$$

Hyperplanet $\{x \mid \langle c, x\rangle = \langle c, x_0\rangle\}$ är i så fall ett stödhyperplan.



Figur 3.3. Ett stödhyperplan till X i punkten x_0

Om ett hyperplan stöder mängden X i punkten x_0, så är stödpunkten x_0 nödvändigtvis en relativ randpunkt till X. För konvexa mängder har vi följande omvändning.

Sats 3.1.4. *Antag att X är en konvex mängd och att $x_0 \in X$ är en relativ randpunkt till X. Då finns det ett stödhyperplan H till X som stöder X i punkten x_0.*

Bevis. Antag först att X har samma dimension som det omgivande rummet $\mathbf{R}^n$. Eftersom x_0 är en randpunkt till mängden X, finns det en följd $(x_n)_1^\infty$ av punkter $x_n \notin \operatorname{cl} X$ som konvergerar mot x_0 då $n \to \infty$, och på grund av sats 3.1.3 finns det för varje $n \geq 1$ ett hyperplan som separerar x_n strikt från X. Sats 3.1.1 ger oss därför en följd $(c_n)_1^\infty$ av vektorer med egenskapen att

$$\langle c_n, x_n \rangle < \langle c_n, x \rangle \qquad \text{för alla } x \in X \tag{3.1}$$

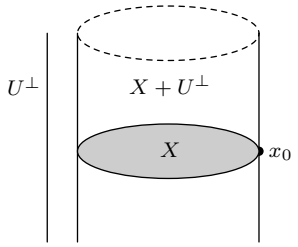
och alla $n \geq 1$. Vi kan vidare uppenbarligen normalisera vektorerna c_n så att $\|c_n\| = 1$ för alla n.

Eftersom enhetssfären $\{x \in \mathbf{R}^n \mid \|x\| = 1\}$ är kompakt, har följden $(c_n)_1^\infty$ enligt Bolzano–Weierstrass sats en delföljd $(c_{n_k})_{k=1}^\infty$, som konvergerar mot en vektor c med längd $\|c\| = 1$. Naturligtvis är $\lim_{k\to\infty} x_{n_k} = x_0$, så det följer genom gränsövergång i olikheten (3.1) att $\langle c, x_0 \rangle \leq \langle c, x \rangle$ för alla $x \in X$.

Mängden X ligger därför i ett av de båda slutna halvrum som bestäms av hyperplanet $H = \{x \in \mathbf{R}^n \mid \langle c, x \rangle = \langle c, x_0 \rangle\}$, och eftersom $\dim X = n$ kan X inte vara en delmängd till H. Hyperplanet H stöder således X i punkten x_0.

Antag härnäst att $\dim X < n$. Då ligger X i en affin delmängd av formen $a + U$, där U är ett linjärt delrum till $\mathbf{R}^n$ och $\dim U = \dim X$. Betrakta mängden $Y = X + U^\perp$, där $U^\perp$ är det ortogonala komplementet till U. Jmf figur 3.4. Y är en "cylinder" med X som "bas", och varje $y \in Y$ har en unik uppdelning på formen $y = x + v$ med $x \in X$ och $v \in U^\perp$.

Y är en konvex mängd av dimension n och x_0 är en randpunkt till Y. Enligt den redan bevisade delen av satsen finns det ett hyperplan som stöder Y i punkten x_0, dvs. det finns en vektor c så att



Figur 3.4. Illustration till beviset för sats 3.1.4.

$$\langle c, x_0\rangle = \inf_{y\in Y}\langle c, y\rangle = \inf_{x\in X,\, v\in U^\perp}\langle c, x+v\rangle = \inf_{x\in X}\langle c, x\rangle + \inf_{v\in U^\perp}\langle c, v\rangle$$

och

$$\langle c, x_0\rangle < \sup_{y\in Y}\langle c, y\rangle = \sup_{x\in X,\, v\in U^\perp}\langle c, x+v\rangle = \sup_{x\in X}\langle c, x\rangle + \sup_{v\in U^\perp}\langle c, v\rangle.$$

Det följer av den första ekvationen att $\inf_{v\in U^\perp}\langle c, v\rangle$ är ett ändligt tal, och eftersom $U^\perp$ är ett vektorrum och därmed för varje $v \in U^\perp$ också innehåller vektorerna tv för alla $t \in \mathbf{R}$, är detta möjligt om och endast om $\langle c, v\rangle = 0$ för alla $v \in U^\perp$. Villkoren ovan reduceras därför till

$$\langle c, x_0\rangle = \inf_{x\in X}\langle c, x\rangle \quad \text{och} \quad \langle c, x_0\rangle < \sup_{x\in X}\langle c, x\rangle,$$

som visar att X har ett stödhyperplan i punkten x_0. □

Vi kan nu visa följande tillräckliga och nödvändiga villkor för separation av konvexa mängder.

Sats 3.1.5. *Två konvexa mängder X och Y kan separeras av ett hyperplan om och endast om deras relativa inre är disjunkta.*

Bevis. Ett hyperplan som separerar två mängder A och B, separerar uppenbarligen också deras slutna höljen $\operatorname{cl} A$ och $\operatorname{cl} B$ och därmed också alla mängder C och D som uppfyller $A \subseteq C \subseteq \operatorname{cl} A$ och $B \subseteq D \subseteq \operatorname{cl} B$.

För att visa att det finns ett hyperplan som separerar de konvexa mängderna X och Y om $\operatorname{rint} X \cap \operatorname{rint} Y = \emptyset$, räcker det därför att visa att det finns ett hyperplan som separerar de konvexa mängderna $A = \operatorname{rint} X$ och $B = \operatorname{rint} Y$, ty $\operatorname{rint} X \subseteq X \subseteq \operatorname{cl}(\operatorname{rint} X) = \operatorname{cl} X$, och motsvarande inklusioner gäller förstås för Y.

Eftersom mängderna A och B är disjunkta, ligger 0 utanför den konvexa mängden $A - B$. Punkten 0 ligger därför antingen utanför $\operatorname{cl}(A-B)$ eller också ligger den i $\operatorname{cl}(A-B)$ och är en relativ randpunkt till $\operatorname{cl}(A-B)$, ty

$$\begin{aligned}\operatorname{cl}(A-B) \setminus (A-B) &\subseteq \operatorname{cl}(A-B) \setminus \operatorname{rint}(A-B)\\ &= \operatorname{rbdry}(A-B) = \operatorname{rbdry}(\operatorname{cl}(A-B)),\end{aligned}$$

I det förstnämnda fallet följer det av sats 3.1.3 att det finns ett hyperplan som separerar 0 och $A-B$ strikt, och i det sistnämnda fallet visar istället sats 3.1.4 att det finns ett hyperplan som separerar 0 från mängden $\operatorname{cl}(A-B)$, och därmed förstås också 0 från $A-B$. Existensen av ett hyperplan som separerar A och B följer sedan av lemma 3.1.2.

Så till omvändningen. Antag att hyperplanet H separerar de konvexa mängderna X och Y. Vi ska visa att $\operatorname{rint} X$ och $\operatorname{rint} Y$ saknar gemensamma punkter. Antag därför att x_0 är en punkt i snittet $X \cap Y$. Då ligger punkten x_0 i hyperplanet H eftersom X och Y ligger i varsitt av de båda slutna halvrum som bestäms av H. Enligt definitionen av separation har minst en av de båda konvexa mängderna, X säg, punkter utanför H. Den av X uppspända affina mångfalden $V = \operatorname{aff} X$ är därför inte en delmängd av H, och den har följaktligen punkter på ömse sidor om H. Om $B = B(x_0; r)$ är en godtycklig öppen boll med centrum i x_0, så innehåller därför också snittet $B \cap V$ punkter från båda sidor av H och därmed säkert punkter som inte tillhör X, vilket betyder att x_0 inte kan vara en relativt inre punkt till X.

Varje punkt i snittet $X \cap Y$ är således en relativ randpunkt till en av de två mängderna X och Y, och snittet $\operatorname{rint} X \cap \operatorname{rint} Y$ är med andra ord tomt. □

Vi övergår nu till att behandla strikt separation. Ett hyperplan som separerar två mängder strikt, separerar uppenbarligen också deras slutna höljen strikt, så det är tillräckligt att undersöka när två slutna konvexa mängder X och Y kan separeras strikt. Ett nödvändigt villkor är förstås att mängderna är disjunkta, dvs. att $0 \notin X - Y$, och lemma 3.1.2 reducerar nu problemet att separera två disjunkta slutna konvexa mängder X och Y strikt till problemet att separera punkten 0 strikt från $X - Y$. Det följer därför omedelbart av sats 3.1.3 att det finns ett strikt separerande hyperplan om mängden $X - Y$ är sluten. Detta ger oss följande sats, där de tillräckliga villkoren följer at sats 2.7.11 och korollarium 2.7.12.

Sats 3.1.6. *Två disjunkta slutna konvexa mängder X och Y kan separeras strikt av ett hyperplan om mängden $X - Y$ är sluten, och ett tillräckligt villkor för att så ska vara fallet är att* $\operatorname{recc} X \cap \operatorname{recc} Y = \{0\}$. *Speciellt kan två disjunkta slutna konvexa mängder separeras strikt om en av mängderna är begränsad.*

Vi avslutar det här avsnittet med ett resultat som visar att äkta konvexa koner är äkta delmängder till koniska halvrum.

Sats 3.1.7. *Låt $X \neq \{0\}$ vara en äkta konvex kon i $\mathbf{R}^n$, där $n \geq 2$. Då är X en äkta delmängd till något slutet koniskt halvrum $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$, vars rand $\{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$ inte innehåller X som delmängd.*

Bevis. 0 är en relativ randpunkt till X, ty ingen punkt på sträckan $]0, -a[$ ligger i X om a är en godtycklig nollskild punkt i X. Enligt sats 3.1.4 finns det därför ett hyperplan $H = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$ genom 0 sådant att X ligger i det slutna halvrummet $K = \{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$ men inte helt i

hyperplanet H. K är ett koniskt halvrum, och den äkta konen X måste vara en äkta delmängd till K, eftersom inga koniska halvrum i $\mathbf{R}^n$ är äkta koner när $n \geq 2$. □

3.2 Dualkonen

Till varje delmängd A av $\mathbf{R}^n$ tillordnar vi en ny delmängd A^+ av $\mathbf{R}^n$ genom att sätta

$$A^+ = \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0 \text{ för alla } a \in A.\}.$$

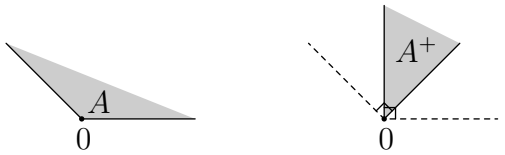
För enpunktsmängder $\{a\}$ är tydligen mängden

$$\{a\}^+ = \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0\}$$

ett koniskt slutet halvrum, och i det allmänna fallet är $A^+ = \bigcap_{a\in A}\{a\}^+$ ett snitt av koniska slutna halvrum. Mängden A^+ är således alltid en sluten konvex kon, och om A är en ändlig mängd är den en polyedrisk kon.

Definition. Den slutna konvexa konen A^+ kallas *dualkonen* till mängden A.

Om $A \subseteq \mathbf{R}^n$ och $n \leq 3$, så har dualkonen A^+ en tydlig geometrisk tolkning – den består av alla vektorer som bildar spetsig eller rät vinkel med alla vektorer i A.



Figur 3.5. Kon A och dess dualkon A^+

Sats 3.2.1. *Låt A och B vara delmängder av $\mathbf{R}^n$. Då gäller:*

(i) $A \subseteq B \;\Rightarrow\; B^+ \subseteq A^+$;

(ii) $A^+ = (\operatorname{con} A)^+$;

(iii) $A^+ = (\operatorname{cl} A)^+$.

Bevis. Egenskap (i) är en omedelbar konsekvens av definitionen av dualkon, och av (i) följer att $(\operatorname{con} A)^+ \subseteq A^+$ och $(\operatorname{cl} A)^+ \subseteq A^+$ eftersom $A \subseteq \operatorname{con} A$ och $A \subseteq \operatorname{cl} A$. För att bevisa (ii) och (iii) återstår det därför bara att bevisa de omvända inklusionerna.

Antag därför att $x \in A^+$. Då är

$$\langle \lambda_1 a_1 + \cdots + \lambda_k a_k, x\rangle = \lambda_1\langle a_1, x\rangle + \cdots + \lambda_k\langle a_k, x\rangle \geq 0$$

för alla koniska kombinationer av element a_i ur A. Detta bevisar implikationen $x \in A^+ \Rightarrow x \in (\operatorname{con} A)^+$, dvs. inklusionen $A^+ \subseteq (\operatorname{con} A)^+$.

För varje $a_0 \in \operatorname{cl} A$ finns det en följd $(a_k)_1^\infty$ av element i A så att $a_k \to a_0$ då $k \to \infty$. Om $\langle a_k, x\rangle \geq 0$ för alla k, så följer genom gränsövergång att $\langle a_0, x\rangle \geq 0$, och följaktligen gäller implikationen $x \in A^+ \Rightarrow x \in (\operatorname{cl} A)^+$. Detta bevisar inklusionen $A^+ \subseteq (\operatorname{cl} A)^+$. □

EXEMPEL 3.2.1. Uppenbarligen är $(\mathbf{R}^n)^+ = \{0\}$ och $\{0\}^+ = \mathbf{R}^n$. □

EXEMPEL 3.2.2. Om $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ är standardbasen i $\mathbf{R}^n$, så är

$$\{\mathbf{e}_j\}^+ = \{x \in \mathbf{R}^n \mid \langle \mathbf{e}_j, x\rangle \geq 0\} = \{x \in \mathbf{R}^n \mid x_j \geq 0\}.$$

Eftersom $\mathbf{R}^n_+ = \operatorname{con}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, följer det att

$$(\mathbf{R}^n_+)^+ = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}^+ = \bigcap_{j=1}^n \{\mathbf{e}_j\}^+ = \{x \in \mathbf{R}^n \mid x_1 \geq 0, \ldots, x_n \geq 0\} = \mathbf{R}^n_+.$$

□

Bidualkonen

Definition. Dualkonen A^+ till en mängd A i $\mathbf{R}^n$ är en ny mängd i $\mathbf{R}^n$, och vi kan därför bilda dualkonen $(A^+)^+$ till A^+. Vi kallar konen $(A^+)^+$ för *bidualkonen* till A och skriver A^{++} istället för $(A^+)^+$.

Sats 3.2.2. *För alla mängder A gäller*

$$A \subseteq \operatorname{con} A \subseteq A^{++}.$$

Bevis. Definitionen av dual- och bidualkon ger implikationerna

$$\begin{aligned} a \in A &\Rightarrow \langle a, x\rangle \geq 0 \text{ för alla } x \in A^+ \\ &\Rightarrow \langle x, a\rangle \geq 0 \text{ för alla } x \in A^+ \\ &\Rightarrow a \in A^{++}. \end{aligned}$$

Följaktligen är $A \subseteq A^{++}$, och eftersom mängden i högerledet är en konvex kon, drar vi slutsatsen att $\operatorname{con} A \subseteq A^{++}$. □

En naturlig fråga i ljuset av ovanstående sats är om $\operatorname{con} A = A^{++}$. Ett nödvändigt villkor för att så skall vara fallet är tydligen att konen $\operatorname{con} A$ är sluten, ty bidualkonen A^{++} är sluten. Nästa sats visar att villkoret också är tillräckligt.

Sats 3.2.3. *Låt X vara en konvex kon. Då är $X^{++} = \operatorname{cl} X$. Följaktligen är $X^{++} = X$ om och endast om konen X är sluten.*

Bevis. Som redan nämnts är bidualkonen X^{++} sluten, så därför följer det av inklusionen $X \subseteq X^{++}$ att $\operatorname{cl} X \subseteq X^{++}$.

För att visa den omvända inklusionen $X^{++} \subseteq \operatorname{cl} X$ antar vi att $x_0 \notin \operatorname{cl} X$. Vi skall visa att $x_0 \notin X^{++}$.

Enligt sats 3.1.3 kan $\operatorname{cl} X$ och x_0 separeras strikt av ett hyperplan. Det finns därför en vektor $c \in \mathbf{R}^n$ och ett reell tal b så att olikheten

$$\langle c, x\rangle \geq b > \langle c, x_0\rangle$$

gäller för alla $x \in X$. För fixt $x \in X$ är därför $t\langle c, x\rangle = \langle c, tx\rangle \geq b$ för alla tal $t \geq 0$, och detta är förstås bara möjligt om $\langle c, x\rangle \geq 0$ och $b \leq 0$.

Följaktligen är $\langle c, x\rangle \geq 0$ för alla $x \in X$ medan $\langle c, x_0\rangle < 0$, vilket visar att $c \in X^+$ och att $x_0 \notin X^{++}$. □

Korollarium 3.2.4. *Om konen X är ändligt genererad så är $X^{++} = X$.*

Bevis. Ty ändligt genererade koner är slutna enligt sats 2.6.11. □

Exempel 3.2.3. Dualkonen till den polyedriska konen

$$X = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \geq 0\}$$

är

$$X^+ = \operatorname{con}\{a_1, a_2, \ldots, a_m\}.$$

Detta följer av korollariet ovan (och sats 3.2.1), ty

$$X = \{a_1, a_2, \ldots, a_m\}^+ = (\operatorname{con}\{a_1, a_2, \ldots, a_m\})^+.$$

Om vi skriver en polyedrisk kon på formen $\{x \in \mathbf{R}^n \mid Ax \geq 0\}$, där A är en $m \times n$-matris, så genereras således dualkonen av *kolonnerna* i den *transponerade matrisen* A^T, dvs.

$$\{x \in \mathbf{R}^n \mid Ax \geq 0\}^+ = \{A^T y \mid y \in \mathbf{R}^m_+\}.$$ □

3.3 Lösbarhet för system av linjära olikheter

Korollarium 3.2.4 kan formuleras om till ett kriterium för lösbarhet hos system av linjära olikheter. Beviset för detta kriterium kommer att utnyttja följande hjälpsats om dualkoner.

Lemma 3.3.1. *Antag att X och Y är slutna konvexa koner i $\mathbf{R}^n$. Då är*

(i) $X \cap Y = (X^+ + Y^+)^+$;

(ii) $X + Y = (X^+ \cap Y^+)^+$, *förutsatt att konen $X + Y$ är sluten.*

Bevis. Det följer av (i) i sats 3.2.1 att $X^+ \subseteq (X \cap Y)^+$ och att $Y^+ \subseteq (X \cap Y)^+$. Följaktligen är $X^+ + Y^+ \subseteq (X \cap Y)^+ + (X \cap Y)^+ \subseteq (X \cap Y)^+$.

Ytterligare en tillämpning av sats 3.2.1 kombinerat med sats 3.2.3 ger därför

$$X \cap Y = (X \cap Y)^{++} \subseteq (X^+ + Y^+)^+.$$

Å andra sidan följer det av inklusionen $X^+ \subseteq X^+ + Y^+$ att

$$(X^+ + Y^+)^+ \subseteq X^{++} = X,$$

och analogt gäller förstås inklusionen $(X^+ + Y^+)^+ \subseteq Y$. Följakligen är

$$(X^+ + Y^+)^+ \subseteq X \cap Y.$$

Detta visar att (i) gäller.

Genom att i (i) byta ut X och Y mot X^+ och Y^+, som är slutna koner, får vi

$$X^+ \cap Y^+ = (X^{++} + Y^{++})^+ = (X + Y)^+,$$

och eftersom konen $X + Y$ antas vara sluten, drar vi slutsatsen att

$$X + Y = (X + Y)^{++} = (X^+ \cap Y^+)^+. \qquad \square$$

Nu kommer det utlovade resultatet om lösbarhet hos ett system av linjära olikheter. Satsen har skräddarsytts för att direkt kunna utnyttjas i beviset för dualitetssatsen i linjär programmering.

Sats 3.3.2. *Antag att U är en ändligt genererad kon i $\mathbf{R}^n$, att V är en ändligt genererad kon i $\mathbf{R}^m$, att A är en $m \times n$-matris samt att c är en $n \times 1$-matris. Då har systemet*

$$\text{(S)} \qquad \begin{cases} Ax \in V^+ \\ \quad x \in U^+ \\ c^T x < 0 \end{cases}$$

en lösning x om och endast om systemet

$$\text{(S}^*\text{)} \qquad \begin{cases} c - A^T y \in U \\ \qquad\quad y \in V \end{cases}$$

saknar lösning y.

Bevis. Att systemet (S*) saknar lösning y är ekvivalent med $c \notin (A^T(V)+U)$. Det är därför naturligt att titta närmare på konen $A^T(V) + U$. Eftersom konerna $A^T(V)$, U och $A^T(V) + U$ är ändligt genererade, är de också slutna. Lemma 3.3.1 kan därför tillämpas och ger till resultat att

$$A^T(V) + U = \big(A^T(V)^+ \cap U^+\big)^+.$$

Villkoret $c \notin (A^T(V) + U)$ är därför ekvivalent med att det finns en vektor $x \in A^T(V)^+ \cap U^+$ som uppfyller $c^T x < 0$, dvs. med att det finns ett x så att

$$(\dagger) \qquad \begin{cases} x \in A^T(V)^+ \\ x \in U^+ \\ c^T x < 0\,. \end{cases}$$

Det återstår nu endast att översätta villkoret $x \in A^T(V)^+$; det är ekvivalent med att

$$\langle y, Ax\rangle = \langle A^T y, x\rangle \geq 0 \qquad \text{för alla } y \in V,$$

dvs. med att $Ax \in V^+$. Systemet (†) är med andra ord ekvivalent med (S), och därmed är satsen bevisad. □

Genom att speciellt välja $U = \{0\}$ och $V = \mathbf{R}^m_+$, vilket ger $U^+ = \mathbf{R}^n$ och $V^+ = \mathbf{R}^m_+$, får vi följande specialfall av sats 3.3.2:

Korollarium 3.3.3 (Farkas lemma)**.** *Låt A vara en $m \times n$-matris och c en $n \times 1$-matris, och betrakta de två systemen:*

$$\text{(S)} \quad \begin{cases} Ax \geq 0 \\ c^T x < 0 \end{cases} \qquad \textit{och} \qquad \text{(S}^*\text{)} \quad \begin{cases} A^T y = c \\ y \geq 0 \end{cases}$$

Systemet (S) *har en lösning om och endast om systemet* (S*) *saknar lösning.*

Exempel 3.3.1. Systemet

$$\begin{cases} x_1 - x_2 + 2x_3 \geq 0 \\ -x_1 + x_2 - x_3 \geq 0 \\ 2x_1 - x_2 + 3x_3 \geq 0 \\ 4x_1 - x_2 + 10x_3 < 0 \end{cases}$$

saknar lösning, ty det duala systemet

$$\begin{cases} y_1 - y_2 + 2y_3 = 4 \\ -y_1 + y_2 - y_3 = -1 \\ 2y_1 - y_2 + 3y_3 = 10 \end{cases}$$

har lösningen $y = (3, 5, 3) \geq 0$. □

EXEMPEL 3.3.2. Systemet

$$\begin{cases} 2x_1 + \ x_2 - \ x_3 \geq 0 \\ \ x_1 + 2x_2 - 2x_3 \geq 0 \\ \ x_1 - \ x_2 + \ x_3 \geq 0 \\ \ x_1 - 4x_2 + 4x_3 < 0 \end{cases}$$

har en lösning, ty det duala systemet

$$\begin{cases} 2y_1 + \ y_2 + y_3 = \ 1 \\ \ y_1 + 2y_2 - y_3 = -4 \\ -y_1 - 2y_2 + y_3 = \ 4 \end{cases}$$

har lösningarna $y = (2-t, -3+t, t)$, $t \in \mathbf{R}$, och inga av dessa är ≥ 0 eftersom $y_1 < 0$ för $t > 2$ och $y_2 < 0$ för $t < 3$. □

I kapitel 10 kommer vi att behöva följande generalisering av exempel 3.2.3.

Sats 3.3.4. *Låt $a_1, a_2, \ldots, a_m$ vara vektorer i $\mathbf{R}^n$, och låt I, J vara en partition av indexmängden $\{1, 2, \ldots, m\}$, dvs. $I \cap J = \emptyset$ och $I \cup J = \{1, 2, \ldots, m\}$. Sätt*

$$X = \bigcap_{i \in I} \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \geq 0\} \cap \bigcap_{i \in J} \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle > 0\},$$

och antag att $X \neq \emptyset$. Då är

$$X^+ = \operatorname{con}\{a_1, a_2, \ldots, a_m\}.$$

Bevis. Sätt

$$Y = \bigcap_{i=1}^{m} \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \geq 0\}.$$

Mängden Y är sluten och innehåller X. Vi skall visa att $Y = \operatorname{cl} X$ genom att visa att varje omgivning av en godtycklig punkt $y \in Y$ innehåller punkter från X.

Så fixera en punkt $x_0 \in X$, och betrakta punkterna $y + tx_0$ för $t > 0$. Eftersom

$$\langle a_i, y + tx_0\rangle = \langle a_i, y\rangle + t\langle a_i, x_0\rangle = \begin{cases} \geq 0 & \text{om } i \in I \\ > 0 & \text{om } i \in J, \end{cases}$$

ligger $y + tx_0$ i X för alla $t > 0$, och eftersom $y + tx_0 \to y$ då $t \to 0$, finns det punkter i X godtyckligt nära y.

Enligt sats 3.2.1 är därför $X^+ = (\operatorname{cl} X)^+ = Y^+$, så påståendet i satsen följer nu av resultatet i exempel 3.2.3. □

Hur avgör man om mängden X i sats 3.3.4 är icke-tom? Om endast en av de m linjära olikheter som definierar X är strikt (dvs. om indexmängden J endast består av ett element), så ges ett nödvändigt och tillräckligt villkor för $X \neq \emptyset$ av Farkas lemma. Generaliseringen till det allmänna fallet lyder så här:

Sats 3.3.5. *Mängden X i sats 3.3.4 är icke-tom om och endast om*

$$\begin{cases} \sum_{i=1}^{m} \lambda_i a_i = 0 \\ \lambda_i \geq 0 \ \textit{för alla } i \end{cases} \quad \Rightarrow \quad \lambda_i = 0 \ \textit{för alla } i \in J.$$

Bevis. Definiera vektorer $\hat{a}_i$ i $\mathbf{R}^{n+1}$ $(= \mathbf{R}^n \times \mathbf{R})$ genom att sätta

$$\hat{a}_i = \begin{cases} (a_i, 0) & \text{för } i \in I \\ (a_i, 1) & \text{för } i \in J. \end{cases}$$

Sätt vidare $\tilde{x} = (x, x_{n+1})$ och

$$\tilde{X} = \bigcap_{i=1}^{m} \{\tilde{x} \in \mathbf{R}^{n+1} \mid \langle \hat{a}_i, \tilde{x} \rangle \geq 0\} = (\text{con}\{\hat{a}_1, \ldots, \hat{a}_m\})^+.$$

Eftersom

$$\langle \hat{a}_i, (x, x_{n+1}) \rangle = \begin{cases} \langle a_i, x \rangle & \text{om } i \in I \\ \langle a_i, x \rangle + x_{n+1} & \text{om } i \in J, \end{cases}$$

och $\langle a_i, x \rangle > 0$ för alla $i \in J$ om och endast om det finns ett negativt reellt tal x_{n+1} så att $\langle a_i, x \rangle + x_{n+1} \geq 0$ för alla $i \in J$, ligger punkten x i X om och endast om det finns ett negativt tal x_{n+1} så att $\langle \hat{a}_i, (x, x_{n+1}) \rangle \geq 0$ för alla i, dvs. om och endast om det finns ett negativt tal x_{n+1} så att $(x, x_{n+1}) \in \tilde{X}$. Detta innebär att mängden X är tom om och endast om implikationen $\tilde{x} \in \tilde{X} \Rightarrow x_{n+1} \geq 0$ gäller. Med hjälp av satserna 3.2.1 och 3.2.3 för dualkoner får vi därför följande kedja av ekvivalenser:

$$\begin{aligned} X = \emptyset \ &\Leftrightarrow \ \tilde{X} \subseteq \mathbf{R}^n \times \mathbf{R}_+ \\ &\Leftrightarrow \ \{0\} \times \mathbf{R}_+ = (\mathbf{R}^n \times \mathbf{R}_+)^+ \subseteq \tilde{X}^+ = \text{con}\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m\} \\ &\Leftrightarrow \ (0,1) \in \text{con}\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m\} \\ &\Leftrightarrow \ \text{det finns tal } \lambda_i \geq 0 \text{ så att } \sum_{i=1}^{m} \lambda_i a_i = 0 \text{ och } \sum_{i \in J} \lambda_i = 1 \\ &\Leftrightarrow \ \text{det finns tal } \lambda_i \geq 0 \text{ så att } \sum_{i=1}^{m} \lambda_i a_i = 0 \text{ och } \lambda_i > 0 \text{ för} \\ &\quad \text{något } i \in J. \end{aligned}$$

(Den sista ekvivalensen beror på att villkoret $\sum_{i=1}^{m} \lambda_i a_i = 0$ är homogent – om det är uppfyllt för en uppsättning icke-negativa λ_i med $\lambda_i > 0$ för något $i \in J$, så kan vi genom att multiplicera med en lämplig konstant åstadkomma att $\sum_{i \in J} \lambda_i = 1$.)

Eftersom ytterleden i ovanstående kedja är ekvivalenta, är också deras negationer ekvivalenta, och detta är detsamma som påståendet i satsen. □

Följande korollarium följer nu omedelbart av satsen ovan.

Korollarium 3.3.6. *Antag att vektorerna $a_1, a_2, \ldots, a_m$ är linjärt oberoende. Då är mängden X i sats 3.3.4 inte tom.*

Genom att i satserna 3.3.4 och 3.3.5 uppfatta vektorerna a_i som rader i två matriser A och C, svarande mot $i \in I$ resp. $i \in J$, får vi följande ekvivalenta matrisversion av sats 3.3.5:

Sats 3.3.7. *Låt A vara en $p \times n$-matris och C en $q \times n$-matris. Då är exakt ett av de båda systemen*

$$\begin{cases} Ax \geq 0 \\ Cx > 0 \end{cases} \qquad \textit{och} \qquad \begin{cases} A^T y + C^T z = 0 \\ y, z \geq 0,\ z \neq 0 \end{cases}$$

lösbart.

I kapitel 6.5 kommer vi att formulera och bevisa en sats om lösbarhet för system av konvexa och affina olikheter, som generaliserar sats 3.3.7.

Övningar

3.1 Ge exempel på två disjunkta slutna konvexa mängder i $\mathbf{R}^2$ som inte kan separeras strikt av ett hyperplan (dvs. en linje).

3.2 Visa att varje sluten konvex mängd i $\mathbf{R}^n$ (utom hela rummet) kan representeras som ett snitt av slutna halvrum, och att varje öppen konvex mängd (utom hela rummet) kan framställas som ett snitt av öppna halvrum.

3.3 Visa följande omvändning till lemma 3.1.2: Om två mängder X och Y kan separeras (strikt), så kan 0 separeras (strikt) från $X - Y$.

3.4 Bestäm dualkonerna till följande koner i $\mathbf{R}^2$:
a) $\mathbf{R}_+ \times \{0\}$ b) $\mathbf{R} \times \{0\}$ c) $\mathbf{R} \times \mathbf{R}_+$ d) $(\mathbf{R}_{++} \times \mathbf{R}_{++}) \cup \{(0,0)\}$
e) $\{x \in \mathbf{R}^2 \mid x_1 + x_2 \geq 0,\ x_2 \geq 0\}$

3.5 Visa att för godtyckliga mängder X och Y är $(X \times Y)^+ = X^+ \times Y^+$.

3.6 Låt $X = \operatorname{con} A$ och bestäm konerna X, X^+ och X^{++} om

a) $A = \{(1,0),(1,1),(-1,1)\}$ b) $A = \{(1,0),(-1,1),(-1,-1)\}$

c) $A = \{x \in \mathbf{R}^2 \mid x_1 x_2 = 1,\ x_1 > 0\}$

3.7 Låt $A = \{a_1, a_2, \ldots, a_m\}$ vara en delmängd av $\mathbf{R}^n$ och antag att $0 \notin A$. Visa att följande tre villkor är ekvivalenta:

(i) $\operatorname{con} A$ är en äkta kon.

(ii) $\sum_{j=1}^m \lambda_j a_j = 0$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \geq 0 \ \Rightarrow \ \lambda = 0$.

(iii) Det finns en vektor c så att $\langle c, a\rangle > 0$ för alla $a \in A$.

3.8 Undersök lösbarheten hos systemet

$$\begin{cases} x_1 - 2x_2 - 7x_3 \geq 0 \\ 5x_1 + x_2 - 2x_3 \geq 0 \\ x_1 + 2x_2 + 5x_3 \geq 0 \\ 18x_1 + 5x_2 - 3x_3 < 0. \end{cases}$$

3.9 Visa att

$$\begin{cases} x_1 + x_2 - x_3 \geq 2 \\ x_1 - x_2 \geq 1 \\ x_1 + x_3 \geq 3 \end{cases} \quad \Rightarrow \quad 6x_1 - 2x_2 + x_3 \geq 11.$$

3.10 Bestäm alla värden på parametern $\alpha \in \mathbf{R}$ som gör följande system av linjära olikheter lösbart:

$$\begin{cases} x_1 + x_2 + \alpha x_3 \geq 0 \\ x_1 + \alpha x_2 + x_3 \geq 0 \\ \alpha x_1 + x_2 + x_3 \geq 0 \\ x_1 + \alpha x_2 + \alpha^2 x_3 < 0\,. \end{cases}$$

3.11 Låt A vara en $m \times n$-matris. Visa att systemet (S) är lösbart om och endast om systemet (S*) inte är lösbart ifall

a) (S) $\begin{cases} Ax = 0 \\ x \geq 0 \\ x \neq 0 \end{cases}$ och (S*) $A^T y > 0$

b) (S) $\begin{cases} Ax = 0 \\ x > 0 \end{cases}$ och (S*) $\begin{cases} A^T y \geq 0 \\ A^T y \neq 0\,. \end{cases}$

3.12 Visa att följande system av linjära olikheter har en lösning:

$$\begin{cases} Ax = 0 \\ x \geq 0 \\ A^T y \geq 0 \\ A^T y + x > 0\,. \end{cases}$$

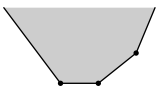
Kapitel 4

Mer om konvexa mängder

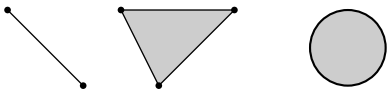
4.1 Extremalpunkter och fasader

Extremalpunkt

Polyedriska mängder, som den i figur 4.1, har hörn. Ett hörn karakteriseras av att det inte är inre punkt till någon sträcka som ligger helt i polyedern. Denna egenskap är meningsful för godtyckliga konvexa mängder.



Figur 4.1. Polyeder med hörn.



Figur 4.2. Extremalpunkter till en sträcka, en triangel och en cirkelskiva.

Definition. En punkt x i en konvex mängd X kallas en *extremalpunkt* till mängden om den inte ligger på någon öppen sträcka vars båda ändpunkter tillhör X, dvs. om

$$a_1,\, a_2 \in X \;\&\; a_1 \neq a_2 \;\;\Rightarrow\;\; x \notin]a_1, a_2[.$$

Mängden av alla extremalpunkter till X kommer att betecknas $\operatorname{ext} X$.

En punkt i det relativa inre av en konvex mängd kan uppenbarligen inte vara en extremalpunkt, utom i det triviala fallet då den konvexa mängden bara består av en enda punkt.[†] Bortsett från detta triviala fall är följaktligen $\operatorname{ext} X$ en delmängd av den relativa randen $\operatorname{rbdry} X$. Öppna konvexa mängder saknar därför extremalpunkter.

[†]Om $X = \{x_0\}$ så är nämligen $\operatorname{rint} X = \{x_0\}$, $\operatorname{rbdry} X = \emptyset$ och $\operatorname{ext} X = \{x_0\}$.

EXEMPEL 4.1.1. En sluten sträckas extremalpunkter är de båda ändpunkterna, och en sluten triangels extremalpunkter är de tre triangelhörnen. Alla punkter på randen $\{x \mid \|x\|_2 = 1\}$ är extremalpunkter till den euklidiska slutna enhetsbollen $\overline{B}(0;1)$ i $\mathbf{R}^n$. □

Extremalstråle

För konvexa koner är extremalpunktsbegreppet ointressant eftersom oäkta koner saknar extremalpunkter och äkta koner har 0 som enda extremalpunkt. För koner handlar det rätta extremalbegreppet istället om strålar, och vi behöver därför först definiera vad som menas med att en stråle ligger mellan två strålar.

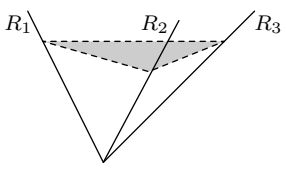
Definition. En stråle $R = \overrightarrow{a}$ säges *ligga mellan* de två strålarna $R_1 = \overrightarrow{a_1}$ och $R_2 = \overrightarrow{a_2}$ om vektorerna a_1 och a_2 är linjärt oberoende och det finns två positiva tal λ_1 och λ_2 så att $a = \lambda_1 a_1 + \lambda_2 a_2$.

Det är lätt att övertyga sig om att begreppet ”ligga mellan” bara beror av strålarna R, R_1 och R_2 och inte av vilka vektorer a, a_1 och a_2 som väljs för att representera dem. Vidare är villkoret att a_1 och a_2 är linjärt oberoende ekvivalent med att strålarna R_1 och R_2 är skilda och inte motsatta, dvs. att $R_1 \neq \pm R_2$,

Definition. En stråle R i en konvex kon kallas en *extremalstråle* till konen, om strålen inte ligger mellan två strålar i konen och den motsatta strålen $-R$ inte ligger i konen.

Mängden av alla extremalstrålar till konen X kommer att betecknas $\operatorname{exr} X$.

Villkoret att den motsatta strålen inte skall ligga i konen är förstås automatiskt uppfyllt för alla äkta koner, och som vi skall se längre fram (sats 4.2.4) medför det att oäkta koner saknar extremalstrålar.



Figur 4.3. Polyedrisk kon i $\mathbf{R}^3$ med tre extremalstrålar.

Av extremalstråle-definitionen följer att inga extremalstrålar till en konvex kon med dimension större än 1 kan innehålla någon punkt från konens relativa inre; eventuella extremalstrålar är med andra ord delmängder till konens relativa rand.

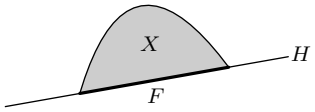
EXEMPEL 4.1.2. För de fyra delkonerna av $\mathbf{R}$ gäller: $\operatorname{exr}\{0\} = \operatorname{exr}\mathbf{R} = \emptyset$, $\operatorname{exr}\mathbf{R}_+ = \mathbf{R}_+$ och $\operatorname{exr}\mathbf{R}_- = \mathbf{R}_-$.

Den oäkta konen $\mathbf{R}\times\mathbf{R}_+$ i $\mathbf{R}^2$ ("övre halvplanet") saknar extremalstrålar eftersom de båda randstrålarna $\mathbf{R}_+ \times \{0\}$ och $\mathbf{R}_- \times \{0\}$ diskvalificeras av kravet att den motsatta strålen inte får ligga i konen. □

Fasad

Definition. En delmängd F till en konvex mängd X kallas en *äkta fasad* till X om $F = X \cap H$ för något stödhyperplan H till X. Dessutom kallas mängden X själv och den tomma mängden $\emptyset$ för *oäkta fasader* till X.‡

Skälet till att inkludera hela mängden och tomma mängden bland fasaderna är att detta förenklar en del satsformuleringar och bevis.



Figur 4.4. Konvex mängd med fasad F.

En konvex mängds fasader är uppenbarligen konvexa mängder. Och för konvexa koner är de äkta fasaderna också koner beroende på att en konvex kons stödhyperplan måste gå genom 0 och således vara ett linjärt delrum.

EXEMPEL 4.1.3. Varje punkt på randen $\{x \mid \|x\|_2 = 1\}$ är en fasad till den slutna enhetsbollen $\overline{B}(0;1)$, eftersom det går ett stödhyperplan genom varje randpunkt och stödhyperplanet inte skär enhetsbollen i någon annan punkt. □

EXEMPEL 4.1.4. En kub i $\mathbf{R}^3$ har 26 äkta fasader: 8 hörnpunkter, 12 kantsträckor och 6 sidoytor. □

‡Det finns en alternativ mer generell definition av begreppet fasad, se övning 4.7. I standardverket *Convex Analysis* av Rockafellar kallas våra äkta fasader för *utsatta fasader*. Alla utsatta fasader är också fasader enligt den alternativa definitionen, men den alternativa definitionen ger för vissa konvexa mängder fasader som inte är utsatta.

Sats 4.1.1. *Om X är en sluten konvex mängd, så är* $\operatorname{rbdry} X = \bigcup F$*, där unionen tas över alla äkta fasader F till X.*

Bevis. Om $F = X \cap H$ är en äkta fasad till X och $x_0 \in F$, så stöder hyperplanet H mängden X i punkten x_0, och eftersom X inte innehålles i H, följer det att x_0 är en relativ randpunkt till X. Detta visar inklusionen $\bigcup F \subseteq \operatorname{rbdry} X$.

Omvänt, om x_0 är en godtycklig relativ randpunkt till X, så finns det ett hyperplan H som stöder mängden X i punkten x_0, och detta innebär att x_0 ligger i den äkta fasaden $X \cap H$. □

Sats 4.1.2. *Snittet av två fasader till en konvex mängd är en fasad till mängden.*

Bevis. Låt F_1 och F_2 vara två fasader till den konvexa mängden X, och sätt $F = F_1 \cap F_2$. Påståendet är trivialt om fasaderna F_1 och F_2 är identiska, eller om de är disjunkta, eller om en av dem är oäkta.

Antag därför att de båda fasaderna F_1 och F_2 är skilda och äkta, dvs. har formen $F_i = X \cap H_i$, där H_1 och H_2 är olika stödhyperplan till mängden X, och att $F \neq \emptyset$. Sätt

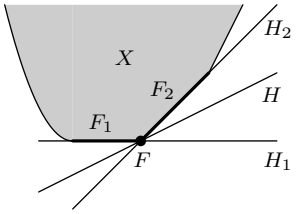
$$H_i = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle = b_i\},$$

där hyperplanens normalvektorer c_i valts så att X ligger i de båda halvrummen $\{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i\}$, och låt $x_1 \in X$ vara en punkt som uppfyller villkoret $\langle c_1, x_1\rangle < b_1$.

Eftersom $X \cap H_1 \cap H_2 = F \neq \emptyset$ kan hyperplanen H_1 och H_2 inte vara parallella, så det följer att $c_2 \neq -c_1$. Vi får därför ett nytt hyperplan

$$H = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = b\}$$

genom att sätta $c = c_1 + c_2$ och $b = b_1 + b_2$. Vi skall visa att H är ett stödhyperplan till X och att $F = X \cap H$, vilket bevisar att snittet F är en fasad till X.



Figur 4.5. Illustration till beviset för sats 4.1.2.

För alla $x \in X$ är

$$\langle c, x\rangle = \langle c_1, x\rangle + \langle c_2, x\rangle \leq b_1 + b_2 = b,$$

för punkten $x_1 \in X$ gäller speciellt att

$$\langle c, x_1\rangle = \langle c_1, x_1\rangle + \langle c_2, x_1\rangle < b_1 + b_2 = b,$$

och för alla $x \in F = X \cap H_1 \cap H_2$ är

$$\langle c, x\rangle = \langle c_1, x\rangle + \langle c_2, x\rangle = b_1 + b_2 = b.$$

Detta visar dels att H är ett stödhyperplan till X, dels att $F \subseteq X \cap H$.

Omvänt, för $x \in X \cap H$ är $\langle c_1, x\rangle \leq b_1$, $\langle c_2, x\rangle \leq b_2$ och

$$\langle c_1, x\rangle + \langle c_2, x\rangle = b_1 + b_2,$$

vilket medför att $\langle c_1, x\rangle = b_1$ och $\langle c_2, x\rangle = b_2$, dvs. $x \in X \cap H_1 \cap H_2 = F$. Således gäller också inklusionen $X \cap H \subseteq F$. □

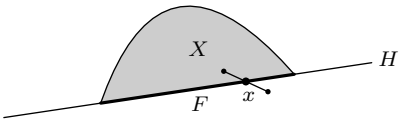
Sats 4.1.3. *(i) Antag att F är en fasad till den konvexa mängden X. En punkt x i fasaden F är en extremalpunkt till F om och endast om x är en extremalpunkt till X.*

(ii) Antag att F är en fasad till den konvexa konen X. En stråle R i fasaden F är en extremalstråle till F om och endast om R är en extremalstråle till X.

Bevis. För oäkta fasader finns det ingenting att visa så vi kan anta att $F = X \cap H$ för något stödhyperplan H till X.

En punkt i ett hyperplan kan inte vara inre punkt i någon sträcka vars båda ändpunkter ligger på samma sida om hyperplanet, såvida inte båda ändpunkterna ligger i hyperplanet, dvs. såvida inte hela sträckan ligger i hyperplanet.

Analogt kan inte någon stråle i ett hyperplan H (genom origo) ligga mellan två strålar i samma slutna halvrum med H som rand, såvida inte båda dessa strålar ligger i hyperplanet H. Och om en stråle ligger i hyperplanet, så ligger förstås också den motsatta strålen i samma hyperplan.



Figur 4.6. Om $x \in H$ är inre punkt på en sträcka, så kan inte sträckans båda ändpunkter ligga i samma öppna halvrum.

(i) Om $x \in F$ är en inre punkt till någon sträcka vars båda ändpunkter tillhör X, så är därför x inre punkt i en sträcka vars båda ändpunkter tillhör fasaden F. Detta visar implikationen $x \notin \operatorname{ext} X \Rightarrow x \notin \operatorname{ext} F$. Den omvända implikationen är trivial eftersom varje sträcka i F ligger i X. Följaktligen gäller ekvivalensen $x \notin \operatorname{ext} X \Leftrightarrow x \notin \operatorname{ext} F$.

(ii) Antag att R är en stråle i F och att R inte är en extremalstråle till konen X. Då ligger antingen R mellan två strålar R_1 och R_2 i X eller också ligger strålen $-R$ i X. I det förstnämnda fallet ligger de båda strålarna R_1 och R_2 nödvändigtvis också i F, och i det andra fallet ligger strålen $-R$ i F. I båda fallen drar vi slutsatsen att R inte heller är en extremalstråle till konen F. Detta visar att implikationen $R \notin \operatorname{exr} X \Rightarrow R \notin \operatorname{exr} F$ gäller.

Den omvända implikationen är trivial, vilket innebär att ekvivalensen $R \notin \operatorname{exr} X \Leftrightarrow R \notin \operatorname{exr} F$ gäller, och detta bevisar påstående (ii). □

4.2 Struktursatser för konvexa mängder

Sats 4.2.1. *Antag att den slutna konvexa mängden X är linjefri och att* $\dim X \geq 2$. *Då är*

$$X = \operatorname{cvx}(\operatorname{rbdry} X).$$

Bevis. Sätt $n = \dim X$. Genom att identifiera mängden X:s affina hölje med $\mathbf{R}^n$ kan vi utan inskränkning anta att X är en konvex mängd i $\mathbf{R}^n$ med maximal dimension, vilket innebär att $\operatorname{rbdry} X = \operatorname{bdry} X$. För att visa likheten i satsen räcker det således att visa att varje punkt i X ligger i det konvexa höljet av randen $\operatorname{bdry} X$, ty inklusionen $\operatorname{cvx}(\operatorname{bdry} X) \subseteq X$ är trivialt uppfylld.

Sätt $C = \operatorname{recc} X$; eftersom mängden X är linjefri, är recessionskonen C en äkta kon, och det följer därför av sats 3.1.7 att det finns ett slutet halvrum

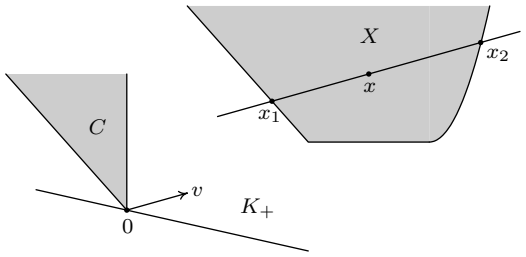
$$K = \{x \in \mathbf{R}^n \mid \langle c, x\rangle \geq 0\}$$

som innehåller C som äkta delmängd. Eftersom C är en sluten kon, måste vidare motsvarande öppna halvrum

$$K_+ = \{x \in \mathbf{R}^n \mid \langle c, x\rangle > 0\}$$

innehålla en vektor v som inte ligger i C. Vektorn $-v$, som ligger i det motsatta öppna halvrummet, kan då heller inte ligga i C. Jmf figur 4.7.

Vi har således producerat två motsatta vektorer $\pm v$ som båda ligger utanför recessionskonen. Om x är en punkt i X, så kommer därför de båda motsatta halvlinjerna $x + \overrightarrow{v}$ och $x - \overrightarrow{v}$ att skära komplementet till X. Snittet mellan X och linjen genom x med riktningen v, som är en sluten konvex mängd, är således antingen en sluten sträcka $[x_1, x_2]$ som innehåller punkten x



Figur 4.7. Illustration till beviset för sats 4.2.1.

och vars båda ändpunkter är randpunkter till X, eller enpunktsmängden $\{x\}$ där då x är randpunkt till X. I det förstnämnda fallet är punkten x förstås en konvex kombination av sträckans ändpunkter, och i båda fallen är därför x en konvex kombination av randpunkter till X, vilket bevisar satsen. □

Vi kan nu ge en fullständig beskrivning av linjefria slutna konvexa mängder i termer av extremalpunkter och recessionskoner.

Sats 4.2.2. *(i) Icke-linjefria slutna konvexa mängder saknar extremalpunkter.*

(ii) Icke-tomma linjefria slutna konvexa mängder har extremalpunkter, och om X är en sådan mängd så är

$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X.$$

Bevis. (i) Låt X vara en icke-linjefri sluten konvex mängd. Att X inte är linjefri innebär att det finns en nollskild vektor y i det recessiva delrummet $\operatorname{lin} X$. För varje punkt $x \in X$ ligger då de båda punkterna $x \pm y$ i X, och eftersom $x \in]x-y, x+y[$, kan x inte vara en extremalpunkt. Mängden $\operatorname{ext} X$ är således tom.

(ii) Att $\operatorname{ext} X \neq \emptyset$ och att $X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$, om mängden X är icke-tom, linjefri, sluten och konvex, visas genom induktion över mängdens dimension n.

För nolldimensionella konvexa mängder, dvs. enpunktsmängder, är påståendet trivialt, och endimensionella konvexa mängder är antingen halvlinjer $a + \overrightarrow{v}$ eller slutna sträckor $[a, b]$ med en resp. två extremalpunkter som uppfyller likheten i (ii).

Antag därför att $n \geq 2$ och att påståendet är sant för alla konvexa mängder av lägre dimension än n, och låt X vara en n-dimensionell sluten konvex mängd. På grund av satserna 4.1.1 och 4.2.1 är

$$X = \operatorname{cvx}\big(\bigcup F\big),$$

där unionen tas över alla äkta fasader F till X. Varje äkta fasad F är en icke-tom linjefri sluten konvex delmängd av något stödhyperplan H och har en dimension som är mindre än eller lika med $n-1$. Induktionsantagandet medför därför att $\operatorname{ext} F \neq \emptyset$ och

$$F = \operatorname{cvx}(\operatorname{ext} F) + \operatorname{recc} F.$$

Eftersom $\operatorname{ext} F \subseteq \operatorname{ext} X$ (sats 4.1.3), följer det att $\operatorname{ext} X \neq \emptyset$. Vidare är $\operatorname{recc} F$ en delmängd till $\operatorname{recc} X$, varför inklusionen $F \subseteq \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$ gäller för varje fasad F. Följaktligen är också unionen $\bigcup F$ inkluderad i den konvexa mängden $\operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$, så det följer att

$$X = \operatorname{cvx}\big(\bigcup F\big) \subseteq \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X \subseteq X + \operatorname{recc} X = X,$$

dvs. $X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{recc} X$, och därmed är induktionen genomförd och satsen bevisad. □

En kompakt mängds recessionskon är lika med nollkonen, så därför är följande resultat ett omedelbart korollarium till sats 4.2.2.

Korollarium 4.2.3. *Varje icke-tom kompakt konvex mängd har extremalpunkter och är lika med konvexa höljet av sina extremalpunkter.*

Vi skall nu ge motsvarigheten till sats 4.2.2 för konvexa koner, och för att förenkla beteckningarna kommer vi att använda oss av följande konvention: Om $\mathcal{A}$ är en familj av strålar, sätter vi

$$\operatorname{con} \mathcal{A} = \operatorname{con}\big(\bigcup_{R \in \mathcal{A}} R\big),$$

dvs. $\operatorname{con} \mathcal{A}$ är den kon som genereras av alla vektorerna på alla strålarna i $\mathcal{A}$. Om vi för varje stråle $R \in \mathcal{A}$ godtyckligt väljer en nollskild vektor a_R på R, och sätter $A = \{a_R \mid R \in \mathcal{A}\}$, så är förstås $\operatorname{con} \mathcal{A} = \operatorname{con} A$.

Om $\mathcal{A}$ är en ändlig mängd av strålar, så är därför konen $\operatorname{con} \mathcal{A}$ ändligt genererad, och som generatorer kan vi välja en nollskild vektor på varje stråle.

Sats 4.2.4. *(i) Oäkta slutna konvexa koner saknar extremalstrålar.*

(ii) Äkta slutna konvexa koner, förutom nollkonen, har extremalstrålar. Om X är en äkta sluten konvex kon, så är

$$X = \operatorname{con}(\operatorname{exr} X).$$

Bevis. (i) Låt X vara en oäkta sluten konvex kon, och låt $R = \overrightarrow{x}$ vara en godtycklig stråle i X. Vi skall visa att R inte kan vara en extremalstråle.

Att X är en oäkta kon innebär att det finns en nollskild vektor a i snittet $X \cap (-X)$. Om $R = \overrightarrow{a}$, så är $-R = \overrightarrow{-a}$ också en stråle i X, så R är inte en extremalstråle i det fallet. Motsvarande gäller förstås för strålen $R = -\overrightarrow{a}$.

Antag därför att $R \neq \pm\overrightarrow{a}$; då är x och a linjärt oberoende, och det följer att strålarna $R_1 = \overrightarrow{x+a}$ och $R_2 = \overrightarrow{x-a}$ är två skilda icke-motsatta strålar i konen X. Eftersom $x = \frac{1}{2}(x+a) + \frac{1}{2}(x-a)$, ligger strålen R mellan R_1 och R_2. R är således inte heller i detta fall en extremalstråle.

(ii) För den triviala nollkonen $X = \{0\}$ gäller likheten $X = \operatorname{con}(\operatorname{exr} X)$ beroende på att $\operatorname{con}\emptyset = \{0\}$. För att bevisa att likheten också gäller för alla icke-triviala äkta slutna konvexa koner och att dessa har extremalstrålar, behöver vi bara göra smärre modifikationer i induktionsbeviset för motsvarande del av sats 4.2.2.

Startsteget är förstås trivialt eftersom en endimensionell äkta kon är en stråle. Antag att påståendet är sant för alla koner av dimension högst lika med $n-1$, och låt X vara en äkta sluten konvex kon av dimension n. Eftersom X speciellt är en linjefri mängd, är $X = \operatorname{cvx}\big(\bigcup F\big)$, där unionen tas över konens alla äkta fasader F. Eftersom $\operatorname{cvx}\big(\bigcup F\big) \subseteq \operatorname{con}\big(\bigcup F\big) \subseteq \operatorname{con} X = X$, följer det nu att

$$X = \operatorname{con}\big(\bigcup F\big).$$

Fasaderna F i denna union är koner, och naturligtvis kan vi ta bort den triviala fasaden $\{0\}$ från unionen utan att likheten upphör att gälla. Varje återstående kon F är en äkta sluten nollskild konvex kon med en dimension som inte överstiger $n-1$.

För varje sådan kon F är på grund av induktionsantagandet mängden $\operatorname{exr} F$ icke-tom och $F = \operatorname{con}(\operatorname{exr} F)$. Eftersom $\operatorname{exr} F \subseteq \operatorname{exr} X$, följer härav dels att mängden $\operatorname{exr} X$ inte är tom, dels att $F \subseteq \operatorname{con}(\operatorname{exr} X)$.

Unionen $\bigcup F$ av fasaderna är således inkluderad i konen $\operatorname{con}(\operatorname{exr} X)$, och av likheten $X = \operatorname{con}\big(\bigcup F\big)$ följer därför att $X \subseteq \operatorname{con}(\operatorname{exr} X)$. Eftersom den omvända inklusionen är trivial, är likheten i (ii) bevisad. □

Eftersom recessionskonen till en linjefri konvex mängd är en äkta kon, är följande representationssats för konvexa mängder en omedelbar konsekvens av satserna 4.2.2 och 4.2.4.

Sats 4.2.5. *Varje linjefri sluten konvex icke-tom mängd X har framställningen*

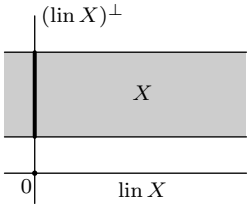
$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{con}(\operatorname{exr}(\operatorname{recc} X)).$$

Studiet av godtyckliga slutna konvexa mängder reduceras till studiet av linjefria sådana med hjälp av följande sats, som innebär att varje icke-linjefri

konvex sluten mängd är en cylinder med en linjefri konvex mängd som bas och med en "axel" som är parallell med det recessiva delrummet $\operatorname{lin} X$.

Sats 4.2.6. *Antag att X är en sluten konvex mängd i $\mathbf{R}^n$. Då är snittet $X \cap (\operatorname{lin} X)^\perp$ en linjefri sluten konvex mängd och*

$$X = \operatorname{lin} X + X \cap (\operatorname{lin} X)^\perp.$$



Figur 4.8. Illustration till sats 4.2.6.

Bevis. Varje $x \in \mathbf{R}^n$ har en unik uppdelning $x = y + z$ med $y \in \operatorname{lin} X$ och $z \in (\operatorname{lin} X)^\perp$. Om $x \in X$, så ligger vektorn z också X beroende på att

$$z = x - y \in X + \operatorname{lin} X = X.$$

Detta bevisar inklusionen $X \subseteq \operatorname{lin} X + X \cap (\operatorname{lin} X)^\perp$, och den omvända inklusionen följer av att $\operatorname{lin} X + X \cap (\operatorname{lin} X)^\perp \subseteq \operatorname{lin} X + X = X$. □

Övningar

4.1 Bestäm $\operatorname{ext} X$ samt avgör om $X = \operatorname{cvx}(\operatorname{ext} X)$ för följande mängder X:

a) $X = \{x \in \mathbf{R}^2_+ \mid x_1 + x_2 \geq 1\}$

b) $X = \big([0,1] \times [0,1[\,\big) \cup \big([0, \tfrac{1}{2}] \times \{1\}\big)$

c) $X = \operatorname{cvx}\big(\{x \in \mathbf{R}^3 \mid (x_1 - 1)^2 + x_2^2 = 1,\, x_3 = 0\} \cup \{(0,0,1), (0,0,-1)\}\big)$.

4.2 Visa att för varje delmängd A av $\mathbf{R}^n$ är $\operatorname{ext}(\operatorname{cvx} A) \subseteq A$.

4.3 Låt $X = \operatorname{cvx} A$ och antag att mängden A är minimal i följande bemärkelse: Om $B \subseteq A$ och $X = \operatorname{cvx} B$, så är $B = A$. Visa att då är $A = \operatorname{ext} X$.

4.4 Låt x_0 vara en punkt i en konvex mängd X. Visa att $x_0 \in \operatorname{ext} X$ om och endast om mängden $X \setminus \{x_0\}$ är konvex.

4.5 Ge exempel på en kompakt konvex delmängd X av $\mathbf{R}^3$ med egenskapen att $\operatorname{ext} X$ inte är sluten.

4.6 En punkt x_0 i en konvex mängd X kallas en *utsatt punkt* om enpunktsmängden $\{x_0\}$ är en fasad, dvs. om det finns ett stödhyperplan H till X så att $X \cap H = \{x_0\}$.

a) Visa att varje utsatt punkt är en extremalpunkt till X.

b) Ge exempel på en sluten konvex mängd i $\mathbf{R}^2$ med en extremalpunkt som inte är utsatt.

4.7 Det finns en mer generell definition av begreppet fasad som lyder som följer:

En *fasad* till en konvex mängd X är en konvex delmängd F av X sådan att varje slutet linjesegment i X, som har någon relativt inre punkt gemensam med F, ligger helt i F, dvs.

$$(a, b \in X \;\&\;]a, b[\cap F \neq \emptyset) \implies a, b \in F.$$

Låt oss kalla fasader enligt denna definition för *generella fasader* för att skilja dem från fasader enligt vår definition, som vi kallar *utsatta fasader* om de är äkta, dvs. inte är fasaderna X eller $\emptyset$.

Den tomma mängden $\emptyset$ och X är uppenbarligen generella fasader till X, och alla extremalpunkter till X är också generella fasader.

Visa följande egenskaper för de generella fasaderna till en konvex mängd X.

a) Varje utsatt fasad är en generell fasad.
b) Det finns en konvex mängd med en generell fasad som inte är en utsatt fasad.
c) Om F är en generell fasad till X och F' är en generell fasad till F, så är F' en generell fasad till X, men motsvarande behöver inte gälla för utsatta fasader.
d) Om F är en generell fasad till X och C är en godtycklig konvex delmängd av X sådan att $F \cap \operatorname{rint} C \neq \emptyset$, så är $C \subseteq F$.
e) Om F är en generell fasad till X, så är $F = X \cap \operatorname{cl} F$. Speciellt är alltså alla generella fasader till en sluten konvex mängd slutna.
f) Om F_1 och F_2 är två generella fasader till X och $\operatorname{rint} F_1 \cap \operatorname{rint} F_2 \neq \emptyset$, så är $F_1 = F_2$.
g) Om F är en generell fasad till X och $F \neq X$, så är $F \subseteq \operatorname{rbdry} X$.

Kapitel 5

Polyedrar

Vi har redan visat en del spridda resultat om polyedrar. Nu skall vi samla ihop dessa trådar och komplettera dem så att vi får en fullständig beskrivning av denna viktiga klass av konvexa mängder.

5.1 Extremalpunkter och extremalstrålar

Polyedrar och extremalpunkter

En polyeder X i $\mathbf{R}^n$, som inte sammanfaller med hela rummet, är ett snitt av ändligt många slutna halvrum och kan därför skrivas på formen

$$X = \bigcap_{j=1}^{m} K_j,$$

med

$$K_j = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \geq b_j\}$$

för lämpliga nollskilda vektorer c_j i $\mathbf{R}^n$ och reella tal b_j. På matrisform fås

$$X = \{x \in \mathbf{R}^n \mid Cx \geq b\};$$

här är C en $m \times n$-matris med ${c_j}^T$ som rader, och $b = \begin{bmatrix} b_1 & b_2 & \dots & b_m \end{bmatrix}^T$.

Vi sätter vidare

$$K_j^\circ = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle > b_j\} = \operatorname{int} K_j,$$
$$H_j = \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle = b_j\} = \operatorname{bdry} K_j.$$

Mängderna K_j° är öppna halvrum, och H_j är hyperplan.

Om $b = 0$, dvs. om alla hyperplanen H_j är linjära delrum, så är X en polyedrisk kon.

Polyedern X ligger uppenbarligen i det slutna halvrummet K_j, som begränsas av hyperplanet H_j. Sätt

$$F_j = X \cap H_j.$$

Om hyperplanet H_j har någon punkt gemensam med polyedern X utan att innehålla hela polyedern, så är det ett stödhyperplan till X, och mängden F_j är i detta fall en äkta fasad till X. Om $X \cap H_j = \emptyset$ eller $X \subseteq H_j$, så är mängden F_j också en fasad på grund av vår konvention beträffande oäkta fasader. Fasaderna F_j är förstås polyedrar.

Alla punkter i fasaderna F_j (äkta såväl som oäkta) är randpunkter till X. Eftersom

$$X = \bigcap_{j=1}^{m} K_j^\circ \cup \bigcup_{j=1}^{m} F_j,$$

och alla punkterna i den öppna mängden $\bigcap_{j=1}^{m} K_j^\circ$ är inre punkter till X, följer det att

$$\operatorname{int} X = \bigcap_{j=1}^{m} K_j^\circ \qquad \text{och} \qquad \operatorname{bdry} X = \bigcup_{j=1}^{m} F_j.$$

Mängden $\operatorname{ext} X$ av extremalpunkter till polyedern X är en delmängd av randen, dvs. av $\bigcup_{j=1}^{m} F_j$. Extremalpunkterna karakteriseras av följande sats.

Sats 5.1.1. *En punkt x_0 i polyedern $X = \bigcap_{j=1}^{m} K_j$ är en extremalpunkt om och endast det finns en delmängd I av indexmängden $\{1, 2, \ldots, m\}$ så att $\bigcap_{j \in I} H_j = \{x_0\}$.*

Bevis. Antag att det finns en sådan indexmängd I. Snittet

$$F = \bigcap_{j \in I} F_j = X \cap \bigcap_{j \in I} H_j = \{x_0\}$$

är på grund av sats 4.1.2 också en fasad till X, och uppenbarligen är x_0 en extremalpunkt till F. Enligt sats 4.1.3 är därför x_0 också en extremalpunkt till X.

Antag omvänt att det inte finns någon sådan indexmängd I, och låt J vara en indexmängd som är maximal med avseende på egenskapen $x_0 \in \bigcap_{j \in J} H_j$. (Om x_0 är en inre punkt i polyedern sätter vi $J = \emptyset$ och tolkar snittet över den tomma indexmängden J som hela rummet $\mathbf{R}^n$.) Snittet $\bigcap_{j \in J} H_j$ är ett affint delrum, som enligt antagande består av mer än en punkt, och det innehåller därför någon hel linje $\{x_0 + tv \mid t \in \mathbf{R}\}$ genom x_0. Hela linjen ligger förstås också i den större mängden $\bigcap_{j \in J} K_j$.

För alla $j \notin J$ och tillräckligt små värden på $|t|$ ligger punkterna $x_0 + tv$ också i halvrummet K_j beroende på att x_0 är en inre punkt i dessa halvrum. Följaktligen finns det ett tal $\delta > 0$ så att sträckan $[x_0 - \delta v, x_0 + \delta v]$ ligger i $X = \bigcap_{j\in J} K_j \cap \bigcap_{j\notin J} K_j$, vilket innebär att x_0 inte är en extremalpunkt. □

Att $\bigcap_{j\in I} H_j = \{x_0\}$ betyder att motsvarande linjära ekvationssystem

$$\langle c_j, x\rangle = b_j, \quad j \in I,$$

med n obekanta, har entydig lösning. Ett nödvändigt villkor för att så skall vara fallet är att indexmängden I innehåller minst n stycken element. Om ekvationssystemet har entydig lösning och fler än n ekvationer, kan vi vidare genom att stryka lämpligt valda ekvationer alltid erhålla ett kvadratiskt delsystem med entydig lösning.

Ett nödvändigt villkor för att polyedern $X = \bigcap_{j=1}^m K_j$ skall ha en extremalpunkt är således att $m \geq n$. (Detta följer också av sats 2.7.7, ty om $m < n$, så är $\dim \operatorname{lin} X = \dim\{x \in \mathbf{R}^n \mid Cx = 0\} = n - \operatorname{rang} C \geq n - m > 0$, varför X inte kan vara linjefri.)

Sats 5.1.1 ger upphov till följande metod för att bestämma samtliga extremalpunkter till X i fallet $m \geq n$:

Lös för varje delmängd J av $\{1, 2, \ldots, m\}$ med n stycken element motsvarande linjära ekvationssystem $\langle c_j, x\rangle = b_j$, $j \in J$. Ett nödvändigt och tillräckligt villkor för att systemet skall ha entydig lösning x_0, dvs. för att $\bigcap_{j\in J} H_j = \{x_0\}$, är att systemets koefficientmatris har rang n. Om så är fallet och lösningen x_0 ligger i X, dvs. satisfierar de resterande linjära olikheterna $\langle c_j, x\rangle \geq b_j$, så är x_0 en extremalpunkt till X.

Eftersom antalet delmängder J med n stycken element är lika med $\binom{m}{n}$, är antalet extremalpunkter till polyedern X begränsat av detta tal, och vi har därmed visat följande sats.

Sats 5.1.2. *En polyeder har ändligt många extremalpunkter.*

Polyedriska koner och extremalstrålar

En polyedrisk kon i $\mathbf{R}^n$ är ett snitt $X = \bigcap_{j=1}^m K_j$ av koniska halvrum K_j som begränsas av hyperplan H_j genom origo, och fasaderna $F_j = X \cap H_j$ är polyedriska koner. Följande sats är en direkt parallell till sats 5.1.1.

Sats 5.1.3. *En punkt x_0 i den polyedriska konen $X = \bigcap_{j=1}^m K_j$ genererar en extremalstråle $R = \overrightarrow{x_0}$ till konen om och endast om $-x_0 \notin X$ och det finns en delmängd I av indexmängden $\{1, 2, \ldots, m\}$ så att $\bigcap_{j\in I} H_j = \{tx_0 \mid t \in \mathbf{R}\}$.*

Bevis. Antag att det finns en sådan indexmängd I och att $-x_0$ inte tillhör konen X; då är

$$\bigcap_{j\in I} F_j = X \cap \bigcap_{j\in I} H_j = R.$$

Detta innebär enligt sats 4.1.2 att R är en fasad till konen X. Strålen R är förstås en extremalstråle till fasaden (konen) R, så det följer därför av sats 4.1.3 att R är en extremalstråle till X.

Om $-x_0$ tillhör X, så är konen X inte äkta, och den saknar därför extremalstrålar enligt sats 4.2.4.

Det återstår att visa att R inte är en extremalstråle i det fall då $-x_0 \notin X$ och det inte finns någon indexmängd I med egenskapen att snittet $\bigcap_{j\in I} H_j$ är lika med linjen genom 0 och x_0. Låt för den skull J vara en indexmängd som är maximal med avseende på egenskapen x_0 ligger i snittet $\bigcap_{j\in J} H_j$. På grund av vårt antagande är snittet ett linjärt delrum av dimension minst lika med två, så det innehåller en vektor v som är linjärt oberoende av x_0. Vektorerna $x_0 + tv$ och $x_0 - tv$ tillhör $\bigcap_{j\in J} H_j$ och därmed $\bigcap_{j\in J} K_j$ för alla reella tal t. För tillräckligt små värden på $|t|$ ligger vektorerna också i K_j för $j \notin J$, ty x_0 är en inre punkt i K_j för dessa index j. Det finns därför ett tal $\delta > 0$ så att vektorerna $x_+ = x_0 + \delta v$ och $x_- = x_0 - \delta v$ båda ligger i konen X. De båda vektorerna x_+ och x_- är linjärt oberoende och $x_0 = \frac{1}{2}x_+ + \frac{1}{2}x_-$, så strålen $R = \overrightarrow{x_0}$ ligger mellan de båda strålarna $\overrightarrow{x_+}$ och $\overrightarrow{x_-}$ i konen X och är därför inte en extremalstråle. □

Man bestämmer således extremalstrålarna till konen

$$X = \bigcap_{j=1}^{m} \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \geq 0\}$$

genom att välja en indexmängd J bestående av $n-1$ stycken element ur mängden $\{1, 2, \ldots, m\}$, vilket kan ske på $\binom{m}{n-1}$ sätt, och sedan lösa motsvarande homogena ekvationssystem $\langle c_j, x\rangle = 0,\ j \in J$. Lösningsmängden är endimensionell och genereras av en vektor x_0 om koefficientmatrisens rang är lika med $n-1$. Om x_0 satisfierar de övriga linjära olikheterna och $-x_0$ inte gör det, så är $R = \overrightarrow{x_0}$ en extremalstråle; om $-x_0$ satisfierar de linjära olikheterna medan x_0 inte gör det är istället $-R$ en extremalstråle.

Det följer att $\binom{m}{n-1}$ är en övre gräns på antalet extremalstrålar. Vi har därför följande korollarium till sats 5.1.3.

Sats 5.1.4. *En polyedriskt kon har ändligt många extremalstrålar.*

5.2 Polyedriska koner

Sats 5.2.1. *En kon är polyedrisk om och endast om den är ändligt genererad.*

Bevis. Vi visar först att varje polyedrisk kon är ändligt genererad.

Det följer av sats 4.2.6 att varje polyedrisk kon X har representationen

$$X = \operatorname{lin} X + X \cap (\operatorname{lin} X)^{\perp},$$

och $X \cap (\operatorname{lin} X)^{\perp}$ är en linjefri, dvs. äkta, polyedrisk kon. Om vi låter mängden B bestå av en punkt på varje extremalstråle till $X \cap (\operatorname{lin} X)^{\perp}$, så är B en ändlig mängd och

$$X \cap (\operatorname{lin} X)^{\perp} = \operatorname{con} B$$

på grund av satserna 5.1.4 och 4.2.4.

Låt $e_1, e_2, \ldots, e_d$ vara en bas för det linjära delrummet $\operatorname{lin} X$ och sätt $e_0 = -(e_1 + e_2 + \cdots + e_d)$; då genereras $\operatorname{lin} X$ som kon av mängden $A = \{e_0, e_1, \ldots, e_d\}$, dvs.

$$\operatorname{lin} X = \operatorname{con} A.$$

Sammanfattningsvis är således

$$X = \operatorname{lin} X + X \cap (\operatorname{lin} X)^{\perp} = \operatorname{con} A + \operatorname{con} B = \operatorname{con}(A \cup B),$$

vilket visar att konen X är ändligt genererad av mängden $A \cup B$.

Antag omvänt att X är en ändligt genererad kon, dvs. att $X = \operatorname{con} A$ för någon ändlig mängd A. Vi börjar med att konstatera att dualkonen X^+ är en polyedrisk kon. Om $A \neq \emptyset$, så är nämligen

$$X^+ = A^+ = \{x \in \mathbf{R}^n \mid \langle x, a\rangle \geq 0 \text{ för alla } a \in A\} = \bigcap_{a \in A} \{x \in \mathbf{R}^n \mid \langle a, x\rangle \geq 0\}$$

ett snitt av ändligt många koniska halvrum, dvs. en polyedrisk kon, och i det triviala fallet $A = \emptyset$ är $X = \{0\}$ och $X^+ = \mathbf{R}^n$.

Den redan bevisade delen av satsen medför därför att dualkonen X^+ är en ändligt genererad kon. Men då är dualkonen till X^+, dvs. bidualkonen X^{++}, också en polyedrisk kon, och bidualkonen X^{++} sammanfaller enligt korollarium 3.2.4 med ursprungskonen X. Vi kan därför dra slutsatsen att X är en polyedrisk kon. □

Vi kan nu bevisa två resultat som vi måste lämna obevisade i kapitel 2.6; jämför med korollarium 2.6.9.

Sats 5.2.2. *(i) Snittet $X \cap Y$ mellan två ändligt genererade koner X och Y är en ändligt genererad kon.*

(ii) Inversa bilden $T^{-1}(X)$ av en ändligt genererad kon X under en linjär avbildning T är en ändligt genererad kon.

Bevis. Snittet av två koniska polyedrar och inversa bilden av en konisk polyeder under en linjär avbildning är uppenbarligen koniska polyedrar. Satsen är därför ett korollarium till sats 5.2.1. □

5.3 Polyederns inre struktur

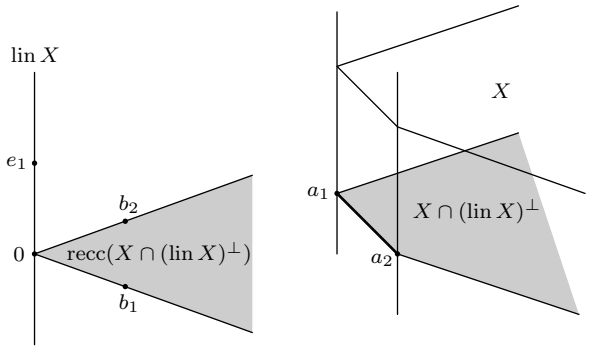
En polyeder är per definition ett snitt av ändligt många slutna halvrum, vilket kan uppfattas som en yttre beskrivning av polyedern. Vi ska nu ge en inre beskrivning av polyedern i termer av extremalpunkter och extremalstrålar. Följande struktursats är det här kapitlets huvudresultat.

Sats 5.3.1. *En icke-tom delmängd X av $\mathbf{R}^n$ är en polyeder om och endast om den kan skrivas på formen*

$$X = \operatorname{cvx} A + \operatorname{con} B,$$

där A och B är ändliga mängder i $\mathbf{R}^n$ och $A \neq \emptyset$.

Konen $\operatorname{con} B$ *är i så fall lika med X:s recessionskon* $\operatorname{recc} X$*. Om polyedern X är linjefri, kan man som A välja* $\operatorname{ext} X$ *och som B en mängd bestående av en vektor från varje extremalstråle i recessionskonen.*



Figur 5.1. Illustration till sats 5.3.1. Den högra delen av figuren föreställer ett utsnitt av en obegränsad polyeder X i $\mathbf{R}^3$. Det recessiva delrummet $\operatorname{lin} X$ är endimensionellt och genereras som kon av e_1 och -e_1. Snittet $X \cap (\operatorname{lin} X)^\perp$, som skuggmarkerats, är en linjefri polyeder med två extremalpunkter a_1 och a_2. Recessionskonen $\operatorname{recc}(X \cap (\operatorname{lin} X)^\perp)$ genereras av b_1 och b_2. Uppdelningen $X = \operatorname{cvx} A + \operatorname{con} B$ gäller för $A = \{a_1, a_2\}$ och $B = \{e_1, -e_1, b_1, b_2\}$.

Bevis. Antag att X är en polyeder, och sätt $Y = X \cap (\operatorname{lin} X)^\perp$. Då är Y en linjefri polyeder, och det följer av satserna 4.2.6 och 4.2.2 att

$$X = \operatorname{lin} X + Y = \operatorname{lin} X + \operatorname{recc} Y + \operatorname{cvx}(\operatorname{ext} Y).$$

De båda polyedriska konerna $\operatorname{lin} X$ och $\operatorname{recc} Y$ genereras enligt sats 5.2.1 av två ändliga mängder B_1 resp. B_2, och deras summa genereras av den ändliga mängden $B = B_1 \cup B_2$. Mängden $\operatorname{ext} Y$ är ändlig på grund av sats 5.1.2, så genom att sätta $A = \operatorname{ext} Y$ erhåller vi den sökta representationen

$$X = \operatorname{cvx} A + \operatorname{con} B.$$

Eftersom mängderna A och B är ändliga, är konen $\operatorname{con} B$ sluten och den konvexa mängden $\operatorname{cvx} A$ kompakt. Det följer därför av korollarium 2.7.13 att $\operatorname{con} B = \operatorname{recc} X$.

För linjefria polyedrar X gäller enligt satserna 4.2.2 och 4.2.4 att

$$X = \operatorname{cvx}(\operatorname{ext} X) + \operatorname{con}(\operatorname{exr}(\operatorname{recc} X)),$$

och vi får den sökta representationen genom att sätta $A = \operatorname{ext} X$ och låta mängden B bestå av en punkt från varje extremalstråle till $\operatorname{recc} X$.

Antag omvänt att $X = \operatorname{cvx} A + \operatorname{con} B$, där mängderna $A = \{a_1, \ldots, a_p\}$ och $B = \{b_1, \ldots, b_q\}$ är ändliga. Betrakta den kon Y i $\mathbf{R}^n \times \mathbf{R}$ som genereras av den ändliga mängden $(A \times \{1\}) \cup (B \times \{0\})$. Konen Y är enligt sats 5.2.1 en polyedrisk kon, vilket innebär att det finns en $m \times (n+1)$-matris C så att

$$(x, x_{n+1}) \in Y \;\Leftrightarrow\; C \begin{bmatrix} x \\ x_{n+1} \end{bmatrix} \geq 0. \tag{5.1}$$

(Här är förstås $\begin{bmatrix} x \\ x_{n+1} \end{bmatrix}$ vektorn $(x_1, \ldots, x_n, x_{n+1})$ skriven som kolonnmatris.)

Låt nu C' beteckna den delmatris av C som består av samtliga kolonner utom den sista, och låt c' beteckna den sista kolonnen i matrisen C; då är

$$C \begin{bmatrix} x \\ x_{n+1} \end{bmatrix} = C'x + x_{n+1}c',$$

vilket innebär att vi kan skriva om ekvivalensen (5.1) på formen

$$(x, x_{n+1}) \in Y \;\Leftrightarrow\; C'x + x_{n+1}c' \geq 0.$$

Av definitionen av Y följer emellertid att vektorn $(x, 1) \in \mathbf{R}^n \times \mathbf{R}$ ligger i Y om och endast om det finns icke-negativa tal $\lambda_1, \lambda_2, \ldots, \lambda_p$ och $\mu_1, \mu_2, \ldots, \mu_q$ så att

$$\begin{cases} x = \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_p a_p + \mu_1 b_1 + \mu_2 b_2 + \cdots + \mu_q b_q \\ 1 = \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{cases}$$

dvs. om och endast om $x \in \operatorname{cvx} A + \operatorname{con} B$. Detta ger oss ekvivalenserna

$$x \in X \;\Leftrightarrow\; (x, 1) \in Y \;\Leftrightarrow\; C'x + c' \geq 0,$$

som innebär att $X = \{x \in \mathbf{R}^n \mid C'x \geq -c'\}$. X är således en polyeder. □

5.4 Polyederbevarande operationer

Sats 5.4.1. *Snittet av ändligt många polyedrar i* $\mathbf{R}^n$ *är en polyeder.*

Bevis. Trivialt. □

Sats 5.4.2. *Antag att* X *är en polyeder i* $\mathbf{R}^n$ *och att* $T\colon \mathbf{R}^n \to \mathbf{R}^m$ *är en affin avbildning. Då är bildmängden* $T(X)$ *en polyeder i* $\mathbf{R}^m$*.*

Bevis. Påståendet är trivialt om polyedern är en tom, så antag att den inte är det och representera den på formen

$$X = \operatorname{cvx} A + \operatorname{con} B,$$

där $A = \{a_1, \ldots, a_p\}$ och $B = \{b_1, \ldots, b_q\}$ är ändliga mängder. Varje $x \in X$ kan då skrivas på formen

$$x = \sum_{j=1}^{p} \lambda_j a_j + \sum_{j=1}^{q} \mu_j b_j = \sum_{j=1}^{p} \lambda_j a_j + \sum_{j=1}^{q} \mu_j b_j - (\sum_{j=1}^{q} \mu_j)0$$

med icke-negativa koefficienter λ_j, μ_j och $\sum_{j=1}^{p} \lambda_j = 1$, dvs. som en affin kombination av element i $A \cup B \cup \{0\}$, och eftersom avbildningen T är affin, är

$$Tx = \sum_{j=1}^{p} \lambda_j Ta_j + \sum_{j=1}^{q} \mu_j Tb_j - (\sum_{j=1}^{q} \mu_j)T0 = \sum_{j=1}^{p} \lambda_j Ta_j + \sum_{j=1}^{q} \mu_j (Tb_j - T0).$$

Detta visar att bildmängden $T(X)$ har formen

$$T(X) = \operatorname{cvx} A' + \operatorname{con} B'$$

för $A' = T(A)$ och $B' = -T0 + T(B) = \{Tb_1 - T0, \ldots, Tb_q - T0\}$. Bildmängden $T(X)$ är därför en polyeder enligt sats 5.3.1. □

Sats 5.4.3. *Antag att* Y *är en polyeder i* $\mathbf{R}^m$ *och att* $T\colon \mathbf{R}^n \to \mathbf{R}^m$ *är en affin avbildning. Då är inversa bildmängden* $T^{-1}(Y)$ *en polyeder i* $\mathbf{R}^n$*.*

Bevis. Antag först att Y är ett slutet halvrum i $\mathbf{R}^m$ (eller hela rummet $\mathbf{R}^m$), dvs. att $Y = \{y \in \mathbf{R}^m \mid \langle c, y\rangle \geq b\}$. (Fallet $Y = \mathbf{R}^m$ fås för $c = 0$ och $b = 0$.) Den affina avbildningen T kan skrivas på formen $Tx = Sx + y_0$, där S är en linjär avbildning och y_0 är en vektor i $\mathbf{R}^m$. Med dessa beteckningar fås

$$T^{-1}(Y) = \{x \in \mathbf{R}^n \mid \langle c, Tx\rangle \geq b\} = \{x \in \mathbf{R}^n \mid \langle S^T c, x\rangle \geq b - \langle c, y_0\rangle\}.$$

Detta är ett slutet halvrum i $\mathbf{R}^n$ om $S^T c \neq 0$, hela rummet $\mathbf{R}^n$ om $S^T c = 0$ och $b \leq \langle c, y_0\rangle$, och tomma mängden $\emptyset$ om $S^T c = 0$ och $b > \langle c, y_0\rangle$.

I det allmänna fallet är $Y = \bigcap_{j=1}^{p} K_j$ ett snitt av ändligt många slutna halvrum. Eftersom $S^{-1}(Y) = \bigcap_{j=1}^{p} S^{-1}(K_j)$, är den inversa bilden $S^{-1}(Y)$ också ett snitt av slutna halvrum, tomma mängden eller hela rummet $\mathbf{R}^n$. $S^{-1}(Y)$ är alltså en polyeder. □

Sats 5.4.4. *Cartesianska produkten $X \times Y$ av två polyedrar X och Y är en polyeder.*

Bevis. Antag att X ligger i $\mathbf{R}^m$ och Y i $\mathbf{R}^n$. Mängden $X \times \mathbf{R}^n$ är en polyeder eftersom den är den inversa bilden av X under projektionen $(x, y) \mapsto x$, och av motsvarande skäl är $\mathbf{R}^m \times Y$ en polyeder. Det följer att $X \times Y$ är en polyeder, ty $X \times Y = (X \times \mathbf{R}^n) \cap (\mathbf{R}^m \times Y)$. □

Sats 5.4.5. *Summan $X + Y$ av två polyedrar i $\mathbf{R}^n$ är en polyeder.*

Bevis. Summan $X + Y$ är lika med bilden av $X \times Y$ under den linjära avbildningen $(x, y) \to x + y$, så satsens påståendet följer av föregående sats och sats 5.4.2. □

5.5 Separation

För polyedrar gäller skarpare separationsresultat än de allmänna resultaten för konvexa mängder i kapitel 3 − jämför följande två satser med sats 3.1.6 och sats 3.1.5.

Sats 5.5.1. *Om X och Y är disjunkta polyedrar, så finns det ett hyperplan som separerar polyedrarna strikt.*

Bevis. Differensen $X - Y$ av två polyedrar X och Y är en sluten mängd, ty den är en polyeder enligt sats 5.4.5. Det följer därför av sats 3.1.6 att det finns ett hyperplan som separerar polyedrarna strikt om de är disjunkta. □

Sats 5.5.2. *Låt X vara en konvex mängd, och låt Y vara en polyeder som är disjunkt från X. Då finns det ett hyperplan som separerar X och Y och som inte innehåller X som delmängd.*

Bevis. Vi visar satsen genom induktion över det omgivande rummets dimension n.

Påståendet i satsen är uppenbarligen sant då $n = 1$, dvs. för delmängder X och Y av $\mathbf{R}$. Så antag att det är sant då dimensionen är $n - 1$, och låt X vara en konvex delmängd av $\mathbf{R}^n$ som är disjunkt från polyedern Y i $\mathbf{R}^n$. Sats 3.1.5 ger oss ett hyperplan H som separerar X och Y och som då inte innehåller båda mängderna som delmängder. Om X inte är en delmängd

av H är saken klar. Antag därför att X är en delmängd av H; polyedern Y ligger då i ett av de båda slutna halvrum som definieras av hyperplanet H. Vi kallar detta slutna halvrum H_+ och låter H_{++} beteckna motsvarande öppna halvrum, och har alltså inklusionen $Y \subseteq H_+$.

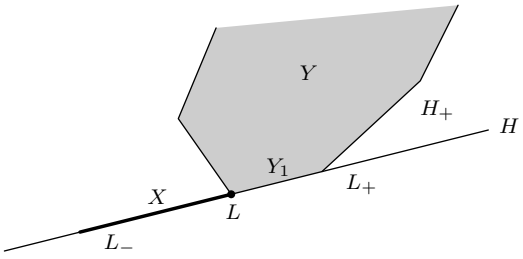
Om nu $Y \subseteq H_{++}$ så är polyedrarna Y och H disjunkta, och på grund av sats 5.5.1 finns det ett hyperplan som separerar Y och H strikt, och detta hyperplan separerar då förstås också Y och X strikt, eftersom X är en delmängd av H.

Detta bevisar satsen i fallet $Y \subseteq H_{++}$. Antag därför att Y är en delmängd av det slutna halvrummet H_+ utan att vara en delmängd av motsvarande öppna halvrum, dvs. betrakta fallet $Y \subseteq H_+$, $Y \cap H \neq \emptyset$. Den icke-tomma polyedern $Y_1 = Y \cap H$ och X kan då på grund av induktionsantagandet separeras i det $(n-1)$-dimensionella hyperplanet H med hjälp av en affin delmängd L av H av dimension $n-2$ och som inte innehåller X som delmängd.

Genom L definieras två slutna halvor L_+ och L_- av hyperplanet H med X som delmängd av L_- och Y_1 som delmängd av L_+. Den mot L_- svarande relativt öppna halvan betecknas L_{--}, dvs. $L_{--} = L_- \setminus L$. Antagandet att X inte är en delmängd av L innebär förstås att $X \cap L_{--} \neq \emptyset$.

Notera att $Y \cap L_- = Y_1 \cap L$. Om $Y_1 \cap L = \emptyset$ så finns det därför enligt sats 5.5.1 ett hyperplan som separerar polyedrarna Y och L_- strikt, och eftersom $X \subseteq L_-$, är beviset för påståendet i satsen klart i detta fall.

Det återstår nu endast att behandla fallet $Y_1 \cap L \neq \emptyset$, och efter en eventuell translation kan vi utan inskränkning anta att origo ligger i $Y_1 \cap L$, vilket medför att L är ett linjärt delrum. Se figur 5.2.



Figur 5.2. Illustration till beviset för sats 5.5.2.

Notera att mängden $H_{++} \cup L_+$ är en kon och att Y är en delmängd av denna kon. Betrakta nu den av polyedern Y genererade konen $\operatorname{con} Y$ och sätt

$$C = L + \operatorname{con} Y.$$

C är förstås också en kon och en delmängd av konen $H_{++} \cup L_+$ eftersom

både Y och L är delmängder av den sistnämnda konen. Konen $\operatorname{con} Y$ är polyedrisk, ty om polyedern Y har representationen $Y = \operatorname{cvx} A + \operatorname{con} B$ med ändliga mängder A och B, så är $\operatorname{con} Y = \operatorname{con}(A \cup B)$ beroende på att $0 \in Y$. Eftersom summan av två polyedriska koner är en polyedrisk kon, är C en polyedrisk kon.

Konen C är disjunkt från mängden L_{--} eftersom mängderna L_{--} och $H_{++} \cup L_{+}$ är disjunkta.

Skriv nu den polyedriska konen C som ett snitt $\bigcap K_i$ av ändligt många slutna halvrum K_i som begränsas av hyperplan H_i genom origo. Varje halvrum K_i är en kon som innehåller såväl Y som L. Om ett givet halvrum K_i också innehåller en punkt från L_{--}, så innehåller det också konen som genereras av punkten och L, dvs. hela L_{-}. Eftersom $C = \bigcap K_i$ och $C \cap L_{--} = \emptyset$, följer det därför att det finns ett halvrum K_i som inte innehåller någon punkt från L_{--}. Motsvarande randhyperplan H_i separerar med andra ord L_{-} och konen C och är disjunkt från L_{--}. Eftersom $X \subseteq L_{-}$, $Y \subseteq C$ och $X \cap L_{--} \neq \emptyset$, separerar H_i mängderna X och Y utan att innehålla X. Därmed är beviset klart. □

Övningar

5.1 Bestäm extremalpunkterna till följande polyedrar:

a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4,\ x \geq 0\}$
d) $X = \{x \in \mathbf{R}^4 \mid x_1 + x_2 + 3x_3 + x_4 \leq 4,\ 2x_2 + 3x_3 \geq 5,\ x \geq 0\}$.

5.2 Bestäm extremalstrålarna till konen

$$X = \{x \in \mathbf{R}^3 \mid x_1 - x_2 + 2x_3 \geq 0,\ x_1 + 2x_2 - 2x_3 \geq 0,\ x_2 + x_3 \geq 0,\ x_3 \geq 0\}.$$

5.3 Bestäm en matris C sådan att

$$\operatorname{con}\{(1,-1,1),(-1,0,1),(3,2,1),(-2,-1,0)\} = \{x \in \mathbf{R}^3 \mid Cx \geq 0\}.$$

5.4 Bestäm ändliga mängder A och B så att $X = \operatorname{con} A + \operatorname{cvx} B$ för följande polyedrar:

a) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \geq 2,\ x_2 \geq -1\}$
b) $X = \{x \in \mathbf{R}^2 \mid -x_1 + x_2 \leq 2,\ x_1 + 2x_2 \leq 2,\ x_2 \geq -1\}$
c) $X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \leq 4,\ x_1 + 2x_2 + x_3 \leq 4,\ x \geq 0\}$
d) $X = \{x \in \mathbf{R}^4 \mid x_1 + x_2 + 3x_3 + x_4 \leq 4,\ 2x_2 + 3x_3 \geq 5,\ x \geq 0\}$.

5.5 Visa att om 0 ligger i polyedern $X = \operatorname{cvx} A + \operatorname{con} B$, där mängderna A och B är ändliga, så är $\operatorname{con} X = \operatorname{con}(A \cup B)$.

Kapitel 6

Konvexa funktioner

6.1 Grundläggande definitioner

Epigraf och subnivåmängd

Definition. Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion med definitionsmängd $X \subseteq \mathbf{R}^n$ och målmängd $\overline{\mathbf{R}}$, dvs. de med ∞ utvidgade reella talen. Mängden

$$\operatorname{epi} f = \{(x,t) \in X \times \mathbf{R} \mid f(x) \leq t\}$$

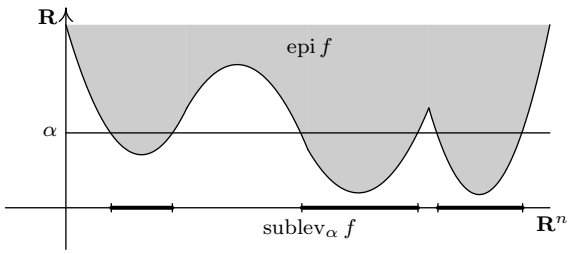
kallas funktionens *epigraf*.

Om α är ett reellt tal kallas mängden

$$\operatorname{sublev}_\alpha f = \{x \in X \mid f(x) \leq \alpha\}$$

en *subnivåmängd* till funktionen, närmare bestämt en *α-subnivåmängd*.

Epigrafen är en delmängd av $\mathbf{R}^{n+1}$; ”epi” betyder ”över” så epigraf betyder ”över grafen”.



Figur 6.1. Epigraf och subnivåmängd

Vi påminner om beteckningen dom f för f:s effektiva domän, dvs. mängden av alla punkter där funktionen $f\colon X \to \overline{\mathbf{R}}$ är ändlig. Uppenbarligen är

$$\operatorname{dom} f = \{x \in X \mid f(x) < \infty\}$$

lika med unionen av funktionens alla subnivåmängder, och dessa bildar en växande följd av mängder, dvs.

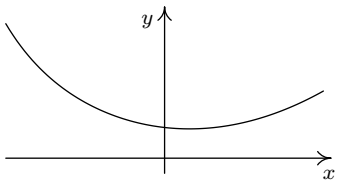
$$\operatorname{dom} f = \bigcup_{\alpha \in \mathbf{R}} \operatorname{sublev}_\alpha f \qquad \text{och} \qquad \alpha < \beta \;\Rightarrow\; \operatorname{sublev}_\alpha f \subseteq \operatorname{sublev}_\beta f.$$

Härav följer specillt att dom f är en konvex mängd om alla subnivåmängderna är konvexa.

Konvexa funktioner

Definition. En funktion $f\colon X \to \overline{\mathbf{R}}$ kallas *konvex* om dess definitionsmängd X och epigraf epi f är konvexa mängder.

En funktion $f\colon X \to \underline{\mathbf{R}}$ kallas *konkav* om funktionen $-f$ är konvex.



Figur 6.2. Grafen till en konvex funktion

Exempel 6.1.1. En affin funktions epigraf är ett slutet halvrum. Alla affina funktioner, och speciellt alla linjära funktioner, är därför både konvexa och konkava. □

Exempel 6.1.2. Exponentialfunktionen e^x med $\mathbf{R}$ som definitionsmängd är en konvex funktion.

Genom att ersätta x med $x - a$ i den elementära olikheten $\mathrm{e}^x \geq x + 1$ erhålls nämligen olikheten $\mathrm{e}^x \geq (x - a)\mathrm{e}^a + \mathrm{e}^a$, som innebär att exponentialfunktionens epigraf kan skrivas som ett snitt

$$\bigcap_{a \in \mathbf{R}} \{(x, y) \in \mathbf{R}^2 \mid y \geq (x - a)\mathrm{e}^a + \mathrm{e}^a\}$$

av slutna halvrum i $\mathbf{R}^2$. Epigrafen är därför konvex. □

Sats 6.1.1. *Om* $f\colon X \to \overline{\mathbf{R}}$ *är en konvex funktion, så är* $\operatorname{dom} f$ *och alla subnivåmängderna* $\operatorname{sublev}_\alpha f$ *konvexa mängder.*

Bevis. Antag att definitionsmängden X är en delmängd av $\mathbf{R}^n$ och betrakta projektionen $P\colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ av $\mathbf{R}^n \times \mathbf{R}$ på första faktorn, dvs. $P(x,t) = x$. Låt vidare K_α beteckna det slutna halvrummet $\{x \in \mathbf{R}^{n+1} \mid x_{n+1} \leq \alpha\}$. Då är $\operatorname{sublev}_\alpha f = P(\operatorname{epi} f \cap K_\alpha)$, ty

$$\begin{aligned} f(x) \leq \alpha &\Leftrightarrow \exists t\colon f(x) \leq t \leq \alpha \Leftrightarrow \exists t\colon (x,t) \in \operatorname{epi} f \cap K_\alpha \\ &\Leftrightarrow x \in P(\operatorname{epi} f \cap K_\alpha). \end{aligned}$$

Snitten $\operatorname{epi} f \cap K_\alpha$ är konvexa mängder, och eftersom konvexitet bevaras under linjära avbildningar, följer det att subnivåmängderna $\operatorname{sublev}_\alpha f$ är konvexa. Deras union $\operatorname{dom} f$ är följaktligen också konvex. □

Kvasikonvexa funktioner

Många viktiga egenskaper hos konvexa funktioner visar sig vara konsekvenser av att subnivåmängderna är konvexa. Detta är en anledning till att studera funktioner med konvexa subnivåmängder och motiverar således följande definition.

Definition. En funktion $f\colon X \to \overline{\mathbf{R}}$ kallas *kvasikonvex* om funktionens definitionsmängd X och samtliga subnivåmängder $\operatorname{sublev}_\alpha f$ är konvexa.

En funktion $f\colon X \to \underline{\mathbf{R}}$ kallas *kvasikonkav* om $-f$ är kvasikonvex.

Konvexa funktioner är kvasikonvexa eftersom deras subnivåmängder är konvexa. Omvändningen gäller inte, ty en funktion f, som är definierad på ett delintervall I av $\mathbf{R}$, är kvasikonvex om den är växande på I, eller om den är avtagande på I, eller mer generellt om det finns en punkt $c \in I$ så att f är avtagande till vänster om c och växande till höger om c. Naturligtvis finns det sådana funktioner som inte är konvexa.

Konvexa utvidgningar

Om funktionen $f\colon X \to \overline{\mathbf{R}}$ är konvex (resp. kvasikonvex), så är funktionens effektiva domän $\operatorname{dom} f$ konvex, och eftersom

$$\begin{aligned} \operatorname{epi} f &= \{(x,t) \in \operatorname{dom} f \times \mathbf{R} \mid f(x) \leq t\} \qquad \text{och} \\ \operatorname{sublev}_\alpha f &= \{x \in \operatorname{dom} f \mid f(x) \leq \alpha\}, \end{aligned}$$

är funktionens restriktion $f|_{\operatorname{dom} f}$ till $\operatorname{dom} f$ en konvex (resp. kvasikonvex) funktion med samma epigraf och α-nivåmängder som f har. Det kan därför

förefalla som att man inte tjänar någonting på att tillåta ∞ som funktionsvärde hos konvexa (resp. kvasikonvexa) funktioner.

Men även om vi förstås primärt är intresserade av funktioner med ändliga värden, så uppstår konvexa funktioner med oändliga värden på ett naturligt sätt då man bildar suprema eller gränsvärden till följder av konvexa funktioner med ändliga värden.

En annan finess med att tillåta ∞ som funktionsvärde hos (kvasi)konvexa funktioner är att vi utan inskränkning kan anta att de är definierade på hela $\mathbf{R}^n$. Om $f\colon X \to \overline{\mathbf{R}}$ är en (kvasi)konvex funktion med en äkta delmängd X av $\mathbf{R}^n$ som definitionsmängd, och vi definierar $\tilde{f}\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ genom att sätta

$$\tilde{f}(x) = \begin{cases} f(x) & \text{om } x \in X \\ \infty & \text{om } x \notin X, \end{cases}$$

så har funktionerna f och $\tilde{f}$ samma epigrafer och samma α-subnivåmängder. Utvidgningen $\tilde{f}$ är därför också (kvasi)konvex. Vidare är förstås $\operatorname{dom} \tilde{f} = \operatorname{dom} f$.

(Kvasi)konkava funktioner har en analog utvidgning till funktioner som tar värden i $\underline{\mathbf{R}} = \mathbf{R} \cup \{-\infty\}$.

Vi kommer något oegentligt säga att de konvexa funktionerna $f\colon X \to \overline{\mathbf{R}}$ och $g\colon Y \to \overline{\mathbf{R}}$ är **lika** och skriva $f = g$, om $\operatorname{dom} f = \operatorname{dom} g$ och $f(x) = g(x)$ för alla $x \in \operatorname{dom} f$.

Alternativ karakterisering av konvexitet

Sats 6.1.2. *En funktion $f\colon X \to \overline{\mathbf{R}}$ med konvex definitionsmängd X är*
(a) konvex om och endast om

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{6.1}$$

för alla $x, y \in X$ och alla $\lambda \in]0,1[$;
(b) kvasikonvex om och endast om

$$f(\lambda x + (1-\lambda)y) \le \max\{f(x), f(y)\} \tag{6.2}$$

för alla $x, y \in X$ och alla $\lambda \in]0,1[$.

Bevis. (a) Antag att funktionen f är konvex, dvs. att epigrafen $\operatorname{epi} f$ är konvex, och låt x och y vara två punkter i $\operatorname{dom} f$. Då ligger punkterna $(x, f(x))$ och $(y, f(y))$ i epigrafen, så konvexiteten hos epigrafen medför att den konvexa kombinationen $\big(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)\big)$ av dessa två punkter

också ligger i epigrafen för $0 < \lambda < 1$. Detta är ekvivalent med att olikhet (6.1) gäller. Om någon av punkterna $x, y \in X$ ligger utanför $\operatorname{dom} f$, så är olikheten trivialt uppfylld eftersom högerledet i så fall är lika med ∞.

För att visa omvändningen antar vi att olikheten (6.1) gäller. Låt (x, s) och (y, t) vara två punkter i epigrafen. Då är $f(x) \le s$ och $f(y) \le t$, och för $0 < \lambda < 1$ följer det därför av olikheten (6.1) att

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) \le \lambda s + (1-\lambda)t,$$

vilket betyder att punkten $(\lambda x + (1-\lambda)y, \lambda s + (1-\lambda)t)$, det vill säga punkten $\lambda(x, s) + (1-\lambda)(y, t)$, ligger i epigrafen, som således är konvex

(b) visas helt analogt. □

En funktion f är uppenbarligen (kvasi)konvex om och endast om funktionens restriktion till varje linje som skär funktionens definitionsmängd X är (kvasi)konvex. Varje sådan linjes ekvation har formen $x = x_0 + tv$, där x_0 är en punkt i X och v är en vektor i $\mathbf{R}^n$, och motsvarande restriktion blir därför en envariabelfunktion $g(t) = f(x_0 + tv)$ (med definitionsmängd $\{t \mid x_0 + tv \in X\}$). Problemet att avgöra om en funktion är (kvasi)konvex kan därför reduceras till att avgöra motsvarande konvexitet för envariabelfunktioner.

Definition. En funktion $f\colon X \to \overline{\mathbf{R}}$, som är definierad på en konvex kon X, kallas

- *subadditiv* om $f(x + y) \le f(x) + f(y)$ för alla $x, y \in X$;
- *positivt homogen* om $f(\alpha x) = \alpha f(x)$ för alla $x \in X$ och alla $\alpha \in \mathbf{R}_+$.

Varje positivt homogen, subadditiv funktion är uppenbarligen konvex. Omvänt är varje konvex, positivt homogen funktion f subadditiv, ty

$$f(x+y) = 2f(\tfrac{1}{2}x + \tfrac{1}{2}y) \le 2(\tfrac{1}{2}f(x) + \tfrac{1}{2}f(y)) = f(x) + f(y).$$

En subadditiv, positivt homogen funktion $f\colon \mathbf{R}^n \to \mathbf{R}$ som dessutom är *symmetrisk*, dvs. uppfyller villkoret

$$f(-x) = f(x) \quad \text{för alla } x \in \mathbf{R}^n,$$

kallas en *seminorm* på $\mathbf{R}^n$.

Symmetri- och homogenitetsvillkoren kan förstås sammanfattas i kravet att $f(\alpha x) = |\alpha| f(x)$ för alla $x \in \mathbf{R}^n$ och alla $\alpha \in \mathbf{R}$.

Om f är en seminorm, så är $f(x) \ge 0$ för alla x; detta följer av räkningen

$$0 = f(0) = f(x - x) \le f(x) + f(-x) = 2f(x).$$

En seminorm f kallas en *norm* om $f(x) = 0$ endast för $x = 0$. För normer används vanligtvis beteckningen $\|\cdot\|$.

Seminormer, och speciellt normer, är förstås konvexa funktioner.

EXEMPEL 6.1.3. Den euklidiska normen och ℓ^1-normen, som definierades i kapitel 1, är specialfall av de s. k. ℓ^p-normerna $\|\cdot\|_p$ på $\mathbf{R}^n$ som för $1 \le p < \infty$ definieras av att

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}.$$

Maxnormen

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|$$

fås vidare som som gränsvärde av $\|x\|_p$ då p går mot oändligheten.

Att ℓ^p-normerna är positivt homogena, symmetriska och lika med 0 endast för $x = 0$ är uppenbart. I fallet $p = 1$ och $p = \infty$ följer subadditiviteten lätt av triangelolikheten $|x + y| \le |x| + |y|$ för reella tal, och i det euklidiska fallet $p = 2$ är subadditiviteten en konsekvens av Cauchy–Schwarz olikhet. I avsnitt 6.4 skall vi visa subadditiviteten för övriga p-värden (sats 6.4.3). □

Strikt konvexitet

Genom att skärpa kravet på olikheterna i den alternativa karakteriseringen av konvexitet får vi följande definitioner.

Definition. En konvex funktion $f\colon X \to \overline{\mathbf{R}}$ kallas *strikt konvex* om

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

för alla par av skilda punkter x, $y \in X$ och alla $\lambda \in]0,1[$.

En kvasikonvex funktion f kallas *strikt kvasikonvex* om olikheten (6.2) i sats 6.1.2 är strikt för alla par av skilda punkter x, $y \in X$ och alla $\lambda \in]0,1[$.

En funktion f kallas *strikt konkav* (resp. *strikt kvasikonkav*) om funktionen $-f$ är strikt konvex (resp. strikt kvasikonvex).

EXEMPEL 6.1.4. En kvadratisk form $q(x) = \langle x, Qx\rangle = \sum_{i,j=1}^{n} q_{ij}x_ix_j$ på $\mathbf{R}^n$ är konvex om och endast om den är positivt semidefinit, och strikt konvex om och endast om den är positivt definit. Detta följer ur identiteten

$$\big(\lambda x_i+(1-\lambda)y_i\big)\big(\lambda x_j+(1-\lambda)y_j\big) = \lambda x_ix_j+(1-\lambda)y_iy_j-\lambda(1-\lambda)(x_i-y_i)(x_j-y_j)$$

som efter multiplikation med q_{ij} och summering ger

$$q(\lambda x + (1-\lambda)y) = \lambda q(x) + (1-\lambda)q(y) - \lambda(1-\lambda)q(x-y).$$

Högerledet är $\leq \lambda q(x) + (1-\lambda)q(y)$ för alla $0 < \lambda < 1$ om och endast om $q(x-y) \geq 0$, vilket gäller för alla $x \neq y$ om och endast om q är positivt semidefinit. För strikt olikhet krävs förstås att q skall vara positivt definit.□

Jensens olikhet

Olikheterna (6.1) och (6.2) utvidgas lätt till att handla om konvexa kombinationer av fler än två punkter.

Sats 6.1.3. *Antag att $x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_m x_m$ är en konvex kombination av punkterna $x_1, x_2, \ldots, x_m$ i definitionsmängden till funktionen f.*
(a) Om funktionen f är konvex, så är

$$f(x) \leq \sum_{j=1}^{m} \lambda_j f(x_j). \qquad \textbf{(Jensens olikhet)} \tag{6.3}$$

Om f är strikt konvex och det råder likhet i olikheten (6.3) *samt $\lambda_j > 0$ för alla j, så är $x_1 = x_2 = \cdots = x_m$.*
(b) Om f är en kvasikonvex funktion, så är

$$f(x) \leq \max_{1 \leq j \leq m} f(x_j). \tag{6.4}$$

Om f är strikt kvasikonvex, det råder likhet i olikheten (6.4) *och $\lambda_j > 0$ för alla j, så är $x_1 = x_2 = \cdots = x_m$.*

Bevis. (a) För att visa Jensens olikhet kan vi utan inskränkning anta att alla koefficienterna λ_j är positiva samt att alla punkterna x_j ligger i $\operatorname{dom} f$, ty högerledet i olikheten är oändligt om någon punkt x_j ligger utanför $\operatorname{dom} f$. Då är

$$\big(x, \sum_{j=1}^{m} \lambda_j f(x_j)\big) = \sum_{j=1}^{m} \lambda_j \big(x_j, f(x_j)\big),$$

och summan i högerledet är en konvex kombination av element i epigrafen $\operatorname{epi} f$ och ligger därför själv i epigrafen. Vänsterledet tillhör därför epigrafen, vilket medför att olikheten (6.3) gäller.

Anta nu att f är strikt konvex samt att det råder likhet i Jensens olikhet för den konvexa kombinationen $x = \sum_{j=1}^{m} \lambda_j x_j$, där alla koefficienter λ_j är positiva och $m \geq 2$. Sätt $y = \sum_{j=2}^{m} \lambda_j (1-\lambda_1)^{-1} x_j$; då är $x = \lambda_1 x_1 + (1-\lambda_1) y$,

och y är en konvex kombination av x_2, x_3, …, x_m, så Jensens olikhet ger

$$\sum_{j=1}^{m} \lambda_j f(x_j) = f(x) \le \lambda_1 f(x_1) + (1-\lambda_1) f(y) \\ \le \lambda_1 f(x_1) + (1-\lambda_1) \sum_{j=2}^{m} \lambda_j (1-\lambda_1)^{-1} f(x_j) = \sum_{j=1}^{m} \lambda_j f(x_j).$$

Eftersom ytterleden är lika, råder det likhet överallt i ovanstående kedja av olikheter. Det följer därför speciellt att $f(x) = \lambda_1 f(x_1) + (1-\lambda_1) f(y)$, och eftersom f är strikt konvex medför detta att $x_1 = y = x$.

Av symmetriskäl följer nu att $x_2 = x$, …, $x_m = x$, vilket bevisar påståendet i satsen.

(b) Antag att f är kvasikonvex, och sätt $\alpha = \max_{1 \le j \le m} f(x_j)$. Om någon av punkterna x_j ligger utanför $\operatorname{dom} f$ finns det inget att visa. I motsatt fall är α ändligt och varje punkt x_j tillhör den konvexa subnivåmängden $\operatorname{sublev}_\alpha f$, så det följer att också punkten x gör det, vilket ger olikheten (6.4). Påståendet om likhet för strikt kvasikonvexa funktioner visas på liknande sätt som motsvarande påstående för strikt konvexa funktioner. □

6.2 Konvexitetsbevarande operationer

Vi skall i det här avsnittet beskriva några sätt att konstruera nya konvexa funktioner av givna konvexa funktioner.

Konisk kombination

Sats 6.2.1. *Antag att funktionerna $f\colon X \to \overline{\mathbf{R}}$ och $g\colon X \to \overline{\mathbf{R}}$ är konvexa och att α och β är icke-negativa reella tal. Då är funktionen $\alpha f + \beta g$ också konvex.*

Bevis. Följer direkt av karakteriseringen av konvexitet i sats 6.1.2. □

De konvexa funktionerna på en given mängd X bildar med andra ord en konvex kon, så det följer att varje konisk kombination $\alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_m f_m$ av konvexa funktioner är konvex.

En summa av kvasikonvexa funktioner behöver däremot inte vara kvasikonvex.

Punktvist gränsvärde

Sats 6.2.2. *Om funktionerna $f_i\colon X \to \overline{\mathbf{R}}$ är konvexa för $i = 1, 2, 3, \ldots$ och gränsvärdet*

$$f(x) = \lim_{i\to\infty} f_i(x)$$

existerar som ändligt värde eller ∞ för alla $x \in X$, så är gränsfunktionen $f\colon X \to \overline{\mathbf{R}}$ också konvex.

Bevis. Antag att x och y är två punkter i X och att $0 < \lambda < 1$. Gränsövergång i olikheten $f_i(\lambda x + (1-\lambda)y) \le \lambda f_i(x) + (1-\lambda)f_i(y)$ ger att

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y),$$

vilket visar att gränsfunktionen f är konvex. □

Med hjälp av sats 6.2.2 utvidgas således omedelbart resultatet i sats 6.2.1 till att gälla för oändliga summor och integraler. Exempelvis är en punktvis konvergent oändlig summa $f(x) = \sum_{i=1}^{\infty} f_i(x)$ av konvexa funktioner konvex.

Och om funktionen $f(x, y)$ är konvex i variabeln x för varje y i någon mängd Y, och α är en icke-negativ funktion som är definierad på mängden Y, så är funktionen g som definieras av att $g(x) = \int_Y \alpha(y) f(x, y)\, dy$ konvex i variabeln x (förutsatt att integralen existerar). Det sistnämnda påståendet följer med hjälp av sats 6.2.2 genom att skriva integralen som ett gränsvärde av Riemannsummor, men visas naturligtvis också enkelt direkt genom integrering av olikheten som karakteriserar konvexiteten hos funktionerna $f(\cdot, y)$.

Sammansättning med affina avbildningar

Sats 6.2.3. *Antag att $A\colon V \to \mathbf{R}^n$ är en affin funktion, att Y är en konvex delmängd av $\mathbf{R}^n$ och att $f\colon Y \to \overline{\mathbf{R}}$ är en konvex funktion. Då är sammansättningen $f \circ A$ konvex på sin definitionsmängd $A^{-1}(Y)$.*

Bevis. Sätt $g = f \circ A$. För $x_1, x_2 \in A^{-1}(Y)$ och $0 < \lambda < 1$ är

$$\begin{aligned} g(\lambda x_1 + (1-\lambda)x_2) &= f(\lambda A x_1 + (1-\lambda)A x_2) \le \lambda f(Ax_1) + (1-\lambda) f(Ax_2) \\ &= \lambda g(x_1) + (1-\lambda) g(x_2), \end{aligned}$$

vilket visar att funktionen g är konvex. □

Sammansättningen $f \circ A$ av en kvasikonvex funktion f och en affin funktion A är på motsvarande vis kvasikonvex.

EXEMPEL 6.2.1. Funktionen $x \mapsto e^{c_1 x_1 + \cdots + c_n x_n}$ är konvex på $\mathbf{R}^n$ eftersom den är sammansatt av en linjär form och den konvexa exponentialfunktionen. □

Punktvist supremum

Sats 6.2.4. *Låt $f_i\colon X \to \overline{\mathbf{R}}$, $i \in I$, vara en familj av funktioner och definiera funktionen $f\colon X \to \overline{\mathbf{R}}$ genom att för $x \in X$ sätta*

$$f(x) = \sup_{i\in I} f_i(x).$$

Då är funktionen f

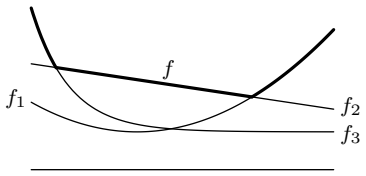
(i) konvex om samtliga funktioner f_i är konvexa;

(ii) kvasikonvex om samtliga funktioner f_i är kvasikonvexa.

Bevis. Enligt definitionen av supremum är $f(x) \le t$ om och endast om $f_i(x) \le t$ för alla $i \in I$, och detta innebär att

$$\operatorname{epi} f = \bigcap_{i\in I} \operatorname{epi} f_i \quad \text{och} \quad \operatorname{sublev}_t f = \bigcap_{i\in I} \operatorname{sublev}_t f_i$$

för alla $t \in \mathbf{R}$. Eftersom snittet av konvexa mängder är konvext, följer nu de båda påståendena i satsen direkt. □



Figur 6.3. $f = \sup f_i$ för en familj av tre funktioner.

Exempel 6.2.2. Ett punktvist maximum av ändligt många affina funktioner, dvs. en funktion av typen

$$f(x) = \max_{1\le i\le m} (\langle c_i, x\rangle + a_i),$$

är en konvex funktion som kallas en konvex *styckvis affin* funktion. □

Exempel 6.2.3. Exempel på styckvis affina konvexa funktioner f på $\mathbf{R}^n$ är:

(a) Beloppet av en vektors i:te koordinat

$$f(x) = |x_i| = \max\{x_i, -x_i\}.$$

(b) Maxnormen

$$f(x) = \|x\|_\infty = \max_{1\le i\le n} |x_i|.$$

(c) Summan av en vektors m största koordinater

$$f(x) = \max\{x_{i_1} + \cdots + x_{i_m} \mid 1 \le i_1 < i_2 < \cdots < i_m \le n\}.$$ □

Sammansättning

Sats 6.2.5. *Antag att funktionen $\phi\colon I \to \overline{\mathbf{R}}$ är definierad på ett reellt intervall I som innehåller värdemängden $f(X)$ till funktionen $f\colon X \to \mathbf{R}$. Sammansättningen $\phi \circ f\colon X \to \overline{\mathbf{R}}$ är konvex*

(i) om f är konvex och ϕ är konvex och växande;

(ii) om f är konkav och ϕ är konvex och avtagande.

Bevis. För $x, y \in X$ och $0 < \lambda < 1$ gäller olikheten

$$\phi\big(f(\lambda x + (1-\lambda)y)\big) \le \phi\big(\lambda f(x) + (1-\lambda)f(y)\big),$$

både om f är konvex och ϕ är växande och om f är konkav och ϕ är avtagande. För konvexa funktioner ϕ är vidare

$$\phi\big(\lambda f(x) + (1-\lambda)f(y)\big) \le \lambda\phi(f(x)) + (1-\lambda)\phi(f(y)),$$

och genom att kombinera de båda olikheterna erhåller vi den sökta slutsatsen. □

Motsvarigheten till sats 6.2.5 för kvasikonvexitet är att sammansättningen $\phi \circ f$ är kvasikonvex om antingen f är kvasikonvex och ϕ är växande, eller f är kvasikonkav och ϕ är avtagande.

Exempel 6.2.4. Exponentialfunktionen är konvex och växande, och positivt semidefinita kvadratiska former är konvexa. Följaktligen är funktionen

$$x \mapsto \mathrm{e}^{x_1^2 + x_2^2 + \cdots + x_k^2},$$

där $1 \le k \le n$, konvex på $\mathbf{R}^n$. □

Exempel 6.2.5. Funktionerna $t \mapsto 1/t$ och $t \mapsto -\ln t$ är konvexa och avtagande på intervallet $]0, \infty[$. För konkava positiva funktioner g är följaktligen funktionen $1/g$ konvex och funktionen $\ln g$ konkav. □

Infimum

Sats 6.2.6. *Låt C vara en konvex delmängd av $\mathbf{R}^{n+1}$, och definiera en funktion g på $\mathbf{R}^n$ genom att för $x \in \mathbf{R}^n$ sätta*

$$g(x) = \inf\{t \in \mathbf{R} \mid (x,t) \in C\},$$

med den sedvanliga konventionen $\inf \emptyset = +\infty$. Antag att det finns en punkt x_0 i det relativa inre av mängden

$$X_0 = \{x \in \mathbf{R}^n \mid g(x) < \infty\} = \{x \in \mathbf{R}^n \mid \exists t \in \mathbf{R}\colon (x,t) \in C\}$$

med ändligt funktionsvärde $g(x_0)$. Då är $g(x) > -\infty$ för alla $x \in \mathbf{R}^n$, och $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ är en konvex funktion med X_0 som effektiv domän.

Bevis. Låt x vara en godtycklig punkt i X_0. För att visa att $g(x) > -\infty$, dvs. att mängden

$$T_x = \{t \in \mathbf{R} \mid (x,t) \in C\}$$

är nedåt begränsad, väljer vi först en punkt $x_1 \in \operatorname{rint} X_0$ sådan att x_0 ligger på den öppna sträckan $]x, x_1[$, och skriver x_0 på formen $x_0 = \lambda x + (1-\lambda)x_1$ med $0 < \lambda < 1$. Vi fixerar sedan ett reellt tal t_1 sådant att $(x_1, t_1) \in C$, och definierar för $t \in T_x$ talet t_0 genom att sätta $t_0 = \lambda t + (1-\lambda)t_1$. Då blir (x_0, t_0) är en konvex kombination av punkterna (x,t) och (x_1,t_1) i C, så konvexitet och g:s definition ger oss olikheten

$$g(x_0) \le t_0 = \lambda t + (1-\lambda)t_1,$$

varav följer att

$$t \ge \frac{1}{\lambda}\big(g(x_0) - (1-\lambda)t_1\big),$$

vilket visar att mängden T_x är nedåt begränsad.

Funktionen g har således $\overline{\mathbf{R}}$ som målmängd och $\operatorname{dom} g = X_0$. Låt nu x_1 och x_2 vara godtyckliga punkter i X_0, och låt λ_1 och λ_2 vara två positiva tal med summa 1. Då finns det för varje $\epsilon > 0$ två reella tal t_1 och t_2 sådana att de båda punkterna (x_1,t_1) och (x_2,t_2) ligger i C och $t_1 < g(x_1) + \epsilon$ och $t_2 < g(x_2) + \epsilon$. Den konvexa kombinationen $(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 t_1 + \lambda_2 t_2)$ av de båda punkterna ligger också i C, och

$$g(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 t_1 + \lambda_2 t_2 \le \lambda_1 g(x_1) + \lambda_2 g(x_2) + \epsilon.$$

Detta betyder att punkten $\lambda_1 x_1 + \lambda_2 x_2$ ligger i X_0, och genom att låta ϵ gå mot 0 erhåller vi olikheten $g(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 g(x_1) + \lambda_2 g(x_2)$. Mängden X_0 är således konvex, och funktionen g är konvex. □

Vi har sett att punktvisa supremum $f(x) = \sup_{i\in I} f_i(x)$ av en godtycklig familj av konvexa funktioner är konvex. Om f är en funktion $X \times Y \to \overline{\mathbf{R}}$ och funktionen $f(\cdot, y)$ är en konvex funktion på X för varje fixt $y \in Y$, så är med andra ord $g(x) = \sup_{y\in Y} f(x,y)$ en konvex funktion på X, och detta oberoende av vad Y är för mängd. Följande korollarium till föregående sats visar att motsvarande infimum är en konvex funktion under förutsättning att f är konvex som funktion på produktmängden $X \times Y$.

Korollarium 6.2.7. *Antag att $f\colon X \times Y \to \mathbf{R}$ är en konvex funktion, och sätt för $x \in X$*

$$g(x) = \inf_{y\in Y} f(x,y).$$

Om $g(x_0) > -\infty$ för någon punkt $x_0 \in \operatorname{rint} X$, så är $g(x) > -\infty$ för alla $x \in X$, och $g\colon X \to \mathbf{R}$ är en konvex funktion.

Bevis. Antag att X är en konvex delmängd av $\mathbf{R}^n$ och sätt

$$C = \{(x,t) \in X \times \mathbf{R} \mid \exists y \in Y\colon f(x,y) \le t\}.$$

Då är C en konvex delmängd av $\mathbf{R}^{n+1}$, ty om (x_1,t_1) och (x_2,t_2) är två punkter i C, finns det två punkter y_1 och y_2 i den konvexa mängden Y sådana att $f(x_i,y_i) \le t_i$ för $i = 1,2$, och för $\lambda_1, \lambda_2 \in]0,1[$ med summa 1 är då

$$f(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 y_1 + \lambda_2 y_2) \le \lambda_1 f(x_1,y_1) + \lambda_2 f(x_2,y_2) \le \lambda_1 t_1 + \lambda_2 t_2,$$

vilket visar att den konvexa kombinationen $\lambda_1(x_1,t_1) + \lambda_2(x_2,t_2)$ ligger i C. Vidare är $g(x) = \inf\{t \in \mathbf{R} \mid (x,t) \in C\}$. Korollariet följer därför omedelbart av sats 6.2.6. □

Perspektiv

Definition. Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion definierad på en kon X i $\mathbf{R}^n$. Funktionen $g\colon X \times \mathbf{R}_{++} \to \overline{\mathbf{R}}$ som definieras av att

$$g(x,s) = sf(x/s)$$

kallas f:s *perspektiv*.

Sats 6.2.8. *Om $f\colon X \to \overline{\mathbf{R}}$ är en konvex funktion med en konvex kon som definitionsmängd, så är f:s perspektivfunktion g också konvex.*

Bevis. Låt (x,s) och (y,t) vara två punkter i $X \times \mathbf{R}_{++}$, och låt α och β vara två positiva tal med summa 1. Då är

$$\begin{aligned} g\big(\alpha(x,s) + \beta(y,t)\big) &= g(\alpha x + \beta y, \alpha s + \beta t) = (\alpha s + \beta t) f\Big(\frac{\alpha x + \beta y}{\alpha s + \beta t}\Big) \\ &= (\alpha s + \beta t) f\Big(\frac{\alpha s}{\alpha s + \beta t} \cdot \frac{x}{s} + \frac{\beta t}{\alpha s + \beta t} \cdot \frac{y}{t}\Big) \\ &\le \alpha s f\Big(\frac{x}{s}\Big) + \beta t f\Big(\frac{y}{t}\Big) = \alpha g(x,s) + \beta g(y,t). \end{aligned}$$ □

Exempel 6.2.6. Om $q(x)$ är en positivt semidefinit kvadratisk form på $\mathbf{R}^{n-1}$, så är alltså funktionen $f(x) = x_n q(x/x_n)$ konvex på $\mathbf{R}^{n-1} \times \mathbf{R}_{++}$. Genom att som kvadratisk form välja den euklidiska normen ser vi att funktionen

$$x \mapsto (x_1^2 + x_2^2 + \cdots + x_{n-1}^2)/x_n$$

är konvex i det öppna halvrummet $\mathbf{R}^{n-1} \times \mathbf{R}_{++}$. □

6.3 Maximum och minimum

Minimipunkter

Att för en godtycklig funktion avgöra om en punkt är en global minimipunkt är ett omöjligt problem − däremot finns det bra numeriska metoder för att hitta lokala minimipunkter om funktionen uppfyller lämpliga regularitetsvillkor. Det är bl. a. därför som konvexitet spelar så stor roll inom optimering. Ett lokalt minimum till en konvex funktion är nämligen automatiskt globalt.

Vi erinrar om att en punkt $x_0 \in X$ är en *lokal minimipunkt* till funktionen $f\colon X \to \overline{\mathbf{R}}$ om det finns en öppen boll $B = B(x_0; r)$ med centrum i x_0 sådan att $f(x) \geq f(x_0)$ för alla $x \in X \cap B$. Punkten är en *(global) minimipunkt* om $f(x) \geq f(x_0)$ för alla $x \in X$.

Sats 6.3.1. *Antag att funktionen $f\colon X \to \overline{\mathbf{R}}$ är konvex och att $x_0 \in \operatorname{dom} f$ är en lokal minimipunkt till f. Då är x_0 en global minimipunkt. Om dessutom f är strikt konvex, så är minimipunkten unik.*

Bevis. Låt $x \in X$ vara en godtycklig punkt skild från x_0.

Eftersom $\lambda x + (1-\lambda)x_0 \to x_0$ då $\lambda \to 0$ och funktionen f är konvex, gäller följande olikheter för $\lambda > 0$ tillräckligt nära 0:

$$f(x_0) \leq f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda) f(x_0)$$

(med strikt olikhet på sista stället om f är strikt konvex). Härav följer nu omedelbart att $f(x) \geq f(x_0)$ (resp. $f(x) > f(x_0)$), vilket visar att x_0 är en global minimipunkt (och att minimum inte antas i några andra punkter än x_0 om konvexiteten är strikt). □

Sats 6.3.2. *För kvasikonvexa funktioner $f\colon X \to \overline{\mathbf{R}}$ är mängden av minimipunkter konvex.*

Bevis. Påståendet är trivialt sant för funktioner som saknar minimipunkt, eftersom den tomma mängden är konvex, och för funktionen som är identiskt lika med ∞ på X. Så antag att f har en minimipunkt $x_0 \in \operatorname{dom} f$ och sätt $m = f(x_0)$. Mängden av minimipunkter är då lika med den konvexa subnivåmängden $\operatorname{sublev}_m f = \{x \in X \mid f(x) \leq m\}$. □

Maximipunkter

Sats 6.3.3. *Antag att $X = \operatorname{cvx} A$. För kvasikonvexa funktioner $f\colon X \to \overline{\mathbf{R}}$ är*

$$\sup_{x \in X} f(x) = \sup_{a \in A} f(a).$$

Om funktionen har något maximum, så finns det en maximipunkt i A.

Bevis. Varje punkt $x \in X$ är en konvex kombination $x = \sum_{j=1}^{m} \lambda_j a_j$ av element $a_j \in A$, varför

$$f(x) = f(\sum_{j=1}^{m} \lambda_j a_j) \leq \max_{1 \leq j \leq m} f(a_j) \leq \sup_{a \in A} f(a).$$

Härav följer att

$$\sup_{x \in X} f(x) \leq \sup_{a \in A} f(a),$$

och eftersom A är en delmängd av X, måste likhet råda.

Om x är en maximipunkt, så följer det vidare av olikheten ovan att $f(x) = \max_{1 \leq j \leq m} f(a_j)$, vilket betyder att maximum måste antas i någon av punkterna $a_j \in A$. □

Om A är en ändlig mängd och $X = \operatorname{cvx} A$, behöver vi således bara jämföra ändligt många funktionsvärden för att bestämma en kvasikonvex funktions maximivärde. För stora mängder A kan detta förstås dock vara praktiskt omöjligt.

Vi påminner om att en kompakt konvex mängd sammanfaller med konvexa höljet av mängdens extremalpunkter. Vi får därför följande korollarium till satsen ovan.

Korollarium 6.3.4. *Om en kvasikonvex funktion med kompakt definitionsmängd har ett maximum, så antas maximum i någon av definitionsmängdens extremalpunkter.*

Exempel 6.3.1. Den kvadratiska formen $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 2x_2^2$ är strikt konvex eftersom den är positivt definit. Maximum av f på triangeln med hörn i punkterna $(1, 1)$, $(-2, 1)$ och $(0, 2)$ antas därför i något av hörnen. Funktionens värden i hörnpunkterna är 5, 2 resp. 8. Maximivärdet är därför lika med 8 och antas för $(x_1, x_2) = (0, 2)$. □

En icke-konstant reellvärd konvex funktion kan inte ha ett maximum i någon inre punkt av definitionsmängden. Vi har nämligen följande sats.

Sats 6.3.5. *Antag att funktionen $f\colon X \to \mathbf{R}$ är konvex och har ett maximum i det relativt inre av X. Då är f konstant på X.*

Bevis. Antag att f har ett maximum i punkten $a \in \operatorname{rint} X$, och låt $x \in X$ vara en godtycklig punkt i X. Eftersom a är en relativt inre punkt, finns det en punkt $y \in X$ så att a ligger på den öppna sträckan $]x, y[$, dvs. $a = \lambda x + (1-\lambda)y$ för något λ med $0 < \lambda < 1$. Eftersom a är en maximipunkt, är $f(y) \leq f(a)$,

och konvexiteten ger oss nu olikheten

$$f(a) = f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \leq \lambda f(x) + (1-\lambda)f(a),$$

med $f(x) \geq f(a)$ som slutsats. Eftersom den omvända olikheten trivialt gäller, är $f(x) = f(a)$. Funktionen är således konstant lika med $f(a)$. □

6.4 Några viktiga olikheter

Många olikheter kan bevisas med hjälp av konvexitetsargument. Vi skall ge tre exempel.

Aritmetiskt och geometriskt medelvärde

Definition. Låt $\theta_1, \theta_2, \ldots, \theta_n$ vara givna positiva tal med summa $\sum_{j=1}^{n} \theta_j = 1$. Det *vägda aritmetiska medelvärdet* A och det *vägda geometriska medelvärdet* G av n stycken positiva tal $a_1, a_2, \ldots, a_n$ med de givna talen $\theta_1, \theta_2, \ldots, \theta_n$ som *vikter* definieras som

$$A = \sum_{j=1}^{n} \theta_j a_j \qquad \text{och} \qquad G = \prod_{j=1}^{n} a_j^{\theta_j}.$$

Vi får förstås de vanliga aritmetiska och geometriska medelvärdena som specialfall genom att låta alla vikter vara lika med $1/n$.

Följande olikhet råder mellan det vägda aritmetiska och det vägda geometriska medelvärdet.

Sats 6.4.1. *För alla positiva tal $a_1, a_2, \ldots, a_n$ är*

$$G \leq A$$

med likhet om och endast om $a_1 = a_2 = \cdots = a_n$.

Bevis. Sätt $x_j = \ln a_j$ så att $a_j = \mathrm{e}^{x_j} = \exp(x_j)$. Olikheten $G \leq A$ blir då ekvivalent med olikheten

$$\exp\Big(\sum_{j=1}^{n} \theta_j x_j\Big) \leq \sum_{j=1}^{n} \theta_j \exp(x_j),$$

som är ett specialfall av Jensens olikhet och som gäller med likhet om och endast om $a_1 = a_2 = \cdots = a_n$ eftersom exponentialfunktionen är strikt konvex. □

EXEMPEL 6.4.1. I många fall kan man använda olikheten mellan det aritmetiska och det geometriska medelvärdet för att lösa maximerings- och minimeringsproblem. Här följer ett allmänt exempel.

Betrakta en funktion f som har formen

$$f(x) = \sum_{i=1}^{m} c_i \big(\prod_{j=1}^{n} x_j^{\alpha_{ij}} \big), \qquad x \in \mathbf{R}^n$$

där $c_i > 0$ och α_{ij} är reella tal för alla i, j.

Ett typiskt exempel är funktionen $g(x) = 16x_1 + 2x_2 + x_1^{-1}x_2^{-2}$, där $n = 2$, $m = 3$, $c = (16, 2, 1)$ och

$$\alpha = [\alpha_{ij}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -2 \end{bmatrix}$$

Antag att vi vill minimera $f(x)$ över mängden $\{x \in \mathbf{R}^n \mid x > 0\}$. Vi kan attackera detta problem på följande sätt. Låt $\theta_1, \theta_2, \ldots, \theta_m$ vara positiva tal med summa 1, och gör omskrivningen

$$f(x) = \sum_{i=1}^{m} \theta_i \big(\frac{c_i}{\theta_i} \prod_{j=1}^{n} x_j^{\alpha_{ij}} \big).$$

Genom att utnyttja olikheten mellan det aritmetiska och det geometriska medelvärdet får vi i detta fall olikheten

$$f(x) \geq \prod_{i=1}^{m} \Big(\big(\frac{c_i}{\theta_i} \big)^{\theta_i} \big(\prod_{j=1}^{n} x_j^{\theta_i \alpha_{ij}} \big) \Big) = C(\theta) \cdot \prod_{j=1}^{n} x_j^{\beta_j}, \tag{6.5}$$

där

$$C(\theta) = \prod_{i=1}^{m} \big(\frac{c_i}{\theta_i} \big)^{\theta_i} \qquad \text{och} \qquad \beta_j = \sum_{i=1}^{m} \theta_i \alpha_{ij}.$$

Om det är möjligt att välja vikterna $\theta_i > 0$ på ett sådant sätt att $\sum_{i=1}^{m} \theta_i = 1$ och

$$\beta_j = \sum_{i=1}^{m} \theta_i \alpha_{ij} = 0 \quad \text{för alla } j,$$

så övergår olikheten (6.5) i

$$f(x) \geq C(\theta),$$

med likhet om och endast om alla produkterna $\dfrac{c_i}{\theta_i} \displaystyle\prod_{j=1}^{n} x_j^{\alpha_{ij}}$ är lika, en egenskap som gör det möjligt att bestämma x. □

Hölders olikhet

Sats 6.4.2 (Hölders olikhet)**.** *Antag att* $1 \le p \le \infty$ *och definiera det duala indexet* $q \in [1, \infty]$ *genom sambandet*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

För alla x, $y \in \mathbf{R}^n$ *gäller då att*

$$|\langle x, y\rangle| = \Big|\sum_{j=1}^{n} x_j y_j\Big| \le \|x\|_p \|y\|_q.$$

För alla x *finns det vidare ett* y *med norm* $\|y\|_q = 1$ *så att* $\langle x, y\rangle = \|x\|_p$.

Anmärkning. Observera att $q = 2$ om $p = 2$, så i fallet $p = 2$ övergår Hölders olikhet i Cauchy–Schwarz olikhet.

Bevis. För $p = \infty$ följer Hölders olikhet direkt av triangelolikheten för summor:

$$\Big|\sum_{j=1}^{n} x_j y_j\Big| \le \sum_{j=1}^{n} |x_j||y_j| \le \sum_{j=1}^{n} \|x\|_\infty \, |y_j| = \|x\|_\infty \|y\|_1.$$

Antag därför att $1 \le p < \infty$. Eftersom $\Big|\sum_1^n x_j y_j\Big| \le \sum_1^n |x_j||y_j|$, vektorn $(|x_1|, \dots, |x_n|)$ har samma ℓ^p-norm som $(x_1, \dots, x_n)$ och vektorn $(|y_1|, \dots, |y_n|)$ har samma ℓ^q-norm som $(y_1, \dots, y_n)$, kan vi i beviset för Hölders olikhet nu utan inskränkning anta att alla talen x_j och y_j är positiva.

Funktionen $t \mapsto t^p$ är konvex på intervallet $[0, \infty[$. För alla positiva tal $t_1, t_2, \dots, t_n$ och alla positiva tal $\lambda_1, \lambda_2, \dots, \lambda_n$ med $\sum_1^n \lambda_j = 1$ är därför

$$\Big(\sum_{j=1}^{n} \lambda_j t_j\Big)^p \le \sum_{j=1}^{n} \lambda_j t_j^p. \tag{6.6}$$

Välj nu speciellt

$$\lambda_j = \frac{y_j^q}{\sum_{j=1}^n y_j^q} \qquad \text{och} \qquad t_j = \frac{x_j y_j}{\lambda_j}.$$

Då är

$$\lambda_j t_j = x_j y_j \qquad \text{och} \qquad \lambda_j t_j^p = \frac{x_j^p y_j^p}{y_j^{(p-1)q}} \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1} = x_j^p \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1}.$$

Insättning i olikheten (6.6) ger nu

$$\Big(\sum_{j=1}^{n} x_j y_j\Big)^p \le \sum_{j=1}^{p} x_j^p \Big(\sum_{j=1}^{n} y_j^q\Big)^{p-1},$$

och vi får Hölders olikhet genom att upphöja båda sidorna till $1/p$.

Genom insättning verifierar man enkelt att likhet råder i Hölders olikhet med $\|y\|_q = 1$ i följande fall.

$x = 0$: Alla y med norm 1.

$x \neq 0,\ 1 \leq p < \infty$: $y_j = \begin{cases} \|x\|_p^{-p/q}|x_j|^p/x_j & \text{om } x_j \neq 0, \\ 0 & \text{om } x_j = 0. \end{cases}$

$x \neq 0,\ p = \infty$: $y_j = \begin{cases} |x_j|/x_j & \text{om } j = j_0, \\ 0 & \text{om } j \neq j_0, \end{cases}$

där j_0 är ett index med $|x_{j_0}| = \|x\|_\infty$. □

Sats 6.4.3 (Minkowskis olikhet)**.** *Antag att $p \geq 1$ och låt x och y vara godtyckliga vektorer i $\mathbf{R}^n$. Då är*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Bevis. Betrakta de linjära formerna $x \mapsto f_a(x) = \langle a, x\rangle$ för vektorer $a \in \mathbf{R}^n$ som uppfyller $\|a\|_q = 1$; enligt Hölders olikhet är

$$f_a(x) \leq \|a\|_q\|x\|_p \leq \|x\|_p.$$

För varje x finns det vidare en vektor a med $\|a\|_q = 1$ som ger likhet i Hölders olikhet, dvs. så att $f_a(x) = \|x\|_p$. Detta innebär att

$$\|x\|_p = \sup\{f_a(x) \mid \|a\|_q = 1\}.$$

Funktionen $f(x) = \|x\|_p$ är därför konvex enligt sats 6.2.4. Uppenbarligen är f också positivt homogen, så det följer att den är subadditiv. Därmed är Minkowskis olikhet bevisad. □

6.5 Lösbarhet för system av konvexa olikheter

I kapitel 3 formulerade och bevisade vi några resultat om lösbarhet för system av linjära olikheter. Nästa sats generaliserar resultatet i sats 3.3.7 och behandlar lösbarheten hos ett system av konvexa och affina olikheter.

Sats 6.5.1. *Låt $f_i\colon \Omega \to \mathbf{R}$, $i = 1, 2, \ldots, m$, vara en familj av konvexa funktioner som är definierade på en konvex delmängd Ω av $\mathbf{R}^n$.*

Låt p vara ett heltal i intervallet $1 \leq p \leq m$, och antag om $p < m$ att funktionerna f_i för $i \geq p + 1$ är restriktioner till Ω av affina *funktioner och att mängden*

$$\{x \in \operatorname{rint}\Omega \mid f_i(x) \leq 0 \text{ för } i = p + 1, \ldots, m\}$$

inte är tom. Då är följande två påståenden ekvivalenta:

(i) Systemet

$$\begin{cases} f_i(x) < 0, & i = 1, 2, \ldots, p \\ f_i(x) \leq 0, & i = p+1, \ldots, m \end{cases}$$

saknar lösning $x \in \Omega$.

(ii) Det finns icke-negativa tal $\lambda_1, \lambda_2, \ldots, \lambda_m$ *sådana att minst ett av talen* $\lambda_1, \lambda_2, \ldots, \lambda_p$ *är nollskilt och*

$$\sum_{i=1}^{m} \lambda_i f_i(x) \geq 0$$

för alla $x \in \Omega$.

Anmärkning. Systemet av olikheter måste alltså innehålla minst en strikt olikhet, men vi tillåter att samtliga olikheter är strikta, dvs. att $p = m$.

Bevis. Att (ii) medför (i) är trivialt, ty om systemet i (i) har en lösning x, så är uppenbarligen summan i (ii) negativ för samma x eftersom åtminstone en term i summan är negativ och övriga är icke-positiva.

För att visa den omvända implikationen antar vi att systemet i (i) saknar lösning och låter M vara mängden av alla $y = (y_1, y_2, \ldots, y_m) \in \mathbf{R}^m$ för vilka systemet

$$\begin{cases} f_i(x) < y_i, & i = 1, 2, \ldots, p \\ f_i(x) = y_i, & i = p+1, \ldots, m \end{cases}$$

har någon lösning $x \in \Omega$.

Mängden M är konvex, ty om y' och y'' är två punkter i M, $0 \leq \lambda \leq 1$, och x', x'' är lösningar i Ω till systemen av olikheter med y' resp. y'' som högerled, så är $x = \lambda x' + (1-\lambda)x'' \in \Omega$ en lösning till systemet med $\lambda y' + (1-\lambda)y''$ som högerled på grund av att funktionerna f_i är konvexa för $i \leq p$ och affina för $i > p$.

Vårt antagande om systemet i (i) innebär vidare att $M \cap \mathbf{R}^m_- = \emptyset$. Eftersom $\mathbf{R}^m_-$ är en polyeder, finns det därför på grund av separationssatsen 5.5.2 en nollskild vektor $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ och ett reellt tal α så att hyperplanet $H = \{y \mid \langle \lambda, y \rangle = \alpha\}$ separerar M och $\mathbf{R}^m_-$ utan att innehålla M som delmängd. Vi kan anta att

$$\lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_m y_m \begin{cases} \geq \alpha & \text{för alla } y \in M, \\ \leq \alpha & \text{för alla } y \in \mathbf{R}^m_-. \end{cases}$$

För $y = 0$ ger detta att $\alpha \geq 0$, och genom att sedan välja $y = t\mathbf{e}_i$, där $\mathbf{e}_i$ är den i:te enhetsvektorn i $\mathbf{R}^m$, och låta t gå mot $-\infty$, drar vi slutsatsen att $\lambda_i \geq 0$ för alla i.

För varje $x \in \Omega$ och $\epsilon > 0$ är

$$y = (f_1(x) + \epsilon, \ldots, f_p(x) + \epsilon, f_{p+1}(x), \ldots, f_m(x))$$

en punkt i M. Följaktligen är

$\lambda_1(f_1(x)+\epsilon)+\cdots+\lambda_p(f_p(x)+\epsilon)+\lambda_{p+1}f_{p+1}(x)+\cdots+\lambda_m f_m(x)\geq\alpha\geq 0$,
och genom att låta ϵ gå mot noll erhåller vi olikheten

$$\lambda_1 f_1(x)+\lambda_2 f_2(x)+\cdots+\lambda_m f_m(x)\geq 0$$

för alla $x\in\Omega$.

I fallet $p=m$ är vi klara, eftersom vektorn $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_m)$ är nollskild, men det återstår att visa att någon av koefficienterna $\lambda_1,\lambda_2,\ldots,\lambda_p$ är nollskild i de övriga fallen, dvs. då $p<m$. Så antag motsatsen, dvs. att $\lambda_1=\lambda_2=\cdots=\lambda_p=0$, och sätt

$$h(x)=\sum_{i=p+1}^{m}\lambda_i f_i(x).$$

Funktionen h är affin och $h(x)\geq 0$ för alla $x\in\Omega$. Enligt förutsättningarna i satsen finns det vidare en punkt x_0 i det relativa inre av Ω sådan att $f_i(x_0)\leq 0$ för alla $i\geq p+1$, vilket medför att $h(x_0)\leq 0$, och följaktligen är $h(x_0)=0$. Restriktionen $h|_\Omega$, som är en konkav funktion eftersom h är affin, har således ett minimum i en relativt inre punkt, och enligt sats 6.3.5 (tillämpad på funktionen $-h|_\Omega$) medför detta att funktionen h är konstant lika med 0 på Ω.

Men för varje $y\in M$ finns det en punkt $x\in\Omega$ sådan att $y_i=f_i(x)$ för $i=p+1,\ldots m$, vilket medför att $\langle\lambda,y\rangle=\sum_{i=p+1}^{m}\lambda_i f_i(x)=h(x)=0$. Detta betyder att $\alpha=0$ och att hyperplanet H innehåller M. Det sistnämnda är en motsägelse, och därmed är satsen bevisad. □

6.6 Kontinuitet

En konvex reellvärd funktion är automatiskt kontinuerlig i alla relativt inre punkter i definitionsmängden. Mer precist gäller följande sats.

Sats 6.6.1. *Antag att $f\colon X\to\overline{\mathbf{R}}$ är en konvex funktion och att a är en punkt i det relativt inre av* dom f*. Då finns det i* dom f *en relativt öppen omgivning U av punkten a och en konstant M så att*

$$|f(x)-f(a)|\leq M\|x-a\|$$

för alla $x\in U$. Speciellt är alltså funktionen f kontinuerlig på det relativt inre av dom f*.*

Bevis. Vi börjar med att visa ett specialfall av satsen och visar sedan hur det allmänna fallet kan reduceras till detta specialfall.

1. Så antag först att X är en öppen delmängd av $\mathbf{R}^n$, att dom $f=X$ så att f i själva verket är en reellvärd konvex funktion, att $a=0$ och att $f(0)=0$.

Vi ska visa att om vi väljer talet $r > 0$ så att den slutna hyperkuben

$$K(r) = \{x \in \mathbf{R}^n \mid \|x\|_\infty \leq r\}$$

ligger helt i X, så finns det en konstant M så att

$$|f(x)| \leq M\|x\| \tag{6.7}$$

för alla $x \in \overline{B}(0;r) = \{x \in \mathbf{R}^n \mid \|x\| \leq r\}$, (där $\|\cdot\|$ är den vanliga euklidiska normen).

Hyperkuben $K(r)$ har 2^n stycken extremalpunkter (hörn). Låt L beteckna det största av f:s funktionsvärden i dessa extremalpunkter. Eftersom konvexa höljet av extremalpunkterna är lika med $K(r)$, följer det då av sats 6.3.3 att

$$f(x) \leq L$$

för alla $x \in K(r)$, och därmed också för alla $x \in \overline{B}(0;r)$, ty $\overline{B}(0;r)$ är en delmängd till $K(r)$.

Vi ska nu skärpa denna olikhet. Låt för den skull x vara en godtycklig punkt i $\overline{B}(0;r)$ skild från medelpunkten 0. Halvlinjen från 0 genom x skär randen av $\overline{B}(0;r)$ i punkten

$$y = \frac{r}{\|x\|}x,$$

och eftersom x ligger på sträckan $[0,y]$, är x en konvex kombination av sträckans ändpunkter, närmare bestämt är $x = \lambda y + (1-\lambda)0$ för $\lambda = \|x\|/r$. På grund av konvexiteten är därför

$$f(x) \leq \lambda f(y) + (1-\lambda)f(0) = \lambda f(y) \leq \lambda L = \frac{L}{r}\|x\|.$$

Ovanstående olikhet gäller för alla $x \in \overline{B}(0;r)$. För att visa att olikheten också gäller med $f(x)$ bytt mot $|f(x)|$ utnyttjar vi att punkten $-x$ ligger i $\overline{B}(0;r)$ om x gör det, samt att $0 = \frac{1}{2}x + \frac{1}{2}(-x)$. På grund av konvexitet är därför

$$0 = f(0) \leq \frac{1}{2}f(x) + \frac{1}{2}f(-x) \leq \frac{1}{2}f(x) + \frac{L}{2r}\|-x\|,$$

vilket ger oss olikheten

$$f(x) \geq -\frac{L}{r}\|-x\| = -\frac{L}{r}\|x\|.$$

Därmed har vi visat att olikheten (6.7) gäller för $x \in \overline{B}(0;r)$ med $M = L/r$.

2. Vi övergår nu till det allmänna fallet. Låt n vara lika med dimensionen hos mängden $\operatorname{dom} f$. Det affina höljet till $\operatorname{dom} f$ kan då skrivas på formen

$a+V$, där V är ett linjärt delrum av dimension n. Eftersom delrummet V är isomorft med $\mathbf{R}^n$ får vi genom att välja ett koordinatsystem i V en bijektiv linjär avbildning $T\colon \mathbf{R}^n \to V$.

Den inversa bilden Y av det relativa inre av $\operatorname{dom} f$ under avbildningen $y \mapsto a + Ty$ av $\mathbf{R}^n$ på $\operatorname{aff}(\operatorname{dom} f)$ är en öppen konvex delmängd av $\mathbf{R}^n$, och Y innehåller punkten 0. Definiera funktionen $g\colon Y \to \mathbf{R}$ genom att sätta

$$g(y) = f(a + Ty) - f(a).$$

Då är g en konvex funktion, eftersom g är sammansatt av en konvex funktion och en affin funktion, och $g(0) = 0$.

För $x = a + Ty \in \operatorname{rint}(\operatorname{dom} f)$ är nu $f(x) - f(a) = g(y)$ och $x - a = Ty$, så för att visa påståendet i satsen ska vi alltså visa att det finns en konstant M sådan att $|g(y)| \leq M\|Ty\|$ för alla y i någon omgivning av 0. Men avbildningen $y \to \|Ty\|$ är en norm på $\mathbf{R}^n$, och eftersom alla normer är ekvivalenta, räcker det att visa att det finns en konstant M sådan att

$$|g(y)| \leq M\|y\|$$

för alla y i någon omgivning av 0, och detta är precis vad vi gjorde i steg 1 av beviset. Därmed är satsen bevisad. □

Eftersom affina mängder saknar relativa randpunkter har sats 6.6.1 följande omedelbara korollarium.

Korollarium 6.6.2. *En konvex funktion $f\colon X \to \mathbf{R}$ med en affin mängd X som definitionsmängd är kontinuerlig.*

För reellvärda funktioner f med ett slutet intervall $I = [a, b]$ som definitionsmängd sätter konvexiteten inga andra restriktioner på funktionsvärdet $f(b)$ än att det måste vara större än eller lika med $\lim_{x\to b-} f(x)$. En konvex funktion behöver alltså inte vara kontinuerlig i ändpunkten b, och motsvarande gäller förstås i vänstra ändpunkten. Exempelvis är funktionen f, som är identiskt noll på $I \setminus \{a, b\}$, konvex om $f(a) \geq 0$ och $f(b) \geq 0$. Jmf övning 7.6.

6.7 Konvexa funktioners recessiva delrum

Exempel 6.7.1. Låt f vara den konvexa funktionen

$$f(x_1, x_2) = x_1 + x_2 + \mathrm{e}^{(x_1-x_2)^2}$$

med $\mathbf{R}^2$ som definitionsmängd. Restriktionerna av funktionen f till linjer med riktningsvektorn $v = (1, 1)$ är affina funktioner, ty

$$f(x + tv) = f(x_1 + t, x_2 + t) = x_1 + x_2 + 2t + \mathrm{e}^{(x_1-x_2)^2} = f(x) + 2t.$$

Låt nu $V = \{x \in \mathbf{R}^2 \mid x_1 = x_2\}$ vara det linjära delrum av $\mathbf{R}^2$ som spänns upp av vektorn v, och betrakta den ortogonala uppdelningen

$$\mathbf{R}^2 = V^\perp + V$$

av $\mathbf{R}^2$. Varje $x \in \mathbf{R}^2$ har en motsvarande entydig uppdelning $x = y + z$ med $y \in V^\perp$ och $z \in V$, nämligen

$$y = \tfrac{1}{2}(x_1 - x_2, x_2 - x_1) \text{ och } z = \tfrac{1}{2}(x_1 + x_2, x_1 + x_2).$$

Eftersom $z = \frac{1}{2}(x_1 + x_2)v = z_1 v$, är vidare

$$f(x) = f(y + z) = f(y) + 2z_1 = f|_{V^\perp}(y) + 2z_1.$$

Så vi har en motsvarande uppdelning av f som en summa av f:s restriktion till $V^\perp$ och en linjär funktion på V. I $\mathbf{R}^2 \times \mathbf{R}$ spänner, som man lätt verifierar, vektorn $(v, 2) = (1, 1, 2)$ upp det recessiva delrummet $\operatorname{lin}(\operatorname{epi} f)$, och V är lika med bilden av $\operatorname{lin}(\operatorname{epi} f)$ under projektionen $P_1 \colon \mathbf{R}^2 \times \mathbf{R} \to \mathbf{R}^2$. □

Resultatet i exemplet ovan kan generaliseras och för att beskriva denna generalisering behöver vi först en definition.

Definition. Låt $f \colon X \to \overline{\mathbf{R}}$ vara en funktion vars definitionsmängd X är en delmängd till $\mathbf{R}^n$. Det linjära delrummet

$$V_f = P_1(\operatorname{lin}(\operatorname{epi} f)),$$

där $P_1 \colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ är projektionen på första faktorn $\mathbf{R}^n$, kallas funktionens *recessiva delrum*.

För konvexa funktioner har vi följande resultat för det recessiva delrummet.

Sats 6.7.1. *Låt f vara en konvex funktion med recessivt delrum V_f.*

(i) En vektor v tillhör V_f om och endast om det finns ett unikt tal α_v sådant att (v, α_v) tillhör det recessiva delrummet $\operatorname{lin}(\operatorname{epi} f)$ *till funktionens epigraf.*

(ii) Avbildningen $g \colon V_f \to \mathbf{R}$, som definieras av att $g(v) = \alpha_v$ för $v \in V_f$, är linjär.

(iii) $\operatorname{dom} f = \operatorname{dom} f + V_f$.

(iv) För alla $x \in \operatorname{dom} f$ och alla $v \in V_f$ är $f(x + v) = f(x) + g(v)$.

(v) Om funktionen f är differentierbar i punkten x så är $g(v) = \langle f'(x), v\rangle$ för alla $v \in V_f$.

(vi) Antag att V är ett linjärt delrum, att $h \colon V \to \mathbf{R}$ är en linjär avbildning, att $\operatorname{dom} f + V \subseteq \operatorname{dom} f$ samt att $f(x + v) = f(x) + h(v)$ för alla $x \in \operatorname{dom} f$ och alla $v \in V$. Då är $V \subseteq V_f$.

Bevis. (i) Per definition gäller att $v \in V_f$ om och endast om det finns ett reellt tal α_v sådant att $(v, \alpha_v) \in \text{lin}(\text{epi}\, f)$. För att visa att talet α_v är entydigt bestämt av vektorn $v \in V_f$ antar vi att paret (v, β) också ligger i $\text{lin}(\text{epi}\, f)$.

För varje $x \in \text{dom}\, f$ och varje $t \in \mathbf{R}$ är $(x + tv, f(x) + t\alpha_v)$ en punkt i epigrafen, dvs.

$$x + tv \in \text{dom}\, f \quad \text{och} \quad f(x + tv) \le f(x) + t\alpha_v. \tag{6.8}$$

Så $(x + tv, f(x + tv))$ är speciellt en punkt i epigrafen, och då är också $(x + tv - tv, f(x + tv) - t\beta)$ en punkt i $\text{epi}\, f$ på grund av vårt antagande $(v, \beta) \in \text{lin}(\text{epi}\, f)$, och detta innebär att

$$f(x) \le f(x + tv) - t\beta \tag{6.9}$$

för alla $t \in \mathbf{R}$. Genom att kombinera de två erhållna olikheterna (6.8) och (6.9) erhålls olikheten $f(x) \le f(x) + (\alpha_v - \beta)t$, som gäller för alla $t \in \mathbf{R}$, vilket medför att $\beta = \alpha_v$. Därmed är entydigheten visad.

(ii) Låt som tidigare P_1 vara projektionen av $\mathbf{R}^n \times \mathbf{R}$ på $\mathbf{R}^n$, och låt på motsvarande sätt P_2 vara projektionen på den andra faktorn $\mathbf{R}$. Entydighetsresultatet (i) innebär att P_1:s restriktion till det linjära delrummet $\text{lin}(\text{epi}\, f)$ är en bijektiv linjär avbildning på V_f, och om Q betecknar inversen till denna avbildning, så är avbildningen g lika med sammansättningen $P_2 \circ Q$ av de två linjära avbildningarna P_2 och Q. Detta betyder att g är en linjär funktion.

(iii) Genom att speciellt välja $t = 1$ i (6.8) fås implikationen

$$x \in \text{dom}\, f \;\&\; v \in V_f \;\Rightarrow\; x + v \in \text{dom}\, f,$$

dvs. inklusionen $\text{dom}\, f + V_f \subseteq \text{dom}\, f$, och den omvända inklusionen är trivial.

(iv) Genom att välja $t = 1$ i olikheterna (6.8) och (6.9) samt använda att $\alpha_v = \beta = g(v)$ erhåller vi de båda olikheterna $f(x + v) \le f(x) + g(v)$ och $f(x) \le f(x + v) - g(v)$, som tillsammans ger likheten i (iv).

(v) Betrakta restriktionen $\phi(t) = f(x + tv)$ av funktionen f till linjen genom punkten $x \in \text{dom}\, f$ med riktningen $v \in V_f$. Enligt (iii) är funktionen ϕ definierad för alla $t \in \mathbf{R}$, och enligt (iv) är $\phi(t) = f(x) + tg(v)$. Speciellt är alltså $\phi'(0) = g(v)$. Men om funktionen f är differentierbar i punkten x, så ger kedjeregeln att $\phi'(0) = \langle f'(x), v \rangle$, och detta bevisar påstående (v).

(vi) Antag att $v \in V$. Om (x, s) är en godtycklig punkt i epigrafen $\text{epi}\, f$, så är $f(x + tv) = f(x) + h(tv) \le s + th(v)$, vilket innebär att $(x + tv, s + th(v))$ ligger i $\text{epi}\, f$ för varje reellt tal t. Detta visar att $(v, h(v))$ tillhör $\text{lin}(\text{epi}\, f)$ och att följaktligen v är en vektor i V_f. □

Enligt nästa sats kan varje konvex funktion skrivas som en summa av en konvex funktion med ett trivialt recessivt delrum och en linjär funktion.

Sats 6.7.2. *Antag att f är en konvex funktion med recessivt delrum V_f. Låt $\tilde{f}$ beteckna restriktionen av funktionen f till* $\operatorname{dom} f \cap V_f^\perp$*, och låt $g\colon V_f \to \mathbf{R}$ vara den linjära funktion som definieras i sats 6.7.1. Då är funktionen $\tilde{f}$:s recessiva delrum $V_{\tilde{f}}$ trivialt, dvs. lika med* $\{0\}$*,* $\operatorname{dom} f = \operatorname{dom} f \cap V_f^\perp + V_f$ *och*

$$f(y+z) = \tilde{f}(y) + g(z)$$

för alla $y \in \operatorname{dom} f \cap V_f^\perp$ och alla $z \in V_f$.

Bevis. Varje $x \in \mathbf{R}^n$ har förstås en unik uppdelning på formen $x = y + z$ med $y \in V_f^\perp$ och $z \in V_f$, och för $x \in \operatorname{dom} f$ gäller speciellt enligt sats 6.7.1 att $y = x - z \in \operatorname{dom} f + V_f = \operatorname{dom} f$, dvs. $y \in \operatorname{dom} f \cap V_f^\perp$. Detta innebär att $\operatorname{dom} f = \operatorname{dom} f \cap V_f^\perp + V_f$.

Uppdelningen $f(y+z) = \tilde{f}(y) + g(z)$ följer nu av (iv) i sats 6.7.1, så det återstår bara att visa att $V_{\tilde{f}} = \{0\}$. Så antag att $v \in V_{\tilde{f}}$ och låt x_0 vara en godtycklig punkt i $\operatorname{dom} \tilde{f}$. Då ligger också $x_0 + v$ i $\operatorname{dom} \tilde{f}$, och eftersom $\operatorname{dom} \tilde{f} \subseteq V_f^\perp$ och $V_f^\perp$ är ett linjärt delrum, följer det att $v = (x_0 + v) - x_0$ är en vektor i $V_f^\perp$. Detta visar inklusionen $V_{\tilde{f}} \subseteq V_f^\perp$.

Sats 6.7.1 ger oss två linjära funktioner $g\colon V_f \to \mathbf{R}$ och $\tilde{g}\colon V_{\tilde{f}} \to \mathbf{R}$ sådana att $f(x+v) = f(x) + g(v)$ för alla $x \in \operatorname{dom} f$ och alla $v \in V_f$, och $\tilde{f}(y+w) = \tilde{f}(y) + \tilde{g}(w)$ för alla $y \in \operatorname{dom} f \cap V_f^\perp$ och alla $w \in V_{\tilde{f}}$.

Låt nu w vara en godtycklig vektor i $V_{\tilde{f}}$ och x vara en godtycklig punkt i $\operatorname{dom} f$, och skriv x på formen $x = y + v$ med $y \in \operatorname{dom} f \cap V_f^\perp$ och $v \in V_f$. Punkten $y + w$ ligger i $\operatorname{dom} f \cap V_f^\perp$, så vi får följande identiteter

$$\begin{aligned} f(x+w) &= f(y+v+w) = f(y+w+v) = f(y+w) + g(v) \\ &= \tilde{f}(y+w) + g(v) = \tilde{f}(y) + \tilde{g}(w) + g(v) \\ &= f(y) + g(v) + \tilde{g}(w) = f(x) + \tilde{g}(w), \end{aligned}$$

och det följer därför av sats 6.7.1 (v) att $V_{\tilde{f}} \subseteq V_f$. Så $V_{\tilde{f}} \subseteq V_f^\perp \cap V_f = \{0\}$, vilket visar att $V_{\tilde{f}} = \{0\}$. □

6.8 Slutna konvexa funktioner

Definition. En konvex funktion kallas *sluten* om funktionens epigraf är en sluten mängd.

Sats 6.8.1. *En konvex funktion $f\colon X \to \overline{\mathbf{R}}$ är sluten om och endast om alla subnivåmängder till f är slutna mängder.*

Bevis. Antag att $X \subseteq \mathbf{R}^n$ och att funktionen f är sluten. Låt

$$X_\alpha = \operatorname{sublev}_\alpha f = \{x \in X \mid f(x) \leq \alpha\}$$

vara en godtycklig icke-tom subnivåmängd till f och sätt

$$Y_\alpha = \operatorname{epi} f \cap \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} \leq \alpha\}.$$

Mängden Y_α är sluten eftersom den är definierad som snittet mellan den slutna epigrafen och ett slutet halvrum. Vidare är $X_\alpha = P(Y_\alpha)$, där P är projektionen $P(x, x_{n+1}) = x$ av $\mathbf{R}^n \times \mathbf{R}$ på $\mathbf{R}^n$.

Recessionskonen $\operatorname{recc} Y_\alpha$ innehåller uppenbarligen ingen nollskild vektor av typen $v = (0, v_{n+1})$, dvs. ingen nollskild vektor som ligger i i nollrummet $\mathcal{N}(P) = \{0\} \times \mathbf{R}$ till projektionen P, så därför är $(\operatorname{recc} Y_\alpha) \cap \mathcal{N}(P) = \{0\}$. Det följer därför av sats 2.7.10 att subnivåmängden X_α är sluten.

Antag för att bevisa omvändningen att alla subnivåmängder är slutna, och låt (x_0, y_0) vara en godtycklig randpunkt till $\operatorname{epi} f$. Låt $\big((x_k, y_k)\big)_1^\infty$ vara en följd av punkter i epigrafen som konvergerar mot (x_0, y_0), och låt $\epsilon > 0$ vara godtyckligt. Eftersom $y_k \to y_0$ då $k \to \infty$, är $f(x_k) \leq y_k \leq y_0 + \epsilon$ för alla tillräckligt stora k, så punkterna x_k tillhör subnivåmängden

$$\{x \in X \mid f(x) \leq y_0 + \epsilon\}$$

för alla tillräckligt stora k. Eftersom subnivåmängden är sluten, ligger gränspunkten x_0 i samma subnivåmängd, dvs. $x_0 \in X$ och $f(x_0) \leq y_0 + \epsilon$, och eftersom $\epsilon > 0$ är godtyckligt, drar vi slutsatsen att $f(x_0) \leq y_0$. Detta innebär att (x_0, y_0) är en punkt i $\operatorname{epi} f$. Epigrafen innehåller således alla sina randpunkter och är därför en sluten mängd. □

Subnivåmängderna till en reellvärd kontinuerlig funktion med sluten definitionsmängd är slutna. Vi får därför följande omedelbara korollarium till sats 6.8.1.

Korollarium 6.8.2. *Kontinuerliga konvexa funktioner $f\colon X \to \mathbf{R}$ med slutna definitionsmängder X är slutna funktioner.*

Sats 6.8.3. *Alla icke-tomma subnivåmängder till en sluten konvex funktion har samma recessionskon och samma recessiva delrum. Om en icke-tom subnivåmängd är begränsad, så är därför samtliga subnivåmängder begränsade.*

Bevis. Låt $f\colon X \to \overline{\mathbf{R}}$ vara en sluten konvex funktion, och antag att x_0 är en punkt i subnivåmängden $X_\alpha = \{x \in X \mid f(x) \leq \alpha\}$. Eftersom X_α och $\operatorname{epi} f$ är slutna konvexa mängder och (x_0, α) är en punkt i $\operatorname{epi} f$, får vi följande

ekvivalenser:

$$\begin{aligned} v \in \operatorname{recc} X_\alpha &\Leftrightarrow x_0 + tv \in X_\alpha \ \ \forall t \in \mathbf{R}_+ \Leftrightarrow f(x_0 + tv) \leq \alpha \ \ \forall t \in \mathbf{R}_+ \\ &\Leftrightarrow (x_0, \alpha) + t(v, 0) = (x_0 + tv, \alpha) \in \operatorname{epi} f \ \ \forall t \in \mathbf{R}_+ \\ &\Leftrightarrow (v, 0) \in \operatorname{recc}(\operatorname{epi} f), \end{aligned}$$

med slutsatsen att recessionskonen

$$\operatorname{recc} X_\alpha = \{v \in \mathbf{R}^n \mid (v, 0) \in \operatorname{recc}(\operatorname{epi} f)\}$$

inte beror av α så länge som $X_\alpha \neq \emptyset$. Naturligtvis gäller då detsamma för det recessiva delrummet

$$\operatorname{lin} X_\alpha = \operatorname{recc} X_\alpha \cap (-\operatorname{recc} X_\alpha) = \{v \in \mathbf{R}^n \mid (v, 0) \in \operatorname{lin}(\operatorname{epi} f)\}.$$

Påståendet rörande begränsade subnivåmängder följer av att en sluten konvex mängd är begränsad om och endast om dess recessionskon är lika med nollkonen $\{0\}$. □

Sats 6.8.4. *Om en konvex funktion f är begränsad på en affin delmängd M, så är funktionen konstant på M.*

Bevis. Sätt $M = a+U$, där U är ett linjärt delrum, och betrakta restriktionen $g = f|_M$ av f till M; det är en kontinuerlig funktion eftersom alla punkter i M är relativt inre punkter, och en sluten funktion eftersom definitionsmängden M är sluten. För $\alpha = \sup\{g(x) \mid x \in M\}$ är $\{x \mid g(x) \leq \alpha\} = M$, så det följer av föregående sats att alla g:s subnivåmängder har samma recessiva delrum som M, dvs. delrummet U.

Låt nu x_0 vara en godtycklig punkt i M. Då har speciellt subnivåmängden $\{x \mid g(x) \leq g(x_0)\}$ hela U som recessivt delrum, så $g(x_0 + u) \leq g(x_0)$ för alla $u \in U$, dvs. $g(x) \leq g(x_0)$ för alla $x \in M$. Punkten x_0 är således en maximipunkt till g, och eftersom x_0 är en godtycklig punkt i M, betyder detta att funktionen g är konstant på M. □

6.9 Stödfunktionen

Definition. Låt A vara en icke-tom delmängd av $\mathbf{R}^n$ och sätt för $x \in \mathbf{R}^n$

$$S_A(x) = \sup\{\langle y, x\rangle \mid y \in A\}$$

(med den sedvanliga konventionen att $S_A(x) = \infty$ om funktionen $y \mapsto \langle y, x\rangle$ är uppåt obegränsad på A). Funktionen $S_A\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ kallas *stödfunktionen* till mängden A.

Sats 6.9.1. *(a) Stödfunktionen S_A är en sluten konvex funktion.*
(b) Om A och B är icke-tomma mängder i $\mathbf{R}^n$, $\alpha > 0$ och $C\colon \mathbf{R}^n \to \mathbf{R}^m$ är en linjär avbildning, så är

(i) $S_A = S_{\operatorname{cvx} A} = S_{\operatorname{cl}(\operatorname{cvx} A)}$
(ii) $S_{\alpha A} = \alpha S_A$
(iii) $S_{A+B} = S_A + S_B$
(iv) $S_{A\cup B} = \max\{S_A, S_B\}$
(v) $S_{C(A)} = S_A \circ C^T$.

Bevis. (a) Eftersom

$$\operatorname{epi} S_A = \{(x,t) \mid \langle y, x\rangle \le t \text{ för alla } y \in A\} = \bigcap_{y\in A}\{(x,t) \mid \langle y, x\rangle \le t\},$$

är stödfunktionens epigraf ett snitt av slutna halvrum i $\mathbf{R}^n \times \mathbf{R}$. Epigrafen är därför en sluten konvex mängd, och det betyder att stödfunktionen är konvex och sluten.

(b) Eftersom linjära former är konvexa, följer det omedelbart av sats 6.3.3 att

$$S_A(x) = \sup\{\langle x, y\rangle \mid y \in A\} = \sup\{\langle x, y\rangle \mid y \in \operatorname{cvx} A\} = S_{\operatorname{cvx} A}(x)$$

för alla $x \in \mathbf{R}^n$. För funktioner f, som är kontinuerliga på slutna höljet till en mängd X, är vidare $\sup_{y\in X} f(y) = \sup_{y\in \operatorname{cl} X} f(y)$. Linjära former är förstås kontinuerliga, så därför följer det att $S_{\operatorname{cvx} A}(x) = S_{\operatorname{cl}(\operatorname{cvx} A)}(x)$ för alla x.

Detta bevisar identiteten (i) och de övriga identiteterna följer av följande räkningar:

$$S_{\alpha A}(x) = \sup_{y\in \alpha A}\langle y, x\rangle = \sup_{y\in A}\langle \alpha y, x\rangle = \alpha \sup_{y\in A}\langle y, x\rangle = \alpha S_A(x).$$

$$\begin{aligned} S_{A+B}(x) &= \sup_{y\in A+B}\langle y, x\rangle = \sup_{y_1\in A,\, y_2\in B}\langle y_1 + y_2, x\rangle \\ &= \sup_{y_1\in A,\, y_2\in B}(\langle y_1, x\rangle + \langle y_2, x\rangle) = \sup_{y_1\in A}\langle y_1, x\rangle + \sup_{y_2\in B}\langle y_2, x\rangle \\ &= S_A(x) + S_B(x). \end{aligned}$$

$$\begin{aligned} S_{A\cup B}(x) &= \sup_{y\in (A\cup B)}\langle y, x\rangle = \max\{\sup_{y\in A}\langle y, x\rangle, \sup_{y\in B}\langle y, x\rangle\} \\ &= \max\{S_A(x), S_B(x)\}. \end{aligned}$$

$$S_{C(A)}(x) = \sup_{y\in C(A)}\langle y, x\rangle = \sup_{z\in A}\langle Cz, x\rangle = \sup_{z\in A}\langle z, C^T x\rangle = S_A(C^T x). \qquad \square$$

EXEMPEL 6.9.1. Stödfunktionen till ett slutet intervall $[a, b]$ på reella axeln är

$$S_{[a,b]}(x) = S_{\{a,b\}}(x) = \max\{ax, bx\}$$

eftersom $[a, b] = \operatorname{cvx}\{a, b\}$. □

EXEMPEL 6.9.2. För att bestämma stödfunktionen till slutna enhetsbollen $\overline{B}_p = \{x \in \mathbf{R}^n \mid \|x\|_p \leq 1\}$ med avseende på ℓ^p-normen använder vi Hölders olikhet och får

$$S_{\overline{B}_p}(x) = \sup\{\langle x, y\rangle \mid \|y\|_p \leq 1\} = \|x\|_q,$$

där sambandet mellan p och q ges av att $1/p + 1/q = 1$. □

Slutna konvexa mängder karakteriseras på grund av följande sats fullständigt av sina stödfunktioner.

Sats 6.9.2. *Antag att X_1 och X_2 är två icke-tomma slutna konvexa delmängder av $\mathbf{R}^n$ med stödfunktioner S_{X_1} och S_{X_2}. Då gäller:*

(a) $$X_1 \subseteq X_2 \;\Leftrightarrow\; S_{X_1} \leq S_{X_2}$$

(b) $$X_1 = X_2 \;\Leftrightarrow\; S_{X_1} = S_{X_2}.$$

Bevis. Påstående (b) är en omedelbar konsekvens av (a), och implikationen $X_1 \subseteq X_2 \Rightarrow S_{X_1} \leq S_{X_2}$ är trivial. Det återstår således enbart att bevisa den omvända implikationen, och denna är logiskt ekvivalent med implikationen $X_1 \not\subseteq X_2 \Rightarrow S_{X_1} \not\leq S_{X_2}$.

Antag för den skull att $X_1 \not\subseteq X_2$, dvs. att det finns en punkt $x_0 \in X_1 \setminus X_2$. Då finns det ett hyperplan som separerar x_0 strikt från den slutna konvexa mängden X_2, och detta innebär att det finns en vektor $c \in \mathbf{R}^n$ och ett tal b så att $\langle x, c\rangle \leq b$ för alla $x \in X_2$ medan $\langle x_0, c\rangle > b$. Följaktligen är

$$S_{X_1}(c) \geq \langle x_0, c\rangle > b \geq \sup\{\langle x, c\rangle \mid x \in X_2\} = S_{X_2}(c),$$

vilket innebär att $S_{X_1} \not\leq S_{X_2}$. □

Genom att kombinera föregående sats med egenskap (i) i sats 6.9.1 får vi följande korollarium.

Korollarium 6.9.3. *Låt A och B vara två icke-tomma delmängder av $\mathbf{R}^n$. Då är $S_A = S_B$ om och endast om* $\operatorname{cl}(\operatorname{cvx} A) = \operatorname{cl}(\operatorname{cvx} B)$.

6.10 Minkowskifunktionalen

Låt X vara en konvex delmängd av $\mathbf{R}^n$ och antag att 0 är en inre punkt i X.

Betrakta mängderna tX för $t \geq 0$. Detta är en växande familj av mängder, vars union utgör hela $\mathbf{R}^n$, dvs.

$$0 \leq s < t \Rightarrow sX \subseteq tX \qquad \text{och} \qquad \bigcup_{t \geq 0} tX = \mathbf{R}^n.$$

Växandet beror på att mängderna tX är konvexa och innehåller 0, ty detta medför att

$$sX = \frac{s}{t}(tX) + (1 - \frac{s}{t})\,0 \subseteq \frac{s}{t}(tX) + (1 - \frac{s}{t})(tX) \subseteq tX.$$

Att mängdernas union är lika med hela $\mathbf{R}^n$ beror på att 0 är en inre punkt i X. Om $\overline{B}(0; r_0)$ är en sluten boll i X med centrum i 0, så ligger nämligen en godtycklig punkt $x \in \mathbf{R}^n$ i mängden $r_0^{-1}\|x\|X$ eftersom punkten $r_0\|x\|^{-1}x$ ligger i $\overline{B}(0; r_0)$.

För varje $x \in \mathbf{R}^n$ är följaktligen mängden $\{t \geq 0 \mid x \in tX\}$ ett obegränsat delintervall till intervallet $[0, \infty[$, och mängden innehåller talet $r_0^{-1}\|x\|$. Vi kan därför definiera en funktion

$$\phi_X \colon \mathbf{R}^n \to \mathbf{R}_+$$

genom att sätta

$$\phi_X(x) = \inf\{t \geq 0 \mid x \in tX\}.$$

Uppenbarligen är

$$\phi_X(x) \leq r_0^{-1}\|x\| \qquad \text{för alla } x.$$

Definition. Funktionen $\phi_X \colon \mathbf{R}^n \to \mathbf{R}_+$ kallas för *Minkowskifunktionalen* till mängden X.

Sats 6.10.1. *Minkowskifunktionalen ϕ_X har följande egenskaper:*

(i) För alla x, $y \in \mathbf{R}^n$ och alla $\lambda \in \mathbf{R}_+$ är
  (a) $\phi_X(\lambda x) = \lambda \phi_X(x)$
  (b) $\phi_X(x + y) \leq \phi_X(x) + \phi_X(y)$.

(ii) Det finns en konstant C sådan att
$$|\phi_X(x) - \phi_X(y)| \leq C\|x - y\|$$
för alla $x, y \in \mathbf{R}^n$.

(iii) $\operatorname{int} X = \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}$ *och* $\operatorname{cl} X = \{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\}$.

Minkowskifunktionalen är med andra ord *positivt homogen*, *subadditiv* och *Lipschitzkontinuerlig*, och speciellt är den alltså konvex.

Bevis. (i) Positiv homogenitet följer av ekvivalensen $x \in tX \Leftrightarrow \lambda x \in \lambda tX$ för $\lambda > 0$ (och av att $\phi_X(0) = 0$).

För att bevisa subadditiviteten väljer vi, givet $\epsilon > 0$, två positiva tal $s < \phi_X(x) + \epsilon$ och $t < \phi_X(y) + \epsilon$ så att $x \in sX$ och $y \in tX$. På grund av konvexitet är

$$\frac{1}{s+t}(x+y) = \frac{s}{s+t}\,\frac{x}{s} + \frac{t}{s+t}\,\frac{y}{t}$$

en punkt i X, så det följer att punkten $x+y$ ligger i mängden $(s+t)X$. Detta innebär att

$$\phi_X(x+y) \leq s + t < \phi_X(x) + \phi_X(y) + 2\epsilon.$$

Eftersom denna olikhet gäller för alla $\epsilon > 0$, blir slutsatsen att

$$\phi_X(x+y) \leq \phi_X(x) + \phi_X(y).$$

(ii) Som vi redan konstaterat gäller olikheten $\phi_X(x) \leq C\|x\|$ för alla x för konstanten $C = r_0^{-1}$. På grund av subadditiviteten är emellertid

$$\phi_X(x) = \phi_X(x - y + y) \leq \phi_X(x - y) + \phi_X(y),$$

varför

$$\phi_X(x) - \phi_X(y) \leq \phi_X(x - y) \leq C\|x - y\|.$$

Av symmetriskäl är

$$\phi_X(y) - \phi_X(x) \leq C\|y - x\| = C\|x - y\|,$$

vilket tillsammans med den föregående olikheten innebär att olikheten i (ii) gäller.

(iii) Mängderna $\{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}$ och $\{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\}$ är öppna resp. slutna beroende på att funktionen ϕ_X är kontinuerlig. För att bevisa påstående (iii) räcker det därför, på grund av karakteriseringen av $\operatorname{int} X$ som den största öppna mängd som ingår i X och $\operatorname{cl} X$ som den minsta slutna mängd som innehåller X, att visa inklusionerna

$$\operatorname{int} X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\} \subseteq X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\} \subseteq \operatorname{cl} X.$$

Antag att $x \in \operatorname{int} X$. Eftersom $tx \to x$ då $t \to 1$, ligger punkterna tx i det inre av X för alla tal t som ligger tillräckligt nära 1. Det finns därför speciellt ett tal $t_0 > 1$ så att $t_0 x \in X$, dvs. så att $x \in t_0^{-1}X$, vilket innebär att $\phi_X(x) \leq t_0^{-1} < 1$. Detta visar att $\operatorname{int} X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) < 1\}$.

Implikationerna $\phi_X(x) < t \Rightarrow x \in tX \Rightarrow \phi_X(x) \leq t$ är direkta konsekvenser av definitionen av $\phi_X(x)$, och genom att speciellt välja $t = 1$ får vi inklusionerna

$$\{x \in \mathbf{R}^n \mid \phi_X(x) < 1\} \subseteq X \subseteq \{x \in \mathbf{R}^n \mid \phi_X(x) \leq 1\}.$$

För att visa den återstående inklusionen räcker det nu att visa inklusionen

$$\{x \in \mathbf{R}^n \mid \phi_X(x) = 1\} \subseteq \operatorname{cl} X.$$

Antag därför att $\phi_X(x) = 1$; då finns det en följd $(t_n)_1^\infty$ av tal > 1 så att $t_n \to 1$ då $n \to \infty$ och $x \in t_n X$ för alla n. Punkterna $t_n^{-1}x$ ligger därför i X för alla n, och eftersom $t_n^{-1}x \to x$ då $n \to \infty$, är x en punkt i det slutna höljet $\operatorname{cl} X$. □

Övningar

6.1 Ge exempel på två kvasikonvexa funktioner f_1, f_2 vars summa $f_1 + f_2$ inte är kvasikonvex.

6.2 Visa att följande funktioner $f\colon \mathbf{R}^3 \to \mathbf{R}$ är konvexa:

a) $f(x) = x_1^2 + 2x_2^2 + 5x_3^2 + 3x_2x_3$

b) $f(x) = 2x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 + 2x_1x_3$

c) $f(x) = \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2 - x_1} + x_3^2 - 2x_3$.

6.3 För vilka värden på det reella talet a är funktionen

$$f(x) = x_1^2 + 2x_2^2 + ax_3^2 - 2x_1x_2 + 2x_1x_3 - 6x_2x_3$$

konvex resp. strikt konvex?

6.4 Visa att funktionen $f(x) = x_1x_2\cdots x_n$ med $\mathbf{R}_+^n$ som definitionsmängd är kvasikonkav, och att funktionen $g(x) = (x_1x_2\cdots x_n)^{-1}$ med $\mathbf{R}_{++}^n$ som definitionsmängd är konvex.

6.5 Givet $x = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$ låter vi $x_{[k]}$ beteckna den k:te största koordinaten hos vektorn x, dvs. $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ är koordinaterna i avtagande ordning. Visa att för varje k är funktionen $f(x) = \sum_{i=1}^k x_{[i]}$ konvex.

6.6 Antag att $f\colon \mathbf{R}_+ \to \mathbf{R}$ är konvex. Bevisa att

$$f(x_1) + f(x_2) + \cdots + f(x_n) \le f(x_1 + x_2 + \cdots + x_n) + (n-1)f(0)$$

för alla x_1, x_2, …, $x_n \ge 0$. Notera specialfallet $f(0) = 0$!

6.7 Funktionen f är definierad på en konvex delmängd av $\mathbf{R}^n$. För varje $c \in \mathbf{R}^n$ är vidare funktionen $f(x) + \langle c, x\rangle$ kvasikonvex. Visa att f är konvex.

6.8 Vi har härlett korollarium 6.2.7 ur sats 6.2.6. Visa omvänt att sats 6.2.6 enkelt följer ur korollarium 6.2.7.

6.9 X är en konvex mängd i $\mathbf{R}^n$ med icke-tomt inre och $f\colon X \to \mathbf{R}$ är en kontinuerlig funktion. Visa att funktionen f är konvex om f:s restriktion till $\operatorname{int} X$ är konvex.

6.10 Antag att funktionen $f\colon X \to \overline{\mathbf{R}}$ är konvex. Visa att

$$\inf\{f(x) \mid x \in X\} = \inf\{f(x) \mid x \in \operatorname{rint}(\operatorname{dom} f)\}.$$

6.11 Använd tekniken i exempel 6.4.1 för att bestämma minimum av funktionen

$$g(x_1, x_2) = 16x_1 + 2x_2 + x_1^{-1}x_2^{-2}$$

över mängden $x_1 > 0$, $x_2 > 0$.

6.12 Bestäm Minkowskifunktionalen till

a) slutna enhetsbollen $\overline{B}(0;1)$ i $\mathbf{R}^n$ med avseende på ℓ^p-normen $\|\cdot\|_p$;

b) halvrummet $\{x \in \mathbf{R}^n \mid x_1 \leq 1\}$.

6.13 Låt X vara en konvex mängd med 0 som inre punkt och antag att mängden är symmetrisk kring 0, dvs. att $x \in X \Rightarrow -x \in X$. Visa att Minkowskifunktionalen ϕ_X är en norm, dvs. att

(i) $\phi_X(x+y) \leq \phi_X(x) + \phi_X(y)$;

(ii) $\phi_X(\lambda x) = \lambda\phi_X(x)$ för alla $\lambda \in \mathbf{R}$;

(iii) $\phi_X(x) = 0 \Leftrightarrow x = 0$.

Kapitel 7

Släta konvexa funktioner

I det här kapitlet skall vi studer släta konvexa funktioner, dvs. konvexa funktioner som är en differentierbara. Deriverbarhet i en punkt förutsätter att funktionen ifråga är definierad och ändlig i en omgivning av punkten. Det är därför bara meningsfullt att studera differentierbarhetsegenskaper för inre punkter i funktionens definitionsmängd. Genom att övergå till funktionens restriktion till det inre av definitionsmängden kan vi därför lika gärna från början anta att definitionsmängden är öppen. Av den anledningen förutsätts i det här kapitlet alla definitionsmängder vara öppna och alla funktionsvärden vara ändliga.

7.1 Konvexa funktioner på R

Vi påminner om att med *högerderivatan* $f'_+(x)$ och *vänsterderivatan* $f'_-(x)$ i punkten $x \in \mathbf{R}$ av en reellvärd funktion f, som är definierad i en omgivning av punkten, menas de enkelsidiga gränsvärdena

$$f'_+(x) = \lim_{t\to 0+} \frac{f(x+t)-f(x)}{t} \quad \text{och} \quad f'_-(x) = \lim_{t\to 0-} \frac{f(x+t)-f(x)}{t}.$$

Funktionen är deriverbar i punkten x om och endast om höger- och vänsterderivatorna båda existerar och är lika, och derivatan $f'(x)$ är i så fall lika med det gemensamma värdet.

Vänsterderivatan av funktionen $f\colon I \to \mathbf{R}$ kan skrivas som en högerderivata med hjälp av funktionen $\check{f}$, definierad som

$$\check{f}(x) = f(-x) \quad \text{för alla } x \in -I.$$

Vi har nämligen

$$f'_-(x) = \lim_{t\to 0+} \frac{f(x-t)-f(x)}{-t} = -\lim_{t\to 0+} \frac{\check{f}(-x+t)-\check{f}(-x)}{t},$$

dvs.

$$f'_-(x) = -\check{f}'_+(-x).$$

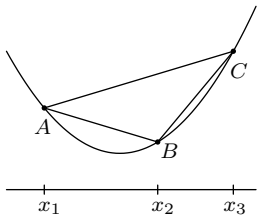
Notera vidare att funktionen $\check{f}$ är konvex om f är konvex.

De grundläggande deriverbarhetsegenskaperna för konvexa funktioner av en variabel följer av följande lemma, som har en uppenbar tolkning i termer av olika kordors lutning. Se figur 7.1.

Lemma 7.1.1. *Antag att f är en reellvärd konvex funktion som är definierad på ett delintervall av $\mathbf{R}$ som innehåller punkterna $x_1 < x_2 < x_3$. Då är*

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_1)}{x_3 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}.$$

Om funktionen är strikt konvex, så är olikheterna ovan strikta.



Figur 7.1. Den geometriska tolkningen av lemma 7.1.1 uttryckt med hjälp av kordornas lutning: $k_{AB} \leq k_{AC} \leq k_{BC}$.

Bevis. Skriv $x_2 = \lambda x_3 + (1 - \lambda)x_1$; då är $\lambda = \dfrac{x_2 - x_1}{x_3 - x_1}$ ett tal i intervallet $]0, 1[$, så konvexitet ger

$$f(x_2) \leq \lambda f(x_3) + (1 - \lambda)f(x_1),$$

vilket förenklat blir $f(x_2) - f(x_1) \leq \lambda(f(x_3) - f(x_1))$, och detta är den vänstra av olikheterna i lemmat.

Eftersom funktionen $\check{f}$ också är konvex, och $-x_3 < -x_2 < -x_1$, medför den redan bevisade olikheten tillämpad på funktionen $\check{f}$ att

$$\frac{f(x_2) - f(x_3)}{x_3 - x_2} = \frac{\check{f}(-x_2) - \check{f}(-x_3)}{-x_2 - (-x_3)} \leq \frac{\check{f}(-x_1) - \check{f}(-x_3)}{-x_1 - (-x_3)} = \frac{f(x_1) - f(x_3)}{x_3 - x_1},$$

som efter multiplikation med -1 ger den högra av olikheterna i lemmat.

Om f är strikt konvex, så är olikheterna ovan strikta. □

Konvexa envariabelfunktioners deriverbarhetsegenskaper ges av följande sats.

Sats 7.1.2. *Antag att funktionen $f\colon I \to \mathbf{R}$ är konvex, där I är ett öppet delintervall av $\mathbf{R}$. Då gäller:*

(a) Funktionen f är höger- och vänsterderiverbar överallt, och för alla $x \in I$ är $f'_-(x) \le f'_+(x)$.

(b) Om $f'_-(x) \le a \le f'_+(x)$, så är

$$f(y) \ge f(x) + a(y - x) \qquad \text{för alla } y \in I.$$

(Om funktionen är strikt konvex, så är olikheten ovan strikt för $y \neq x$.)

(c) Om $x < y$, så är $f'_+(x) \le f'_-(y)$ (med strikt olikhet om f är strikt konvex).

(d) Funktionerna $f'_+\colon I \to \mathbf{R}$ och $f'_-\colon I \to \mathbf{R}$ är växande, och de är strikt växande om f är strikt konvex.

(e) Mängden av $x \in I$ där funktionen f inte är deriverbar, är ändlig eller uppräknelig.

Bevis. Fixera $x \in I$ och sätt

$$F(t) = \frac{f(x+t) - f(x)}{t}.$$

Funktionen F:s definitionsmängd J_x är öppet intervall med punkten 0 borttagen.

Vi börjar med att notera att om $s, t, u \in J_x$ och $u < 0 < t < s$ så är

$$F(u) \le F(t) \le F(s) \tag{7.1}$$

(och olikheterna är strikta om f är strikt konvex).

Den högra olikheten $F(t) \le F(s)$ följer direkt av den vänstra olikheten i lemma 7.1.1 genom att välja $x_1 = x$, $x_2 = x + t$ och $x_3 = x + s$, och den vänstra olikheten $F(u) \le F(t)$ följer av olikheten mellan ytterleden i samma lemma genom att istället välja $x_1 = x + u$, $x_2 = x$ och $x_3 = x + t$.

Av olikheten (7.1) följer att funktionen $F(t)$ är växande för $t > 0$ (strikt växande om f är strikt konvex) och nedåt begränsad av $F(u_0)$, där u_0 är ett godtyckligt negativt tal i F:s definitionsmängd. Följaktligen existerar

$$f'_+(x) = \lim_{t \to 0+} F(t),$$

och för alla $t > 0$ i F:s definitionsmängd är

$$F(t) \ge f'_+(x)$$

(med strikt olikhet om f är strikt konvex). Genom att i denna olikhet ersätta t med $y - x$, ser vi att för $a \leq f'_+(x)$ gäller implikationen

$$y > x \Rightarrow f(y) - f(x) \geq f'_+(x)(y - x) \geq a(y - x) \tag{7.2}$$

(med strikt olikheten om f är strikt konvex).

Samma resonemang applicerat på funktionen $\check{f}$ och punkten $-x$ visar att $\check{f}'_+(-x)$ existerar, och att

$$-y > -x \Rightarrow \check{f}(-y) - \check{f}(-x) \geq -a(-y - (-x))$$

om $-a \leq \check{f}'_+(-x)$. Eftersom $f'_-(x) = -\check{f}'_+(-x)$, betyder detta i sin tur att vänsterderivatan $f'_-(x)$ existerar och att

$$y < x \Rightarrow f(y) - f(x) \geq a(y - x) \tag{7.3}$$

för alla konstanter a som uppfyller $a \geq f'_-(x)$. Båda implikationerna (7.2) och (7.3) är uppfyllda om $f'_-(x) \leq a \leq f'_+(x)$, vilket bevisar påstående (b).

Av olikheten (7.1), som speciellt ger att $F(-t) \leq F(t)$ för små positiva t, följer nu

$$f'_-(x) = \lim_{t \to 0+} F(-t) \leq \lim_{t \to 0+} F(t) = f'_+(x),$$

och därmed är också påstående (a) fullständigt bevisat.

Påstående (b) ger oss nu de båda olikheterna

$$\begin{aligned} f(y) - f(x) &\geq f'_+(x)(y - x) \qquad \text{och} \\ f(x) - f(y) &\geq f'_-(y)(x - y), \end{aligned}$$

vilket efter division med $y - x$ resulterar i implikationen

$$y > x \Rightarrow f'_+(x) \leq \frac{f(y) - f(x)}{y - x} \leq f'_-(y).$$

(Om f är strikt konvex gäller strikt olikhet på båda ställena.) Därmed är också påstående (c) bevisat.

Genom att kombinera (c) med olikheten i (a) fås implikationen

$$x < y \Rightarrow f'_+(x) \leq f'_-(y) \leq f'_+(y),$$

som visar att högerderivatan f'_+ är växande. Analogt visas att vänsterderivatan är växande. (I det strikt konvexa fallet blir olikheten ovan strikt med slutsatsen att derivatorna är strikt växande.)

För att visa påstående (e) sätter vi $I_x =]f'_-(x), f'_+(x)[$. Det öppna intervallet I_x är tomt om derivatan $f'(x)$ existerar och icke-tomt om derivatan

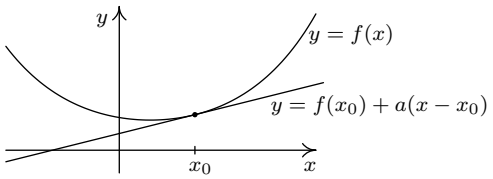
inte existerar, och intervall I_x och I_y som hör till olika punkter x och y är disjunkta på grund av påstående (c). Välj nu för varje punkt x där derivatan inte existerar ett rationellt tal r_x i intervallet I_x. Eftersom intervallen är parvis disjunkta, blir de valda talen garanterat skilda, och eftersom mängden av rationella tal är uppräknelig, kan det därför finnas högst uppräkneligt många punkter x där derivatan saknas. □

Definition. Linjen $y = f(x_0) + a(x - x_0)$ kallas en *stödlinje* till funktionen $f\colon I \to \mathbf{R}$ i punkten $x_0 \in I$ om

$$f(x) \geq f(x_0) + a(x - x_0) \tag{7.4}$$

för alla $x \in I$.

En stödlinje i punkten x_0 är en linje genom punkten $(x_0, f(x_0))$ som ligger under eller på på funktionskurvan $y = f(x)$. En stödlinje är med andra ord ett (endimensionellt) stödhyperplan till f:s epigraf. I nästa kapitel kommer vi att generalisera begreppet för funktioner av flera variabler.



Figur 7.2. Stödlinje.

Påstående (b) i föregående sats visar att en konvex reellvärd funktion med ett öppet intervall som definitionsmängd har en stödlinje i varje punkt i definitionsmängden, samt att tangenten är stödlinje i de punkter där derivatan existerar. Nästa sats visar att existensen av stödlinjer till en funktion också är ett tillräckligt villkor för konvexitet.

Sats 7.1.3. *Antag att funktionen $f\colon I \to \mathbf{R}$, där I är ett öppet intervall, har en stödlinje i varje punkt i intervallet I. Då är funktionen f konvex.*

Bevis. Antag att $x, y \in I$ och $0 < \lambda < 1$, och låt a vara den konstant som hör till punkten $x_0 = \lambda x + (1 - \lambda)y$ i definitionen (7.4) av stödlinje. Då är dels $f(x) \geq f(x_0) + a(x - x_0)$, dels $f(y) \geq f(x_0) + a(y - x_0)$. Multiplicera den första olikheten med λ och den andra med $(1 - \lambda)$, samt addera; detta ger

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(x_0) + a\big(\lambda x + (1 - \lambda)y - x_0\big) = f(x_0).$$

Funktionen f är således konvex.

Notera också att om olikheten (7.4) är strikt för alla $x \neq x_0$ och för alla punkter x_0 i funktionens definitionsmängd, så är f strikt konvex. □

För deriverbara funktioner får vi nu följande nödvändiga och tillräckliga villkor för konvexitet.

Sats 7.1.4. *En deriverbar funktion $f\colon I \to \mathbf{R}$ är konvex om och endast om derivatan f' är växande. Den är strikt konvex om och endast om derivatan är strikt växande.*

Bevis. Påstående (d) i sats 7.1.2 visar att derivatan till en (strikt) konvex funktion är (strikt) växande.

Antag omvänt att derivatan f' är växande. Det följer då av medelvärdessatsen att för $x \neq x_0 \in I$ är

$$\frac{f(x)-f(x_0)}{x-x_0} = f'(\xi) \begin{cases} \geq f'(x_0) & \text{om } x > x_0, \\ \leq f'(x_0) & \text{om } x < x_0, \end{cases}$$

beroende på att ξ är någon punkt mellan x_0 och x. Genom att multiplicera med $x - x_0$ får vi i båda fallen olikheten

$$f(x) - f(x_0) \geq f'(x_0)(x - x_0),$$

som visar att $y = f(x_0) + f'(x_0)(x - x_0)$ är en stödlinje till funktionen i punkten x_0. Det följer därför av sats 7.1.3 att funktionen f är konvex. För strikt växande derivator erhåller vi strikta olikheter, så funktionen f är strikt konvex i det fallet. □

För två gånger deriverbara funktioner får vi följande korollarium till föregående sats.

Korollarium 7.1.5. *En två gånger deriverbar funktion $f\colon I \to \mathbf{R}$ är konvex om och endast om $f''(x) \geq 0$ för alla $x \in I$. Funktionen är strikt konvex om $f''(x) > 0$ för alla $x \in I$.*

Bevis. Om andraderivatan f'' är icke-negativ (resp. positiv), så är derivatan f' växande (resp. strikt växande), och om derivatan f' är växande, så är andraderivatan icke-negativ överallt. □

Anmärkning. En kontinuerlig funktion $f\colon J \to \mathbf{R}$ med ett icke-öppet intervall J som definitionsmängd är konvex om (och endast om) funktionens restriktion till det inre av J är konvex. Om derivatan existerar och är växande i det inre av J, eller om andraderivatan existerar och $f''(x) \geq 0$ för alla inre punkter i intervallet, så är således funktionen konvex på J. Se övning 7.7.

EXEMPEL 7.1.1. Envariabelfunktionerna $x \mapsto \mathrm{e}^x$, $x \mapsto -\ln x$ och $x \mapsto x^p$, där $p > 1$, är strikt konvexa på sina respektive definitionsintervall $\mathbf{R}$, $]0, \infty[$ och $[0, \infty[$, ty deras förstaderivator är strikt växande. □

7.2 Differentierbara konvexa funktioner

En deriverbar envariabelfunktion f är konvex om och endast om derivatan är växande. För att kunna generalisera detta resultat till funktioner av flera variabler behöver vi kunna uttrycka att derivatan är växande på ett generaliserbart sätt. Vi noterar därför att derivatan f' är växande på ett intervall om och endast om $(f'(y) - f'(x)) \cdot (y - x) \geq 0$ för alla tal x och y i intervallet, och denna olikhet är meningsfull också för funktioner av flera variabler om vi tolkar $\cdot$ som en skalärprodukt. Olikheten som generaliserar att derivatan av en flervariabelfunktion f är växande i en öppen konvex mängd X kommer därför att skrivas $\langle f'(y) - f'(x), y - x\rangle \geq 0$ för alla $x, y \in X$, eller ekvivalent med $v = y - x$ som $Df(x + v)[v] \geq Df(x)[v]$ för alla $x, x + v \in X$.

Vi har nu följande karakterisering av konvexitet.

Sats 7.2.1. *Låt X vara en öppen konvex delmängd av $\mathbf{R}^n$, och antag att funktionen $f\colon X \to \mathbf{R}$ är differentierbar. Då är följande tre villkor ekvivalenta:*

(i) Funktionen f är konvex.

(ii) För alla $x, x + v \in X$ är $f(x + v) \geq f(x) + Df(x)[v]$.

(iii) För alla $x, x + v \in X$ är $Df(x + v)[v] \geq Df(x)[v]$.

Funktionen f är strikt konvex om och endast om olikheterna i (ii) resp. (iii) gäller strikt för $v \neq 0$.

Bevis. Betrakta för givna punkter x och $x + v$ i X restriktionen $\phi_{x,v}$ av f till linjen genom x med riktning v, dvs. envariabelfunktionen

$$\phi_{x,v}(t) = f(x + tv)$$

med det öppna intervallet $I_{x,v} = \{t \in \mathbf{R} \mid x + tv \in X\}$ som definitionsmängd. Funktionerna $\phi_{x,v}$ är deriverbara med derivata

$$\phi'_{x,v}(t) = Df(x + tv)[v],$$

och f är konvex om och endast om samtliga restriktioner $\phi_{x,v}$ är konvexa.

(i) $\Rightarrow$ (ii) Om f är konvex, så är alltså funktionerna $\phi_{x,v}$ konvexa, och det följer därför av sats 7.1.2 (b) att $\phi_{x,v}(t) \geq \phi_{x,v}(0) + \phi'_{x,v}(0)t$ för alla $t \in I_{x,v}$, vilket betyder att

$$f(x + tv) \geq f(x) + Df(x)[v]\, t$$

för alla t med $x + tv \in X$, och vi får olikheten i (ii) genom att välja $t = 1$.

(ii) ⇒ (iii) Olikheten i (iii) erhålls genom addition av olikheterna

$$f(x+v) \geq f(x) + Df(x)[v] \quad \text{och} \quad f(x) \geq f(x+v) + Df(x+v)[-v].$$

(iii) ⇒ (i) Antag att (iii) gäller, och sätt $y = x + sv$ och $w = (t-s)v$. För $t > s$ är

$$\begin{aligned}\phi'_{x,v}(t) - \phi'_{x,v}(s) &= Df(x+tv)[v] - Df(x+sv)[v] \\ &= Df(y+w)[v] - Df(y)[v] \\ &= \frac{1}{t-s}\big(Df(y+w)[w] - Df(y)[w]\big) \geq 0,\end{aligned}$$

vilket betyder att derivatan $\phi'_{x,v}$ är växande. Funktionerna $\phi_{x,v}$ är följaktligen konvexa.

Därmed är ekvivalenserna mellan (i), (ii) och (iii) bevisade, och genom att ersätta samtliga olikheter i beviset med strikta olikheter fås motsvarande påståenden för strikt konvexa funktioner. □

I en lokal minimipunkt till en differentierbar funktion är derivatan lika med noll. För konvexa funktioner gäller också omvändningen.

Sats 7.2.2. *Antag att $f\colon X \to \mathbf{R}$ är en differentierbar konvex funktion. Då är $\hat{x} \in X$ en global minimipunkt om och endast om $f'(\hat{x}) = 0$.*

Bevis. Att derivatan är lika med noll i en minimipunkt gäller som sagt generellt, och omvändningen följer av (ii) i föregående sats; om $f'(\hat{x}) = 0$, så är $f(x) \geq f(\hat{x}) + Df(\hat{x})[x - \hat{x}] = f(\hat{x})$ för alla $x \in X$. □

Konvexitet kan också uttryckas med hjälp av villkor på andraderivatan. Envariabelvillkoret $f''(x) \geq 0$ blir för flervariabelfunktioner villkoret att andraderivatan skall vara positivt semidefinit.

Sats 7.2.3. *Låt X vara en öppen, konvex delmängd av $\mathbf{R}^n$, och antag att funktionen $f\colon X \to \mathbf{R}$ är två gånger differentierbar. Då är f konvex om och endast om andraderivatan $f''(x)$ är positivt semidefinit för alla $x \in X$.*

Om $f''(x)$ är positivt definit för alla $x \in X$, så är f strikt konvex.

Bevis. Envariabelfunktionerna $\phi_{x,v}(t) = f(x+tv)$ är nu två gånger deriverbara med andraderivata

$$\phi''_{x,v}(t) = D^2 f(x+tv)[v,v] = \langle v, f''(x+tv)v\rangle.$$

Eftersom funktionen f är konvex om och endast om alla funktionerna $\phi_{x,v}$ är konvexa, är f konvex om och endast om alla andraderivatorna $\phi''_{x,v}$ är icke-negativa funktioner.

Om nu andraderivatan $f''(x)$ är positivt semidefinit för alla $x \in X$, så är $\phi''_{x,v}(t) = \langle v, f''(x+tv)v\rangle \geq 0$ för alla $x \in X$ och alla $v \in \mathbf{R}^n$, vilket innebär att andraderivatorna $\phi''_{x,v}$ är icke-negativa funktioner. Omvänt, om andraderivatorna är icke-negativa, så är speciellt $\langle v, f''(x)v\rangle = \phi''_{x,v}(0) \geq 0$ för alla $x \in X$ och alla $v \in \mathbf{R}^n$, vilket innebär att andraderivatan $f''(x)$ är positivt semidefinit överallt.

Om andraderivatorna $f''(x)$ är positivt definita, så är $\phi''_{x,v}(t) > 0$ för $v \neq 0$, vilket medför att funktionerna $\phi_{x,v}$ är strikt konvexa, och då är också f strikt konvex. □

7.3 Stark konvexitet

Funktionsytan till en konvex funktion "kröker uppåt", men det finns ingen nedre positiv gräns för krökningen. Genom att införa en sådan gräns får vi begreppet stark konvexitet.

Definition. Låt μ vara ett positivt tal. En funktion $f\colon X \to \overline{\mathbf{R}}$ säges vara *μ-starkt konvex* om funktionen $f(x) - \frac{1}{2}\mu\|x\|^2$ är konvex, och funktionen f kallas *starkt konvex* om den är μ-starkt konvex för något positivt tal μ.

Sats 7.3.1. *En differentierbar funktion $f\colon X \to \mathbf{R}$ med konvex definitionsmängd X är μ-starkt konvex om och endast om följande två inbördes ekvivalenta olikheter gäller för alla $x, x+v \in X$:*

(i) $$Df(x+v)[v] \geq Df(x)[v] + \mu\|v\|^2$$

(ii) $$f(x+v) \geq f(x) + Df(x)[v] + \tfrac{1}{2}\mu\|v\|^2.$$

Bevis. Sätt $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$ och notera att $g'(x) = f'(x) - \mu x$ och att följaktligen $Df(x)[v] = Dg(x)[v] + \mu\langle x, v\rangle$.

Om f är μ-starkt konvex, så är funktionen g konvex, och det följer därför av sats 7.2.1 att

$$\begin{aligned} Df(x+v)[v] - Df(x)[v] &= Dg(x+v)[v] - Dg(x)[v] + \mu\langle x+v, v\rangle - \mu\langle x, v\rangle \\ &\geq \mu\langle v, v\rangle = \mu\|v\|^2, \end{aligned}$$

dvs. olikheten (i) gäller.

(i) ⇒ (ii): Antag att (i) gäller, och sätt för $0 \leq t \leq 1$

$$\Phi(t) = f(x+tv) - f(x) - Df(x)[v]\,t.$$

Då är

$$\Phi'(t) = Df(x+tv)[v] - Df(x)[v] = t^{-1}\big(Df(x+tv)[tv] - Df(x)[tv]\big),$$

så det följer av olikheten (i) att

$$\Phi'(t) \geq t^{-1}\mu\|tv\|^2 = \mu\|v\|^2\, t.$$

Genom att integrera den sistnämnda olikheten över intervallet $[0,1]$ erhålles

$$\Phi(1) = \Phi(1) - \Phi(0) \geq \tfrac{1}{2}\mu\|v\|^2,$$

vilket är detsamma som olikheten (ii).

Om olikheten (ii) gäller, så är

$$\begin{aligned} g(x+v) &= f(x+v) - \tfrac{1}{2}\mu\|x+v\|^2 \geq f(x) + Df(x)[v] + \tfrac{1}{2}\mu\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + \tfrac{1}{2}\mu\|x\|^2 + Dg(x)[v] + \mu\langle x, v\rangle + \tfrac{1}{2}\mu\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + Dg(x)[v]. \end{aligned}$$

Funktionen g är därför konvex enligt sats 7.2.1, och $f(x) = g(x) + \frac{1}{2}\mu\|x\|^2$ är följaktligen μ-starkt konvex. □

Sats 7.3.2. *En två gånger differentierbar funktion $f\colon X \to \mathbf{R}$ med konvex definitionsmängd är μ-starkt konvex om och endast om*

$$\langle v, f''(x)v\rangle = D^2 f(x)[v,v] \geq \mu\|v\|^2 \tag{7.5}$$

för alla $x \in X$ och alla $v \in \mathbf{R}^n$.

Anmärkning. För en symmetrisk operator A är

$$\min_{v\neq 0} \frac{\langle v, Av\rangle}{\|v\|^2} = \lambda_{\min},$$

där $\lambda_{\min}$ är operatorns minsta egenvärde. En två gånger differentierbar funktion f med konvex definitionsmängd är således μ-starkt konvex om och endast om egenvärdena till hessianen $f''(x)$ är större än eller lika med μ i alla punkter x.

Bevis. Sätt $\phi_{x,v}(t) = f(x+tv)$. Om villkoret (7.5) gäller, så är

$$\phi''_{x,v}(t) = D^2 f(x+tv)[v,v] \geq \mu\|v\|^2$$

för alla t i funktionens definitionsmängd. Det följer därför av Taylors formel med restterm att

$$\phi_{x,v}(t) = \phi_{x,v}(0) + \phi'_{x,v}(0)t + \tfrac{1}{2}\phi''_{x,v}(\xi)t^2 \geq \phi_{x,v}(0) + \phi'_{x,v}(0)t + \tfrac{1}{2}\mu\|v\|^2\, t^2.$$

För $t = 1$ är detta olikheten (ii) i sats 7.3.1, så funktionen f är μ-starkt konvex.

Omvänt, om f är μ-starkt konvex, så är på grund av (i) i samma sats

$$\frac{\phi'_{x,v}(t) - \phi'_{x,v}(0)}{t} = \frac{Df(x+tv)[tv] - Df(x)[tv]}{t^2} \geq \mu\|v\|^2.$$

Gränsövergång då $t \to 0$ ger $D^2 f(x)[v,v] = \phi''_{x,v}(0) \geq \mu\|v\|^2$. □

7.4 Konvexa funktioner med Lipschitzkontinuerlig derivata

Konvergenshastigheten hos klassiska iterativa algoritmer för minimering av funktioner är beroende av derivatans variation − ju mer derivatan varierar i en omgivning av minimipunkten desto långsammare konvergens, och för funktioner med Lipschitzkontinuerlig derivata är storleken på Lipschitzkonstanten ett mått på variationen. Vi startar därför med ett resultat som för godtyckliga funktioner kopplar ihop Lipschitzkontinuitet hos derivatan med begränsningar på andraderivatan.

Sats 7.4.1. *Antag att f är en två gånger differentierbar funktion och att X är en konvex delmängd av funktionens definitionsmängd.*

(i) Om $\|f''(x)\| \leq L$ för alla $x \in X$, så är derivatan f' Lipschitzkontinuerlig på mängden X med Lipschitzkonstant L.

(ii) Om derivatan f' är Lipschitzkontinuerlig på mängden X med konstant L, så är $\|f''(x)\| \leq L$ för alla $x \in \operatorname{int} X$.

Bevis. (i) Antag att $\|f''(x)\| \leq L$ för alla $x \in X$, och låt x och y vara två godtyckliga punkter i X. Sätt $v = y - x$, låt w vara en godtycklig vektor med $\|w\| = 1$, och definiera funktionen ϕ för $0 \leq t \leq 1$ genom

$$\phi(t) = Df(x + tv)[w] = \langle f'(x + tv, w\rangle.$$

Då är funktionen ϕ deriverbar med derivata

$$\phi'(t) = D^2 f(x + tv)[w, v] = \langle w, f''(x + tv)v\rangle,$$

och det följer av Cauchy–Schwarz olikhet att

$$|\phi'(t)| \leq \|w\|\|f''(x + tv)v\| \leq \|f''(x + tv)\|\|v\| \leq L\|v\|,$$

eftersom $x + tv$ är en punkt i X. Enligt medelvärdessatsen är $\phi(1) - \phi(0) = \phi'(s)$ för någon punkt $s \in]0, 1[$, och följaktligen är

$$|\langle f'(y) - f'(x), w\rangle| = |\phi(1) - \phi(0)| = |\phi'(s)| \leq L\|y - x\|.$$

Eftersom w är en godtycklig vektor av norm 1 drar vi slutsatsen att

$$\|f'(y) - f'(x)\| = \sup_{\|w\|=1} \langle f'(y) - f'(x), w\rangle \leq L\|y - x\|,$$

dvs. derivatan f' är Lipschitzkontinuerlig på mängden X med konstant L.

(ii) Antag omvänt att förstaderivatan f' är Lipschitzkontinuerlig med konstant L på mängden X. Låt x vara en godtycklig punkt i det inre av X, och låt v och w vara godtyckliga vektorer med norm 1. För alla t i en omgivning av 0 är då funktionen

$$\phi(t) = Df(x+tv)[w] = \langle f'(x+tv, w\rangle$$

definierad och deriverbar och ligger punkten $x+tv$ i X. Det följer att

$$\begin{aligned}|\phi(t) - \phi(0)| &= |\langle f'(x+tv) - f'(x), w\rangle| \le \|f'(x+tv) - f'(x)\|\|w\| \\ &\le L\|tv\| = L|t|,\end{aligned}$$

och division med t och gränsövergång ger

$$|\langle w, f''(x)v\rangle| = |\phi'(0)| \le L$$

med slutsatsen att

$$\|f''(x)\| = \sup_{\|v\|=1} \|f''(x)v\| = \sup_{\|v\|,\|w\|=1} \langle w, f''(x)v\rangle \le L. \qquad \square$$

Definition. Låt μ och L vara två positiva tal, och låt X vara en öppen konvex delmängd av $\mathbf{R}^n$. Funktionsklassen $\mathcal{S}_{\mu,L}(X)$ består av alla differentierbara μ-starkt konvexa funktioner $f\colon X \to \mathbf{R}$ vars derivator f' är Lipschitzkontinuerliga med Lipschitzkonstant L. Värdet $Q = L/\mu$ kallas funktionsklassens *konditionstal*.

En differentierbar funktion f med konvex definitionsmängd X tillhör på grund av sats 7.3.1 klassen $\mathcal{S}_{\mu,L}(X)$ om och endast om det båda olikheterna

$$\langle f'(x+v) - f'(x), v\rangle \ge \mu\|v\|^2 \quad \text{och} \quad \|f'(x+v) - f'(x)\| \le L\|v\|$$

gäller för alla $x, x+v \in X$. Av den första olikheten följer med hjälp av Cauchy–Schwarz olikhet att $\mu\|v\| \le \|f'(x+v) - f'(x)\|$. Följaktligen är $\mu \le L$ och $Q \ge 1$.

Exempel 7.4.1. Strikt konvexa kvadratiska funktioner

$$f(x) = \tfrac{1}{2}\langle x, Px\rangle + \langle q, x\rangle + r$$

tillhör klassen $\mathcal{S}_{\lambda_{\min},\lambda_{\max}}(\mathbf{R}^n)$, där $\lambda_{\min}$ och $\lambda_{\max}$ betecknar den positivt definita matrisen P:s minsta respektive största egenvärde.

Eftersom $f'(x) = Px + q$ och $f''(x) = P$, är nämligen

$$\begin{aligned}D^2f(x)[v,v] = \langle v, Pv\rangle &\ge \lambda_{\min}\|v\|^2 \qquad \text{och} \\ \|f'(x+v) - f'(x)\| = \|Pv\| &\le \|P\|\,\|v\| = \lambda_{\max}\|v\|.\end{aligned}$$

Funktionens konditionstal är följaktligen lika med kvoten $\lambda_{\max}/\lambda_{\min}$ mellan det största och det minsta egenvärdet. $\square$

Subnivåmängderna $\{x \mid f(x) \leq \alpha\}$ till en strikt konvex kvadratisk funktion är ellipsoider för alla värden på α som är större än funktionens minimivärde, och förhållandet mellan ellipsoidernas längsta och kortaste axel är lika med $\sqrt{\lambda_{\max}/\lambda_{\min}}$, dvs. roten ur funktionens konditionstal Q. Detta förhållande är uppenbarligen också lika med förhållandet mellan radierna i den minsta boll som innehåller en av dessa ellipsoider och den största boll som innehålles i densamma. Som vi nu ska se gäller något liknande generellt för subnivåmängderna till funktionerna i klassen $\mathcal{S}_{\mu,L}(X)$.

Sats 7.4.2. *Låt f vara en funktion i klassen $\mathcal{S}_{\mu,L}(X)$ med minimipunkt $\hat{x}$, och låt α vara ett tal som är större än funktionens minimivärde $f(\hat{x})$. Då är*

$$B(\hat{x}; r) \subseteq \{x \in X \mid f(x) \leq \alpha\} \subseteq B(\hat{x}; R),$$

där $r = \sqrt{2L^{-1}(\alpha - f(\hat{x}))}$ och $R = \sqrt{2\mu^{-1}(\alpha - f(\hat{x}))}$.

Anmärkning. Notera att $R/r = \sqrt{L/\mu} = \sqrt{Q}$.

Bevis. Eftersom $f'(\hat{x}) = 0$ får vi följande olikheter av satserna 1.1.2 och 7.3.1 (genom att byta x mot $\hat{x}$ och v mot $x - \hat{v}$)

$$f(\hat{x}) + \tfrac{1}{2}\mu\|x - \hat{x}\|^2 \leq f(x) \leq f(\hat{x}) + \tfrac{1}{2}L\|x - \hat{x}\|^2.$$

För $x \in S = \{x \in X \mid f(x) \leq \alpha\}$ är därför

$$\tfrac{1}{2}\mu\|x - \hat{x}\|^2 \leq f(x) - f(\hat{x}) \leq \alpha - f(\hat{x}) = \tfrac{1}{2}\mu R^2,$$

vilket innebär att $\|x - \hat{x}\| \leq R$ och bevisar inklusionen $S \subseteq B(\hat{x}; R)$.

För $x \in B(\hat{x}; r)$ är istället $f(x) \leq f(\hat{x}) + \frac{1}{2}Lr^2 = \alpha$, vilket betyder att $x \in S$ och bevisar inklusionen $B(\hat{x}; r) \subseteq S$. □

Konvexa funktioner på $\mathbf{R}^n$ med Lipschitzkontinuerlig derivata karakteriseras av följande sats.

Sats 7.4.3. *En differentierbar funktion $f\colon \mathbf{R}^n \to \mathbf{R}$ är konvex och dess derivata är Lipschitzkontinuerlig med Lipschitzkonstant L om och endast om följande inbördes ekvivalenta olikheter är uppfyllda för alla $x, v \in \mathbf{R}^n$:*

(i) $$f(x) + Df(x)[v] \leq f(x+v) \leq f(x) + Df(x)[v] + \frac{L}{2}\|v\|^2$$

(ii) $$f(x+v) \geq f(x) + Df(x)[v] + \frac{1}{2L}\|f'(x+v) - f'(x)\|^2$$

(iii) $$Df(x+v)[v] \geq Df(x)[v] + \frac{1}{L}\|f'(x+v) - f'(x)\|^2.$$

Bevis. Att olikheten (i) gäller för konvexa funktioner med Lipschitzkontinuerlig derivata följer av satserna 1.1.2 och 7.2.1.

(i) ⇒ (ii): Sätt $w = f'(x+v) - f'(x)$ och tillämpa den högra olikheten i (i) med x bytt mot $x+v$ och v bytt mot $-L^{-1}w$; detta ger oss olikheten

$$f(x+v-L^{-1}w) \le f(x+v) - L^{-1}Df(x+v)[w] + \tfrac{1}{2}L^{-1}\|w\|^2.$$

Den vänstra olikheten i (i) med $v - L^{-1}w$ istället för v ger

$$f(x+v-L^{-1}w) \ge f(x) + Df(x)[v-L^{-1}w].$$

Genom att kombinera de båda erhållna olikheterna fås

$$\begin{aligned} f(x+v) &\ge f(x) + Df(x)[v-L^{-1}w] + L^{-1}Df(x+v)[w] - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\big(Df(x+v)[w] - Df(x)[w]\big) - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\langle f'(x+v) - f'(x), w\rangle - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + L^{-1}\langle w, w\rangle - \tfrac{1}{2}L^{-1}\|w\|^2 \\ &= f(x) + Df(x)[v] + \tfrac{1}{2}L^{-1}\|w\|^2, \end{aligned}$$

och detta är olikheten i (ii).

(ii) ⇒ (iii): Addera olikheten i (ii) till den olikhet som fås genom att byta x mot $x+v$ och v mot $-v$. Detta ger (iii).

Antag slutligen att olikheten (iii) gäller. Att funktionen f är konvex följer av sats 7.2.1, och genom att kombinera (iii) med Cauchy–Schwarz olikhet får vi olikheten

$$\begin{aligned} \frac{1}{L}\|f'(x+v) - f'(x)\|^2 \le Df(x+v)[v] - Df(x)[v] &= \langle f'(x+v) - f'(x), v\rangle \\ &\le \|f'(x+v) - f'(x)\| \cdot \|v\|, \end{aligned}$$

som efter division med $\|f'(x+v) - f'(x)\|$ ger den önskade slutsatsen att derivatan är Lipschitzkontinuerlig med Lipschitzkonstant L. □

Sats 7.4.4. *Om $f \in \mathcal{S}_{\mu,L}(\mathbf{R}^n)$, så gäller olikheten*

$$Df(x+v)[v] \ge Df(x)[v] + \frac{\mu L}{\mu+L}\|v\|^2 + \frac{1}{\mu+L}\|f'(x+v) - f'(x)\|^2$$

för alla $x, v \in \mathbf{R}^n$.

Bevis. Sätt $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$; då är funktionen g konvex, och eftersom $Dg(x)[v] = Df(x)[v] - \mu\langle x, v\rangle$, följer det av sats 1.1.2 att

$$\begin{aligned} g(x+v) &= f(x+v) - \tfrac{1}{2}\mu\|x+v\|^2 \\ &\le f(x) + Df(x)[v] + \tfrac{1}{2}L\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + \tfrac{1}{2}\mu\|x\|^2 + Dg(x)[v] + \mu\langle x, v\rangle + \tfrac{1}{2}L\|v\|^2 - \tfrac{1}{2}\mu\|x+v\|^2 \\ &= g(x) + Dg(x)[v] + \tfrac{1}{2}(L-\mu)\|v\|^2. \end{aligned}$$

Detta visar att g uppfyller villkoret (i) i sats 7.4.3 med L bytt mot $L-\mu$, så derivatan g' är Lipschitzkontinuerlig med konstant $L-\mu$. Samma sats ger oss därför olikheten

$$Dg(x+v)[v] \ge Dg(x)[v] + \frac{1}{L-\mu}\|g'(x+v) - g'(x)\|^2,$$

som efter förenkling ger olikheten i satsen. □

Övningar

7.1 Visa att följande funktioner är konvexa på de angivna mängderna.

a) $f(x_1, x_2) = \mathrm{e}^{x_1} + \mathrm{e}^{x_2} + x_1x_2, \quad x_1 + x_2 > 0$

b) $f(x_1, x_2) = \sin(x_1 + x_2), \quad -\pi < x_1 + x_2 < 0$

c) $f(x_1, x_2) = -\sqrt{\cos(x_1 + x_2)}, \quad -\frac{\pi}{2} < x_1 + x_2 < \frac{\pi}{2}$.

7.2 Är funktionen

$$f(x_1, x_2) = \frac{x_1^2}{x_2} + \frac{x_2^2}{x_1}$$

konvex i första kvadranten $x_1 > 0$, $x_2 > 0$?

7.3 Visa att funktionen

$$f(x) = \sum_{j=1}^{n-1} x_j^2/x_n$$

är konvex i halvrummet $x_n > 0$.

7.4 Visa att följande funktion f är konkav på mängden $[0,1[\times[0,1[\times[0,1[$:

$$\begin{aligned} f(x_1, x_2, x_3) = \ln(1-x_1) + \ln(1-x_2) + \ln(1-x_3) \\ - (x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3). \end{aligned}$$

7.5 Låt I vara ett intervall och antag att funktionen $f\colon I \to \mathbf{R}$ är konvex. Visa att antingen är f växande på intervallet eller avtagande på intervallet eller också finns det en punkt $c \in I$ så att f är avtagande till vänster om c och växande till höger om c.

7.6 Antag att funktionen $f\colon]a,b[\to \mathbf{R}$ är konvex.

a) Visa att de båda enkelsidiga gränsvärdena $\lim_{x\to a+} f(x)$ och $\lim_{x\to b-} f(x)$ existerar (som ändliga tal eller $\pm\infty$).

b) Visa att om intervallet är ändligt och funktionen utvidgas till det slutna intervallet $[a,b]$ genom definitionen $f(a) = \alpha$ och $f(b) = \beta$, så är den utvidgade funktionen konvex om och endast om $\alpha \geq \lim_{x\to a+} f(x)$ och $\beta \geq \lim_{x\to b-} f(x)$.

7.7 Visa att en kontinuerlig funktion $f\colon [a,b] \to \mathbf{R}$ är konvex om dess restriktion till det öppna intervallet $]a,b[$ är konvex.

7.8 $\mathcal{F}$ är en familj av differentierbara funktioner på $\mathbf{R}^n$ med följande två egenskaper:

(i) $f \in \mathcal{F} \;\Rightarrow\; f + g \in \mathcal{F}$ för alla affina funktioner $g\colon \mathbf{R}^n \to \mathbf{R}$.
(ii) Om $f \in \mathcal{F}$ och $f'(x_0) = 0$, så är x_0 en minimipunkt till f.

Visa att alla funktionerna i $\mathcal{F}$ är konvexa.

7.9 Antag att funktionen $f\colon X \to \mathbf{R}$ är konvex och två gånger differentierbar. Visa att funktionens recessiva delrum V_f är en delmängd av $\mathcal{N}(f''(x))$ för varje $x \in X$.

7.10 Låt $f\colon X \to \mathbf{R}$ vara en differentierbar funktion med konvex definitionsmängd X. Visa att funktionen är kvasikonvex om och endast om

$$f(x+v) \leq f(x) \;\Rightarrow\; Df(x)[v] \leq 0$$

för alla $x, x+v \in X$.
[Ledning: Det räcker att visa påståendet för funktioner på $\mathbf{R}$; det generella resultatet följer sedan genom att ta restriktionen till linjer.]

7.11 Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar funktion med konvex definitionsmängd X. Visa följande påståenden:

a) Om f är kvasikonvex, så gäller implikationen

$$Df(x)[v] = 0 \;\Rightarrow\; D^2f(x)[v,v] \geq 0$$

för alla $x \in X$ och alla $v \in \mathbf{R}^n$.

b) Om

$$Df(x)[v] = 0 \;\Rightarrow\; D^2f(x)[v,v] > 0$$

för alla $x \in X$ och alla $v \neq 0$, så är funktionen f kvasikonvex.

[Ledning: Det räcker att visa resultaten för funktioner definierade på $\mathbf{R}$.]

7.12 Visa att funktionen $\alpha_1 f_1 + \alpha_2 f_2$ är $(\alpha_1\mu_1 + \alpha_2\mu_2)$-starkt konvex om f_1 är μ_1-starkt konvex, f_2 är μ_2-starkt konvex och $\alpha_1, \alpha_2 > 0$.

7.13 Visa att om en differentierbar μ-starkt konvex funktion $f\colon X \to \mathbf{R}$ har minimum i punkten $\hat{x}$, så är $\|x - \hat{x}\| \leq \mu^{-1}\|f'(x)\|$ för alla $x \in X$.

Kapitel 8

Subdifferentialen

I det här kapitlet skall vi generalisera ett antal resultat från förra kapitlet till konvexa funktioner som inte nödvändigtvis är differentierbara överallt. Konvexa reellvärda funktioner kan dock inte vara speciellt oregelbundna; de är som vi redan har sett kontinuerliga i alla inre punkter av definitionsmängden, och de har också riktningsderivator i alla inre punkter.

8.1 Subdifferentialen

För differentierbara funktioner f är $y = f(a) + \langle f'(a), x - a\rangle$ ekvationen för ett hyperplan som tangerar ytan $y = f(x)$ i punkten $(a, f(a))$, det s. k. tangentplanet. Om funktionen är konvex, är $f(x) \geq f(a) + \langle f'(a), x - a\rangle$ för alla x i definitionsmängden (sats 7.2.1), vilket betyder att tangentplanet ligger under funktionsytan och är ett stödhyperplan till epigrafen.

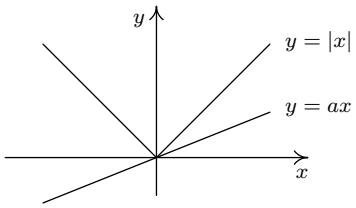
En godtycklig konvex funktions epigraph är per definition konvex, så genom varje randpunkt i epigrafen går det ett stödhyperplan. För konvexa envariabelsfunktioner f, som är definierade på ett öppet intervall, ges stödhyperplanen av sats 7.1.2, som säger att linjen $y = f(x_0)+a(x-x_0)$ stöder epigrafen i punkten $(x_0, f(x_0))$ om (och endast om) $f'_-(x_0) \leq a \leq f'_+(x_0)$.

Eftersom existensen av stödhyperplan karakteriserar konvexitet ska vi nu studera epigrafers stödyperplan närmare.

Definition. Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion som är definierad på en delmängd X av $\mathbf{R}^n$. En vektor $c \in \mathbf{R}^n$ kallas en *subgradient till f i punkten* $a \in X$ om olikheten

$$f(x) \geq f(a) + \langle c, x - a\rangle \tag{8.1}$$

gäller för alla $x \in X$. Mängden av alla subgradienter till f i punkten a kallas *subdifferentialen* till f i a och betecknas $\partial f(a)$.



Figur 8.1. Linjen $y = ax$ är en stödlinje till funktionen $f(x) = |x|$ i origo om $-1 \leq a \leq 1$.

Anmärkning. Om x är en punkt i $X \setminus \operatorname{dom} f$ så är olikheten (8.1) trivialt uppfylld av alla punkter $a \in X$ och alla vektorer $c \in \mathbf{R}^n$. Därför är c en subgradient till f i punkten a om olikheten gäller för alla $x \in \operatorname{dom} f$.

Om a är en punkt i $X \setminus \operatorname{dom} f$ och x är en punkt i $\operatorname{dom} f$, så gäller inte olikheten (8.1) för någon vektor c. Därför är $\partial f(a) = \emptyset$ för alla $a \in X \setminus \operatorname{dom} f$, utom i det triviala fallet $\operatorname{dom} f = \emptyset$, dvs. då f är konstant lika med ∞ på X. Då är istället $\partial f(a) = \mathbf{R}^n$ för alla $a \in X$ eftersom olikheten (8.1) nu är trivialt uppfylld av alla $a, x \in X$ och alla $c \in \mathbf{R}^n$.

Exempel 8.1.1. För funktionen $f\colon \mathbf{R} \to \mathbf{R}$, $f(x) = |x|$, är

$$\partial f(a) = \begin{cases} \{-1\} & \text{om } a < 0, \\ [-1, 1] & \text{om } a = 0, \\ \{1\} & \text{om } a > 0. \end{cases}$$ □

Sats 8.1.1. *En godtycklig funktions subdifferentialer är slutna och konvexa.*

Bevis. För $a \in \operatorname{dom} f$ kan subdifferentialen $\partial f(a)$ till funktionen $f\colon X \to \overline{\mathbf{R}}$ skrivas på formen $\partial f(a) = \bigcap_{x \in \operatorname{dom} f} \{c \in \mathbf{R}^n \mid \langle c, x - a\rangle \leq f(x) - f(a)\}$, dvs. som ett snitt av slutna halvrum, och den är därför en sluten konvex mängd.

För $a \in X \setminus \operatorname{dom} f$ är $\partial f(a) = \emptyset$ eller, om $\operatorname{dom} f = \emptyset$, $\partial f(a) = \mathbf{R}^n$. □

En funktions subdifferential $\partial f(a)$ kan naturligtvis vara tom även för punkter $a \in \operatorname{dom} f$.

Sats 8.1.2. *Punkten $a \in X$ är en minimipunkt till funktionen $f\colon X \to \overline{\mathbf{R}}$ om och endast om $0 \in \partial f(a)$.*

Bevis. Påståendet följer direkt ur subgradientdefinitionen. □

För differentierbara funktioner finns det på grund av följande sats bara en subgradientkandidat, nämligen derivatan, och geometriskt innebär detta att inga andra hyperplan än tangentplanen kan vara stödhyperplan till epigrafen.

Sats 8.1.3. *Antag att funktionen $f\colon X \to \overline{\mathbf{R}}$ är differentierbar i punkten $a \in \operatorname{dom} f$. Då är antingen $\partial f(a) = \{f'(a)\}$ eller $\partial f(a) = \emptyset$.*

Bevis. Antag att $c \in \partial f(a)$. Enligt definitionen av differentierbarhet är

$$f(a+v) - f(a) = \langle f'(a), v\rangle + r(v)$$

med en restterm som uppfyller

$$\lim_{v\to 0} \frac{r(v)}{\|v\|} = 0,$$

och enligt definitionen av subgradient är $f(a+v) - f(a) \geq \langle c, v\rangle$ för alla v för vilka punkten $a+v$ ligger i X. Det följer att

$$\frac{\langle c, v\rangle}{\|v\|} \leq \frac{\langle f'(a), v\rangle + r(v)}{\|v\|} \tag{8.2}$$

för alla v med tillräckligt liten norm $\|v\|$.

För den j:te enhetsvektorn $\mathbf{e}_j$ är $\langle c, \mathbf{e}_j\rangle = c_j$ och $\langle f'(a), \mathbf{e}_j\rangle = \dfrac{\partial f}{\partial x_j}(a)$. Genom att speciellt välja $v = t\mathbf{e}_j$ i olikheten (8.2) och sedan låta $t \to 0$ från höger resp. från vänster får vi därför, eftersom $\|t\mathbf{e}_j\| = |t|$, de två olikheterna

$$c_j \leq \frac{\partial f}{\partial x_j}(a) \quad \text{och} \quad -c_j \leq -\frac{\partial f}{\partial x_j}(a),$$

som innebär att $c_j = \dfrac{\partial f}{\partial x_j}(a)$. Därmed har vi visat att $c = f'(a)$ och att följaktligen $\partial f(a) = \{f'(a)\}$, om subdifferentialen är en icke-tom mängd. □

Sats 7.2.1 innebär att en differentierbar funktion f med öppen konvex definitionsmängd är konvex om och endast om funktionen har en subgradient (som då är lika med $f'(x)$) i varje punkt x i definitionsmängden. Vi skall nu generalisera detta resultat.

Sats 8.1.4. *Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion med konvex definitionsmängd X.*

(a) Om $\operatorname{dom} f$ är en konvex mängd och $\partial f(x) \neq \emptyset$ för alla $x \in \operatorname{dom} f$, så är funktionen f konvex.

(b) Om funktionen är konvex, så är är $\partial f(x) \neq \emptyset$ för alla $x \in \operatorname{rint}(\operatorname{dom} f)$.

Bevis. (a) Låt x och y vara två godtyckliga punkter i $\operatorname{dom} f$ och betrakta punkten $z = \lambda x + (1-\lambda)y$, där $0 < \lambda < 1$. Enligt förutsättningarna har f en subgradient c i punkten z. Genom att använda olikheten (8.1) i punkten $a = z$ två gånger, ena gången med x bytt mot y, erhåller man därför olikheten

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq \lambda\big(f(z) + \langle c, x-z\rangle\big) + (1-\lambda)\big(f(z) + \langle c, y-z\rangle\big)\\ &= f(z) + \langle c, \lambda x + (1-\lambda)y - z\rangle = f(z) + \langle c, 0\rangle = f(z),\end{aligned}$$

som visar att restriktionen av f till $\operatorname{dom} f$ är konvex, och då är förstås f själv konvex.

(b) Antag omvänt att funktionen f är konvex, och låt a vara en punkt i det relativa inre av $\operatorname{dom} f$. Vi skall visa att subdifferentialen $\partial f(a)$ är icke-tom.

Punkten $(a, f(a))$ är en relativ randpunkt till den konvexa mängden $\operatorname{epi} f$. Det finns därför ett stödhyperplan

$$H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \langle c, x - a\rangle + c_{n+1}(x_{n+1} - f(a)) = 0\}$$

till $\operatorname{epi} f$ genom punkten $(a, f(a))$, och vi kan välja normalvektorn (c, c_{n+1}) så att

$$\langle c, x - a\rangle + c_{n+1}(x_{n+1} - f(a)) \geq 0 \tag{8.3}$$

för alla punkter $(x, x_{n+1}) \in \operatorname{epi} f$. Vi skall visa att detta medför att $c_{n+1} > 0$.

Genom att tillämpa olikheten (8.3) på punkten $(a, f(a) + 1)$, som förstås ligger i epigrafen, ser vi för det första att $c_{n+1} \geq 0$.

Antag nu att $c_{n+1} = 0$, och sätt $L = \operatorname{aff}(\operatorname{dom} f)$. Eftersom $\operatorname{epi} f \subseteq L \times \mathbf{R}$ och stödhyperplanet $H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \langle c, x - a\rangle = 0\}$ per definition inte innehåller $\operatorname{epi} f$ som delmängd, innehåller det förstås inte heller $L \times \mathbf{R}$, och vi drar slutsatsen att det finns en punkt $y \in L$ med egenskapen att $\langle c, y - a\rangle \neq 0$. Betrakta nu punkterna $y_\lambda = (1 - \lambda)a + \lambda y$ för $\lambda \in \mathbf{R}$; dessa ligger i den affina mängden L, och $y_\lambda \to a$ då $\lambda \to 0$. Eftersom a är en punkt i det relativa inre av $\operatorname{dom} f$, ligger punkterna y_λ i $\operatorname{dom} f$ för all λ med tillräckligt små belopp. Detta medför att olikheten (8.3) inte kan gälla för samtliga punkter av typen $(y_\lambda, f(y_\lambda))$ i epigrafen, ty uttrycket $\langle c, y_\lambda - a\rangle$ $(= \lambda\langle c, y - a\rangle)$ antar olika tecken beroende på om λ är positivt eller negativt.

Vi har därmed erhållit en motsägelse som visar att $c_{n+1} > 0$ och genom att dividera olikheten (8.3) med c_{n+1} samt sätta $d = -(1/c_{n+1})c$ får vi olikheten $x_{n+1} \geq f(a) + \langle d, x - a\rangle$, som gäller för alla $(x, x_{n+1}) \in \operatorname{epi} f$. Speciellt är alltså $f(x) \geq f(a) + \langle d, x - a\rangle$ för alla $x \in \operatorname{dom} f$, vilket innebär att d är en subgradient till f i punkten a. □

Av sats 8.1.4 följer förstås speciellt att en funktion $f\colon X \to \mathbf{R}$ med en öppen konvex definitionsmängd X är konvex om och endast om $\partial f(x) \neq \emptyset$ för alla $x \in X$.

Sats 8.1.5. *Antag att f är en konvex funktion och att a är en inre punkt i $\operatorname{dom} f$. Då är subdifferentialen $\partial f(a)$ en kompakt icke-tom mängd.*

Bevis. Enligt sats 8.1.1 är subdifferentialen $\partial f(a)$ sluten och enligt sats 8.1.4 är den icke-tom, så det återstår bara att visa att den är begränsad.

Sats 6.6.1 ger två positiva konstanter M och δ med egenskapen att den slutna bollen $\overline{B}(a;\delta)$ ligger i $\operatorname{dom} f$ och

$$|f(x) - f(a)| \le M\|x - a\| \quad \text{för } x \in \overline{B}(a;\delta).$$

Antag nu att $c \in \partial f(a)$ samt att $c \neq 0$. Genom att välja $x = a + \delta c/\|c\|$ i olikheten (8.1) drar vi slutsatsen att

$$\delta\|c\| = \langle c, x - a\rangle \le f(x) - f(a) \le M\|x - a\| = \delta M$$

med begränsningen $\|c\| \le M$ som följd, och detta visar att subdifferentialen är en begränsad mängd. □

Sats 8.1.6. *Antag att $f\colon X \to \overline{\mathbf{R}}$ är en starkt konvex funktion. Då är alla subnivåmängder till f begränsade mängder.*

Bevis. Antag att f är μ-starkt konvex, och låt x_0 vara en relativt inre punkt i $\operatorname{dom} f$. Eftersom funktionen $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$ per definition är konvex, har g enligt sats 8.1.4 en subgradient c i punkten x_0. För alla x i subnivåmängden $S = \{x \in X \mid f(x) \le \alpha\}$ är nu

$$\begin{aligned}\alpha \ge f(x) &= g(x) + \tfrac{1}{2}\mu\|x\|^2 \ge g(x_0) + \langle c, x - x_0\rangle + \tfrac{1}{2}\mu\|x\|^2\\ &= f(x_0) - \tfrac{1}{2}\mu\|x_0\|^2 + \langle c, x - x_0\rangle + \tfrac{1}{2}\mu\|x\|^2\\ &= f(x_0) + \tfrac{1}{2}\mu\big(\|x + \mu^{-1}c\|^2 - \|x_0 + \mu^{-1}c\|^2\big),\end{aligned}$$

vilket innebär att

$$\|x + \mu^{-1}c\|^2 \le \|x_0 + \mu^{-1}c\|^2 + 2\mu^{-1}(\alpha - f(x_0)).$$

Subnivåmängden S är således inkluderad i en sluten boll med centrum i punkten $-\mu^{-1}c$ och radie $R = \sqrt{\|x_0 + \mu^{-1}c\|^2 + 2\mu^{-1}(\alpha - f(x_0))}$. □

Korollarium 8.1.7. *Om en kontinuerlig, starkt konvex funktion har en icke-tom sluten subnivåmängd, så har den en unik minimipunkt.*

Speciellt har varje starkt konvex funktion $f\colon \mathbf{R}^n \to \mathbf{R}$ en unik minimipunkt.

Bevis. Låt f vara en kontinuerlig, starkt konvex funktion med en icke-tom sluten subnivåmängd S. Då är S en kompakt mängd enligt föregående sats, så restriktionen av f till S antar ett minimum i någon punkt i S, och denna punkt är uppenbarligen en global minimipunkt till f. Eftersom starkt konvexa funktioner är strikt konvexa, är minimipunkten unik.

En starkt konvex funktion $f\colon \mathbf{R}^n \to \mathbf{R}$ är automatiskt kontinuerlig, och kontinuerliga funktioner på $\mathbf{R}^n$ är slutna. Alla subnivåmängderna till f är därför slutna, så det följer av den redan bevisade delen av satsen att det finns en unik minimipunkt. □

8.2 Slutna konvexa funktioner

I det här avnittet ska vi använda subdifferentialen för att komplettera resultaten om slutna konvexa funktioner i kapitel 6.8 med några nya resultat. Vi börjar med en alternativ karakterisering av slutna konvexa funktioner.

Sats 8.2.1. *En konvex funktion $f\colon X \to \overline{\mathbf{R}}$ är sluten om och endast om det för alla konvergenta följder $(x_k)_1^\infty$ i $\operatorname{dom} f$ med gränsvärde x_0 gäller att*

$$\varliminf_{k\to\infty} f(x_k) \begin{cases} \geq f(x_0) & \text{om } x_0 \in \operatorname{dom} f, \\ = +\infty & \text{om } x_0 \in \operatorname{cl}(\operatorname{dom} f) \setminus \operatorname{dom} f. \end{cases} \tag{8.4}$$

Bevis. Antag att f är sluten, dvs. att $\operatorname{epi} f$ är en sluten mängd, och låt $(x_k)_1^\infty$ vara en följd i $\operatorname{dom} f$ som konvergerar mot en punkt $x_0 \in \operatorname{cl}(\operatorname{dom} f)$, och sätt

$$L = \varliminf_{k\to\infty} f(x_k).$$

Om a är en godtycklig punkt i det relativa inre av $\operatorname{dom} f$ och c är en subgradient till f i punkten a, så är $f(x_k) \geq f(a) + \langle c, x_k - a\rangle$ för alla k, och eftersom högerledet konvergerar (mot $f(a) + \langle c, x_0 - a\rangle$) då $k \to \infty$, följer det att följden $(f(x_k))_1^\infty$ är nedåt begränsad. Följdens minsta hopningspunkt, dvs. dess undre limes L, är därför ett reellt tal eller $+\infty$.

Olikheten (8.4) är trivialt uppfylld om $L = +\infty$, så antag att L är ett ändligt tal, och låt $(x_{k_j})_{j=1}^\infty$ vara en delföljd till den givna följden med egenskapen att $f(x_{k_j}) \to L$ då $j \to \infty$. Då konvergerar punkterna $(x_{k_j}, f(x_{k_j}))$ i epigrafen mot punkten (x_0, L), och eftersom epigrafen är sluten ligger gränspunkten också i epigrafen, dvs. vi har $x_0 \in \operatorname{dom} f$ och $f(x_0) \leq L$.

Ifall gränspunkten x_0 inte ligger i $\operatorname{dom} f$ utan i $\operatorname{cl}(\operatorname{dom} f) \setminus \operatorname{dom} f$, så är därför nödvändigtvis $L = +\infty$. Därmed har vi visat att (8.4) gäller.

Antag omvänt att (8.4) gäller för alla konvergenta följder. Låt $((x_k, t_k))_1^\infty$ vara en följd av punkter i $\operatorname{epi} f$ som konvergerar mot en punkt (x_0, t_0). Då konvergerar $(x_k)_1^\infty$ mot x_0 och $(t_k)_1^\infty$ mot t_0, och eftersom $f(x_k) \leq t_k$ för alla k är

$$\varliminf_{k\to\infty} f(x_k) \leq \varliminf_{k\to\infty} t_k = t_0.$$

Därför är säkert $\varliminf_{k\to\infty} f(x_k) < +\infty$, så det följer av olikheten (8.4) att dels $x_0 \in \operatorname{dom} f$, dels $f(x_0) \leq t_0$. Detta betyder att gränspunkten (x_0, t_0) ligger i $\operatorname{epi} f$, som därför är en sluten mängd. □

Korollarium 8.2.2. *Antag att funktionen $f\colon X \to \overline{\mathbf{R}}$ är konvex och att $\operatorname{dom} f$ är en relativt öppen mängd. Då är funktionen f sluten om och endast om $\lim_{k\to\infty} f(x_k) = +\infty$ för alla följder $(x_k)_1^\infty$ i $\operatorname{dom} f$ som konvergerar mot en relativ randpunkt till $\operatorname{dom} f$.*

Bevis. Eftersom en konvex funktion är kontinuerlig i alla punkter som ligger i det relativa inre av den effektiva domänen, är i detta fall $\lim_{k\to\infty} f(x_k) = f(x_0)$ för alla följder $(x_k)_1^\infty$ i $\operatorname{dom} f$ som konvergerar mot en punkt x_0 i $\operatorname{dom} f$. Villkoret (8.4) i föregående sats är därför uppfyllt om och endast om $\lim_{k\to\infty} f(x_k) = +\infty$ för alla följder $(x_k)_1^\infty$ i $\operatorname{dom} f$ som konvergerar mot en punkt i $\operatorname{rbdry}(\operatorname{dom} f)$. □

Eftersom affina mängder saknar relativa randpunkter, är speciellt alla reellvärda konvexa funktioner med affina mängder som definitionsmängder slutna (och kontinuerliga).

EXEMPEL 8.2.1. Den konvexa funktionen $f(x) = -\ln x$ med $\mathbf{R}_{++}$ som definitionsmängd är sluten eftersom $\lim\limits_{x\to 0} f(x) = +\infty$. □

Sats 8.2.3. *Om funktionen $f\colon X \to \overline{\mathbf{R}}$ är konvex och sluten, så är*

$$f(x) = \lim_{\lambda\to 1^-} f(\lambda x + (1-\lambda)y)$$

för alla $x, y \in \operatorname{dom} f$.

Bevis. För alla konvexa funktioner f är

$$\overline{\lim_{\lambda\to 1^-}}\, f(\lambda x + (1-\lambda)y) \le \overline{\lim_{\lambda\to 1^-}}\, (\lambda f(x) + (1-\lambda)f(y)) = f(x),$$

medan olikheten

$$\underline{\lim_{\lambda\to 1^-}}\, f(\lambda x + (1-\lambda)y) \ge f(x)$$

gäller för slutna konvexa funktioner på grund av sats 8.2.1. □

Sats 8.2.4. *Antag att f och g är två slutna konvexa funktioner sådana att*

$$\operatorname{rint}(\operatorname{dom} f) = \operatorname{rint}(\operatorname{dom} g)$$

och

$$f(x) = g(x)$$

för alla $x \in \operatorname{rint}(\operatorname{dom} f)$. Då är $f = g$.

Bevis. Vi påminner om att likheten $f = g$ ska tolkas så att $\operatorname{dom} f = \operatorname{dom} g$ och $f(x) = g(x)$ för alla punkter x i den gemensamma effektiva domänen.

Om $\operatorname{rint}(\operatorname{dom} f) = \emptyset$, så är $\operatorname{dom} f = \operatorname{dom} g = \emptyset$, och då finns det ingenting att visa, så antag att x_0 är en punkt i $\operatorname{rint}(\operatorname{dom} f)$. För varje $x \in \operatorname{dom} f$ och $0 < \lambda < 1$ ligger då punkterna $\lambda x + (1-\lambda)x_0$ också i $\operatorname{rint}(\operatorname{dom} f)$, så det följer av våra antaganden och sats 8.2.3 att

$$g(x) = \lim_{\lambda\to 1^-} g(\lambda x + (1-\lambda)x_0) = \lim_{\lambda\to 1^-} f(\lambda x + (1-\lambda)x_0) = f(x),$$

vilket innebär att $g(x) = f(x)$ för alla $x \in \operatorname{dom} f$ och att $\operatorname{dom} f \subseteq \operatorname{dom} g$. Av symmetriskäl gäller också den omvända inklusionen, så $\operatorname{dom} f = \operatorname{dom} g$. □

Sats 8.2.5. *Om $f\colon X \to \overline{\mathbf{R}}$ och $g\colon Y \to \overline{\mathbf{R}}$ är två slutna konvexa funktioner och $X \cap Y \neq \emptyset$, så är summan $f+g\colon X \cap Y \to \overline{\mathbf{R}}$ en sluten konvex funktion.*

Bevis. Satsen följer av karakteriseringen av slutenhet i sats 8.2.1. Låt $(x_k)_1^\infty$ vara en konvergent följd av punkter i $\operatorname{dom}(f+g)$ med gränsvärde x_0. Om x_0 ligger i $\operatorname{dom}(f+g)$ $(= \operatorname{dom} f \cap \operatorname{dom} g)$, så är nämligen

$$\varliminf_{k\to\infty} (f(x_k) + g(x_k)) \geq \varliminf_{k\to\infty} f(x_k) + \varliminf_{k\to\infty} g(x_k) \geq f(x_0) + g(x_0),$$

och om x_0 ligger utanför $\operatorname{dom}(f+g)$, så använder vi den triviala inklusionen

$$\operatorname{cl}(A \cap B) \setminus A \cap B \subseteq (\operatorname{cl} A \setminus A) \cup (\operatorname{cl} B \setminus B),$$

med $A = \operatorname{dom} f$ och $B = \operatorname{dom} g$, och drar slutsatsen att summan $f(x_k)+g(x_k)$ går mot $+\infty$, ty åtminstone en av följderna $(f(x_k))_1^\infty$ och $(g(x_k))_1^\infty$ går mot $+\infty$, och den andra följden har ett ändligt undre limes såvida den inte också går mot $+\infty$. □

Tillslutningen

Definition. Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion som är definierad på en delmängd av $\mathbf{R}^n$ och definiera $(\operatorname{cl} f)(x)$ för $x \in \mathbf{R}^n$ genom att sätta

$$(\operatorname{cl} f)(x) = \inf\{t \mid (x,t) \in \operatorname{cl}(\operatorname{epi} f)\}.$$

Funktionen $\operatorname{cl} f\colon \mathbf{R}^n \to \overline{\underline{\mathbf{R}}}$ kallas *tillslutningen av f*.

Tillslutningen av en konvex funktion har följande egenskaper.

Sats 8.2.6. *Antag att f är en konvex funktion vars effektiva domän är en icke-tom delmängd av $\mathbf{R}^n$. Då gäller:*

(i) *Tillslutningen* $\operatorname{cl} f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ *är en konvex funktion.*
(ii) $\operatorname{dom} f \subseteq \operatorname{dom}(\operatorname{cl} f) \subseteq \operatorname{cl}(\operatorname{dom} f)$.
(iii) $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$.
(iv) $(\operatorname{cl} f)(x) \leq f(x)$ *för alla* $x \in \operatorname{dom} f$.
(v) $(\operatorname{cl} f)(x) = f(x)$ *för alla* $x \in \operatorname{rint}(\operatorname{dom} f)$.
(vi) $\operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$.

Bevis. (i) Låt x_0 vara en godtycklig punkt i $\operatorname{rint}(\operatorname{dom} f)$, och låt c vara en subgradient till f i punkten x_0. Då är $f(x) \geq f(x_0) + \langle c, x - x_0\rangle$ för alla $x \in \operatorname{dom} f$, vilket innebär att epigrafen $\operatorname{epi} f$ ligger i den slutna mängden $K = \{(x,t) \in \operatorname{cl}(\operatorname{dom} f) \times \mathbf{R} \mid \langle c, x - x_0\rangle + f(x_0) \leq t\}$. Det följer att $\operatorname{cl}(\operatorname{epi} f) \subseteq K$, så

$$\begin{aligned}(\operatorname{cl} f)(x) &= \inf\{t \mid (x,t) \in \operatorname{cl}(\operatorname{epi} f)\} \geq \inf\{t \mid (x,t) \in K\} \\ &= f(x_0) + \langle c, x - x_0\rangle > -\infty\end{aligned}$$

för alla $x \in \mathbf{R}^n$. Funktionen $\operatorname{cl} f$ har därför $\overline{\mathbf{R}}$ som målmängd, och att den är konvex följer nu av sats 6.2.6 eftersom mängden $\operatorname{cl}(\operatorname{epi} f)$ är konvex.

(ii), (iv) och (v) Av inklusionen $\operatorname{epi} f \subseteq \operatorname{cl}(\operatorname{epi} f) \subseteq K$ följer att

$$(\operatorname{cl} f)(x) \begin{cases} \leq \inf\{t \mid (x,t) \in \operatorname{epi} f\} = f(x) < +\infty & \text{om } x \in \operatorname{dom} f, \\ \geq \inf\{t \mid (x,t) \in K\} = \inf \emptyset = +\infty & \text{om } x \notin \operatorname{cl}(\operatorname{dom} f), \end{cases}$$

$$(\operatorname{cl} f)(x_0) \geq \inf\{t \mid (x_0,t) \in K\} = f(x_0).$$

Detta visar att $\operatorname{dom} f \subseteq \operatorname{dom}(\operatorname{cl} f) \subseteq \operatorname{cl}(\operatorname{dom} f)$, att $(\operatorname{cl} f)(x) \leq f(x)$ för alla $x \in \operatorname{dom} f$, samt att $(\operatorname{cl} f)(x_0) = f(x_0)$, och eftersom x_0 är en godtycklig punkt i $\operatorname{rint}(\operatorname{dom} f)$ drar vi slutsatsen att $(\operatorname{cl} f)(x) = f(x)$ för alla $x \in \operatorname{rint}(\operatorname{dom} f)$.

(iii) Eftersom $\operatorname{rint}(\operatorname{cl} X) = \operatorname{rint} X$ för godtyckliga konvexa mängder X, följer det speciellt av (ii) att

$$\operatorname{rint}(\operatorname{dom} f) \subseteq \operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) \subseteq \operatorname{rint}(\operatorname{cl}(\operatorname{dom} f)) = \operatorname{rint}(\operatorname{dom} f),$$

med slutsatsen att $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$.

(vi) Implikationerna

$$(x,t) \in \operatorname{cl}(\operatorname{epi} f) \Rightarrow (\operatorname{cl} f)(x) \leq t \Rightarrow (x,t) \in \operatorname{epi}(\operatorname{cl} f)$$

följer omedelbart av definitionerna av $\operatorname{cl} f$ och epigraf. Antag omvänt att (x,t) är en punkt i $\operatorname{epi}(\operatorname{cl} f)$, dvs. att $(\operatorname{cl} f)(x) \leq t$, och låt $U \times I$ vara en öppen omgivning av (x,t). I omgivningen I av t finns det då ett tal s sådant att $(x,s) \in \operatorname{cl}(\operatorname{epi} f)$, och eftersom $U \times I$ också är en öppen omgivning av (x,s), är $\operatorname{epi} f \cap (U \times I) \neq \emptyset$. Detta visar att $(x,t) \in \operatorname{cl}(\operatorname{epi} f)$, och således gäller också implikationen

$$(x,t) \in \operatorname{epi}(\operatorname{cl} f) \Rightarrow (x,t) \in \operatorname{cl}(\operatorname{epi} f).$$

Därmed har vi också visat att $\operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$. □

Sats 8.2.7. *Om f är en sluten konvex funktion, så är $\operatorname{cl} f = f$.*

Bevis. Enligt sats 8.2.6 är $\operatorname{rint}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{rint}(\operatorname{dom} f)$ och $(\operatorname{cl} f)(x) = f(x)$ för alla $x \in \operatorname{rint}(\operatorname{dom} f)$. Det följer därför av sats 8.2.4 att $\operatorname{cl} f = f$. □

8.3 Konjugatfunktionen

Definition. Låt $f\colon X \to \overline{\underline{\mathbf{R}}}$ vara en godtycklig funktion som är definierad på en delmängd X av $\mathbf{R}^n$. Funktionen f^*, som är definierad på hela $\mathbf{R}^n$ av att

$$f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in X\}$$

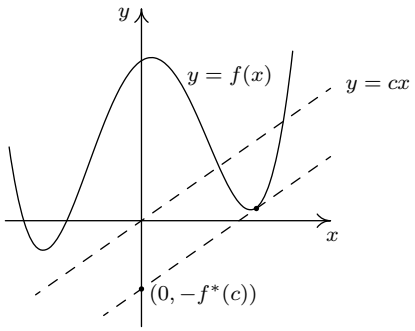
för alla $y \in \mathbf{R}^n$, kallas *konjugatfunktionen* eller *Fencheltransformen* till f.

Vi använder vidare den kortare beteckningen f^{**} för konjugatfunktionen till f^*, dvs. $f^{**} = (f^*)^*$.

För funktioner $f\colon X \to \overline{\mathbf{R}}$ med icke-tom effektiv domän är konjugatfunktionen f^* uppenbarligen en funktion $\mathbf{R}^n \to \overline{\mathbf{R}}$, och

$$f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in \operatorname{dom} f\}.$$

I det triviala fallet då $f\colon X \to \overline{\mathbf{R}}$ är en funktion med tom effektiv domän, är förstås $f^*(y) = -\infty$ för alla $y \in \mathbf{R}^n$, och i det andra extremfallet att $f\colon X \to \overline{\underline{\mathbf{R}}}$ antar värdet $-\infty$ i någon punkt, är $f^*(y) = +\infty$ för alla $y \in \mathbf{R}^n$.



Figur 8.2. Då f är en funktion $\mathbf{R} \to \mathbf{R}$ är konjugatfunktionsvärdet $f^*(c)$ lika med det maximala vertikala avståndet mellan linjen $y = cx$ och funktionskurvan $y = f(x)$. Om funktionen f är deriverbar, är $f^*(c) = cx_0 - f(x_0)$ för någon punkt x_0 där $f'(x_0) = c$.

Exempel 8.3.1. Stödfunktionerna, som definierades i avsnitt 6.9, är exempel på konjugatfunktioner. Definiera, givet en godtycklig delmängd A av $\mathbf{R}^n$, funktionen $\chi_A\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ genom att sätta

$$\chi_A(x) = \begin{cases} 0 & \text{om } x \in A, \\ +\infty & \text{om } x \notin A. \end{cases}$$

Funktionen χ_A kallas *indikatorfunktionen* till mängden A, och den är konvex om mängden A är konvex. Uppenbarligen är

$$\chi_A^*(y) = \sup\{\langle y, x\rangle \mid x \in A\} = S_A(y)$$

för alla $y \in \mathbf{R}^n$, så stödfunktionen till mängden A sammanfaller med konjugatfunktionen χ_A^* till A:s indikatorfunktion. □

Vi är här primärt intresserade av konjugatfunktioner till konvexa funktioner $f\colon X \to \overline{\mathbf{R}}$, men vi startar med några generella resultat.

Sats 8.3.1. *Konjugatfunktionen f^* till en funktion $f\colon X \to \overline{\mathbf{R}}$ med icke-tom effektiv domän är konvex och sluten.*

Bevis. Epigrafen epi f^* består av alla punkter $(y, t) \in \mathbf{R}^n \times \mathbf{R}$ som uppfyller olikheterna $\langle x, y\rangle - t \leq f(x)$ för alla $x \in \operatorname{dom} f$, vilket betyder att den är lika med ett snitt av slutna halvrum i $\mathbf{R}^n \times \mathbf{R}$. Så epi f^* är en sluten konvex mängd, och konjugatfunktionen f^* är därför en sluten konvex funktion. □

Sats 8.3.2 (Fenchels olikhet)**.** *Låt $f\colon X \to \overline{\mathbf{R}}$ vara en funktion med icke-tom effektiv domän. För alla $x \in X$ och alla $y \in \mathbf{R}^n$ är*

$$\langle x, y\rangle \leq f(x) + f^*(y),$$

och likhet gäller för ett givet $x \in \operatorname{dom} f$ om och endast om $y \in \partial f(x)$.

Bevis. Olikheten är trivialt uppfylld för $x \in X \setminus \operatorname{dom} f$, och för $x \in \operatorname{dom} f$ och godtyckligt $y \in \mathbf{R}^n$ följer den omedelbart av definitionen av $f^*(y)$ som supremum. Definitionen av subgradient ger oss vidare för $x \in \operatorname{dom} f$ ekvivalenserna

$$\begin{aligned} y \in \partial f(x) &\Leftrightarrow f(z) - f(x) \geq \langle y, z - x\rangle \quad \text{för alla } z \in \operatorname{dom} f \\ &\Leftrightarrow \langle y, z\rangle - f(z) \leq \langle y, x\rangle - f(x) \quad \text{för alla } z \in \operatorname{dom} f \\ &\Leftrightarrow f^*(y) \leq \langle y, x\rangle - f(x) \\ &\Leftrightarrow f(x) + f^*(y) \leq \langle x, y\rangle, \end{aligned}$$

och genom att kombinera detta med den redan bevisade Fenchels olikhet erhåller vi ekvivalensen $y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = \langle x, y\rangle$. □

För alla punkter y i mängden $\bigcup\{\partial f(x) \mid x \in \operatorname{dom} f\}$ är således

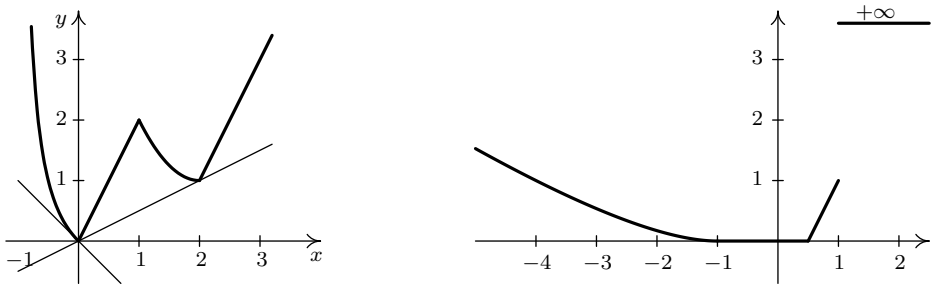
$$f^*(y) = \langle x_y, y\rangle - f(x_y),$$

där x_y är en punkt sådan att $y \in \partial f(x_y)$. För differentierbara funktioner f erhålls punkterna x_y som lösningar till ekvationen $f'(x) = y$. Här följer ett konkret exempel.

EXEMPEL 8.3.2. Låt $f\colon]-1,\infty[\to \mathbf{R}$ vara funktionen

$$f(x) = \begin{cases} -x(x+1)^{-1} & \text{om } -1 < x \le 0, \\ 2x & \text{om } 0 \le x < 1, \\ (x-2)^2 + 1 & \text{om } 1 \le x < 2, \\ 2x - 3 & \text{om } x \ge 2, \end{cases}$$

vars graf visas i den vänstra delen av figur 8.3.



Figur 8.3. Till vänster grafen till en funktion $f\colon]-1,\infty[\to \mathbf{R}$ och till höger grafen till konjugatfunktionen $f^*\colon \mathbf{R} \to \overline{\mathbf{R}}$.

En titt på figuren ger vid handen att funktionskurvan $y = f(x)$ ligger ovanför alla linjer som tangerar kurvan i en punkt (x, y) med $-1 < x < 0$, ovanför alla linjer genom origo med en lutning mellan $f'_-(0) = -1$ och lutningen hos kordan mellan origo och punkten $(2, 1)$ på kurvan, och ovanför alla linjer genom punkten $(2, 1)$ med lutning mellan $\frac{1}{2}$ och $f'_+(2) = 2$. Detta innebär att

$$\begin{aligned} \bigcup\{\partial f(x)\} &= \bigcup_{-1<x<0} \{f'(x)\} \cup \partial f(0) \cup \partial f(2) \cup \bigcup_{x>2} \{f'(x)\} \\ &=]-\infty, -1[\cup [-1, \tfrac{1}{2}] \cup [\tfrac{1}{2}, 2] \cup \{2\} =]-\infty, 2]. \end{aligned}$$

Ekvationen $f'(x) = c$ har för $c < -1$ lösningen $x = -1 + \sqrt{-1/c}$ i intervallet $-1 < x < 0$. Med

$$x_c = \begin{cases} -1 + \sqrt{-1/c} & \text{för } c < -1, \\ 0 & \text{för } -1 \le c \le \frac{1}{2}, \\ 2 & \text{för } \frac{1}{2} \le c \le 2 \end{cases}$$

gäller nu $c \in \partial f(x_c)$, så det följer av påpekandet efter sats 8.3.2 att

$$f^*(c) = cx_c - f(x_c) = \begin{cases} -c - 2\sqrt{-c} + 1 & \text{för } c < -1, \\ 0 & \text{för } -1 \le c \le \frac{1}{2}, \\ 2c - 1 & \text{för } \frac{1}{2} \le c \le 2. \end{cases}$$

För $c > 2$ är uppenbarligen

$$f^*(c) = \sup_{x>-1}\{cx - f(x)\} \ge \sup_{x\ge 2}\{cx - f(x)\} = \sup_{x\ge 2}\{(c-2)x + 3\} = +\infty,$$

så $\operatorname{dom} f^* =]-\infty, 2]$. Grafen till konjugatfunktionen f^* visas i den högra delen av figur 8.3. □

Sats 8.3.3. *Låt $f\colon X \to \overline{\mathbf{R}}$ vara en godtycklig funktion. Då är*

$$f^{**}(x) \le f(x)$$

för alla $x \in X$. Vidare är $f^{**}(x) = f(x)$ om $x \in \operatorname{dom} f$ och $\partial f(x) \neq \emptyset$.

Bevis. Om $f(x) = +\infty$ för alla $x \in X$, så är $f^* \equiv -\infty$ och $f^{**} \equiv +\infty$ enligt anmärkningarna efter definitionen, så olikheten gäller med likhet för alla $x \in X$ i detta triviala fall.

Antag därför att $\operatorname{dom} f \neq \emptyset$. För $x \in X$ är då $\langle x, y\rangle - f^*(y) \le f(x)$ för alla $y \in \operatorname{dom} f^*$ på grund av Fenchels olikhet, så det följer genom supremumbildning att $f^{**}(x) = \sup\{\langle x, y\rangle - f^*(y) \mid y \in \operatorname{dom} f^*\} \le f(x)$.

Om $\partial f(x) \neq \emptyset$, så får vi vidare genom att välja $y \in \partial f(x)$ likhet i Fenchels olikhet, vilket betyder att $f(x) = \langle x, y\rangle - f^*(y) \le f^{**}(x)$ och medför att $f(x) = f^{**}(x)$. □

Eftersom en konvex funktion har en subgradient i alla punkter i det relativa inre av den effektiva domänen, har vi följande omedelbara korollarium till sats 8.3.3.

Korollarium 8.3.4. *Om f är en konvex funktion, så är $f^{**}(x) = f(x)$ för alla x i det relativa inre av* $\operatorname{dom} f$.

Vi ska visa att $f^{**} = \operatorname{cl} f$ för alla konvexa funktioner f, och för den skull behöver vi följande lemma.

Lemma 8.3.5. *Antag att f är en konvex funktion och att (x_0, t_0) är en punkt i $\mathbf{R}^n \times \mathbf{R}$ som inte tillhör* $\operatorname{cl}(\operatorname{epi} f)$. *Då finns det en vektor $c \in \mathbf{R}^n$ och ett reellt tal d så att det "icke-vertikala" hyperplanet*

$$H = \{(x, x_{n+1}) \mid x_{n+1} = \langle c, x\rangle + d\}$$

separerar punkten (x_0, t_0) strikt från $\operatorname{cl}(\operatorname{epi} f)$.

Bevis. Det finns enligt separationssatsen 3.1.3 ett hyperplan

$$H = \{(x, x_{n+1}) \mid c_{n+1}x_{n+1} = \langle c, x\rangle + d\}$$

som separerar punkten strikt från $\operatorname{cl}(\operatorname{epi} f)$. Om $c_{n+1} \neq 0$, så kan vi utan inskränkning anta att $c_{n+1} = 1$, och vi har ingenting att visa. Så antag att $c_{n+1} = 0$, och välj tecken på c och d så att $\langle c, x_0\rangle + d > 0$ och $\langle c, x\rangle + d < 0$ för alla $x \in \operatorname{dom} f$.

Genom att använda oss av subgradienten c' i en punkt i det relativa inre av $\operatorname{dom} f$ får vi en affin funktion $\langle c', x\rangle + d'$ sådan att $f(x) \geq \langle c', x\rangle + d'$ för alla $x \in \operatorname{dom} f$. För $x \in \operatorname{dom} f$ och alla positiva tal λ är

$$f(x) \geq \langle c', x\rangle + d' + \lambda(\langle c, x\rangle + d) = \langle c' + \lambda c, x\rangle + d' + \lambda d,$$

medan

$$\langle c' + \lambda c, x_0\rangle + d' + \lambda d = \langle c', x_0\rangle + d' + \lambda(\langle c, x_0\rangle + d) > t_0$$

för alla tillräckligt stora tal λ. Om λ är tillräckligt stort ligger således epigrafen $\operatorname{epi} f$ ovanför hyperplanet

$$H_\lambda = \{(x, x_{n+1}) \mid x_{n+1} = \langle c' + \lambda c, x\rangle + d' + \lambda d\}.$$

och punkten (x_0, t_0) strikt under samma hyperplan, och vi behöver nu bara parallellförskjuta detta hyperplan en smula nedåt för att erhålla ett icke-vertikalt hyperplan som separerar (x_0, t_0) strikt från $\operatorname{cl}(\operatorname{epi} f)$. □

Lemma 8.3.6. *Om $f\colon X \to \overline{\mathbf{R}}$ är en konvex funktion, så är*

$$\operatorname{rint}(\operatorname{dom} f^{**}) = \operatorname{rint}(\operatorname{dom} f).$$

Bevis. Eftersom $\operatorname{rint}(\operatorname{dom} f) = \operatorname{rint}(\operatorname{cl}(\operatorname{dom} f))$, räcker det att visa inklusionen

$$\operatorname{dom} f \subseteq \operatorname{dom} f^{**} \subseteq \operatorname{cl}(\operatorname{dom} f).$$

Den vänstra inklusionen följer omedelbart av olikheten i sats 8.3.3. För att visa den högra inklusionen antar vi att $x_0 \notin \operatorname{cl}(\operatorname{dom} f)$ och ska visa att detta medför att $x_0 \notin \operatorname{dom} f^{**}$.

Det följer av vårt antagande att punkterna (x_0, t_0) inte tillhör $\operatorname{cl}(\operatorname{epi} f)$ för något tal t_0. Så givet $t_0 \in \mathbf{R}$ finns det enligt föregående lemma ett hyperplan

$$H = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} = \langle c, x\rangle + d\}$$

som separarerar (x_0, t_0) strikt från $\operatorname{cl}(\operatorname{epi} f)$, och detta medför speciellt att $t_0 < \langle c, x_0\rangle + d$ och att $\langle c, x\rangle + d < f(x)$ för alla $x \in \operatorname{dom} f$. Det följer att

$$-d \geq \sup\{\langle c, x\rangle - f(x) \mid x \in \operatorname{dom} f\} = f^*(c),$$

varför

$$t_0 < \langle c, x_0\rangle + d \leq \langle c, x_0\rangle - f^*(c) \leq f^{**}(x_0).$$

Eftersom detta gäller för alla reella tal t_0 är $f^{**}(t_0) = +\infty$, vilket innebär att $x_0 \notin \operatorname{dom} f^{**}$. □

Sats 8.3.7. *Om f är en konvex funktion, så är $f^{**} = \operatorname{cl} f$.*

Bevis. Det följer av lemma 8.3.6 och (iii) i sats 8.2.6 att $\operatorname{rint}(\operatorname{dom} f^{**}) = \operatorname{rint}(\operatorname{dom}(\operatorname{cl} f))$, och av sats 8.3.4 och (v) i sats 8.2.6 att $f^{**}(x) = (\operatorname{cl} f)(x)$ för alla $x \in \operatorname{rint}(\operatorname{dom} f^{**})$. Eftersom de båda funktionerna f^{**} och $\operatorname{cl} f$ är slutna och konvexa, är de därför lika enligt sats 8.2.4. □

Korollarium 8.3.8. *Om f är en sluten konvex funktion, så är $f^{**} = f$.*

8.4 Riktningsderivatan

Definition. Antag att funktionen $f\colon X \to \mathbf{R}$ är definierad i en omgivning av x och låt v vara en godtycklig vektor i $\mathbf{R}^n$. Gränsvärdet

$$f'(x;v) = \lim_{t\to 0+} \frac{f(x+tv) - f(x)}{t}$$

kallas, förutsatt att det existerar, *riktningsderivatan* till f i punkten x med riktningen v.

Om funktionen f är differentierbar i punkten x, så är förstås $f'(x;v) = Df(x)[v]$.

Exempel 8.4.1. För envariabelfunktioner f är tydligen

$$f'(x;v) = \begin{cases} f'_+(x)v & \text{om } v > 0, \\ 0 & \text{om } v = 0, \\ f'_-(v)v & \text{om } v < 0. \end{cases}$$

Riktningsderivatan generaliserar således begreppen vänster- och högerderivata. □

Sats 8.4.1. *Om $f\colon X \to \mathbf{R}$ är en konvex funktion med öppen definitionsmängd, så existerar riktningsderivatan $f'(x;v)$ för alla $x \in X$ och alla riktningar v, och*

$$f(x+v) \geq f(x) + f'(x;v)$$

om $x + v$ ligger i X.

Bevis. Sätt $\phi(t) = f(x + tv)$; då är $f'(x; v) = \phi'_+(0)$, som existerar eftersom konvexa envariabelfunktioner har högerderivator i varje punkt enligt sats 7.1.2. Vidare är

$$\phi(t) \geq \phi(0) + \phi'_+(0)\, t$$

för alla t i funktionens definitionsmängd, och vi får olikheten i satsen genom att välja $t = 1$. □

Sats 8.4.2. *Riktningsderivatan $f'(x; v)$ till en konvex funktion är positivt homogen och konvex i den andra variabeln v, dvs.*

$$f'(x; \alpha v) = \alpha f'(x; v) \quad \textit{om } \alpha \geq 0$$
$$f'(x; \alpha v + (1-\alpha)w) \leq \alpha f'(x; v) + (1-\alpha) f'(x; w) \quad \textit{om } 0 \leq \alpha \leq 1.$$

Bevis. Homogeniteten följer direkt ur definitionen (för godtyckliga funktioner). För konvexa funktioner är vidare

$$\begin{aligned} f(x + t(\alpha v + (1-\alpha)w)) - f(x) &= f(\alpha(x+tv) + (1-\alpha)(x+tw)) - f(x) \\ &\leq \alpha\big(f(x+tv) - f(x)\big) + (1-\alpha)\big(f(x+tw) - f(x)\big). \end{aligned}$$

Division med $t > 0$ och gränsövergång då $t \to 0+$ ger nu den sökta konvexitetsolikheten. □

För konvexa envariabelfunktioner f ges sambandet mellan subgradient och riktningsderivata av sats 7.1.2 − talet c är en subgradient i punkten x om och endast om $f'_-(x) \leq c \leq f'_+(x)$, och subdifferentialen $\partial f(x)$ är med andra ord lika med intervallet $[f'_-(x), f'_+(x)]$.

Låt oss uttrycka detta samband med hjälp av subdifferentialens stödfunktion. Vi påminner då om att stödfunktionen S_X till en mängd X i $\mathbf{R}^n$ definieras av att $S_X(x) = \sup\{\langle y, x\rangle \mid y \in X\}$. I envariabelfallet är

$$\begin{aligned} S_{\partial f(x)}(v) = S_{[f'_-(x), f'_+(x)]}(v) = \max\{f'_+(x)v, f'_-(x)v\} &= \begin{cases} f'_+(x)v & \text{om } v > 0, \\ 0 & \text{om } v = 0, \\ f'_-(x)v & \text{om } v < 0 \end{cases} \\ &= f'(x; v), \end{aligned}$$

och det är denna likhet som vi skall generalisera. Vi behöver för den skull betrakta subgradienterna till funktionen $v \mapsto f'(x; v)$, och denna funktions subdifferential i punkten v_0 kommer att betecknas $\partial_2 f'(x; v_0)$.

Om funktionen $f\colon X \to \mathbf{R}$ är konvex med öppen definitionsmängd X, så är funktionerna $v \mapsto f'(x; v)$ också konvexa enligt föregående sats, och följaktligen är subdifferentialerna $\partial_2 f'(x; v)$ icke-tomma för alla $x \in X$ och alla $v \in \mathbf{R}^n$.

Lemma 8.4.3. *För konvexa funktioner $f\colon X \to \mathbf{R}$ med öppen definitionsmängd X och punkter $x \in X$ gäller:*

(i) $\quad c \in \partial_2 f'(x;0) \Leftrightarrow f'(x;v) \geq \langle c, v\rangle \quad$ *för alla* $v \in \mathbf{R}^n$

(ii) $\quad \partial_2 f'(x;v) \subseteq \partial_2 f'(x;0) \quad$ *för alla* $v \in \mathbf{R}^n$

(iii) $\quad c \in \partial_2 f'(x;v) \Rightarrow f'(x;v) = \langle c, v\rangle$

(iv) $\quad \partial f(x) = \partial_2 f'(x;0).$

Bevis. Ekvivalensen (i) följer direkt av definitionen av subgradient beroende på att $f'(x;0) = 0$.

(ii) och (iii): Antag att $c \in \partial_2 f'(x;v)$ och låt $w \in \mathbf{R}^n$ vara en godtycklig vektor. För $t \geq 0$ får vi då på grund av homogenitet och definitionen av subgradient:

$$tf'(x;w) = f'(x;tw) \geq f'(x;v) + \langle c, tw - v\rangle = f'(x;v) + t\langle c, w\rangle - \langle c, v\rangle,$$

och detta är möjligt för alla $t > 0$ endast om $f'(x;w) \geq \langle c, w\rangle$. Det följer därför av (i) att $c \in \partial_2 f'(x;0)$, och därmed är inklusionen i (ii) bevisad. För $t = 0$ får vi istället den omvända olikheten $f'(x;v) \leq \langle c, v\rangle$, så $f'(x;v) = \langle c, v\rangle$. Därmed är också implikationen (iii) bevisad.

(iv) Antag att $c \in \partial_2 f'(x;0)$. På grund av (i) och sats 8.4.1 är

$$f(y) \geq f(x) + f'(x; y - x) \geq f(x) + \langle c, y - x\rangle$$

för alla $y \in X$, vilket visar att c är en subgradient till f i punkten x och ger oss inklusionen $\partial_2 f'(x;0) \subseteq \partial f(x)$.

Antag omvänt att $c \in \partial f(x)$; då är $f(x+tv) - f(x) \geq \langle c, tv\rangle = t\langle c, v\rangle$ för alla tillräckligt små t. Division med $t > 0$ och gränsövergång då $t \to 0+$ ger olikheten $f'(x;v) \geq \langle c, v\rangle$, och av (i) följer därför $c \in \partial_2 f'(x;0)$. Därmed är också den omvända inklusionen $\partial f(x) \subseteq \partial_2 f'(x;0)$ visad. □

Sats 8.4.4. *Antag att $f\colon X \to \mathbf{R}$ är en konvex funktion med öppen definitionsmängd. Då är*

$$f'(x;v) = S_{\partial f(x)}(v) = \max\{\langle c, v\rangle \mid c \in \partial f(x)\}$$

för alla $x \in X$ och alla $v \in \mathbf{R}^n$.

Bevis. Det följer av (i) och (iv) i föregående lemma att $\langle c, v\rangle \leq f'(x;v)$ för alla $c \in \partial f(x)$, och av (ii), (iii) och (iv) i samma lemma att $\langle c, v\rangle$ antar värdet $f'(x;v)$ för alla subgradienter c i den icke-tomma delmängden $\partial_2 f'(x;v)$ av $\partial f(x)$. □

8.5 Subdifferentieringsregler

Sats 8.5.1. *Antag att funktionerna $f_i\colon X \to \mathbf{R}$ är konvexa med öppen definitionsmängd X och talen α_i är icke-negativa för $i = 1, 2 \ldots, m$, samt definiera*

$$f = \sum_{i=1}^{m} \alpha_i f_i.$$

Då är

$$\partial f(x) = \sum_{i=1}^{m} \alpha_i \partial f_i(x).$$

Bevis. En summa av kompakta, konvexa mängder är kompakt och konvex, så mängden $\sum_{i=1}^{m} \alpha_i \partial f_i(x)$ är liksom mängden $\partial f(x)$ sluten och konvex. Det räcker därför på grund av sats 6.9.2 att visa att de båda mängderna har samma stödfunktion. Men på grund av satserna 8.4.4 och 6.9.1 är

$$S_{\partial f(x)}(v) = f'(x;v) = \sum_{i=1}^{m} \alpha_i f_i'(x;v) = \sum_{i=1}^{m} \alpha_i S_{\partial f_i}(x) = S_{\sum_{i=1}^{m} \alpha_i \partial f_i(x)}(v). \quad \square$$

Sats 8.5.2. *Antag att funktionerna $f_i\colon X \to \mathbf{R}$ är konvexa med öppen definitionsmängd X för $i = 1, 2, \ldots, m$, och definiera*

$$f = \max_{1 \le i \le m} f_i.$$

För $x \in X$ är då

$$\partial f(x) = \operatorname{cvx}\Big(\bigcup_{i \in I(x)} \partial f_i(x) \Big),$$

där $I(x) = \{i \mid f_i(x) = f(x)\}$.

Bevis. Funktionerna f_i är kontinuerliga och $f_j(x) < f(x)$ för alla $j \notin I(x)$. För alla tillräckligt små t är därför

$$f(x+tv) - f(x) = \max_{i \in I(x)} f_i(x+tv) - f(x) = \max_{i \in I(x)} (f_i(x+tv) - f_i(x)),$$

och det följer därför, efter division med t och gränsövergång, att

$$f'(x;v) = \max_{i \in I(x)} f_i'(x;v).$$

Enligt sats 6.9.1 är följaktligen

$$\begin{aligned} S_{\partial f(x)}(v) = f'(x;v) &= \max_{i \in I(x)} f_i'(x;v) = \max_{i \in I(x)} S_{\partial f_i(x)}(v) = S_{\bigcup_{i \in I(x)} \partial f_i(x)}(v) \\ &= S_{\operatorname{cvx}(\bigcup_{i \in I(x)} \partial f_i(x))}(v), \end{aligned}$$

och påståendet i satsen följer nu av sats 6.9.2. $\square$

Nästa sats visar hur sammansättning med affina funktioner påverkar subdifferentialen.

Sats 8.5.3. *Antag att C är en linjär avbildning från $\mathbf{R}^n$ till $\mathbf{R}^m$, att b är en vektor i $\mathbf{R}^m$ och att g är en konvex funktion med öppen definitionsmängd i $\mathbf{R}^m$ och definiera funktionen f genom att sätta $f(x) = g(Cx + b)$. För varje $x \in \operatorname{dom} f$ är då*

$$\partial f(x) = C^T(\partial g(Cx + b)).$$

Bevis. Mängderna $\partial f(x)$ och $C^T(\partial g(Cx + b))$ är konvexa och kompakta, så det räcker därför att visa att de har samma stödfunktioner. För varje vektor v är

$$\begin{aligned} f'(x;v) &= \lim_{t\to 0+} \frac{g(C(x+tv)+b) - g(Cx+b)}{t} \\ &= \lim_{t\to 0+} \frac{g(Cx+b+t\,Cv) - g(Cx+b)}{t} = g'(Cx+b;Cv), \end{aligned}$$

så det följer med hjälp av sats 6.9.1 att

$$S_{\partial f(x)}(v) = f'(x;v) = g'(Cx+b;Cv) = S_{\partial g(Cx+b)}(Cv) = S_{C^T(\partial g(Cx+b))}(v). \quad \square$$

Karush–Kuhn–Tuckers sats

Som tillämpning av räknereglerna för subgradienter visar vi en variant av en sats av Karush–Kuhn–Tucker om minimering av konvexa funktioner med konvexa bivillkor. Vi återkommer med en utförlig behandling av detta tema i kapitlen 10 och 11.

Sats 8.5.4. *Antag att funktionerna f, $g_1, g_2, \ldots, g_m$ är konvexa och definierade på en öppen konvex mängd Ω, och sätt*

$$X = \{x \in \Omega \mid g_i(x) \le 0 \ \textit{för}\ i = 1, 2, \ldots, m.\}$$

Antag vidare att det finns en punkt $\overline{x} \in \Omega$ så att $g_i(\overline{x}) < 0$ för $i = 1, 2, \ldots, m$ (Slaters villkor).

Då är en punkt $\hat{x} \in X$ en minimipunkt till restriktionen $f|_X$ om och endast om det för $i = 1, 2, \ldots, m$ existerar subgradienter $c_i \in \partial g_i(\hat{x})$ och skalärer $\hat{\lambda}_i \in \mathbf{R}_+$ med följande egenskaper:

(i) $$-\sum_{i=1}^{m} \hat{\lambda}_i c_i \in \partial f(\hat{x}) \qquad \textit{och}$$

(ii) $$\hat{\lambda}_i g_i(\hat{x}) = 0 \quad \textit{för}\ i = 1, 2, \ldots, m.$$

Anmärkning. För differentierbara funktioner får villkoret (i) formen:

$$\nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) = 0.$$

Jämför med sats 11.2.1.

Bevis. Låt $\hat{x}$ vara en punkt i X och betrakta den konvexa funktionen

$$h(x) = \max\{f(x) - f(\hat{x}), g_1(x), \ldots, g_m(x)\}$$

med Ω som definitionsmängd. Tydligen är $h(\hat{x}) = 0$. Om vi definierar

$$I(\hat{x}) = \{i \mid g_i(\hat{x}) = 0\},$$

så är $I(\hat{x}) = \{i \mid g_i(\hat{x}) = h(\hat{x})\}$, och det följer därför av sats 8.5.2 att

$$\partial h(\hat{x}) = \operatorname{cvx}\big(\partial f(\hat{x}) \cup \bigcup\{\partial g_i(\hat{x}) \mid i \in I(\hat{x})\}\big).$$

Antag nu att $\hat{x}$ är en minimipunkt till restriktionen $f|_X$. För $x \in X$ är då $h(x) = f(x) - f(\hat{x}) \geq 0$ med likhet för $x = \hat{x}$, och för $x \notin X$ är $h(x) > 0$ beroende på $g_i(x) > 0$ för något i. Följaktligen är $\hat{x}$ en global minimipunkt till h.

Omvänt, om $\hat{x}$ är en global minimipunkt till h, så är $h(x) \geq 0$ för alla $x \in \Omega$, och för $x \in X$ betyder detta att $h(x) = f(x) - f(\hat{x}) \geq 0$, så $\hat{x}$ är också en minimipunkt till restriktionen $f|_X$.

Med hjälp av sats 8.1.2 får vi därför följande ekvivalenser:

$$\begin{aligned}
\hat{x} \text{ är minimipunkt till } f|_X \quad &\Leftrightarrow \quad \hat{x} \text{ är minimipunkt till } h \\
&\Leftrightarrow \quad 0 \in \partial h(\hat{x}) \\
&\Leftrightarrow \quad 0 \in \operatorname{cvx}\big(\partial f(\hat{x}) \cup \bigcup\{\partial g_i(\hat{x}) \mid i \in I(\hat{x})\}\big) \\
&\Leftrightarrow \quad 0 = \lambda_0 c_0 + \sum_{i \in I(\hat{x})} \lambda_i c_i \\
&\Leftrightarrow \quad \lambda_0 c_0 = -\sum_{i \in I(\hat{x})} \lambda_i c_i, \qquad (8.5)
\end{aligned}$$

där $c_0 \in \partial f(\hat{x})$, $c_i \in \partial g_i(\hat{x})$ för $i \in I(\hat{x})$, och skalärerna λ_i är icke-negativa tal med summa 1.

Vi påstår nu att $\lambda_0 > 0$. Antag nämligen motsatsen; då är $\sum_{i \in I(\hat{x})} \lambda_i c_i = 0$ och det följer att

$$\sum_{i \in I(\hat{x})} \lambda_i g_i(\overline{x}) \geq \sum_{i \in I(\hat{x})} \lambda_i \big(g_i(\hat{x}) + \langle c_i, \overline{x} - \hat{x}\rangle\big) = \langle \sum_{i \in I(\hat{x})} \lambda_i c_i, \overline{x} - \hat{x}\rangle = 0,$$

vilket är en motsägelse eftersom $g_i(\overline{x}) < 0$ för alla i och $\lambda_i > 0$ för något $i \in I(\hat{x})$.

Vi kan därför dividera likheten i (8.5) med λ_0, och villkoren (i) och (ii) i satsen blir nu uppfyllda om vi sätter $\hat{\lambda}_i = \lambda_i/\lambda_0$ för $i \in I(\hat{x})$ och $\hat{\lambda}_i = 0$ för övriga i, samt väljer godtyckliga subgradienter $c_i \in \partial g_i(\hat{x})$ för $i \notin I(\hat{x})$. □

Övningar

8.1 Antag att funktionen $f\colon \mathbf{R}^n \to \mathbf{R}$ är starkt konvex. Visa att

$$\lim_{\|x\|\to\infty} f(x) = \infty.$$

8.2 Bestäm $\partial f(-1,1)$ för funktionen $f(x_1,x_2) = \max(|x_1|,|x_2|)$.

8.3 Bestäm subdifferentialen $\partial f(0)$ i origo till funktionen $f\colon \mathbf{R}^n \to \mathbf{R}$ om

a) $f(x) = \|x\|_2$ b) $f(x) = \|x\|_\infty$ c) $f(x) = \|x\|_1$.

8.4 Bestäm konjugatfunktionen till följande funktioner:

a) $f(x) = ax + b$, $\operatorname{dom} f = \mathbf{R}$
b) $f(x) = -\ln x$, $\operatorname{dom} f = \mathbf{R}_{++}$
c) $f(x) = \mathrm{e}^x$, $\operatorname{dom} f = \mathbf{R}$
d) $f(x) = x\ln x$, $\operatorname{dom} f = \mathbf{R}_{++}$
e) $f(x) = 1/x$, $\operatorname{dom} f = \mathbf{R}_{++}$.

8.5 Utnyttja sambandet mellan stödfunktion S_A och indikatorfunktion χ_A och det faktum att $S_A = S_{\mathrm{cl}(\mathrm{cvx}\,A)}$ för att visa korollarium 6.9.3, dvs att

$$\mathrm{cl}(\mathrm{cvx}\,A) = \mathrm{cl}(\mathrm{cvx}\,B) \;\Leftrightarrow\; S_A = S_B.$$

Del II

Optimering – grundläggande teori

Kapitel 9

Optimering

Det latinska ordet *optimum* betyder 'det yppersta'. Det optimala alternativet bland ett antal olika alternativ är det som är bäst i någon mening. I vid bemärkelse är följaktligen *optimering* konsten att bestämma det bästa alternativet.

Optimeringsproblem förekommer inte bara inom olika områden av mänsklig planering, utan även många fenomen i naturen kan förklaras utifrån enkla optimeringsprinciper. Exempel på detta är ljusets utbredning och brytning i olika medier, värmeledning och kemisk jämvikt.

I vardagliga optimeringsproblem är det ofta svårt, för att inte säga omöjligt, att jämföra och värdera olika alternativ på ett meningsfullt sätt. I den här framställningen lämnar vi den svårigheten åt sidan, ty den kan ändå inte lösas med matematiska metoder. Vår utgångspunkt är att alternativen är värderade med hjälp av en funktion, t. ex. en vinst- eller kostnadsfunktion, och att det alternativ som ger det största eller minsta funktionsvärdet är bäst.

I det här kapitlet kommer vi alltså att inleda vårt studium av problemet att minimera eller maximera en funktion över mängder som är givna av ett antal bivillkor.

9.1 Optimeringsproblem

Grundläggande beteckningar

För problemet att minimera en funktion $f\colon \Omega \to \overline{\mathbf{R}}$ över en delmängd X av funktionens definitionsmängd Ω kommer vi att använda beteckningssättet

$$\begin{aligned} &\min \quad f(x) \\ &\text{då} \quad x \in X. \end{aligned}$$

Elementen i mängden X kallas optimeringsproblemets *tillåtna punkter* eller *tillåtna lösningar*. Funktionen f är problemets *målfunktion*.

Observera att vi tillåter ∞ som funktionsvärde hos målfunktionen i ett minimeringsproblem.

Minimeringsproblemets *(optimala) värde* $v_{\min}$ definieras som

$$v_{\min} = \begin{cases} \inf\{f(x) \mid x \in X\} & \text{om } X \neq \emptyset, \\ \infty & \text{om } X = \emptyset. \end{cases}$$

Det optimala värdet är alltså ett reellt tal om målfunktionen är nedåt begränsad och inte identiskt lika med ∞ på mängden X, värdet är $-\infty$ om funktionen inte är nedåt begränsad på X, och värdet är slutligen ∞ om målfunktionen är identiskt lika med ∞ på X eller om $X = \emptyset$.

Vi kommer naturligtvis också att studera maximeringsproblem, och problemet att maximera en funktion $f\colon \Omega \to \underline{\mathbf{R}}$ över X skrivs förstås

$$\begin{array}{ll} \max & f(x) \\ \text{då} & x \in X. \end{array}$$

Maximeringsproblemets *(optimala) värde* $v_{\max}$ definieras som

$$v_{\max} = \begin{cases} \sup\{f(x) \mid x \in X\} & \text{om } X \neq \emptyset, \\ -\infty & \text{om } X = \emptyset. \end{cases}$$

På så sätt blir alltså ett minimerings- eller maximeringsproblems värde alltid definierat som ett reellt tal eller som någon av symbolerna $-\infty$ eller ∞, dvs. som ett element i den utvidgade tallinjen $\overline{\underline{\mathbf{R}}}$. Om värdet är ett reellt tal säger vi att optimeringsproblemet har ett *ändligt* värde.

Vi kallar en tillåten punkt x_0 i ett optimeringsproblem med målfunktion f för en *optimal punkt* eller *optimal lösning* om problemets värde är ändligt och lika med $f(x_0)$. En optimal lösning till ett minimeringsproblem är med andra ord detsamma som en global minimipunkt (med ändligt funktionsvärde). Naturligtvis kan problem med ändliga värden sakna optimala lösningar.

Ur matematisk synvinkel är det ingen principiell skillnad mellan maximeringsproblem och minimeringsproblem, ty mellan de optimala värdena $v_{\max}$ och $v_{\min}$ för problemen

$$\begin{array}{ll} \max & f(x) \\ \text{då} & x \in X \end{array} \qquad \text{resp.} \qquad \begin{array}{ll} \min & -f(x) \\ \text{då} & x \in X \end{array}$$

råder sambandet $v_{\max} = -v_{\min}$, och x_0 är en maximipunkt till f om och endast om x_0 är en minimipunkt till $-f$. Av den anledningen kommer vi oftast att nöja oss med att formulera resultat för minimeringsproblem.

Slutligen en kommentar till varför vi tillåter ∞ och $-\infty$ som funktionsvärden hos målfunktionerna i minimerings- och maximeringsproblem eftersom detta kan tyckas komplicera framställningen. Den viktigaste anledningen är att vi ibland kommer att behöva funktioner som är definierade som punktvisa suprema av en oändlig familj av funktioner, och även om samtliga i familjen ingående funktioner är reellvärda, kan supremumfunktionen anta oändliga värden. Alternativet till att tillåta funktioner med värden på den utvidgade reella tallinjen skulle i så fall vara att inskränka definitionsmängden för dessa supremumfunktioner, och detta blir varken enklare eller elegantare.

Allmänna synpunkter

Det finns några allmänna och kanske helt självklara synpunkter som kan anläggas på varje optimeringsproblem.

Existens av tillåtna punkter

Denna punkt kan förefalla trivial, ty om ett problem saknar tillåtna punkter finns det inte så mycket mer att tillägga. Man skall emellertid komma ihåg att mängden av tillåtna punkter sällan är given explicit. Ofta definieras den i stället av ett system av ekvationer och olikheter, som kanske inte är konsistent. Även om det bakomliggande verkliga problemet har tillåtna punkter, kan förenklingar och defekter i den matematiska modellen leda till att det studerade matematiska problemet saknar tillåtna punkter.

Existens av optimal lösning

En förutsättning för att kunna bestämma den optimala lösningen är förstås att det finns någon sådan. Många teoretiska resultat har formen "Om x_0 är en optimal punkt, så uppfyller x_0 det eller det villkoret", men även om detta vanligtvis begränsar antalet potentiella kandidater, så visar det inte att det finns någon optimal punkt.

Ur praktisk synvinkel är det däremot kanske inte så viktigt att ett problem har en optimal lösning och – om så är fallet – att hitta den exakta optimala lösningen; praktikern nöjer sig ofta med en tillåten lösning som har ett målfunktionsvärde som är tillräckligt bra.

Entydighet

Är den optimala lösningen, om nu en sådan existerar, unik? För tillämparen, som är ute efter den "bästa" lösningen, är svaret förmodligen ointressant, ty hon borde vara nöjd med att ha funnit en bästa lösning även om det finns flera

lika bra. Och om hon nu skulle tycka att någon av de optimala lösningarna är bättre än de andra, så kan vi bara konstatera att optimeringsproblemet inte är riktigt ställt från början eftersom målfunktionen tydligen inte inkluderar allt som krävs för att sortera fram den bästa lösningen.

För teoretikern kan det dock vara av intresse att veta att en optimal lösning är unik – entydighet kan ofta användas för att dra intressanta slutsatser om lösningen.

Parameterberoende och känslighet

Ibland, och i synnerhet i problem som kommer direkt från "verkligheten", innehåller målfunktion och bivillkor parametrar, som bara är givna med en viss noggrannhet och som i värsta fall är mer eller mindre grova uppskattningar. I sådana fall räcker det inte att bestämma den optimala lösningen, utan det är minst lika viktigt att veta hur lösningen ändras då parametrarna ändras. Om en liten störning av någon parameter förändrar den optimala lösningen mycket, finns det anledning att betrakta lösningen med stor skepsis.

Kvalitativa aspekter

Det är naturligtvis bara för en liten klass av optimeringsproblem som man kan ange den optimala lösningen i exakt form, eller där lösningen kan beskrivas med en algoritm som terminerar efter ändligt många iterationer. Den matematiska lösningen på ett optimeringsproblem består ofta av ett antal nödvändiga och/eller tillräckliga villkor, som den optimala lösningen måste uppfylla. I bästa fall kan dessa ligga till grund för användbara numeriska algoritmer, och i andra fall kan de kanske bara användas för kvalitativa utsagor om de optimala lösningarna, vilket emellertid i många situationer kan vara väl så intressant.

Algoritmer

Det finns ingen numerisk algoritm som löser alla optimeringsproblem även om vi begränsar oss till problem av typen att minimera funktioner över mängder som beskrivs av olikheter och likheter. Däremot finns det mycket effektiva numeriska algoritmer för delklasser av optimeringsproblem, och många viktiga tillämpade optimeringproblem råkar tillhöra just dessa klasser. Vi kommer att studera ett par sådana algoritmer i del III och del IV av den här boken.

För optimeringsläran har utvecklingen inom algoritmområdet varit minst lika viktig som hårdvaruutvecklingen inom datorområdet, och mycket av utvecklingen har skett så sent som under de senaste decennierna.

9.2 Klassificering av optimeringsproblem

För att kunna säga någonting vettigt om minimeringsproblemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & x \in X \end{array}$$

måste vi göra diverse antaganden om målfunktionen $f\colon \Omega \to \overline{\mathbf{R}}$ och om mängden X av tillåtna punkter.

Vi kommer alltid att anta att Ω är någon delmängd av $\mathbf{R}^n$ och att mängden X kan uttryckas som lösningsmängd till ett antal olikheter och likheter, dvs. att

$$X = \{x \in \Omega \mid g_1(x) \leq 0,\, \ldots,\, g_p(x) \leq 0,\, g_{p+1}(x) = 0,\, \ldots,\, g_m(x) = 0\},$$

där $g_1, g_2, \ldots, g_m$ är reellvärda funktioner definierade på Ω.

Vi utesluter naturligtvis inte möjligheten att samtliga bivillkor är likheter ($p = 0$) eller att samtliga bivillkor är olikheter ($p = m$) eller att det saknas bivillkor överhuvudtaget ($m = 0$).

Eftersom en likhet $h(x) = 0$ kan ersättas av de två olikheterna $\pm h(x) \leq 0$, skulle vi utan inskränkning generellt kunna anta att samtliga bivillkor är olikheter, men det är praktiskt att kunna formulera resultat för optimeringsproblem med likheter bland bivillkoren utan att först behöva göra sådana omskrivningar.

Om $\hat{x}$ är en tillåten punkt och $g_i(\hat{x}) = 0$, säger man att det i:te bivillkoret är *aktivt* i punkten $\hat{x}$. Samtliga bivillkor i form av likheter ($i \geq p + 1$) är naturligtvis aktiva i samtliga tillåtna punkter.

Villkoret $x \in \Omega$ är (i fallet $\Omega \neq \mathbf{R}^n$) förstås också ett slags bivillkor, men det spelar en annan roll än de övriga bivillkoren. Vi kommer ibland att kalla det för det *implicita* bivillkoret för att särskilja det från de övriga *explicita* bivillkoren. Om Ω ges som lösningsmängd till ett antal olikheter av typen $h_i(x) \leq 0$ och funktionerna h_i, målfunktionen och de explicita bivillkorsfunktionerna är definierade på hela rummet $\mathbf{R}^n$, kan vi naturligtvis inkludera olikheterna $h_i(x) \leq 0$ bland de explicita villkoren och helt stryka det implicita villkoret.

Definitionsmängden Ω kommer ofta att vara underförstådd eller framgå av sammanhanget, och den skrivs då inte ut explicit i problemformuleringen utan optimeringsproblemet (P) skrivs som

$$\begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m. \end{cases} \end{array}$$

Linjär programmering

Problemet att maximera eller minimera en linjär form över en polyeder, som är given i form av ett snitt av slutna halvrum, kallas *linjär programmering*, förkortat LP. Problemet (P) är med andra ord ett LP-problem om målfunktionen f är linjär och mängden X är given som lösningsmängd till ett ändligt antal linjära likheter och olikheter.

Vi kommer att studera LP-problem utförligt i kapitel 12.

Konvex optimering

Minimeringsproblemet

$$\begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

med implicit bivillkor $x \in \Omega$ kallas *konvext*, om mängden Ω är konvex, målfunktionen $f\colon \Omega \to \overline{\mathbf{R}}$ är konvex, och bivillkorsfunktionerna g_i är konvexa för $i = 1, 2, \ldots, p$ och affina för $i = p+1, \ldots, m$.

De affina villkoren $g_{p+1}(x) = 0, \ldots, g_m(x) = 0$ i ett konvext problem kan förstås sammanfattas på formen $Ax = b$, där A är en $(m-p) \times n$-matris.

I ett konvext minimeringsproblem är mängden X av tillåtna punkter konvex, ty omskrivningen

$$X = \bigcap_{i=1}^{p} \{x \in \Omega \mid g_i(x) \leq 0\} \cap \bigcap_{i=p+1}^{m} \{x \mid g_i(x) = 0\}$$

framställer X som ett snitt av subnivåmängder till konvexa funktioner och hyperplan.

Ett maximeringsproblem

$$\begin{array}{ll} \max & f(x) \\ \text{då} & x \in X \end{array}$$

kallas konvext om motsvarande ekvivalenta minimeringsproblem

$$\begin{array}{ll} \min & -f(x) \\ \text{då} & x \in X \end{array}$$

är konvext, vilket betyder att målfunktionen f skall vara konkav.

LP-problem är förstås konvexa optimeringsproblem. Allmänna konvexa optimeringsproblem studeras i kapitel 11.

Konvex kvadratisk programmering

Vi får ett specialfall av konvex optimering om X är en polyeder och målfunktionen f är en summa av en linjär form och en positivt semidefinit kvadratisk form, dvs. har utseendet $f(x) = \langle c, x\rangle + \langle x, Qx\rangle$, där Q är en positivt semidefinit matris. Problemet (P) kallas då för *konvex kvadratisk programmering.* LP-problemen utgör förstås en delklass av de konvext kvadratiska problemen.

Icke-linjär optimering

Med *icke-linjär optimering* menas behandlingen av optimeringsproblem som inte förutsätts vara LP-problem. Eftersom icke-linjär optimering inkluderar nästan allt, finns det förstås ingen generell teori som kan tillämpas på ett godtyckligt icke-linjärt optimeringsproblem.

Om f är en differentierbar funktion och X är ett "hyggligt" område i $\mathbf{R}^n$, kan man naturligtvis använda differentialkalkyl för att studera minimeringsproblemet (P). Vi påminner i detta sammanhang om Lagranges sats, som ger ett nödvändigt villkor för minimum (och maximum) i det fall då

$$X = \{x \in \mathbf{R}^n \mid g_1(x) = g_2(x) = \cdots = g_m(x) = 0\}.$$

En motsvarighet till Lagranges sats för optimeringsproblem med bivillkor i form av olikheter ges i kapitel 10.

Heltalsprogrammering

Om somliga eller samtliga variabler i ett optimeringsproblem bara får anta heltalsvärden, talar man om *heltalsprogrammering.* Problemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & x \in X \cap (\mathbf{Z}^m \times \mathbf{R}^{n-m}) \end{array}$$

där $\langle c, x\rangle$ är en linjär form och X är en polyeder i $\mathbf{R}^n$ kallas *linjär heltalsprogrammering.* Många problem som handlar om flöden i nätverk, t. ex. varudistributionsproblem och maxflödesproblem, är linjära heltalsprogrammeringsproblem och kan lösas med speciella algoritmer.

Simultan optimering

Vi skall slutligen kortfattat diskutera en typ av problem som egentligen inte är optimeringsproblem i den tidigare bemärkelsen.

I många situationer kan en individ genom sitt agerande påverka något utfall utan att för den skull ha full kontroll över situationen. Somliga variabler kan ligga i händerna på andra individer med helt olika önskemål vad utfallet

beträffar, medan andra variabler kan vara styrda av slumpen. Problemet att i någon mening optimera utfallet skulle kunna kallas *simultan optimering*.

Simultan optimering behandlas i *spelteorin*, som är en teori för agerande i konfliktsituationer. Spelteoretiska begrepp och resultat har visat sig vara mycket fruktbara i olika sammanhang, t ex inom matematisk ekonomi.

9.3 Ekvivalenta problemformuleringar

Låt oss informellt kalla två optimeringsproblem *ekvivalenta* om man på ett enkelt sätt kan bestämma en optimal lösning till det ena problemet givet en optimal lösning till det andra och vice versa.

Ett trivialt exempel på ekvivalenta problem som redan nämnts är problemen

$$\begin{array}{ll} \max & f(x) \\ \text{då} & x \in X \end{array} \qquad \text{och} \qquad \begin{array}{ll} \min & -f(x). \\ \text{då} & x \in X \end{array}$$

Vi skall nu beskriva några användbara transformationer som leder till ekvivalenta optimeringsproblem.

Elimination av likheter

Betrakta problemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m. \end{cases} \end{array}$$

Om det är möjligt att lösa systemet av likheter och uttrycka systemets lösningar på formen $x = h(y)$ med hjälp en parameter y som beskriver någon delmängd av $\mathbf{R}^d$, så kan vi eliminera likheterna ur problemet (P) och skriva det som

$$\text{(P}') \qquad \begin{array}{ll} \min & f(h(y)) \\ \text{då} & g_i(h(y)) \le 0, \quad i = 1, 2, \ldots, p \end{array}$$

Om $\hat{y}$ är en optimal lösning till (P$'$), så är förstås $h(\hat{y})$ en lösning till (P). Omvänt, om $\hat{x}$ är en optimal lösning till (P), så är $\hat{x} = h(\hat{y})$ för något parametervärde $\hat{y}$, och detta parametervärde är då en optimal lösning till (P$'$).

Eliminationen är alltid möjlig (med en enkel algoritm) om samtliga bivillkorslikheter är affina, dvs. om systemet kan skrivas på formen $Ax = b$ för någon $(m - p) \times n$-matris A. Förutsatt att systemet är konsistent är

lösningsmängden ett affint delrum av dimension $d = n - \operatorname{rang} A$, och det finns en $n \times d$-matris C av rang d och en lösning x_0 till systemet så att $Ax = b$ om och endast om $x = Cy + x_0$ för något $y \in \mathbf{R}^d$. Problemet (P) är i detta fall således ekvivalent med problemet

$$\begin{array}{ll} \min & f(Cy + x_0) \\ \text{då} & g_i(Cy + x_0) \le 0, \qquad i = 1, 2, \ldots, p \end{array}$$

(med implicit bivillkor $Cy + x_0 \in \Omega$).

I konvexa optimeringsproblem, och speciellt i LP-problem, kan vi således i princip alltid eliminera alla bivillkorslikheter och ersätta problemet med ett ekvivalent optimeringsproblem med d variabler och oförändrat antal bivillkorsolikheter.

Slackvariabler

En olikhet $g(x) \le 0$ gäller om och endast om det finns ett tal $s \ge 0$ så att $g(x) + s = 0$. Genom att på detta sätt ersätta samtliga olikheter i problemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m. \end{cases} \end{array}$$

erhåller vi följande ekvivalenta problem

$$\text{(P}') \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) + s_i = 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \\ s_i \ge 0, & i = 1, 2, \ldots, p \end{cases} \end{array}$$

med $n + p$ variabler, m bivillkor i form av likheter och p olikhetsbivillkor. De nya variablerna s_i kallas *slackvariabler*.

Om $\hat{x}$ är en optimal lösning till (P), så får vi optimal lösning $(\hat{x}, \hat{s})$ till (P$'$) genom att sätta $\hat{s}_i = -g_i(\hat{x})$. Omvänt, om $(\hat{x}, \hat{s})$ är en optimal lösning till det sistnämnda problemet, så är förstås $\hat{x}$ en optimal lösning till ursprungsproblemet.

Om de ursprungliga bivillkoren är affina, så är också samtliga nya villkor affina. Transformationen överför således LP-problem i LP-problem.

Olikheter av typen $g(x) \ge 0$ kan naturligtvis på liknande sätt skrivas som likheter $g(x) - s = 0$ med icke-negativa variabler s, som då brukar kallas *surplusvariabler*.

Icke-negativa variabler

Varje reellt tal kan skrivas som en differens mellan två icke-negativa tal. I ett optimeringsproblem kan vi således ersätta en oinskränkt variabel x_i, dvs. en variabel som apriori får anta alla reella värden, med två icke-negativa variabler x_i' och x_i'' genom att sätta

$$x_i = x_i' - x_i'', \quad x_i' \geq 0, \; x_i'' \geq 0.$$

För varje oinskränkt variabel som ersätts ökar antalet variabler med ett och antalet olikhetsvillkor med två, men transformationen leder uppenbarligen till ett ekvivalent problem. Transformationen överför vidare konvexa problem till konvexa problem och LP-problem till LP-problem.

Exempel 9.3.1. LP-problemet

$$\begin{array}{ll} \min & x_1 + 2x_2 \\ \text{då} & \begin{cases} x_1 + x_2 \geq 2 \\ 2x_1 - x_2 \leq 3 \\ x_1 \geq 0 \end{cases} \end{array}$$

transformeras med hjälp av två slack/surplus-variabler och genom att ersätta den oinskränkta variabeln med en differens av två icke-negativa variabler till följande ekvivalenta LP-problem, där samtliga variabler är icke-negativa och alla övriga bivillkor är likheter.

$$\begin{array}{ll} \min & x_1 + 2x_2' - 2x_2'' + 0s_1 + 0s_2 \\ \text{då} & \begin{cases} x_1 + x_2' - x_2'' - s_1 = 2 \\ 2x_1 - x_2' + x_2'' + s_2 = 3 \\ x_1, x_2', x_2'', s_1, s_2 \geq 0. \end{cases} \end{array}$$

□

Epigrafformulering

Varje optimeringsproblem kan ersättas av ett ekvivalent problem med linjär målfunktionen, och tricket som åstadkommer detta består i att betrakta den ursprungliga målfunktionens epigraf. De båda problemen

$$\begin{array}{llll} \text{(P)} & \min \; f(x) & \text{och} & \text{(P}') \quad \min \; t \\ & \text{då} \quad x \in X & & \quad\;\; \text{då} \quad f(x) \leq t, \; x \in X \end{array}$$

är nämligen ekvivalenta, och målfunktionen i (P′) är linjär. Om $\hat{x}$ är en optimal lösning till (P), så är $(\hat{x}, f(\hat{x}))$ en optimal lösning till (P′), och om $(\hat{x}, \hat{t})$ är en optimal lösning till (P′), så är $\hat{x}$ en optimal lösning till (P).

Om problemet (P) är konvext, dvs. har formen

$$\begin{array}{ll}\min & f(x)\\ \text{då} & \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases}\end{array}$$

med konvexa funktioner f och g_i för $1 \le i \le p$, och affina funktionerna g_i för $i \ge p+1$, så är också epigrafvarianten

$$\begin{array}{ll}\min & t\\ \text{då} & \begin{cases} f(x) - t \le 0, & \\ \quad g_i(x) \le 0, & i = 1, 2, \ldots, p \\ \quad g_i(x) = 0, & i = p+1, \ldots, m \end{cases}\end{array}$$

ett konvext problem.

Då man studerar allmänna egenskaper hos konvexa optimeringsproblem kan man således utan inskränkning anta att målfunktionen är linjär.

Styckvis affina målfunktioner

Antag att X är en polyeder (given som ett snitt av slutna halvrum) och betrakta det konvexa optimeringsproblemet

$$\text{(P)} \qquad \begin{array}{ll}\min & f(x)\\ \text{då} & x \in X\end{array}$$

där målfunktionen $f(x)$ är styckvis affin och ges som

$$f(x) = \max\{\langle c_i, x\rangle + b_i \mid i = 1, 2, \ldots, m\}.$$

Epigraftransformationen resulterar först i det ekvivalenta konvexa problemet

$$\begin{array}{ll}\min & t\\ \text{då} & \begin{cases} \max\limits_{1 \le i \le m} (\langle c_i, x\rangle + b_i) \le t \\ \qquad\qquad\qquad x \in X, \end{cases}\end{array}$$

och eftersom $\max_{1 \le i \le m} \alpha_i \le t$ om och endast om $\alpha_i \le t$ för samtliga i, är detta problem i sin tur ekvivalent med LP-problemet

$$\text{(P}') \qquad \begin{array}{ll}\min & t\\ \text{då} & \begin{cases} \langle c_i, x\rangle - t + b_i \le 0, & i = 1, 2, \ldots, m \\ \qquad x \in X. \end{cases}\end{array}$$

LP-problemets bivillkorsmängd är en polyeder i $\mathbf{R}^n \times \mathbf{R}$.

Om målfunktionen i problemet (P) istället är en summa

$$f(x) = f_1(x) + f_2(x) + \cdots + f_k(x)$$

av styckvis affina funktioner f_i, så är problemet (P) ekvivalent med det konvexa problemet

$$\begin{array}{ll} \min & t_1 + t_2 + \cdots + t_k \\ \text{då} & \begin{cases} f_i(x) \le t_i & i = 1, 2, \ldots, k \\ x \in X \end{cases} \end{array}$$

och detta problem blir ett LP-problem om varje olikhet $f_i(x) \le t_i$ uttrycks som ett system av linjära olikheter på liknande sätt som ovan.

9.4 Några modellexempel

Dietproblemet

Vi börjar med ett klassiskt LP-problem som formulerades och studerades i större skala i linjärprogrammeringens barndom. I handeln finns n livsmedel $L_1, L_2, \ldots, L_n$ till en kostnad av $c_1, c_2, \ldots, c_n$ kr per enhet. Livsmedlen innehåller olika näringsämnen $N_1, N_2, \ldots, N_m$ (proteiner, kolhydrater, fetter, vitaminer, etc). Antal enheter näringsämne per enhet livsmedel framgår av nedanstående tabell:

	L_1	L_2	$\ldots$	L_n
N_1	a_{11}	a_{12}	$\ldots$	a_{1n}
N_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
N_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Om man köper $x_1, x_2, \ldots, x_n$ enheter av livsmedlen erhåller man således

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

enheter av näringsämnet N_i till en kostnad av

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n.$$

Antag att dagsbehovet av de olika näringsämnena är $b_1, b_2, \ldots, b_m$ och att det inte är skadligt att få för mycket av något ämne. Problemet att tillgodose

dagsbehovet till lägsta möjlig kostnad kallas *dietproblemet*. Matematiskt har det formen

$$\begin{array}{ll} \min & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{då} & \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geq b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geq b_m \\ x_1,\, x_2, \ldots, x_n \geq 0. \end{cases} \end{array}$$

Dietproblemet är således ett LP-problem. Förutom att bestämma den optimala dieten och kostnaden för denna är det av intresse att kunna besvara följande frågor:

1. Hur påverkas den optimala dieten och kostnaden av en prisändring på ett eller flera av livsmedlen?
2. Hur påverkas den optimala dieten om dagsbehovet av något näringsämne ändras?
3. Antag att det finns rena näringsämnen på marknaden. Vad får de kosta för att det skall löna sig att tillgodose sitt näringsbehov genom att köpa rena näringsämnen och bara äta sådana? Knappast något smakligt alternativ för en gourmet men fullt tänkbart vid djurutfodring.

Antag att kostnaden för den optimala dieten är z, och att kostnaden för den optimala dieten då behovet av näringsämne N_1 ändras från b_1 till $b_1 + \Delta b_1$, allt annat oförändrat, är $z + \Delta z$. Det är självklart att kostnaden inte kan minska när behovet ökar, så därför medför $\Delta b_1 > 0$ att $\Delta z \geq 0$. Om det är möjligt att köpa näringsämnet N_1 i helt ren form till priset p_1, så är det ekonomiskt fördelaktigt att tillgodose det ökade behovet genom att inta näringsämnet i ren form förutsatt att $p_1\Delta b_1 \leq \Delta z$. Det maximala pris på N_1 som gör näringsämnet i ren form till ett ekonomiskt alternativ är därför $\Delta z/\Delta b_1$, och gränsvärdet då $\Delta b_1 \to 0$, dvs. den partiella derivatan $\frac{\partial z}{\partial b_1}$, kallas i ekonomiska sammanhang för det *duala priset* eller *skuggpriset*.

Man kan beräkna näringsämnenas skuggpriserna genom att lösa ett med dietproblemet relaterat LP-problem. Antag som ovan att marknaden tillhandahåller näringsämnena i ren form och att priserna är $y_1, y_2, \ldots, y_m$. Eftersom en enhet av livsmedel L_i innehåller $a_{1i}, a_{2i}, \ldots, a_{mi}$ enheter av respektive näringsämne, kan vi "tillverka" en enhet av livsmedlet L_i genom att köpa precis denna uppsättning näringsämnen. Det är därför ekonomiskt fördelaktigt att ersätta alla livsmedel med rena näringsämnen om

$$a_{1i}y_1 + a_{2i}y_2 + \cdots + a_{mi}y_m \leq c_i$$

för $i = 1, 2, \ldots, n$. Under dessa villkor blir kostnaden för den erforderliga

dagsransonen $b_1, b_2, \ldots, b_m$ högst lika med maximivärdet för LP-problemet

$$\begin{array}{ll} \max & b_1y_1 + b_2y_2 + \cdots + b_my_m \\ \text{då} & \begin{cases} a_{11}y_1 + a_{21}y_2 + \ldots + a_{m1}y_m \leq c_1 \\ a_{12}y_1 + a_{22}y_2 + \ldots + a_{m2}y_m \leq c_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}y_1 + a_{2n}y_2 + \ldots + a_{mn}y_m \leq c_n \\ \qquad\qquad\qquad\qquad y_1,\, y_2, \ldots, y_m \geq 0. \end{cases} \end{array}$$

Vi kommer att visa att detta s. k. *duala problem* har samma optimala värde som det ursprungliga dietproblemet och att skuggpriserna ger den optimala lösningen.

Produktionsplanering

Många problem med anknytning till produktionsplanering kan formuleras som LP-problem, och en pionjär inom området var den ryske matematikern och ekonomen Leonid Kantorovich som studerade och löste sådana problem i slutet av 1930-talet. Här följer ett typiskt sådant problem.

Vid en fabrik kan man tillverka olika varor $V_1, V_2, \ldots, V_n$. För detta behövs olika insatsvaror (råvaror och halvfabrikat) samt olika typer av arbetsinsatser, något som vi med ett gemensamt namn kallar produktionsfaktorer $P_1, P_2, \ldots, P_m$. Dessa finns tillgängliga i begränsade kvantiteter $b_1, b_2, \ldots, b_m$. För att tillverka, marknadsföra och försälja en enhet av respektive vara åtgår det produktionsfaktorer i en omfattning som ges av följande tabell:

	V_1	V_2	$\ldots$	V_n
P_1	a_{11}	a_{12}	$\ldots$	a_{1n}
P_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
P_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Varje tillverkad vara V_j kan säljas med en vinst som är c_j kr/enhet. Man vill nu planera produktionen $x_1, x_2, \ldots, x_n$ av de olika varorna så att vinsten maximeras.

Genom att tillverka $x_1, x_2, \ldots, x_n$ enheter av varorna erhåller man en vinst som är lika med $c_1x_1 + c_2x_2 + \cdots + c_nx_n$. Därvid förbrukas $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$ enheter av produktionsfaktor P_i. Det optimeringsproblem som vi behöver lösa är således LP-problemet

$$\max\ c_1x_1 + c_2x_2 + \cdots + c_nx_n$$
$$\text{då}\quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \le b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \le b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \le b_m \\ \qquad\qquad\qquad x_1,\, x_2, \ldots, x_n \ge 0. \end{cases}$$

Här är det rimligt att ställa motsvarande frågor som för dietproblemet, dvs. hur påverkas den optimala lösningen och den optimala vinsten av

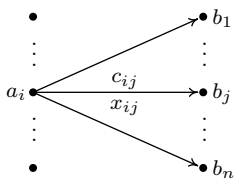
1. förändrad prissättning $c_1, c_2, \ldots, c_n$;
2. förändrad resurstilldelning.

Om vi ökar någon resurs P_i som redan utnyttjas fullt ut, så ökar (normalt) vinsten. Vad skall priset på denna resurs vara för att utökningen skall löna sig? Det kritiska priset kallas skuggpriset, och det kan tolkas som en partiell derivata och som lösningen på ett dualt problem.

Transportproblemet

Ett annat klassiskt LP-problem, som i litet större skala uppställdes och löstes innan simplexalgoritmen formulerats, är det s. k. transportproblemet.

En vara (t. ex. bensin) finns lagrad i m lager $L_1, L_2, \ldots, L_m$ på olika orter och efterfrågas vid n förbrukningsplatser $F_1, F_2, \ldots, F_n$. Att frakta 1 enhet från lagringsställe L_i till förbrukningsplats F_j kostar c_{ij} kr. Vid L_i finns a_i enheter och vid F_j efterfrågas b_j enheter. Det totala lagret, dvs. $\sum_{i=1}^{m} a_i$, antas motsvara den totala efterfrågan $\sum_{j=1}^{n} b_j$, så det är möjligt att tillgodose efterfrågan genom att distribuera x_{ij} enheter från L_i till F_j. Problemet att



Figur 9.1. Transportproblemet.

göra detta till lägsta transportkostnad ger tydligen upphov till LP-problemet

$$\min \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$

$$\text{då} \quad \begin{cases} \sum_{j=1}^{n} x_{ij} = a_i, & i = 1, 2, \ldots, m \\ \sum_{i=1}^{m} x_{ij} = b_j, & j = 1, 2, \ldots, n \\ x_{ij} \geq 0, & \text{alla } i, j. \end{cases}$$

Ett investeringsproblem

En investerare har 1 milj kr, som han tänker investera i olika projekt. Det finns m intressanta projekt $P_1, P_2, \ldots, P_m$ att investera pengarna i. Avkastningen kommer att bero dels på projekten, dels på den kommande ekonomiska konjunkturen. Han tycker sig kunna identifiera n olika konjunkturlägen $K_1, K_2, \ldots, K_n$, men vilken konjunktur som kommer att råda under det kommande året, efter vilket han avser att ta hem vinsten, går det inte att ha någon uppfattning om. Däremot kan man med säkerhet bedöma avkastningen av de olika projekten under de olika konjunkturerna; varje investerad miljon kr i projekt P_i ger en avkastning av a_{ij} milj kr om konjunkturläge K_j råder. Vi har med andra ord följande tabell över avkastningen för olika projekt och konjunkturer:

	K_1	K_2	$\ldots$	K_n
P_1	a_{11}	a_{12}	$\ldots$	a_{1n}
P_2	a_{21}	a_{22}	$\ldots$	a_{2n}
$\vdots$				
P_m	a_{m1}	a_{m2}	$\ldots$	a_{mn}

Vår investerare avser att satsa $x_1, x_2, \ldots, x_m$ milj kr på respektive projekt. Om nu konjunkturläge K_j kommer att råda blir alltså hans avkastning

$$a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{mj}x_m$$

milj kr. Eftersom vår investerare är en mycket försiktig person, vill han gardera sig mot värsta tänkbara utfall. Sämsta möjliga utfall, om han investerar $x_1, x_2, \ldots, x_m$, är lika med

$$\min_{1\leq j\leq n} \sum_{i=1}^{m} a_{ij}x_i.$$

Han vill därför maximera detta utfall, vilket han gör genom att lösa proble-

met

$$\begin{aligned} &\max \ \min_{1\le j\le n} \sum_{i=1}^{m} a_{ij}x_i \\ &\text{då} \quad x \in X \end{aligned}$$

där X är mängden $\{(x_1, x_2, \ldots, x_m) \in \mathbf{R}_+^m \mid \sum_{i=1}^m x_i = 1\}$ av alla möjliga sätt att fördela miljonen på de olika projekten.

Som problemet är formulerat är det ett konvext maximeringsproblem med en styckvis affin konkav målfunktion, och vi kan transformera det till ett ekvivalent LP-problem genom att använda oss av en hypografformulering och utnyttja tekniken i föregående avsnitt. Investerarens problem är således ekvivalent med LP-problemet

$$\begin{aligned} &\max \ v \\ &\text{då} \ \begin{cases} a_{11}x_1 + a_{21}x_2 + \ldots + a_{m1}x_m \ge v \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{m2}x_m \ge v \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}x_1 + a_{2n}x_2 + \ldots + a_{mn}x_m \ge v \\ x_1 + \quad x_2 + \ldots + \quad x_m = 1 \\ x_1,\, x_2, \ldots, x_m \ge 0. \end{cases} \end{aligned}$$

Tvåpersoners nollsummespel

Två personer, radspelaren Rulle och kolonnspelaren Kalle, väljer oberoende av varandra var sitt heltal. Rulle väljer ett tal i i intervallet $1 \le i \le m$ och Kalle ett tal j i intervallet $1 \le j \le n$. Om de väljer paret (i, j) vinner Rulle a_{ij} kronor av Kalle, varvid ett negativt belopp skall tolkas som att Rulle istället betalar beloppet $-a_{ij}$ till Kalle.

Talen m, n och a_{ij} förutsätts vara kända av båda spelarna, som strävar efter att vinna så mycket som möjligt (eller ekvivalent, att förlora så lite som möjligt). I allmänhet finns det inget optimalt val för någon av spelarna, så spelarna får istället inrikta sig på att maximera sina förväntade spelvinster genom att välja rad- resp. kolonnindex slumpvis med en viss sannolikhetsfördelning.

Antag att Rulle väljer talet i med sannolikheten x_i och att Kalle väljer talet j med sannoliheten y_j. Alla sannolikheterna är förstås icke-negativa tal, och $\sum_{i=1}^m x_i = \sum_{j=1}^n y_j = 1$. Vi sätter

$$X = \{x \in \mathbf{R}_+^m \mid \sum_{i=1}^{m} x_i = 1\} \quad \text{och} \quad Y = \{y \in \mathbf{R}_+^n \mid \sum_{j=1}^{n} y_j = 1\}.$$

Elementen i X brukar kallas radspelarens *blandade strategier*, och elementen i Y är kolonnspelarens blandade strategier.

Under förutsättning att talen i och j väljs oberoende av varandra kommer utfallet (i, j) att inträffa med sannolikheten $x_i y_j$. Utbetalningen till Rulle blir därför en stokastisk variabel med det förväntade värdet

$$f(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_i y_j.$$

Radspelare Rulle kan nu tänkas argumentera så här. Det sämsta som kan hända mig om jag väljer sannolikhetsfördelningen x, är att min motspelare Kalle råkar välja en sannolikhetsfördelning y som minimerar min förväntade vinst $f(x, y)$. I så fall får Rulle nämligen beloppet

$$g(x) = \min_{y \in Y} f(x, y) = \min_{y \in Y} \sum_{j=1}^{n} y_j \big(\sum_{i=1}^{m} a_{ij} x_i \big).$$

Summan $\sum_{j=1}^{n} y_j \big(\sum_{i=1}^{m} a_{ij} x_i \big)$ är ett vägt aritmetiskt medelvärde med vikter $y_1, y_2, \ldots, y_n$ av de n stycken talen $\sum_{i=1}^{m} a_{ij} x_i,\ j = 1, 2, \ldots, n$, och ett sådant medelvärde är alltid större eller lika med det minsta av de n talen, och likhet fås om man lägger hela vikten på detta tal. Därför är

$$g(x) = \min_{1 \le j \le n} \sum_{i=1}^{m} a_{ij} x_i.$$

Rulle, som vill maximera sin utdelning, bör därför välja att maximera $g(x)$, dvs. Rulles problem blir

$$\begin{array}{ll} \max & g(x) \\ \text{då} & x \in X. \end{array}$$

Detta är precis samma problem som investerarens problem. Rulles optimala strategi, dvs. val av sannolikheter $(x_1, x_2, \ldots, x_m)$, är lösningar till LP-problemet

$$\begin{array}{ll} \max & v \\ \text{då} & \left\{ \begin{array}{l} a_{11}x_1 + a_{21}x_2 + \ldots + a_{m1}x_m \ge v \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{m2}x_m \ge v \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{1n}x_1 + a_{2n}x_2 + \ldots + a_{mn}x_m \ge v \\ x_1 + x_2 + \ldots + x_m = 1 \\ x_1,\, x_2, \ldots, x_m \ge 0. \end{array} \right. \end{array}$$

Kolonnspelarens problem är analogt, men han vill förstås minimera den maximala förväntade utdelningen $f(x, y)$. Kalle skall således lösa problemet

$$\begin{aligned} &\min \quad \max_{1\le i\le m} \sum_{j=1}^{n} a_{ij}y_j \\ &\text{då} \quad y \in Y \end{aligned}$$

för att hitta sin optimala strategi, och detta problem är ekvivalent med LP-problemet

$$\begin{aligned} &\min \quad u \\ &\text{då} \quad \begin{cases} a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n \le u \\ a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n \le u \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}y_1 + a_{m2}y_2 + \ldots + a_{mn}y_n \le u \\ y_1 + y_2 + \ldots + y_n = 1 \\ y_1,\, y_2, \ldots, y_n \ge 0. \end{cases} \end{aligned}$$

De båda spelarnas problem är exempel på duala problem. Det följer av resultat som vi skall visa i kapitel 12 att de båda spelarnas problem har samma optimala värde.

Konsumentteori

I den gren av nationalekonomin som kallas mikroteori studerar man hur enskilda individer agerar. Vi antar att det finns n varor $V_1, V_2, \ldots, V_n$ på marknaden och att priset på dessa varor ges av prisvektorn $p = (p_1, p_2, \ldots, p_n)$. En varukorg x bestående av $x_1, x_2, \ldots, x_n$ enheter av de olika varorna kostar således $\langle p, x\rangle = p_1x_1 + p_2x_2 + \cdots + p_nx_n$ kr.

En konsument värderar sin nytta av varukorgen x med hjälp av en subjektiv s. k. *nyttofunktion* f, varvid $f(x) > f(y)$ betyder att han föredrar varukorgen x framför varukorgen y. Ett rimligt antagande om nyttofunktionen är att varje konvex kombination $\lambda x + (1-\lambda)y$ av två varukorgar skall värderas såsom varande minst lika bra som den sämsta av de två varukorgarna x och y, dvs. att $f(\lambda x + (1-\lambda)y) \ge \min\big(f(x), f(y)\big)$. Nyttofunktionen f antas med andra ord vara kvasikonkav, och ett starkare antagande, som man ofta gör i ekonomisk litteratur och som vi gör här, är att den är konkav.

Antag nu att vår konsuments inkomst är I, att hela inkomsten är disponibel för konsumtion, och att han vill maximera sin nytta. Det problem som

han då skall lösa är det konvexa optimeringsproblemet

$$\begin{array}{ll} \max & f(x) \\ \text{då} & \begin{cases} \langle p, x \rangle \leq I \\ x \geq 0. \end{cases} \end{array}$$

Att empiriskt bestämma hur en konsuments nyttofunktion ser ut är naturligtvis ogörligt, så mikroteorin är knappast användbar för kvantitativa beräkningar. Däremot kan man göra en kvalitativ analys och besvara frågor av typen: Hur förändras konsumentens beteende vid en inkomstökning? och Hur förändras köpmönstret då de relativa priserna på varorna ändras?

Portföljval

En person tänker köpa aktier i n olika bolag $B_1, B_2, \ldots, B_n$ för s kr. En satsad krona i bolaget B_j ger en utdelning av X_j kr, där X_j är en stokastisk variabel med känt väntevärde

$$\mu_j = \mathrm{E}[X_j].$$

Vidare antas kovarianserna

$$\sigma_{ij} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

vara kända.

Om personen satsar x_j kr i bolag B_j, $j = 1, 2, \ldots, n$, så blir den förväntade totala utdelningen

$$e(x) = \mathrm{E}\big[\sum_{j=1}^{n} x_j X_j\big] = \sum_{j=1}^{n} \mu_j x_j,$$

och variansen för den totala utdelningen blir

$$v(x) = \mathrm{Var}\,\big[\sum_{j=1}^{n} x_j X_j\big] = \sum_{i,j=1}^{n} \sigma_{ij} x_i x_j.$$

Observera att $v(x)$ är en positivt semidefinit kvadratisk form.

Personen kan inte maximera den totala utdelningen eftersom den är en stokastisk variabel, dvs. beror av slumpen. Däremot kan han maximera den förväntade utdelningen under lämpliga riskvillkor, dvs. krav på variansen. Alternativt kan han minimera risken med placeringen givet vissa krav på en förväntad utdelning. Det finns alltså flera möjliga strategier, och vi skall formulera tre stycken sådana.

(i) Strategin att maximera den förväntade totala utdelningen, givet en övre gräns b på variansen, leder till det konvexa optimeringsproblemet

$$\begin{array}{ll} \max & e(x) \\ \text{då} & \begin{cases} v(x) \leq b \\ x_1 + x_2 + \cdots + x_n = s \\ x \geq 0. \end{cases} \end{array}$$

(ii) Strategin att minimera variansen av den totala utdelningen, givet en undre gräns b för väntevärdet, ger det konvex-kvadratiska programmeringsproblemet

$$\begin{array}{ll} \min & v(x) \\ \text{då} & \begin{cases} e(x) \geq b \\ x_1 + x_2 + \cdots + x_n = s \\ x \geq 0. \end{cases} \end{array}$$

(iii) De två strategierna kan vägas ihop på följande sätt. Låt $\epsilon \geq 0$ vara en (subjektiv) parameter och betrakta det konvex-kvadratiska programmeringsproblemet

$$\begin{array}{ll} \min & \epsilon v(x) - e(x) \\ \text{då} & \begin{cases} x_1 + x_2 + \cdots + x_n = s \\ x \geq 0. \end{cases} \end{array}$$

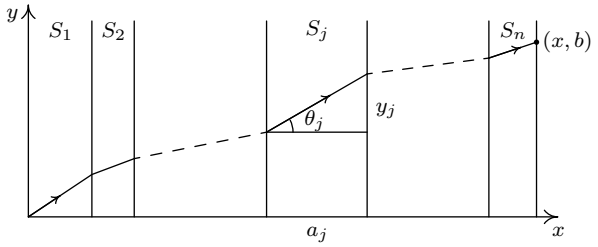
Låt $x(\epsilon)$ vara en optimal lösning till problemet. Vi lämnar som övning att visa att

$$v(x(\epsilon_1)) \geq v(x(\epsilon_2)) \qquad \text{och} \qquad e(x(\epsilon_1)) \geq e(x(\epsilon_2))$$

om $0 \leq \epsilon_1 \leq \epsilon_2$. Parametern ϵ är alltså ett mått på personens risktagande; ju mindre ϵ desto större risk (= varians) men också desto större förväntad utdelning.

Snells brytningslag

Vi skall studera en ljusstråles väg genom n stycken parallella skivor. Den j:te skivan S_j antas vara a_j enheter bred och bestå av ett homogent medium i vilket ljushastigheten är v_j. Välj ett koordinatsystem som i figur 9.2 och betrakta en ljusstråle på dess väg från origo i ytan på den första skivan till en punkt med y-koordinaten b på ytan av den sista skivan.



Figur 9.2. Ljusstrålens väg genom skivor med olika brytningsindex.

Enligt Fermats princip väljer ljuset den snabbaste vägen. Strålens väg bestäms allså av den optimala lösningen till det konvexa optimeringsproblemet

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} v_j^{-1} \sqrt{y_j^2 + a_j^2} \\ \text{då} \quad & \sum_{j=1}^{n} y_j = b. \end{aligned}$$

Genom att lösa problemet erhåller man *Snells brytningslag:*

$$\frac{\sin \theta_i}{\sin \theta_j} = \frac{v_i}{v_j}.$$

Överbestämda ekvationssystem.

Om ett linjärt ekvationssystem

$$Ax = b$$

med n obekanta och m ekvationer är inkonsistent, dvs. saknar lösning, vill man kanske ändå bestämma den ”bästa approximativa lösningen”, dvs. den n-tipel $x = (x_1, x_2, \ldots, x_n)$ som gör att felet blir så litet som möjligt. Med felet menas differensen $Ax - b$ mellan höger- och vänsterled, och som mått på felets storlek används $\|Ax - b\|$ för någon lämpligt vald norm.

Funktionen $x \mapsto \|Ax - b\|$ är konvex, så oberoende av vilken norm som används är problemet att minimera $\|Ax - b\|$ över alla $x \in \mathbf{R}^n$ ett konvext problem, men den optimala lösningen beror naturligtvis på valet av norm.

Låt i fortsättningen a_{ij} beteckna elementet på plats i, j i matrisen A och sätt $b = (b_1, b_2, \ldots, b_m)$.

1. Om den euklidiska normen $\|\cdot\|_2$ används får man den s. k. *minsta kvadratlösningen*. Eftersom

$$\|Ax-b\|_2^2 = \sum_{i=1}^{m}(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i)^2,$$

är minimeringsproblemet ekvivalent med det konvex-kvadratiska problemet

$$\text{minimera } \sum_{i=1}^{m}(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i)^2.$$

I den optimala lösningen är gradienten för målfunktionen lika med noll, vilket betyder att den optimala lösningen fås som lösning till det linjära ekvationssystemet

$$A^TAx = A^Tb.$$

2. Genom att istället använda $\|\cdot\|_\infty$-normen erhåller man den lösning som ger den minsta maximala avvikelsen mellan vänster- och högerleden i ekvationssystemet $Ax = b$. Eftersom

$$\|Ax-b\|_\infty = \max_{1\le i\le m}|a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i|,$$

är målfunktionen i optimeringsproblemet nu styckvis affin, och problemet är därför ekvivalent med LP-problemet

$$\begin{array}{ll}\min & t \\ \text{då} & \left\{\begin{array}{c} \pm(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1) \le t \\ \vdots \\ \pm(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n - b_m) \le t. \end{array}\right.\end{array}$$

3. Ett alternativ till att minimera kvadratsumman av avvikelserna mellan vänster- och högerled är förstås att minimera summan av beloppen på avvikelserna, dvs. att använda $\|\cdot\|_1$-normen. Eftersom målfunktionen

$$\|Ax-b\|_1 = \sum_{i=1}^{m}|a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i|$$

nu är en summa av konvexa styckvis affina funktioner, blir det konvexa minimeringsproblemet i detta fall ekvivalent med LP-problemet

$$\begin{array}{ll}\min & t_1 + t_2 + \cdots + t_m \\ \text{då} & \left\{\begin{array}{c} \pm(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1) \le t_1 \\ \vdots \\ \pm(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n - b_m) \le t_m. \end{array}\right.\end{array}$$

Största inskrivna bollen

En konvex mängd X med icke-tomt inre är given i $\mathbf{R}^n$, och vi vill bestämma en boll $B(x;r)$ i X (med avseende på en given norm) med största möjliga radie r. Vi antar att X kan beskrivas som lösningsmängden till ett system av olikheter, dvs. att

$$X = \{x \in \mathbf{R}^n \mid g_i(x) \leq 0,\ i = 1, 2, \ldots, m\},$$

med konvexa funktioner g_i. Då ligger bollen $B(x;r)$ i X om och endast om $g_i(x + ry) \leq 0$ för alla y med $\|y\| \leq 1$ och $i = 1, 2, \ldots, m$, vilket gör det naturligt att betrakta funktionerna

$$h_i(x, r) = \sup_{\|y\| \leq 1} g_i(x + ry).$$

Funktionerna h_i är konvexa eftersom de definieras som supremum av konvexa funktioner i variablerna x och r.

Problemet att bestämma bollen med största möjlig radie har nu transformerats till det konvexa optimeringsproblemet

$$\begin{aligned} &\max\ r \\ &\text{då} \quad h_i(x, r) \leq 0, \quad i = 1, 2, \ldots, m. \end{aligned}$$

För allmänna konvexa mängder X kan man naturligtvis inte bestämma funktionerna h_i explicit, men om X är en polyeder, $g_i(x) = \langle c_i, x\rangle - b_i$, och normen ifråga är ℓ^p-normen, så är på grund av Hölders olikhet

$$h_i(x, r) = \sup_{\|y\|_p \leq 1} (\langle c_i, x\rangle + r\langle c_i, y\rangle - b_i) = \langle c_i, x\rangle + r\|c_i\|_q - b_i$$

för $r \geq 0$, där $\|\cdot\|_q$ betecknar den duala normen.

Problemet att bestämma centrum x och radie r i den största boll som ligger helt i polyedern

$$X = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i,\ i = 1, 2, \ldots, m\}$$

har därmed reducerats till LP-problemet

$$\begin{aligned} &\max\ r \\ &\text{då} \quad \langle c_i, x\rangle + r\|c_i\|_q \leq b_i, \quad i = 1, 2, \ldots, m. \end{aligned}$$

Övningar

9.1 I en kemisk fabrik kan man använda fyra olika processer P_1, ..., P_4 för att tillverka produkterna V_1, V_2 och V_3. Producerade kvantiteter av de olika produkterna mätt i ton per timme under de olika processerna framgår av följande tabell:

	P_1	P_2	P_3	P_4
V_1	-1	2	2	1
V_2	4	1	0	2
V_3	3	1	2	1

(Process P_1 förbrukar således 1 ton av V_1 per timme.) Att köra processerna P_1, P_2, P_3 och P_4 kostar 5 000, 4 000, 3 000 resp. 4 000 kr per timme. Fabriken skall framställa 16, 40 resp. 24 ton av produkterna V_1, V_2, V_3 till lägsta kostnad. Formulera problemet att bestämma ett optimalt produktionsschema.

9.2 Klodvig har problem med väderleken. Vädret förekommer i de tre tillstånden ösregn, duggregn och solsken. Klodvig äger en regnrock och ett paraply och är något rädd om sin kostym. Regnrocken är besvärlig att medföra, och detsamma gäller − ehuru i mindre grad − paraplyet; det senare är dock inte fullt tillfredsställande i fall av ösregn. Följande tabell avslöjar hur pass nöjd Klodvig anser sig vara i de olika situationer som kan uppstå (siffrorna är relaterade till hans blodtryck, varvid 0 får anses motsvara hans normaltillstånd).

	Ösregn	Duggregn	Solsken
Regnrock	2	1	-2
Paraply	1	2	-1
Bara kavajen	-4	-2	2

På morgonen, när Klodvig går till jobbet, vet han inte hur vädret kommer att vara när han skall gå hem, och han vill därför välja den persedelvalsstrategi som optimerar hans sinnelag under hempromenaden. Formulera Klodvigs problem som ett LP-problem.

9.3 Betrakta följande tvåpersonersspel, där vardera spelaren har tre alternativ och där utbetalningen till radspelaren ges av följande matris

	1	2	3
1	1	0	5
2	3	3	4
3	2	4	0

I detta fall är det uppenbart vilka alternativ de båda spelarna skall välja. Hur skall de spela?

9.4 Kalle och Rulle har tre spelkort vardera. Båda har ruter ess och spader ess. Kalle har dessutom ruter 2 och Rulle har spader 2. Spelarna spelar samtidigt var sitt kort. Kalle vinner om båda dessa kort har samma färg och förlorar i motsatt fall. Vinnaren erhåller i betalning värdet av sitt vinnande kort från förloraren, varvid ess räknas som 1. Skriv upp utbetalningsmatrisen för detta tvåpersonersspel, och formulera kolonnspelare Kalles problem att optimera den förväntade vinsten som ett LP-problem.

9.5 Det överbestämda ekvationssystemt

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 - x_2 = 0 \\ 3x_1 + 2x_2 = 4 \end{cases}$$

saknar lösning.

a) Bestäm minsta kvadratlösningen.
b) Formulera problemet att bestämma den lösning som minimerar den maximala avvikelsen mellan vänster- och högerled.
c) Formulera problemet att bestämma den lösning som minimerar summan av beloppen på avvikelserna mellan vänster- och högerled.

9.6 Formulera problemet att bestämma

a) den största cirkelskiva,
b) den största kvadrat vars sidor är parallella med koordinataxlarna,

som ryms inom triangeln som begränsas av linjerna $x_1 - x_2 = 0$, $x_1 - 2x_2 = 0$ och $x_1 + x_2 = 1$ i planet.

Kapitel 10

Lagrangefunktionen

10.1 Lagrangefunktionen och det duala problemet

Lagrangefunktionen

Till minimeringsproblemet

$$\text{(P)} \qquad \begin{aligned} &\min \ f(x) \\ &\text{då} \ \begin{cases} g_i(x) \le 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{aligned}$$

med $x \in \Omega$ som implicit villkor och m stycken explicita bivillkor, varav de p första i form av olikheter, skall vi associera ett s. k. dualt maximeringsproblem, och verktyget för att åstadkomma detta är Lagrangefunktionen som definieras nedan. För att undvika trivialiteter förutsätter vi att $\operatorname{dom} f \neq \emptyset$, dvs. att målfunktionen $f \colon \Omega \to \overline{\mathbf{R}}$ inte är identiskt lika med ∞ på Ω.

X betecknar som tidigare mängden av tillåtna punkter i problemet (P), dvs. $X = \{x \in \Omega \mid g_1(x) \le 0, \ldots, g_p(x) \le 0, g_{p+1}(x) = 0, \ldots, g_m(x) = 0\}$, och $v_{\min}(P)$ är problemets optimala värde.

Definition. Givet minimeringsproblemet (P) sätter vi

$$\Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}$$

och definierar funktionen $L \colon \Omega \times \Lambda \to \overline{\mathbf{R}}$ genom att sätta

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x).$$

Funktionen L kallas *Lagrangefunktionen* till minimeringsproblemet (P), och variablerna $\lambda_1, \lambda_2, \ldots, \lambda_m$ kallas *Lagrangemultiplikatorer*.

För fixt $x \in \operatorname{dom} f$ är uttrycket $L(x, \lambda)$ summan av ett reellt tal och en linjär form i $\lambda_1, \lambda_2, \ldots, \lambda_m$, så funktionen $\lambda \mapsto L(x, \lambda)$ är affin (eller rättare sagt restriktionen till Λ av en affin funktion på $\mathbf{R}^m$). Lagrangefunktionen är alltså speciellt *konkav i variabeln* λ för varje fixt $x \in \operatorname{dom} f$.

För $x \in \Omega \setminus \operatorname{dom} f$ är uppenbarligen $L(x, \lambda) = \infty$ för alla $\lambda \in \Lambda$, så det följer att

$$\inf_{x \in \Omega} L(x, \lambda) = \inf_{x \in \operatorname{dom} f} L(x, \lambda) < \infty$$

för alla $\lambda \in \Lambda$.

Definition. För $\lambda \in \Lambda$ sätter vi

$$\phi(\lambda) = \inf_{x \in \Omega} L(x, \lambda)$$

och kallar funktionen $\phi \colon \Lambda \to \underline{\mathbf{R}}$ för den till minimeringsproblemet (P) hörande *duala funktionen*.

Det kan naturligtvis hända att den duala funktionens effektiva domän

$$\operatorname{dom} \phi = \{\lambda \in \Lambda \mid \phi(\lambda > -\infty\}$$

är den tomma mängden; detta inträffar om funktionen $x \mapsto L(x, \lambda)$ inte är nedåt begränsad på Ω för något $\lambda \in \Lambda$.

Sats 10.1.1. *Den duala funktionen ϕ till minimeringsproblemet (P) är konkav och*

$$\phi(\lambda) \leq v_{\min}(P)$$

för alla $\lambda \in \Lambda$.

Speciellt är alltså $\operatorname{dom} \phi = \emptyset$ om målfunktionen i (P) inte är nedåt begränsad på bivillkorsmängden, dvs. om $v_{\min}(P) = -\infty$.

Bevis. Funktionerna $\lambda \to L(x, \lambda)$ är konkava för $x \in \operatorname{dom} f$, så funktionen ϕ fås som ett infimum av en familj av konkava funktioner. Det följer därför av sats 6.2.4 att ϕ är konkav.

Antag att $\lambda \in \Lambda$ och $x \in X$; då är $\lambda_i g_i(x) \leq 0$ för $i \leq p$ och $\lambda_i g_i(x) = 0$ för $i > p$, så det följer att

$$L(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) \leq f(x),$$

och att följaktligen

$$\phi(\lambda) = \inf_{x \in \Omega} L(x, \lambda) \leq \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in X} f(x) = v_{\min}(P). \qquad \square$$

Följande optimalitetskriterium är nu en omedelbar konsekvens av föregående sats.

Sats 10.1.2 (Optimalitetskriteriet)**.** *Antag att $\hat{x}$ är en tillåten punkt i minimeringsproblemet (P) och att det finns en punkt $\hat{\lambda} \in \Lambda$ sådan att*

$$\phi(\hat{\lambda}) = f(\hat{x}).$$

Då är $\hat{x}$ en optimal lösning.

Bevis. Det gemensamma värdet $f(\hat{x})$ ligger i snittet $\overline{\mathbf{R}} \cap \underline{\mathbf{R}} = \mathbf{R}$ av f:s och ϕ:s målmängder och är således ett reellt tal, och på grund av sats 10.1.1 är $f(\hat{x}) \leq v_{\min}(P)$, vilket förstås medför att $f(\hat{x}) = v_{\min}(P)$. □

Exempel 10.1.1. Betrakta det enkla problemet

$$\begin{array}{ll} \min & f(x) = x_1^2 - x_2^2 \\ \text{då} & x_1^2 + x_2^2 \leq 1. \end{array}$$

Problemets Lagrangefunktion är

$$\begin{aligned} L(x_1, x_2, \lambda) &= x_1^2 - x_2^2 + \lambda(x_1^2 + x_2^2 - 1) \\ &= (\lambda + 1)x_1^2 + (\lambda - 1)x_2^2 - \lambda \end{aligned}$$

med $(x_1, x_2) \in \mathbf{R}^2$ och $\lambda \in \mathbf{R}_+$.

För fixt $\lambda \in \mathbf{R}_+$ är Lagrangefunktionen nedåt begränsad om och endast om $\lambda \geq 1$, så $\operatorname{dom} \phi = [1, +\infty[$, och för fixt $\lambda \geq 1$ har funktionen uppenbarligen minimivärdet $-\lambda$, som antas då $x_1 = x_2 = 0$. Detta innebär att problemets duala funktion är

$$\phi(\lambda) = \begin{cases} -\infty\,, & \text{om } 0 \leq \lambda < 1 \\ -\lambda\,, & \text{om } \lambda \geq 1. \end{cases}$$

Vi noterar slutligen att optimalitetsvillkoret $\phi(\hat{\lambda}) = f(\hat{x})$ uppfylls av punkten $\hat{x} = (0, 1)$ och Lagrangemultiplikatorn $\hat{\lambda} = 1$. Följaktligen är $(0, 1)$ en optimal lösning. □

Optimalitetskriteriet ger ett tillräckligt villkor för optimalitet, men det är inte nödvändigt, som följande triviala exempel visar.

Exempel 10.1.2. Betrakta problemet

$$\begin{array}{ll} \min & f(x) = x \\ \text{då} & x^2 \leq 0. \end{array}$$

Problemet har bara en tillåten punkt, $\hat{x} = 0$, som därför är den optimala lösningen. Lagrangefunktionen $L(x, \lambda) = x + \lambda x^2$ är nedåt begränsad för $\lambda > 0$ och

$$\phi(\lambda) = \inf_{x \in \mathbf{R}} (x + \lambda x^2) = \begin{cases} -1/4\lambda, & \text{om } \lambda > 0 \\ -\infty, & \text{om } \lambda = 0. \end{cases}$$

För alla $\lambda \in \Lambda = \mathbf{R}_+$ är $\phi(\lambda) < 0 = f(\hat{x})$. Optimalitetsvillkoret i sats 10.1.2 är därför inte uppfyllt i den optimala punkten. □

För att omvändningen till sats 10.1.2 skall gälla behövs det således något tilläggsvillkor. Vi skall beskriva ett sådant i kapitel 11.1.

Det duala problemet

För att erhålla bästa möjliga undre uppskattning av det optimala värdet i minimeringsproblemet (P) bör vi i ljuset av sats 10.1.1 maximera den duala funktionen. Detta föranleder följande definition.

Definition. Optimeringsproblemet

$$\text{(D)} \qquad \begin{aligned} &\max \ \phi(\lambda) \\ &\text{då} \quad \lambda \in \Lambda \end{aligned}$$

kallas det *duala problemet* till minimeringsproblemet (P).

Eftersom den duala funktionen är konkav, är det duala problemet ett konvext problem, oberoende av om problemet (P) är konvext eller ej. Det duala problemets värde kommer att betecknas $v_{\max}(D)$ med de sedvanliga konventionerna för $\pm\infty$-värden.

Nästa resultat är nu ett omedelbart korrelarium till sats 10.1.1.

Sats 10.1.3 (Svag dualitet)**.** *Följande olikhet gäller för de optimala värdena till problemet (P) och det duala problemet (D):*

$$v_{\max}(D) \leq v_{\min}(P).$$

Olikheten i satsen ovan går under namnet *svag dualitet*. Om likhet råder mellan de optimala värdena, dvs. om

$$v_{\max}(D) = v_{\min}(P)$$

säges *stark dualitet* råda för problemet (P).

Svag dualitet råder således alltid, medan stark dualitet bara gäller för speciella typer av problem. Stark dualitet föreligger naturligtvis om optimalitetskriteriet i sats 10.1.2 är uppfyllt.

EXEMPEL 10.1.3. Betrakta minimeringsproblemet

$$\begin{array}{ll} \min & x_1^3 + 2x_2 \\ \text{då} & x_1^2 + x_2^2 \leq 1. \end{array}$$

Det är lätt att se att minimum antas för $x = (0, -1)$ och att problemets optimala värde således är $v_{\min}(P) = -2$. Lagrangefunktionen

$$\begin{aligned} L(x_1, x_2, \lambda) &= x_1^3 + 2x_2 + \lambda(x_1^2 + x_2^2 - 1) \\ &= x_1^3 + \lambda x_1^2 + 2x_2 + \lambda x_2^2 - \lambda \end{aligned}$$

går för fixt $\lambda \geq 0$ mot $-\infty$ då $x_2 = 0$ och $x_1 \to -\infty$. Lagrangefunktionen är med andra ord nedåt obegränsad på $\mathbf{R}^2$ för alla λ, så $\phi(\lambda) = -\infty$ för alla $\lambda \in \Lambda$. Det duala problemets värde är därför $v_{\max}(D) = -\infty$, så stark dualitet gäller inte i det här problemet. □

Lagrangefunktionen, den duala funktionen och det duala problemet till ett minimeringsproblem av typ (P) har definierats med hjälp av de ingående bivillkorsfunktionerna. Det kan därför vara värt att påpeka att problem som är ekvivalenta i den meningen att de har samma målfunktion f och samma mängd X av tillåtna punkter inte behöver ha ekvivalenta duala problem. Således kan stark dualitet gälla för det ena sättet att presentera problemet men inte för det andra. Se övning 10.2.

EXEMPEL 10.1.4. Vi skall bestämma det duala problemet till LP-problemet

$$\text{(LP-P)} \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & \begin{cases} Ax \geq b \\ x \geq 0. \end{cases} \end{array}$$

Här är A en $m \times n$-matris, c en vektor i $\mathbf{R}^n$ och b en vektor i $\mathbf{R}^m$. Låt oss istället presentera problemet på formen

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & \begin{cases} b - Ax \leq 0 \\ x \in \mathbf{R}_+^n \end{cases} \end{array}$$

med $x \in \mathbf{R}_+^n$ som implicit bivillkor. Matrisolikheten $b - Ax \leq 0$ består av m stycken linjära olikheter, så Lagrangefunktionen är definierad på produktmängden $\mathbf{R}_+^n \times \mathbf{R}_+^m$ och ges av att

$$L(x, \lambda) = \langle c, x \rangle + \langle \lambda, b - Ax \rangle = \langle c - A^T\lambda, x \rangle + \langle b, \lambda \rangle.$$

För fixt λ är $L(x,\lambda)$ nedåt begränsad på mängden $\mathbf{R}^n_+$ om och endast om $c - A^T\lambda \geq 0$ och minimivärdet fås i så fall för $x = 0$ och är lika med $\langle b, \lambda\rangle$. Den duala funktionen $\phi\colon \mathbf{R}^m_+ \to \underline{\mathbf{R}}$ ges således av att

$$\phi(\lambda) = \begin{cases} \langle b, \lambda\rangle\,, & \text{om } A^T\lambda \leq c \\ -\infty\,, & \text{för övrigt,} \end{cases}$$

och det duala problemet till LP-problemet (LP-P) är därför också ett LP-problem, nämligen (efter namnbyte på parametern λ) problemet

$$\text{(LP-D)} \qquad \begin{aligned} &\max\ \langle b, y\rangle \\ &\text{då}\ \begin{cases} A^T y \leq c \\ \quad y \geq 0. \end{cases} \end{aligned}$$

Notera den vackra symmetrin mellan de båda problemen.

Eftersom svag dualitet alltid föreligger, kan vi redan nu med säkerhet säga att maximeringsproblemets optimala värde är mindre än eller lika med minimeringsproblemets optimala värde. Som vi skall se längre fram råder också stark dualitet, dvs. de båda problemen har samma optimala värde, förutsatt att åtminstone ett av problemen har tillåtna punkter. □

Vi återvänder nu till det allmänna minimeringsproblemet

$$\text{(P)} \qquad \begin{aligned} &\min\ f(x) \\ &\text{då}\ \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{aligned}$$

med X som mängd av tillåtna punkter, Lagrangefunktion $L\colon \Omega \times \Lambda \to \overline{\mathbf{R}}$ och dual funktion ϕ. Nästa sats visar att optimalitetsvillkoret i sats 10.1.2 kan uttryckas som ett sadelpunktsvillkor för Lagrangefunktionen.

Sats 10.1.4. *Antag att $(\hat{x}, \hat{\lambda}) \in \Omega \times \Lambda$. Följande tre villkor är då ekvivalenta för optimeringsproblemet (P):*

(i) $\hat{x} \in X$ och $f(\hat{x}) = \phi(\hat{\lambda})$, dvs. optimalitetskriteriet är uppfyllt.

(ii) För alla $(x, \lambda) \in \Omega \times \Lambda$ är

$$L(\hat{x}, \lambda) \leq L(\hat{x}, \hat{\lambda}) \leq L(x, \hat{\lambda}),$$

dvs. $(\hat{x}, \hat{\lambda})$ är en sadelpunkt *till Lagrangefunktionen.*

(iii) $\hat{x} \in X$, $\hat{x}$ minimerar funktionen $x \mapsto L(x, \hat{\lambda})$ då x genomlöper Ω, och

$$\hat{\lambda}_i g_i(\hat{x}) = 0$$

för $i = 1, 2, \ldots, p$.

Speciellt är alltså $\hat{x}$ en optimal lösning till problemet (P) om något av de ekvivalenta villkoren (i)–(iii) är uppfyllt.

Villkoret i (iii) att $\hat{\lambda}_i g_i(\hat{x}) = 0$ för $i = 1, 2, \ldots, p$ kallas *komplementaritet*. Ett ekvivalent sätt att uttrycka detta på, och som förklarar namnet, är att

$$\hat{\lambda}_i = 0 \quad \text{eller} \quad g_i(\hat{x}) = 0.$$

Ett bivillkor med positiv Lagrangemultiplikator är alltså aktivt i punkten $\hat{x}$.

Bevis. (i) $\Rightarrow$ (ii): För $\hat{x} \in X$ och godtyckligt $\lambda \in \Lambda$ ($= \mathbf{R}_+^p \times \mathbf{R}^{n-p}$) gäller att

$$L(\hat{x}, \lambda) = f(\hat{x}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{x}) = f(\hat{x}) + \sum_{i=1}^{p} \lambda_i g_i(\hat{x}) \leq f(\hat{x}),$$

beroende på att $\lambda_i \geq 0$ och $g_i(\hat{x}) \leq 0$ för $i = 1, 2, \ldots, p$. Vidare är

$$\phi(\hat{\lambda}) = \inf_{z \in \Omega} L(z, \hat{\lambda}) \leq L(x, \hat{\lambda}) \qquad \text{för alla } x \in \Omega.$$

Om $f(\hat{x}) = \phi(\hat{\lambda})$, så är följaktligen

$$L(\hat{x}, \lambda) \leq f(\hat{x}) = \phi(\hat{\lambda}) \leq L(x, \hat{\lambda})$$

för alla $(x, \lambda) \in \Omega \times \Lambda$, och genom att speciellt välja $x = \hat{x}$, $\lambda = \hat{\lambda}$ i olikheten ser vi att $f(\hat{x}) = L(\hat{x}, \hat{\lambda})$. Detta visar att sadelpunktsolikheten i (ii) gäller med $L(\hat{x}, \hat{\lambda}) = f(\hat{x})$.

(ii) $\Rightarrow$ (iii): Uppenbarligen minimerar $\hat{x}$ funktionen $L(\,\cdot\,, \hat{\lambda})$ om och endast om den högra sadelpunktsolikheten gäller. Minimivärdet är vidare ändligt (på grund av vårt stående antagande att dom $f \neq \emptyset$), så $f(\hat{x})$ är ett ändligt tal.

Den vänstra sadelpunktsolikheten betyder att för alla $\lambda \in \Lambda$ är

$$f(\hat{x}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{x}) \leq f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(\hat{x}),$$

vilket är ekvivalent med att för alla $\lambda \in \Lambda$ är

$$\sum_{i=1}^{m} (\lambda_i - \hat{\lambda}_i) g_i(\hat{x}) \leq 0.$$

Fixera att index k och välj i olikheten ovan λ så att $\lambda_i = \hat{\lambda}_i$ för alla i utom $i = k$. Det följer att

$$(\lambda_k - \hat{\lambda}_k) g_k(\hat{x}) \leq 0 \tag{10.1}$$

för alla sådana λ. Om $k > p$ väljer vi nu $\lambda_k = \hat{\lambda}_k \pm 1$ med slutsatsen att $\pm g_k(\hat{x}) \leq 0$, dvs. att $g_k(\hat{x}) = 0$.

För $k \leq p$ väljer vi istället $\lambda_k = \hat{\lambda}_k + 1$, med slutsatsen att $g_k(\hat{x}) \leq 0$ för sådana k. Detta visar att $\hat{x}$ satisfierar bivillkoren, dvs. $\hat{x} \in X$.

För $k \leq p$ väljer vi slutligen $\lambda_k = 0$ resp. $\lambda_k = 2\hat{\lambda}_k$ i olikheten (10.1) med $\pm\hat{\lambda}_k g_k(\hat{x}) \leq 0$ som resultat. Detta betyder att $\hat{\lambda}_k g_k(\hat{x}) = 0$ för $k \leq p$, och därmed är implikationen (ii) $\Rightarrow$ (iii) fullständigt bevisad.

(iii) $\Rightarrow$ (i): Av (iii) följer omedelbart att

$$\phi(\hat{\lambda}) = \inf_{x\in\Omega} L(x,\hat{\lambda}) = L(\hat{x},\hat{\lambda}) = f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(\hat{x}) = f(\hat{x}),$$

vilket är (i). □

Om mål- och bivillkorsfunktionerna f, $g_1, g_2, \ldots, g_m$ är differentierbara, så är också Lagrangefunktionen $L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$ differentierbar, och vi kommer att använda $L'_x(x_0,\lambda)$ som beteckning för värdet av derivatan till funktionen $x \mapsto L(x,\lambda)$ i punkten x_0, dvs.

$$L'_x(x_0,\lambda) = f'(x_0) + \sum_{i=1}^{m} \lambda_i g'_i(x_0).$$

Om den differentierbara funktionen $x \mapsto L(x,\lambda)$ har ett minimum i en inre punkt x_0 i Ω, så är $L'_x(x_0,\lambda) = 0$. Vi får därför omedelbart följande korollarium till implikationen (i) $\Rightarrow$ (iii) i sats 10.1.4.

Korollarium 10.1.5. *Antag att $\hat{x}$ är en optimal lösning till minimeringsproblemet (P), att $\hat{x}$ är en inre punkt i definitionsmängden Ω, att mål- och bivillkorsfunktionerna är differentierbara i punkten $\hat{x}$, samt att optimalitetsvillkoret $f(\hat{x}) = \phi(\hat{\lambda})$ är uppfyllt för någon uppsättning av Lagrangemultiplikatorer $\hat{\lambda} \in \Lambda$. Då är*

$$\text{(KKT)} \qquad \begin{cases} L'_x(\hat{x},\hat{\lambda}) = 0 & \text{och} \\ \hat{\lambda}_i g_i(\hat{x}) = 0 & \text{för } i = 1,2,\ldots,p. \end{cases}$$

Systemet (KKT) kallas *Karush–Kuhn–Tuckers villkor.*

Likheten $L'_x(\hat{x},\hat{\lambda}) = 0$ betyder att

$$f'(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g'_i(\hat{x}) = 0,$$

vilket fullt utskrivet blir ekvationssystemet:

$$\begin{cases} \dfrac{\partial f}{\partial x_1}(\hat{x}) + \displaystyle\sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i}{\partial x_1}(\hat{x}) = 0 \\ \qquad\qquad \vdots \\ \dfrac{\partial f}{\partial x_n}(\hat{x}) + \displaystyle\sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i}{\partial x_n}(\hat{x}) = 0. \end{cases}$$

EXEMPEL 10.1.5. I exempel 10.1.1 fann vi att $\hat{x} = (0, 1)$ är en optimal lösning till minimeringsproblemet

$$\begin{array}{ll} \min & x_1^2 - x_2^2 \\ \text{då} & x_1^2 + x_2^2 \leq 1 \end{array}$$

samt att optimalitetsvillkoret är uppfyllt för $\hat{\lambda} = 1$. Lagrangefunktionen är $L(x, \lambda) = x_1^2 - x_2^2 + \lambda(x_1^2 + x_2^2 - 1)$, och mycket riktigt satisfierar också $x = (0, 1)$ och $\lambda = 1$ KKT-systemet

$$\begin{cases} \dfrac{\partial L(x,\lambda)}{\partial x_1} = 2(\lambda + 1)x_1 = 0 \\ \dfrac{\partial L(x,\lambda)}{\partial x_1} = 2(\lambda - 1)x_2 = 0 \\ \qquad\qquad \lambda(x_1^2 + x_2^2 - 1) = 0. \end{cases}$$

□

10.2 Johns sats

Villkor, som garanterar att KKT-villkoren är uppfyllda i en optimal punkt, brukar kallas *kvalificerande villkor* (eng. constraint qualification), och i nästa kapitel kommer vi att beskriva ett sådant villkor för konvexa problem. I det här avsnittet skall vi studera ett annat kvalificerande villkor, Johns villkor, för allmänna optimeringsproblem med bivillkor i form av olikheter.

Betrakta därför ett problem av typen

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & g_i(x) \leq 0, \qquad i = 1, 2, \ldots, m \end{array}$$

med Ω som definitionsmängd för målfunktion och bivillkorsfunktioner.

Huruvida ett bivillkor är aktivt eller ej i en optimal punkt spelar stor roll, och eventuella affina villkor är därvid enklare att hantera än övriga, så därför inför vi följande beteckningar:

$$\begin{aligned} I_{\text{aff}}(x) &= \{i \mid \text{funktionen } g_i \text{ är affin och } g_i(x) = 0\}, \\ I_{\text{övr}}(x) &= \{i \mid \text{funktionen } g_i \text{ är inte affin och } g_i(x) = 0\}, \\ I(x) &= I_{\text{aff}}(x) \cup I_{\text{övr}}(x). \end{aligned}$$

$I_{\text{aff}}(x)$ består med andra ord av index för alla i punkten x aktiva affina bivillkor, $I_{\text{övr}}(x)$ är index för alla övriga i punkten aktiva bivillkor, och $I(x)$ är index för samtliga i punkten aktiva bivillkor.

Sats 10.2.1 (Johns sats)**.** *Antag att $\hat{x}$ är en lokal minimipunkt i problemet (P), att $\hat{x}$ är en inre punkt i Ω och att funktionerna f och $g_1, g_2, \ldots, g_m$ är differentierbara i punkten $\hat{x}$. Om det finns en vektor $z \in \mathbf{R}^n$ sådan att*

$$\text{(J)} \qquad \begin{cases} \langle g_i'(\hat{x}), z\rangle \geq 0 & \text{för alla } i \in I_{\text{aff}}(\hat{x}) \\ \langle g_i'(\hat{x}), z\rangle > 0 & \text{för alla } i \in I_{\text{övr}}(\hat{x}), \end{cases}$$

så finns det Lagrangeparametrar $\hat{\lambda} \in \mathbf{R}_+^m$ så att

$$\text{(KKT)} \qquad \begin{cases} L_x'(\hat{x}, \hat{\lambda}) = 0 \\ \hat{\lambda}_i g_i(\hat{x}) = 0 \qquad \text{för } i = 1, 2, \ldots, m. \end{cases}$$

Anmärkning 1. Enligt sats 3.3.5 är systemet (J) lösbart om och endast om

$$\text{(J}') \qquad \begin{cases} \sum_{i \in I(\hat{x})} u_i g_i'(\hat{x}) = 0 \\ u \geq 0 \end{cases} \Rightarrow \quad u_i = 0 \quad \text{för alla } i \in I_{\text{övr}}(\hat{x}).$$

Speciellt är således systemet (J) lösbart om gradientvektorerna $\nabla g_i(\hat{x})$ är linjärt oberoende för $i \in I(\hat{x})$.

Anmärkning 2. Om $I_{\text{övr}}(\hat{x}) = \emptyset$, så är (J) trivialt uppfyllt av $z = 0$.

Bevis. Låt Z beteckna lösningsmängden till systemet (J). Den första delen av beviset består i att visa att Z är en delmängd av det koniska halvrummet $\{z \in \mathbf{R}^n \mid -\langle f'(\hat{x}), z\rangle \geq 0\}$.

Antag därför att $z \in Z$ och betrakta halvlinjen $\hat{x} - tz$ för $t \geq 0$. Vi påstår att $\hat{x} - tz \in X$ för alla tillräckligt små $t > 0$.

För affina funktioner g, dvs. funktioner på formen $g(x) = \langle c, x\rangle + b$, är $g'(x) = c$, så $g(x + y) = \langle c, x + y\rangle + b = \langle c, x\rangle + b + \langle c, y\rangle = g(x) + \langle g'(x), y\rangle$ för alla vektorer x och y. För alla index $i \in I_{\text{aff}}(\hat{x})$ är följaktligen

$$g_i(\hat{x} - tz) = g_i(\hat{x}) - t\langle g_i'(\hat{x}), z\rangle = -t\langle g_i'(\hat{x}), z\rangle \leq 0$$

för alla $t \geq 0$.

För index $i \in I_{\text{övr}}(\hat{x})$ får vi istället med hjälp av kedjeregeln

$$\frac{d}{dt}\, g_i(\hat{x} - tz)|_{t=0} = -\langle g_i'(\hat{x}), z\rangle < 0.$$

Funktionen $t \mapsto g_i(\hat{x} - tz)$ är med andra ord avtagande i punkten $t = 0$, varför $g_i(\hat{x} - tz) < g_i(\hat{x}) = 0$ för alla tillräckligt små $t > 0$.

För inaktiva villkor, dvs. för $i \notin I(\hat{x})$, gäller slutligen $g_i(\hat{x}) < 0$. Det följer därför av kontinuitetsskäl att $g_i(\hat{x} - tz) < 0$ för alla tillräckligt små $t > 0$.

Vi har därmed visat att punkterna $\hat{x} - tz$ tillhör bivillkorsmängden X bara $t > 0$ är tillräckligt litet. Eftersom $\hat{x}$ är en lokal minimipunkt till f, följer det att $f(\hat{x} - tz) \geq f(\hat{x})$ för alla tillräckligt små $t > 0$. Följaktligen är

$$-\langle f'(\hat{x}), z\rangle = \frac{d}{dt}\, f(\hat{x} - tz)\big|_{t=0} = \lim_{t\to 0^+} \frac{f(\hat{x} - tz) - f(\hat{x})}{t} \geq 0.$$

Därmed har vi visat den påstådda inklusionen

$$Z \subseteq \{z \in \mathbf{R}^n \mid -\langle f'(\hat{x}), z\rangle \geq 0\} = \{-f'(\hat{x})\}^+ = \big(\mathrm{con}\{-f'(\hat{x})\}\big)^+,$$

och det följer nu av sats 3.2.1, korollarium 3.2.4 och sats 3.3.4 att

$$\mathrm{con}\{-f'(\hat{x})\} \subseteq Z^+ = \mathrm{con}\{g_i'(\hat{x}) \mid i \in I(\hat{x})\}.$$

Vektorn $-f'(\hat{x})$ tillhör därför konen i högerledet ovan, dvs. det finns icke-negativa tal $\hat{\lambda}_i$, $i \in I(\hat{x})$, så att

$$-f'(\hat{x}) = \sum_{i\in I(\hat{x})} \hat{\lambda}_i g_i'(\hat{x}).$$

Definiera slutligen $\hat{\lambda}_i = 0$ för alla $i \notin I(\hat{x})$; då är

$$f'(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i'(\hat{x}) = 0$$

och $\hat{\lambda}_i g_i(\hat{x}) = 0$ för $i = 1, 2, \ldots, m$. Detta innebär att KKT-villkoren är uppfyllda. □

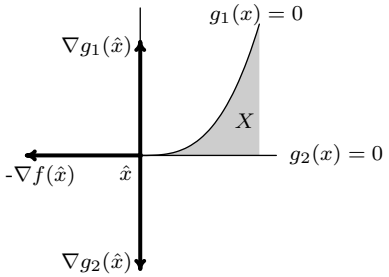
Villkoret i Johns sats att systemet (J) har en lösning kan ersättas med andra kvalificerande villkor men inte helt strykas utan att slutsatsen går förlorad. Detta framgår av följande exempel.

Exempel 10.2.1. Problemet

$$\begin{array}{ll} \min & f(x) = x_1 \\ \text{då} & \begin{cases} g_1(x) = -x_1^3 + x_2 \leq 0 \\ g_2(x) = \qquad -x_2 \leq 0 \end{cases} \end{array}$$

har den unika optimala lösningen $\hat{x} = (0, 0)$. Problemets Lagrangefunktion är $L(x, \lambda) = x_1 + \lambda_1(x_2 - x_1^3) - \lambda_2 x_2$, och systemet $L_x'(\hat{x}, \lambda) = 0$, dvs.

$$\begin{cases} \qquad 1 = 0 \\ \lambda_1 - \lambda_2 = 0, \end{cases}$$



Figur 10.1. Illustration till exempel 10.2.1: Vektorn $-\nabla f(\hat{x})$ tillhör inte konen som genereras av gradienterna $\nabla g_1(\hat{x})$ och $\nabla g_2(\hat{x})$.

saknar lösningar. Detta förklaras av att systemet (J), dvs.

$$\begin{cases} -z_2 \geq 0 \\ z_2 > 0 \end{cases}$$

saknar lösningar. □

Exempel 10.2.2. Vi skall lösa problemet

$$\begin{array}{ll} \min & x_1x_2 + x_3 \\ \text{då} & \begin{cases} 2x_1 - 2x_2 + x_3 + 1 \leq 0 \\ x_1^2 + x_2^2 - x_3 \leq 0 \end{cases} \end{array}$$

med hjälp av Johns sats. Observera då först att bivillkoren definierar en kompakt mängd X, ty olikheterna

$$x_1^2 + x_2^2 \leq x_3 \leq -2x_1 + 2x_2 - 1$$

medför att $(x_1+1)^2+(x_2-1)^2 \leq 1$. Följaktligen är $-2 \leq x_1 \leq 0$, $0 \leq x_2 \leq 2$ och $0 \leq x_3 \leq 7$. Eftersom målfunktionen är kontinuerlig, finns det en optimal lösning.

Vi undersöker nu först om systemet (J) är lösbart och använder oss därvid av den ekvivalenta versionen (J′) i anmärkningen efter satsen. Observera då först att gradienterna till funktionerna i bivillkoren aldrig är noll. Villkoret (J′) är således uppfyllt i de punkter där endast ett av bivillkoren är aktivt.

Antag därför att x är en punkt där $I(x) = \{1, 2\}$, dvs. där båda villkoren är aktiva, och att $u_1(2, -2, 1) + u_2(2x_1, 2x_2, -1) = (0, 0, 0)$. Om $u_2 > 0$ så är $u_1 = u_2$, $x_1 = -1$ och $x_2 = 1$. Detta strider emellertid mot att båda bivillkoren är aktiva, ty insättning i de båda bivillkoren ger om likhet råder $x_3 = 3$ resp. $x_3 = 2$, vilket är motsägelsefullt. Således är $u_2 = 0$, dvs. villkoret (J′) är uppfyllt i alla tillåtna punkter.

Den optimala punkten satisfierar därför KKT-villkoret, som i detta fall är

$$\begin{cases} x_2 + 2\lambda_1 + 2x_1\lambda_2 = 0 & \text{(i)} \\ x_1 - 2\lambda_1 + 2x_2\lambda_2 = 0 & \text{(ii)} \\ 1 + \lambda_1 - \lambda_2 = 0 & \text{(iii)} \\ \lambda_1(2x_1 - 2x_2 + x_3 + 1) = 0 & \text{(iv)} \\ \lambda_2(x_1^2 + x_2^2 - x_3) = 0 & \text{(v)} \end{cases}$$

Den fortsatta undersökningen delas upp på två fall.

$\lambda_1 = 0$: Ekvation (iii) medför att $\lambda_2 = 1$, varför (i) och (ii) ger $x_1 = x_2 = 0$, och av (v) följer nu $x_3 = 0$. Detta är en falsk lösning, ty $(0,0,0) \notin X$.

$\lambda_1 > 0$: Ekvation (iv) medför i detta fall att

$$2x_1 - 2x_2 + x_3 + 1 = 0. \qquad \text{(vi)}$$

Av (i) och (ii) följer $(x_1 + x_2)(1 + 2\lambda_2) = 0$, så på grund av att $\lambda_2 \geq 0$ är

$$x_1 + x_2 = 0. \qquad \text{(vii)}$$

Enligt (iii) är $\lambda_2 > 0$. Villkoret (v) medför därför att

$$x_1^2 + x_2^2 - x_3 = 0. \qquad \text{(viii)}$$

Ekvationerna (vi), (vii), (viii) har två lösningar, nämligen

$$\hat{x} = (-1+\sqrt{1/2}, 1-\sqrt{1/2}, 3-2\sqrt{2}) \text{ och } \overline{x} = (-1-\sqrt{1/2}, 1+\sqrt{1/2}, 3+2\sqrt{2}).$$

Med hjälp av (i) och (iii) beräknar vi motsvarande λ, och får då

$$\hat{\lambda} = (-1/2 + \sqrt{1/2}, 1/2 + \sqrt{1/2}) \text{ resp. } \overline{\lambda} = (-1/2 - \sqrt{1/2}, 1/2 - \sqrt{1/2}).$$

Observera att $\hat{\lambda} \geq 0$ och $\overline{\lambda} < 0$. Systemet KKT har således en unik lösning (x, λ) med $\lambda \geq 0$, nämligen $x = \hat{x}$, $\lambda = \hat{\lambda}$. Enligt Johns sats är därför $\hat{x}$ minimeringsproblemets unika optimala lösning. Problemets optimala värde är $3/2 - \sqrt{2}$. □

Övningar

10.1 Bestäm den duala funktionen i optimeringsproblemet

$$\begin{array}{ll} \min & x_1^2 + x_2^2 \\ \text{då} & x_1 + x_2 \geq 2 \end{array}$$

och visa att $(1,1)$ är en optimal lösning genom att visa att optimalitetskriteriet är uppfyllt för $\hat{\lambda} = 2$. Visa också att KKT-villkoren är uppfyllda i den optimala punkten.

10.2 Betrakta de båda minimeringsproblemen

$$(P_a)\quad \begin{array}{l}\min\ e^{-x_1}\\ x_1^2/x_2 \le 0\end{array} \qquad \text{och} \qquad (P_b)\quad \begin{array}{l}\min\ e^{-x_1}\\ |x_1| \le 0\end{array}$$

båda med $\Omega = \{(x_1, x_2) \mid x_2 > 0\}$ som definitionsområde. Båda problemen har samma mängd $X = \{(0, x_2) \mid x_2 > 0\}$ av tillåtna punkter och samma optimala värde $v_{\min} = 1$. Bestäm de båda problemens duala funktioner och duala problem, samt visa att stark dualitet gäller för (P_b) men inte för (P_a).

10.3 Antag att funktionen $f\colon X \times Y \to \mathbf{R}$ har två sadelpunkter $(\hat{x}_1, \hat{y}_1)$ och $(\hat{x}_2, \hat{y}_2)$. Visa att

a) $f(\hat{x}_1, \hat{y}_1) = f(\hat{x}_2, \hat{y}_2)$;

b) $(\hat{x}_1, \hat{y}_2)$ och $(\hat{x}_2, \hat{y}_1)$ också är sadelpunkter till funktionen.

10.4 Låt $f\colon X \times Y \to \mathbf{R}$ vara en godtycklig funktion.

a) Visa att

$$\sup_{y\in Y} \inf_{x\in X} f(x,y) \le \inf_{x\in X} \sup_{y\in Y} f(x,y).$$

b) Antag att det finns en punkt $(\hat{x}, \hat{y}) \in X \times Y$ så att

$$\sup_{y\in Y} \inf_{x\in X} f(x,y) = \inf_{x\in X} f(x,\hat{y}) \quad \text{och} \quad \inf_{x\in X} \sup_{y\in Y} f(x,y) = \sup_{y\in Y} f(\hat{x},y).$$

Visa att $(\hat{x}, \hat{y})$ är en sadelpunkt till funktionen f om och endast om

$$\inf_{x\in X} f(x,\hat{y}) = \sup_{y\in Y} f(\hat{x},y),$$

och att det gemensamma värdet i så fall är $f(\hat{x}, \hat{y})$.

10.5 Betrakta ett minimeringsproblem

$$\begin{array}{ll}\min & f(x)\\ \text{då} & g_i(x) \le 0, \qquad i = 1, 2, \ldots, m\end{array}$$

med *konvexa* differentierbara bivillkorsfunktioner $g_1, g_2, \ldots, g_m$, och antag att det finns en punkt $x_0 \in X = \{x \mid g_1(x) \le 0, \ldots, g_m(x) \le 0\}$ som uppfyller alla icke-affina bivillkor med strikt olikhet. Visa att systemet (J) är lösbart i alla punkter $\hat{x} \in X$.

[Ledning: Visa att $z = \hat{x} - x_0$ satisfierar (J).]

10.6 Lös följande optimeringsproblem

a) $\min\ x_1^3 + x_1x_2^2$ då $\begin{cases} x_1^2 + 2x_2^2 \le 1 \\ x_2 \ge 0 \end{cases}$

b) $\max\ x_1^2 + x_2^2 + \arctan x_1x_2$ då $\begin{cases} x_1^2 + x_2^2 \le 2 \\ 0 \le x_1 \le x_2 \end{cases}$

c) $\min\ x_1x_2$ då $\begin{cases} x_1^2 + x_1x_2 + 4x_2^2 \le 1 \\ x_1 + 2x_2 \ge 0 \end{cases}$

d) $\max\ x_1^2x_2x_3$ då $\begin{cases} 2x_1 + x_1x_2 + x_3 \le 1 \\ x_1,\ x_2,\ x_3 \ge 0. \end{cases}$

Kapitel 11

Konvex optimering

11.1 Stark dualitet

Vi påminner om att minimeringsproblemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

kallas konvext om

- den implicita definitionsmängden Ω är konvex,
- målfunktionen f är konvex,
- bivillkorsfunktionerna g_i är konvexa för $i = 1, 2, \ldots, p$ och affina för $i = p+1, \ldots, m$.

I ett konvext problem är mängden X av tillåtna punkter konvex, och Lagrangefunktionen

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

är konvex i variabeln x för varje fixt $\lambda \in \Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}$ eftersom den är en konisk kombination av konvexa funktioner.

Vi har redan konstaterat att optimalitetskriteriet i sats 10.1.2 inte behöver vara uppfyllt i en optimal punkt, inte ens i konvexa problem, ty det triviala problemet i exempel 10.1.2 är konvext. För att kriteriet skall vara uppfyllt behövs det något tilläggsvillkor, och ett svagt sådant ges i nästa definition.

Definition. Problemet (P) uppfyller *Slaters villkor* om det finns en tillåten punkt $\overline{x}$ i det relativa inre av Ω sådan att $g_i(\overline{x}) < 0$ för alla *icke-affina* bivillkorsfunktioner g_i.

Slaters villkor är förstås tomt uppfyllt om samtliga bivillkor är affina.

För konvexa problem som uppfyller Slaters villkor, är optimalitetskriteriet både tillräckligt och nödvändigt för optimalitet. Vi har nämligen följande resultat.

Sats 11.1.1 (Dualitetssatsen). *Antag att problemet (P) är konvext och uppfyller Slaters villkor samt att problemets optimala värde $v_{\min}$ är ändligt, och låt $\phi\colon \Lambda \to \underline{\mathbf{R}}$ beteckna problemets duala funktion. Då finns det en punkt $\hat{\lambda} \in \Lambda$ sådan att*

$$\phi(\hat{\lambda}) = v_{\min}.$$

Bevis. Antag först att samtliga bivillkor är olikheter, dvs. att $p = m$, och numrera bivillkoren så att funktionerna g_i är konvexa för $i = 1, 2, \ldots, k$ och affina för $i = k+1, \ldots, m$.

På grund av Slaters villkor har systemet

$$\begin{cases} g_i(x) < 0, & i = 1, 2, \ldots, k \\ g_i(x) \leq 0, & i = k+1, \ldots, m \end{cases}$$

en lösning i det relativa inre av Ω, medan systemet

$$\begin{cases} f(x) - v_{\min} < 0 & \\ g_i(x) < 0, & i = 1, 2, \ldots, k \\ g_i(x) \leq 0, & i = k+1, \ldots, m \end{cases}$$

saknar lösning i Ω på grund av definitionen av $v_{\min}$. Det följer därför av sats 6.5.1 att det finns icke-negativa skalärer $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_m$ sådana att

$$\hat{\lambda}_0(f(x) - v_{\min}) + \hat{\lambda}_1 g_1(x) + \hat{\lambda}_2 g_2(x) + \cdots + \hat{\lambda}_m g_m(x) \geq 0$$

för alla $x \in \Omega$, och där någon av koefficienterna $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_k$ är positiv. Här måste vidare koefficienten $\hat{\lambda}_0$ vara positiv, ty om $\hat{\lambda}_0 = 0$ är

$$\hat{\lambda}_1 g_1(x) + \cdots + \hat{\lambda}_m g_m(x) \geq 0$$

för alla $x \in \Omega$, vilket enligt sats 6.5.1 strider mot att det förstnämnda systemet av olikheter har en lösning i Ω. Vi kan därför, genom att vid behov dividera med $\hat{\lambda}_0$, anta att $\hat{\lambda}_0 = 1$, och får då olikheten

$$L(x, \hat{\lambda}) = f(x) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(x) \geq v_{\min}$$

för alla $x \in \Omega$. Det följer att

$$\phi(\hat{\lambda}) = \inf_{x \in \Omega} L(x, \hat{\lambda}) \geq v_{\min},$$

vilket i kombination med sats 10.1.1 ger den sökta likheten $\phi(\hat{\lambda}) = v_{\min}$.

Om problemet har affina likheter, dvs. om $p < m$, ersätter vi varje likhet $g_i(x) = 0$ med de två olikheterna $\pm g_i(x) \leq 0$, och det följer av det redan bevisade fallet av satsen att det finns icke-negativa Lagrangemultiplikatorer $\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{\mu}_{p+1}, \ldots, \hat{\mu}_m, \hat{\nu}_{p+1}, \ldots, \hat{\nu}_m$, så att

$$f(x) + \sum_{i=1}^{p} \hat{\lambda}_i g_i(x) + \sum_{i=p+1}^{m} (\hat{\mu}_i - \hat{\nu}_i) g_i(x) \geq v_{\min}$$

för alla $x \in \Omega$, Genom att sätta $\hat{\lambda}_i = \hat{\mu}_i - \hat{\nu}_i$ för $i = p+1, \ldots, m$ erhålls en punkt $\hat{\lambda} \in \Lambda = \mathbf{R}_+^p \times \mathbf{R}^{m-p}$ som uppfyller $\phi(\hat{\lambda}) \geq v_{\min}$, och därmed är beviset komplett. □

Genom att kombinera sats 11.1.1 med sats 10.1.2 erhåller vi följande korollarium.

Korollarium 11.1.2. *Om problemet (P) är konvext och uppfyller Slaters villkor, så är en tillåten punkt $\hat{x}$ optimal om och endast om optimalitetskriteriet är uppfyllt, dvs. om och endast om det finns ett $\hat{\lambda} \in \Lambda$ så att $\phi(\hat{\lambda}) = f(\hat{x})$.*

11.2 Karush–Kuhn–Tuckers sats

Varianter av följande sats visades först av Karush och Kuhn–Tucker, och satsen brukar därför kallas Karush–Kuhn–Tuckers sats.

Sats 11.2.1. *Låt*

$$(P) \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

vara ett konvext problem, och antag att mål- och bivillkorsfunktionerna är differentierbara i den tillåtna punkten $\hat{x}$.

(i) Om $\hat{\lambda}$ är en punkt i Λ och paret $(\hat{x}, \hat{\lambda})$ uppfyller KKT-villkoren

$$\begin{cases} L'_x(\hat{x}, \hat{\lambda}) = 0 \\ \hat{\lambda}_i g_i(\hat{x}) = 0 \qquad \text{för } i = 1, 2, \ldots, p \end{cases}$$

så föreligger starkt dualitet; $\hat{x}$ är en optimal lösning till problemet (P) och $\hat{\lambda}$ är en optimal lösning till det duala problemet.

(ii) Omvänt, om Slaters villkor är uppfyllt och $\hat{x}$ är en optimal lösning, så finns det Lagrangemultiplikatorer $\hat{\lambda} \in \Lambda$ så att $(\hat{x}, \hat{\lambda})$ satisfierar KKT-villkoren.

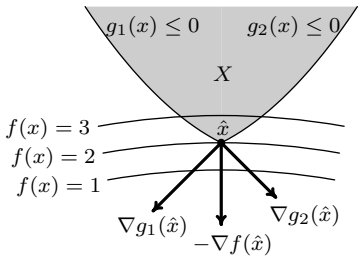
Bevis. (i) KKT-villkoren innebär att $\hat{x}$ är en stationär punkt till den konvexa funktionen $x \mapsto L(x, \hat{\lambda})$, och en inre stationär punkt till en konvex funktion är en minimipunkt enligt sats 7.2.2. Villkoret (iii) i sats 10.1.4 är således uppfyllt, och detta betyder att optimalitetskriteriet är uppfyllt av paret $(\hat{x}, \hat{\lambda})$.

(ii) Omvänt, om Slaters villkor är uppfyllt och $\hat{x}$ är en optimal lösning, så är enligt sats 11.1.1 optimalitetskriteriet $f(\hat{x}) = \phi(\hat{\lambda})$ uppfyllt för något $\hat{\lambda} \in \Lambda$. KKT-villkoren är därför uppfyllda på grund av korollarium 10.1.5. □

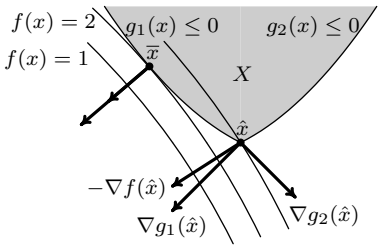
KKT-villkoren har en naturlig geometrisk tolkning. Antag för enkelhets skull att samtliga bivillkor är olikheter, dvs. att $p = m$, och låt $I(\hat{x})$ beteckna indexmängden för de i den optimala punkten $\hat{x}$ aktiva bivillkoren. KKT-villkoren innebär att $\hat{\lambda}_i = 0$ för alla index $i \notin I(\hat{x})$ och att

$$-\nabla f(\hat{x}) = \sum_{i \in I(\hat{x})} \hat{\lambda}_i \nabla g_i(\hat{x}),$$

där samtliga i summan förekommande koefficienter $\hat{\lambda}_i$ är icke-negativa. Den geometriska betydelsen av likheten ovan är att vektorn $-\nabla f(\hat{x})$ ligger i konen som genereras av gradienterna $\nabla g_i(\hat{x})$ till de aktiva olikhetsbivillkoren. Jmf figur 11.1 och figur 11.2.



Figur 11.1. Punkten $\hat{x}$ är optimal eftersom båda bivillkoren är aktiva i punkten och $-\nabla f(\hat{x}) \in \text{con}\{\nabla g_1(\hat{x}), \nabla g_2(\hat{x})\}$.



Figur 11.2. Här är punkten $\hat{x}$ inte optimal eftersom $-\nabla f(\hat{x}) \notin \text{con}\{\nabla g_1(\hat{x}), \nabla g_2(\hat{x})\}$. Optimum antas istället i en punkt $\overline{x}$, där $-\nabla f(\overline{x}) = \lambda_1 \nabla g_1(\overline{x})$ för något $\lambda_1 > 0$.

Exempel 11.2.1. Betrakta problemet

$$\begin{aligned} &\min\ \mathrm{e}^{x_1 - x_3} + \mathrm{e}^{-x_2} \\ &\begin{cases} (x_1 - x_2)^2 - x_3 \le 0 \\ x_3 - 4 \le 0. \end{cases} \end{aligned}$$

Målfunktionen och funktionerna i bivillkoren är konvexa. Slaters villkor är uppfyllt eftersom t. ex. $(1,1,1)$ satisfierar båda bivillkoren med strikt olikhet. Enligt sats 11.2.1 är därför x en optimal lösning till problemet om och endast om x löser systemet

$$\begin{cases} \mathrm{e}^{x_1-x_3} + 2\lambda_1(x_1-x_2) = 0 & \text{(i)} \\ -\mathrm{e}^{-x_2} - 2\lambda_1(x_1-x_2) = 0 & \text{(ii)} \\ -\mathrm{e}^{x_1-x_3} - \lambda_1 + \lambda_2 = 0 & \text{(iii)} \\ \lambda_1\big((x_1-x_2)^2 - x_3\big) = 0 & \text{(iv)} \\ \lambda_2(x_3-4) = 0 & \text{(v)} \\ \lambda_1,\, \lambda_2 \geq 0 & \text{(vi)} \end{cases}$$

Sambanden (i) och (vi) medför att $\lambda_1 > 0$, medan (iii) och (vi) ger att $\lambda_2 > 0$. Av (iv) och (v) följer därför $x_3 = 4$ och $x_1 - x_2 = \pm 2$. På grund av (i) och (vi) är emellertid $x_1 - x_2 < 0$, varför $x_1 - x_2 = -2$. Genom att jämföra (i) och (ii) ser vi att $x_1 - x_3 = -x_2$, dvs. $x_1 + x_2 = 4$. Det följer att $x = (1,3,4)$ och $\lambda = (\mathrm{e}^{-3}/4, 5\mathrm{e}^{-3}/4)$ är systemets unika lösning. Problemet har därför en unik optimal lösning, nämligen $(1,3,4)$. Det optimala värdet är lika med $2\mathrm{e}^{-3}$. □

11.3 Tolkning av Lagrangemultiplikatorerna

I det här avsnittet skall vi studera hur det optimala värdet $v_{\min}(b)$ i ett godtyckligt minimeringsproblem av typen

$$(\mathrm{P}_b) \qquad \begin{aligned} &\min\ f(x) \\ &\text{då } \begin{cases} g_i(x) \leq b_i, & i = 1,2,\ldots,p \\ g_i(x) = b_i, & i = p+1,\ldots,m \end{cases} \end{aligned}$$

beror av parametrarna $b_1, b_2, \ldots, b_m$ i bivillkorens högerled. Funktionerna f och $g_1, g_2, \ldots, g_m$ är som tidigare definierade på någon delmängd Ω av $\mathbf{R}^n$, $b = (b_1, \ldots, b_m)$ är en vektor i $\mathbf{R}^m$, och

$$X(b) = \{x \in \Omega \mid g_i(x) \leq b_i \text{ för } 1 \leq i \leq p \text{ och } g_i(x) = b_i \text{ för } p < i \leq m\}$$

är mängden av tillåtna punkter.

Till minimeringsproblemet (P_b) hör Lagrange- och dualfunktioner, som betecknas L_b resp. ϕ_b. Per definition är

$$L_b(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i(g_i(x) - b_i),$$

och sambandet mellan Lagrangefunktionerna L_b och $L_{\overline{b}}$ hörande till två skilda parameteruppsättningar b och $\overline{b}$ ges därför av ekvationen

$$L_b(x,\lambda) = L_{\overline{b}}(x,\lambda) + \sum_{i=1}^{m} \lambda_i(\overline{b}_i - b_i) = L_{\overline{b}}(x,\lambda) + \langle \lambda, \overline{b} - b\rangle.$$

Genom att bilda infimum över $x \in \Omega$ får vi omedelbart följande samband för dualfunktionerna:

$$\phi_b(\lambda) = \phi_{\overline{b}}(\lambda) + \langle \lambda, \overline{b} - b\rangle. \tag{11.1}$$

Följande sats ger oss nu en tolkning av Lagrangeparametrarnas betydelse för problem där optimalitetskriteriet i sats 10.1.2 är uppfyllt, och således speciellt för konvexa problem som uppfyller Slaters villkor.

Sats 11.3.1. *Antag att minimeringsproblemet $(P_{\overline{b}})$ har en optimal lösning $\overline{x}$ och att optimalitetskriteriet är uppfyllt i punkten, dvs. att det finns Lagrangemultiplikatorer $\overline{\lambda}$ så att $\phi_{\overline{b}}(\overline{\lambda}) = f(\overline{x})$. Då gäller:*

(i) Funktionen f är nedåt begränsad på mängden $X(b)$ för varje $b \in \mathbf{R}^m$, så det optimala värdet $v_{\min}(b)$ i problemet (P_b) är ändligt om mängden $X(b)$ av tillåtna punkter inte är tom, och lika med $+\infty$ om $X(b) = \emptyset$.

(ii) Vektorn $-\overline{\lambda}$ är en subgradient till värdefunktionen $v_{\min}\colon \mathbf{R}^m \to \overline{\mathbf{R}}$ i punkten $\overline{b}$.

(iii) Om optimalitetskriteriet är uppfyllt i problemet (P_b) för alla b i någon öppen konvex mängd U, så är funktionen $v_{\min}$ konvex på U.

Bevis. Genom att utnyttja att svag dualitet säkert gäller för problemet (P_b), sambandet (11.1) och optimalitetskriteriet för problemet $(P_{\overline{b}})$, får vi direkt

$$\begin{aligned} v_{\min}(b) &= \inf_{x\in X(b)} f(x) \geq \phi_b(\overline{\lambda}) = \phi_{\overline{b}}(\overline{\lambda}) + \langle \overline{\lambda}, \overline{b} - b\rangle = f(\overline{x}) + \langle \overline{\lambda}, \overline{b} - b\rangle \\ &= v_{\min}(\overline{b}) - \langle \overline{\lambda}, b - \overline{b}\rangle. \end{aligned}$$

Härav följer dels att det optimala värdet $v_{\min}(b)$ inte kan vara $-\infty$, dels att $-\overline{\lambda}$ är en subgradient till funktionen $v_{\min}$ i punkten $\overline{b}$.

Om optimalitetskriteriet är uppfyllt för alla $b \in U$, så har därför funktionen $v_{\min}$ en subgradient i alla punkter i U, och en sådan funktion är konvex. □

Om funktionen $v_{\min}$ är differentierbar i punkten $\overline{b}$, så är enligt sats 8.1.3 subgradienten i punkten unik och lika med gradienten, och det följer därför av (ii) i satsen ovan att $v'_{\min}(\overline{b}) = -\overline{\lambda}$. För små tillskott Δb_j är således

$$v_{\min}(\overline{b}_1 + \Delta b_1, \ldots, \overline{b}_m + \Delta b_m) \approx v_{\min}(\overline{b}_1, \ldots, \overline{b}_m) - \overline{\lambda}_1 \Delta b_1 \cdots - \overline{\lambda}_m \Delta b_m.$$

Lagrangemultiplikatorerna ger med andra ord information om hur det optimala värdet påverkas av små förändringar av parametrarna.

EXEMPEL 11.3.1. Som illustration till sats 11.3.1 skall vi lösa det konvexa problemet

$$\begin{aligned} &\min\ \ x_1^2 + x_2^2 \\ &\text{då } \begin{cases} x_1 + 2x_2 \le b_1 \\ 2x_1 + \ x_2 \le b_2. \end{cases} \end{aligned}$$

Eftersom det handlar om att minimera avståndet i kvadrat från origo till en polyeder, finns det säkert en optimal lösning för varje högerled b, och eftersom bivillkoren är affina, följer det av Karush–Kuhn–Tuckers sats att de optimala lösningarna satisfierar KKT-systemet, som i föreliggande fall är

$$\begin{cases} 2x_1 + \ \lambda_1 + 2\lambda_2 = 0 & \text{(i)} \\ 2x_2 + 2\lambda_1 + \ \lambda_2 = 0 & \text{(ii)} \\ \lambda_1(x_1 + 2x_2 - b_1) = 0 & \text{(iii)} \\ \lambda_2(2x_1 + x_2 - b_2) = 0 & \text{(iv)} \\ \lambda_1,\, \lambda_2 \ge 0. & \end{cases}$$

För att lösa systemet gör vi en uppdelning på fyra fall:

$\lambda_1 = \lambda_2 = 0$: I detta fall är $x_1 = x_2 = 0$ den unika lösningen till KKT-systemet. Punkten $(0, 0)$ är således optimal, förutsatt att den är tillåten, och så är fallet om och endast om $b_1 \ge 0$ och $b_2 \ge 0$. För sådana parametervärden b är således det optimala värdet $v_{\min}(b) = 0$.

$\lambda_1 > 0$, $\lambda_2 = 0$: Av (i) och (ii) följer först att $x_2 = 2x_1 = -\lambda_1$, och (iii) ger sedan $x = \frac{1}{5}(b_1, 2b_1)$. Denna punkt är tillåten om $2x_1 + x_2 = \frac{4}{5}b_1 \le b_2$, och för att Lagrangemultiplikatorn $\lambda_1 = -\frac{2}{5}b_1$ skall vara > 0 krävs dessutom att $b_1 < 0$. Punkten $x = \frac{1}{5}(b_1, 2b_1)$ är således optimal om $b_1 < 0$ och $4b_1 \le 5b_2$ och motsvarande värde är $v_{\min}(b) = \frac{1}{5}b_1^2$.

$\lambda_1 = 0$, $\lambda_2 > 0$: Av (i) och (ii) följer nu att $x_1 = 2x_2 = -\lambda_2$, vilket insatt i (iv) ger $x = \frac{1}{5}(2b_2, b_2)$. Detta är en tillåten punkt om $x_1 + 2x_2 = \frac{4}{5}b_2 \le b_1$. Lagrangemultiplikatorn $\lambda_2 = -\frac{2}{5}b_2$ är positiv om $b_2 < 0$. För $b_2 < 0$ och $4b_2 \le 5b_1$ är således punkten $x = \frac{1}{5}(2b_2, b_2)$ optimal, och det optimala värdet är $v(b) = \frac{1}{5}b_2^2$.

$\lambda_1 > 0$, $\lambda_2 > 0$: Genom att lösa det system som fås av (iii) och (iv) får man $x = \frac{1}{3}(2b_2 - b_1, 2b_1 - b_2)$, och (i) och (ii) ger sedan $\lambda = \frac{2}{9}(4b_2 - 5b_1, 4b_1 - 5b_2)$. De båda Lagrangemultiplikatorerna är positiva om $\frac{5}{4}b_1 < b_2 < \frac{4}{5}b_1$. För dessa parametervärden är punkten x optimal, och $v_{\min}(b) = \frac{1}{9}(5b_1^2 - 8b_1b_2 + 5b_2^2)$.

Resultatet av vår undersökning kan sammanfattas i följande tabell:

	$v_{\min}(b)$	$-\lambda_1 = \dfrac{\partial v}{\partial b_1}$	$-\lambda_2 = \dfrac{\partial v}{\partial b_2}$
$b_1 \geq 0,\ b_2 \geq 0$	0	0	0
$b_1 < 0,\ b_2 \geq \frac{4}{5}b_1$	$\frac{1}{5}b_1^2$	$\frac{2}{5}b_1$	0
$b_2 < 0,\ b_2 \leq \frac{5}{4}b_1$	$\frac{1}{5}b_2^2$	0	$\frac{2}{5}b_2$
$\frac{5}{4}b_1 < b_2 < \frac{4}{5}b_1$	$\frac{1}{9}(5b_1^2 - 8b_1b_2 + 5b_2^2)$	$\frac{2}{9}(5b_1 - 4b_2)$	$\frac{2}{9}(5b_2 - 4b_1)$

□

Övningar

11.1 Låt $b > 0$ och betrakta följande triviala konvexa optimeringsproblem

$$\begin{array}{ll} \min & x^2 \\ \text{då} & x \geq b. \end{array}$$

Slaters villkor är uppfyllt och det optimala värdet antas i punkten $\hat{x} = b$, så enligt sats 11.1.1 är optimalitetsvillkoret uppfyllt för något $\hat{\lambda}$. Bestäm $\hat{\lambda}$.

11.2 Verifiera i föregående övning att $v'(b) = \hat{\lambda}$.

11.3 Betrakta minimeringsproblemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

med $x \in \Omega$ som implicit bivillkor och den ekvivalenta epigrafformuleringen

$$\text{(P}') \qquad \begin{array}{ll} \min & t \\ \text{då} & \begin{cases} f(x) - t \leq 0, & \\ \quad g_i(x) \leq 0, & i = 1, 2, \ldots, p \\ \quad g_i(x) = 0, & i = p+1, \ldots, m \end{cases} \end{array}$$

av problemet med $(t, x) \in \mathbf{R} \times \Omega$ som implicit bivillkor.

a) Visa att (P′) uppfyller Slaters villkor om och endast om (P) gör det.

b) Bestäm sambandet mellan de båda problemens Lagrangefunktioner och sambandet mellan deras duala funktioner.

c) Visa att de båda duala problemen har samma optimala värde och att optimalitetskriteriet är uppfyllt i minimeringsproblemet (P) om och endast om det är uppfyllt i problemet (P′).

11.4 Visa att ett konvext problem uppfyller Slaters villkor om det för varje icke-affint olikhetsbivillkor $g_i(x) \leq 0$ finns en tillåten punkt $\overline{x}_i$ i det relativa inre av Ω sådan att $g_i(\overline{x}_i) < 0$.

11.5 Låt

$$(\mathrm{P}_b) \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & \begin{cases} g_i(x) \le b_i, & i = 1, 2, \ldots, p \\ g_i(x) = b_i, & i = p+1, \ldots, m \end{cases} \end{array}$$

vara ett konvext problem och antag att problemets optimala värde $v_{\min}(b)$ är $> -\infty$ för alla högerled b som tillhör någon konvex delmängd U av $\mathbf{R}^m$. Visa att restriktionen av $v_{\min}$ till U är en konvex funktion.

11.6 Lös följande konvexa optimeringsproblem.

a) $\begin{array}{ll} \min & \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2} - x_1 \\ \text{då} & x \in \mathbf{R}^2 \end{array}$

b) $\begin{array}{ll} \min & \mathrm{e}^{x_1 - x_2} + \mathrm{e}^{x_2} - x_1 \\ \text{då} & \begin{cases} x_1^2 + x_2^2 \le 1 \\ x_1 + x_2 \ge -1 \end{cases} \end{array}$

c) $\begin{array}{ll} \min & -x_1 - 2x_2 \\ \text{då} & \begin{cases} \mathrm{e}^{x_1} + x_2 \le 1 \\ x_2 \ge 0 \end{cases} \end{array}$

d) $\begin{array}{ll} \min & x_1 + 2x_2 \\ \text{då} & \begin{cases} x_1^2 + x_2^2 \le 5 \\ x_1 - x_2 \le 1 \end{cases} \end{array}$

e) $\begin{array}{ll} \min & x_1 - x_2 \\ \text{då} & \begin{cases} 0 < x_1 \le 2 \\ 0 \le x_2 \le \ln x_1 \end{cases} \end{array}$

f) $\begin{array}{ll} \min & \mathrm{e}^{x_1} + \mathrm{e}^{x_2} + x_1 x_2 \\ \text{då} & \begin{cases} x_1 + x_2 \ge 1 \\ x_1,\, x_2 \ge 0 \end{cases} \end{array}$

11.7 Lös det konvexa optimeringsproblemet

$$\begin{array}{ll} \min & x_1^2 + x_2^2 - \ln(x_1 + x_2) \\ \text{då} & \begin{cases} (x_1 - 1)^2 + x_2^2 \le 9 \\ x_1 + x_2 \ge 2 \\ x_1,\, x_2 \ge 0. \end{cases} \end{array}$$

11.8 Lös det konvexa optimeringsproblemet

$$\begin{array}{ll} \min & \displaystyle\sum_{j=1}^{n} v_j^{-1} \sqrt{y_j^2 + a_j^2} \\ \text{då} & \begin{cases} \sum_{j=1}^{n} y_j = b \\ y \in \mathbf{R}^n \end{cases} \end{array}$$

som förekom i vår diskussion av ljusbrytning i avsnitt 9.4, och verifiera därigenom *Snells brytningslag:*

$$\frac{\sin\theta_i}{\sin\theta_j} = \frac{v_i}{v_j},$$

där $\theta_j = \arctan y_j / a_j$.

11.9 Lisa har ärvt 1 miljon kronor som hon ämnar investera genom att köpa aktier i tre bolag: A, B och C. Bolaget A tillverkar mobiltelefoner, B tillverkar antenner till mobiltelefoner, och C tillverkar glass. Den årliga avkastningen av en investering i bolagen är en stokastisk variabel, och den förväntade avkastningen för respektive bolag uppskattas till

	A	B	C
Förväntad avkastning:	20%	12%	4%

Om Lisa investerar x_1, x_2, x_3 milj kronor i de tre bolagen blir således hennes förväntade avkastning lika med

$$0.2\,x_1 + 0.12\,x_2 + 0.04\,x_3.$$

Med en investerings risk menas avkastningens varians. För att kunna beräkna denna behöver man känna variansen för varje enskilt bolags avkastning samt korrelationen mellan avkastningarna i de olika bolagen. Av uppenbara skäl finns det en stark korrelation mellan försäljningen i bolagen A och B, medan försäljningen i bolaget C enbart beror på om sommarvädret är vackert eller ej och inte på hur många mobiltelefoner som säljs. Den s. k. kovariansmatrisen har därför i vårt fall följande utseende:

$$\begin{bmatrix} 50 & 40 & 0 \\ 40 & 40 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

För den som behärskar sannolikhetsteorins grunder är det nu enkelt att beräkna risken − den ges av uttrycket

$$50x_1^2 + 80x_1x_2 + 40x_2^2 + 10x_3^2.$$

Lisa, som är en försiktig person, vill minimera sin investerings risk men hon vill också ha en förväntad avkastning på minst 12%. Formulera och lös Lisas optimeringsproblem.

11.10 Betrakta konsumentproblemet

$$\begin{array}{ll} \max & f(x) \\ \text{då} & \begin{cases} \langle p, x\rangle \le I \\ x \ge 0 \end{cases} \end{array}$$

som diskuterades i avsnitt 9.4. Här är $f(x)$ konsumentens nyttofunktion, som förutsätts vara konkav och differentierbar, I är hennes disponibla inkomst, $p = (p_1, p_2, \ldots, p_n)$ är prisvektorn och $x = (x_1, x_2, \ldots, x_n)$ betecknar varukorgar.

Antag att $\hat{x}$ är en optimal lösning. Den optimala nyttan v beror liksom $\hat{x}$ förstås av inkomsten I; låt oss anta $v = v(I)$ är en deriverbar funktion. Visa

under dessa förutsättningar att

$$\begin{aligned} \hat{x}_j,\, \hat{x}_k > 0 \;\; &\Rightarrow \;\; \frac{1}{p_j}\frac{\partial f}{\partial x_j}\Big|_{\hat{x}} = \frac{1}{p_k}\frac{\partial f}{\partial x_k}\Big|_{\hat{x}} = \frac{dv}{dI} \\ \hat{x}_j = 0,\, \hat{x}_k > 0 \;\; &\Rightarrow \;\; \frac{1}{p_j}\frac{\partial f}{\partial x_j}\Big|_{\hat{x}} \leq \frac{1}{p_k}\frac{\partial f}{\partial x_k}\Big|_{\hat{x}}. \end{aligned}$$

I ord betyder detta:

I den optimala lösningen är kvoten mellan en varas marginella nytta och pris densamma för alla varor som faktiskt köps och lika med den marginella nyttoökningen vid en inkomstökning. För varor som inte köps är motsvarande kvot inte större.

Slutsatsen är tämligen självklar, ty om $x_k > 0$ och $\dfrac{1}{p_j}\dfrac{\partial f}{\partial x_j} > \dfrac{1}{p_k}\dfrac{\partial f}{\partial x_k}$, tjänar konsumenten på att byta ut en liten kvantitet ϵ/p_k av vara nr k mot kvantiteten ϵ/p_j av vara nr j.

Kapitel 12

Linjär programmering

I det här kapitlet skall vi beskriva den grundläggande matematiska teorin för linjär programmering och framförallt studera det mycket viktiga dualitetsbegreppet.

12.1 Optimala lösningar

I kapitel 9.1 definierade vi vad som menas med ett optimeringsproblems värde. Speciellt har förstås varje LP-problem

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & x \in X \end{array}$$

ett optimalt värde, som vi i det här avsnittet betecknar $v_{\min}(c)$ för att kunna beskriva värdets beroende av målfunktionen.

LP-problem med ändliga värden har alltid optimala lösningar. Om polyedern av tillåtna punkter är begränsad, dvs. kompakt, så följer förstås existensen av maximum och minimum direkt av att målfunktionen är kontinuerlig. För godtyckliga LP-problem utnyttjar vi istället representationssatsen för polyedrar för att bevisa existens av optimala lösningar.

Sats 12.1.1. *Antag att polyedern X av tillåtna lösningar i LP-problemet (P) inte är tom och en delmängd av $\mathbf{R}^n$. Då gäller:*

(i) Värdefunktionen $v_{\min}\colon \mathbf{R}^n \to \underline{\mathbf{R}}$ är konkav med effektiv domän

$$\operatorname{dom} v_{\min} = (\operatorname{recc} X)^+.$$

Målfunktionen $\langle c, x \rangle$ är med andra ord nedåt begränsad på X om och endast om c ligger i dualkonen till X:s recessionskon.

(ii) För varje $c \in (\operatorname{recc} X)^+$ har problemet optimala lösningar, och mängden av optimala lösningar är en polyeder.

(iii) Om polyedern X är linjefri och $c \in (\operatorname{recc} X)^+$, så antas optimum i någon av polyederns extremalpunkter.

Bevis. Det optimala värdets definition $v_{\min}(c) = \inf\{\langle c, x\rangle \mid x \in X\}$ innebär att värdefunktionen är ett punktvist infimum av en familj av konkava funktioner, nämligen de linjära funktionerna $c \mapsto \langle c, x\rangle$, då x genomlöper X. Det följer därför av sats 6.2.4 att värdefunktionen är konkav.

Låt oss nu bestämma värdefunktionens effektiva domän, dvs. mängden av c för vilka $v_{\min}(c) > -\infty$. Enligt struktursatsen för polyedrar (sats 5.3.1) finns det en ändlig icke-tom mängd A så att $X = \operatorname{cvx} A + \operatorname{recc} X$, där $A = \operatorname{ext} X$ ifall polyedern är linjefri. För LP-problemets optimala värde $v_{\min}(c)$ gäller därför

$$\begin{aligned} v_{\min}(c) &= \inf\{\langle c, y+z\rangle \mid y \in \operatorname{cvx} A,\ z \in \operatorname{recc} X\} \\ &= \inf\{\langle c, y\rangle \mid y \in \operatorname{cvx} A\} + \inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\} \\ &= \min\{\langle c, y\rangle \mid y \in A\} + \inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\}, \end{aligned} \tag{12.1}$$

där likheten $\inf\{\langle c, y\rangle \mid y \in \operatorname{cvx} A\} = \min\{\langle c, y\rangle \mid y \in A\}$ gäller på grund av sats 6.3.3 eftersom linjära funktioner speciellt är konkava.

Om c ligger i dualkonen $(\operatorname{recc} X)^+$, så är $\langle c, z\rangle \geq 0$ för alla $z \in \operatorname{recc} X$ med likhet för $z = 0$. Det följer därför av likheten (12.1) att

$$v_{\min}(c) = \min\{\langle c, y\rangle \mid y \in A\} > -\infty.$$

Detta visar inklusionen $(\operatorname{recc} X)^+ \subseteq \operatorname{dom} v_{\min}$, samt att det optimala värdet antas i en punkt i A, och då speciellt i en extremalpunkt till X om polyedern X är linjefri.

Om $c \notin (\operatorname{recc} X)^+$, så finns det en vektor $z_0 \in \operatorname{recc} X$ med $\langle c, z_0\rangle < 0$. Eftersom $tz_0 \in \operatorname{recc} X$ för $t > 0$ och $\lim_{t\to\infty}\langle c, tz_0\rangle = -\infty$, följer det att

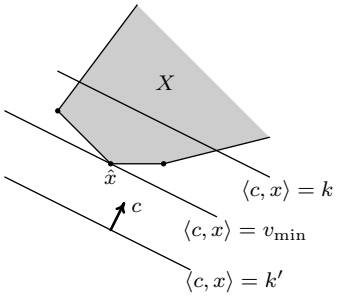
$$\inf\{\langle c, z\rangle \mid z \in \operatorname{recc} X\} = -\infty,$$

och (12.1) medför således i detta fall att $v_{\min}(c) = -\infty$. Därmed är likheten $\operatorname{dom} v_{\min} = (\operatorname{recc} X)^+$ bevisad.

Mängden av minimipunkter till ett LP-problem med ändligt värde är lika med snittet

$$X \cap \{x \in \mathbf{R}^n \mid \langle c, x\rangle = v_{\min}\}$$

mellan polyedern X och ett hyperplan, och den är följaktligen en polyeder. □

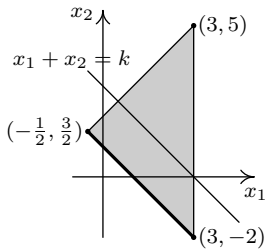


Figur 12.1. Minimum av $\langle c, x\rangle$ över en linjefri polyeder X antas i en extremalpunkt.

EXEMPEL 12.1.1. I LP-problemet

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{då} & \begin{cases} x_1 - x_2 \geq -2 \\ x_1 + x_2 \geq \ \ 1 \\ -x_1 \qquad \geq -3 \end{cases} \end{array}$$

har polyedern X av tillåtna punkter tre extremalpunkter, nämligen $(3, 5)$, $(-\frac{1}{2}, \frac{3}{2})$ och $(3, -2)$. I dessa punkter antar målfunktionen $f(x) = x_1 + x_2$ värdena $f(3, 5) = 8$ och $f(-\frac{1}{2}, \frac{3}{2}) = f(3, -2) = 1$. Det minsta av dessa är 1, vilket är problemets optimala värde 1. Minimivärdet antas i två extremalpunkter, $(\frac{1}{2}, \frac{3}{2})$ och $(3, -2)$, och därmed också i alla punkter på sträckan mellan dessa två punkter.



Figur 12.2. Illustration till exempel 12.1.1. □

Antag att $X = \{x \in \mathbf{R}^n \mid Ax \geq b\}$ är en linjefri polyeder. För att bestämma det optimala värdet till LP-problemet att minimera en linjär funktion över X behöver man, förutsatt att funktionen är nedåt begränsad på X, enligt föregående sats bara beräkna målfunktionens värden i de ändligt många extremalpunkterna till X. I teorin är detta enkelt, men i praktiken kan det vara ett oöverstigligt problem beroende på att antalet extremalpunkter

kan vara ohyggligt stort. Om A är en $m \times n$-matris, så är antalet potentiella extremalpunkter lika med $\binom{m}{n}$, vilket för $m = 100$ och $n = 50$ är ett tal som är större än 10^{29}. Simplexalgoritmen, som vi skall studera i kapitel 13, bygger på att man inte behöver söka igenom samtliga extremalpunkter; algoritmen genererar istället en följd $x_1, x_2, x_3, \ldots$ av extremalpunkter med avtagande målfunktionsvärden $\langle c, x_1\rangle \geq \langle c, x_2\rangle \geq \langle c, x_3\rangle \geq \ldots$ till dess att minimipunkten hittats. Antalet extremalpunkter som behöver undersökas blir därför i allmänhet relativt litet.

Känslighetsanalys

Låt oss skriva polyedern av tillåtna punkter i LP-problemet

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & x \in X \end{array}$$

på formen

$$X = \operatorname{cvx} A + \operatorname{con} B$$

med ändliga mängder A och B. Av föregående sats och dess bevis följer då att en tillåten punkt $\overline{x}$ är en optimal lösning till LP-problemet om och endast om

$$\begin{cases} \langle c, a\rangle \geq \langle c, \overline{x}\rangle & \text{för alla } a \in A \\ \langle c, b\rangle \geq 0 & \text{för alla } b \in B, \end{cases}$$

och dessa olikheter definierar en konvex kon $C_{\overline{x}}$ i variabeln c. Mängden av alla c för vilka en given tillåten punkt i problemet (P) är optimal är således en konvex kon.

Antag nu att $\overline{x}$ är en optimal lösning till LP-problemet (P). Hur mycket kan vi ändra koefficienterna i målfunktionen utan att ändra den optimala lösningen? Studiet av denna fråga är ett exempel på *känslighetsanalys.*

Uttryckt i termer av konen $C_{\overline{x}}$ är svaret enkelt: Om vi ändrar målfunktionskoefficienterna till $c+\Delta c$, så är $\overline{x}$ en optimal lösning också till det störda LP-problemet

$$\text{(P}') \qquad \begin{array}{ll} \min & \langle c + \Delta c, x\rangle \\ \text{då} & x \in X \end{array}$$

om och endast om $c+\Delta c$ ligger i konen $C_{\overline{x}}$, dvs. om och endast om Δc ligger i polyedern $-c + C_{\overline{x}}$. Sammanfattningsvis har vi således kommit fram till följande slutsatser.

Sats 12.1.2. *(i) Mängden av alla c för vilka en given tillåten punkt är optimal i LP-problemet (P) är en konvex kon.*

(ii) Om $\overline{x}$ är en optimal lösning till problemet (P), så finns det en polyeder så att $\overline{x}$ också är en optimal lösning till det störda LP-problemet (P′) för alla Δc i polyedern.

Mängden

$$\{\Delta c_k \mid \Delta c \text{ tillhör polyedern } -c + C_{\overline{x}} \text{ och } \Delta c_j = 0 \text{ för } j \neq k\}$$

är ett (eventuellt obegränsat) slutet intervall $[-d_k, e_k]$ kring 0. En optimal lösning till problemet (P) är således också optimal till det störda problem som erhålls genom att endast variera målkoefficienten c_k, förutsatt att störningen Δc_k ligger i intervallet $-d_k \leq \Delta c_k \leq e_k$. Många datorprogram för LP-problem levererar förutom optimalt värde och optimal lösning också automatiskt information om just dessa intervall.

Vi kommer att studera känslighetsanalysfrågor i samband med simplexalgoritmen i kapitel 13.7.

EXEMPEL 12.1.2. Då ett LP-problem med $c = (20, 30, 40, \dots)$ löstes med hjälp av ett datorprogram innehöll utskriften bl. a. följande information:

Optimalt värde: 4000 **Optimal lösning:** $\overline{x} = (50, 40, 10, \dots)$

Känslighetsrapport:

Variabel	Värde	Mål-koeff.	Tillåten minskning	Tillåten ökning
x_1	50	20	15	5
x_2	40	30	10	10
x_3	10	40	15	20
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Använd utskriften för att bestämma den optimala lösningen och det optimala värdet om koefficienterna c_1, c_2 och c_3 ändras till 17, 35 resp. 45 och övriga målkoefficienter lämnas oförändrade.

Lösning: Kolumnerna "Tillåten minskning" och "Tillåten ökning" visar att polyedern av förändringar Δc som inte påverkar den optimala lösningen bl. a. innehåller punkterna $(-15, 0, 0, 0, \dots)$, $(0, 10, 0, 0, \dots)$ och $(0, 0, 20, 0, \dots)$. Eftersom

$$(-3, 5, 5, 0, \dots) = \tfrac{1}{5}(-15, 0, 0, 0, \dots) + \tfrac{1}{2}(0, 10, 0, 0, \dots) + \tfrac{1}{4}(0, 0, 20, 0, \dots)$$

och $\frac{1}{5} + \frac{1}{2} + \frac{1}{4} = \frac{19}{20} < 1$, är $\Delta c = (-3, 5, 5, 0, \dots)$ en konvex kombination av förändringar som inte påverkar den optimala lösningen, nämligen av de tre nämnda förändringarna och $(0, 0, 0, 0, \dots)$. Lösningen $\overline{x} = (50, 40, 10, \dots)$ är således fortfarande optimal i problemet då $c = (17, 35, 45, \dots)$. Däremot är förstås det nya optimala värdet $4000 - 20 \cdot 3 + 30 \cdot 5 + 40 \cdot 5 = 4290$. □

12.2 Dualitet

Duala problem

Genom att beskriva polyedern X i ett linjärt minimeringsproblem

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & x \in X \end{array}$$

som lösningsmängd till ett system av linjära olikheter får vi ett problem med en tillhörande Lagrangefunktion och följaktligen också en dual funktion och ett dualt problem. Beskrivningen av X som lösningsmängd är naturligtvis inte unik, så det duala problemet är inte entydigt bestämt av X som polyeder, men oavsett vilken beskrivning vi väljer får vi ett dualt problem där stark dualitet föreligger, beroende på sats 11.1.1 och att Slaters villkor är uppfyllt för konvexa problem med affina bivillkor.

I det här avsnittet skall vi beskriva det duala problemets utseende för några vanligen förekommande polyederbeskrivningar samt ge ett alternativt bevis för dualitetssatsen. Vår utgångspunkt är att polyedern X ges på formen

$$X = \{x \in U^+ \mid Ax - b \in V^+\},$$

där

- U och V är ändligt genererade koner i $\mathbf{R}^n$ resp. $\mathbf{R}^m$;
- A är en $m \times n$-matris;
- b är en vektor i $\mathbf{R}^m$.

Som vanligt identifierar vi vektorer med kolonnmatriser och matriser med linjära avbildningar. Mängden X är förstås en polyeder, ty omskrivningen

$$X = U^+ \cap A^{-1}(b + V^+)$$

framställer X som ett snitt av två polyedrar − den koniska polyedern U^+ och inversa bilden $A^{-1}(b + V^+)$ av polyedern $b + V^+$ under den linjära avbildningen A.

LP-problemet att minimera $\langle c, x\rangle$ över polyedern X med ovanstående beskrivning kommer nu att skrivas

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax - b \in V^+,\ x \in U^+ \end{array}$$

och för att bilda ett lämpligt dualt problem kommer vi att uppfatta villkoret $x \in U^+$ som ett implicit bivillkor och uttrycka det andra villkoret $Ax - b \in V^+$ som ett system av linjära olikheter. Antag därför att den ändligt genererade konen V genereras av kolonnerna i $m \times k$-matrisen D, dvs. att

$$V = \{Dz \mid z \in \mathbf{R}^k_+\}.$$

För dualkonen V^+ gäller då att

$$V^+ = \{y \in \mathbf{R}^m \mid D^T y \geq 0\},$$

och bivillkoret $Ax - b \in V^+$ kan nu uttryckas som ett system av olikheter, nämligen $D^T Ax - D^T b \geq 0$.

Vårt LP-problem (P) har således bringats på formen

$$\begin{aligned} &\min \ \langle c, x\rangle \\ &\text{då} \quad D^T b - D^T Ax \leq 0,\ x \in U^+ \end{aligned}$$

Till problemet hör en Lagrangefunktion $L\colon U^+ \times \mathbf{R}^k_+ \to \mathbf{R}$, som definieras av att

$$L(x,\lambda) = \langle c, x\rangle + \langle \lambda, D^T b - D^T Ax\rangle = \langle c - A^T D\lambda, x\rangle + \langle b, D\lambda\rangle,$$

och en dualfunktion $\phi\colon \mathbf{R}^k_+ \to \underline{\mathbf{R}}$, som ges av att

$$\phi(\lambda) = \inf_{x\in U^+} L(x,\lambda) = \begin{cases} \langle b, D\lambda\rangle\,, & \text{om } c - A^T D\lambda \in U \\ -\infty\,, & \text{annars.} \end{cases}$$

Det duala problemet får därför formen

$$\begin{aligned} &\max \ \langle b, D\lambda\rangle \\ &\text{då} \quad c - A^T D\lambda \in U,\ \lambda \in \mathbf{R}^k_+. \end{aligned}$$

Eftersom $D\lambda$ beskriver konen V då λ genomlöper $\mathbf{R}^k_+$, kan vi genom att sätta $y = D\lambda$ formulera om det duala problemet till

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{då} \quad c - A^T y \in U,\ y \in V. \end{aligned}$$

Det är därför naturligt att definiera dualitet för LP-problem på formen (P) på följande sätt.

Definition. Givet LP-problemet

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{då} \quad Ax - b \in V^+,\ x \in U^+, \end{aligned}$$

som vi kallar det *primala* problemet, kallar vi problemet

$$\text{(D)} \qquad \begin{aligned} &\max \ \langle b, y\rangle \\ &\text{då} \quad c - A^T y \in U,\ y \in V \end{aligned}$$

för det *duala* LP-problemet. De optimala värdena för respektive problem kommer att betecknas $v_{\min}(P)$ resp. $v_{\max}(D)$, och respektive polyeder av tillåtna punkter kommer att kallas X resp. Y.

Observera noga sambandet mellan å ena sidan konerna U^+ och V^+ i det primala problemet och de därmed duala konerna U och V i det duala problemet.

EXEMPEL 12.2.1. Olika val av konerna U och V ger oss olika konkreta duala problem (P) och (D). Vi exemplifierar med fyra viktiga specialfall.

1. För $U = \{0\}$, $U^+ = \mathbf{R}^n$ och $V = V^+ = \mathbf{R}^m_+$ fås det duala paret:

$$\begin{array}{llll} (\mathrm{P}_1) & \min \ \langle c, x\rangle & \text{och} & (\mathrm{D}_1) \quad \max \ \langle b, y\rangle \\ & \text{då} \quad Ax \geq b & & \text{då} \quad A^T y = c, \ y \geq 0. \end{array}$$

Eftersom varje polyeder kan skrivas som ett snitt av halvrum, dvs. på formen $Ax \geq b$, kan varje LP-problem ges formuleringen (P_1).

2. För $U = U^+ = \mathbf{R}^n_+$ och $V = V^+ = \mathbf{R}^m_+$ får vi istället det duala paret:

$$\begin{array}{llll} (\mathrm{P}_2) & \min \ \langle c, x\rangle & \text{och} & (\mathrm{D}_2) \quad \max \ \langle b, y\rangle \\ & \text{då} \quad Ax \geq b, \ x \geq 0 & & \text{då} \quad A^T y \leq c, \ y \geq 0. \end{array}$$

Detta är den mest symmetriska formuleringen av dualitet, och den naturliga formuleringen för många tillämpningsproblem med variabler som står för fysiska kvantiteter eller priser som förstås är icke-negativa. Dietproblemet och produktionsplaneringsproblemet i kapitel 9.4 är exempel på sådana problem.

3. För $U = U^+ = \mathbf{R}^n_+$, $V = \mathbf{R}^m$ och $V^+ = \{0\}$ är det duala paret:

$$\begin{array}{llll} (\mathrm{P}_3) & \min \ \langle c, x\rangle & \text{och} & (\mathrm{D}_3) \quad \max \ \langle b, y\rangle \\ & \text{då} \quad Ax = b, \ x \geq 0 & & \text{då} \quad A^T y \leq c. \end{array}$$

Formuleringen (P_3) är den naturliga utgångspunkten för simplexalgoritmen.

4. $U = \{0\}$, $U^+ = \mathbf{R}^n$, $V = \mathbf{R}^m$ och $V^+ = \{0\}$ ger slutligen paret

$$\begin{array}{llll} (\mathrm{P}_4) & \min \ \langle c, x\rangle & \text{och} & (\mathrm{D}_4) \quad \max \ \langle b, y\rangle \\ & \text{då} \quad Ax = b & & \text{då} \quad A^T y = c. \end{array}$$

□

EXEMPEL 12.2.2. Ett trivialt exempel på duala LP-problem i en variabel är

$$\begin{array}{lll} \min \ 5x & \text{och} & \max \ 4y \\ \text{då} \quad 2x \geq 4 & & \text{då} \quad 2y = 5, \ y \geq 0 \end{array}$$

Båda problemens optimala värde är 10. □

EXEMPEL 12.2.3. Problemen

$$\min\ x_1 + x_2 \quad \text{då} \quad \begin{cases} x_1 - x_2 \geq -2 \\ x_1 + x_2 \geq 1 \\ -x_1 \geq -3 \end{cases} \qquad \text{och} \qquad \max\ -2y_1 + y_2 - 3y_3 \quad \text{då} \quad \begin{cases} y_1 + y_2 - y_3 = 1 \\ -y_1 + y_2 = 1 \\ y_1, y_2, y_3 \geq 0 \end{cases}$$

är duala. I exempel 12.1.1 bestämde vi samtliga optimala lösningar till minimeringsproblemet samt problemets optimala värde 1. De tillåtna punkterna till maximeringsproblemet har formen $y = (t, 1+t, 2t)$, där $t \geq 0$, och motsvarande värde hos målfunktionen är $1 - 7t$. Maximum antas därför för $t = 0$ i punkten $(0, 1, 0)$, och maximivärdet är 1. □

Dualitetssatsen

Att de duala problemen i exemplen 12.2.2 och 12.2.3 har samma optimala värde är naturligtvis ingen tillfällighet utan följer av dualitetssatsen, som formuleras nedan och är ett specialfall av dualitetssatsen för konvexa problem (sats 11.1.1). I det här avsnittet skall vi ge ett alternativt bevis för denna viktiga sats och börjar med det triviala resultatet om svag dualitet.

Sats 12.2.1 (Svag dualitet)**.** *De optimala värdena till de duala LP-problemen (P) och (D) satisfierar olikheten*

$$v_{\max}(D) \leq v_{\min}(P).$$

Bevis. Olikheten är trivialt uppfylld om någon av polyedrarna X och Y av tillåtna punkter är tom; om $Y = \emptyset$ är definitionsmässigt $v_{\max}(D) = -\infty$, och om $X = \emptyset$ är $v_{\min}(P) = +\infty$.

Antag därför att båda problemen har tillåtna punkter. För $x \in X$ och $y \in Y$ gäller per definition dels $y \in V$ och $(Ax - b) \in V^+$, dels $(c - A^T y) \in U$ och $x \in U^+$. Följaktligen är $\langle Ax - b, y\rangle \geq 0$ och $\langle c - A^T y, x\rangle \geq 0$. Vi får därför olikheten

$$\begin{aligned} \langle b, y\rangle &\leq \langle b, y\rangle + \langle c - A^T y, x\rangle = \langle b, y\rangle + \langle c, x\rangle - \langle y, Ax\rangle \\ &= \langle c, x\rangle + \langle b, y\rangle - \langle Ax, y\rangle = \langle c, x\rangle - \langle Ax - b, y\rangle \leq \langle c, x\rangle. \end{aligned}$$

Målfunktionen $\langle b, y\rangle$ i maximeringsproblemet (D) är med andra ord uppåt begränsad på Y av $\langle c, x\rangle$ för varje $x \in X$, varför

$$v_{\max}(D) = \sup_{y \in Y} \langle b, y\rangle \leq \langle c, x\rangle.$$

Målfunktionen $\langle c, x\rangle$ i minimeringsproblemet är därför nedåt begränsad på X av $v_{\max}(D)$. Detta medför att $v_{\max}(D) \leq v_{\min}(P)$, och därmed är satsen bevisad. □

Den svaga dualiteten ger oss följande kriterium för optimalitet.

Sats 12.2.2 (Optimalitetskriteriet)**.** *Antag att $\hat{x}$ är en tillåten punkt i minimeringsproblemet (P), att $\hat{y}$ är en tillåten punkt i det duala maximeringsproblemet (D) samt att*

$$\langle c, \hat{x} \rangle = \langle b, \hat{y} \rangle.$$

Då är $\hat{x}$ och $\hat{y}$ optimala lösningar till respektive problem.

Bevis. Antagandena om $\hat{x}$ och $\hat{y}$ ger i kombination med sats 12.2.1 följande kedja av olikheter

$$v_{\max}(D) \geq \langle b, \hat{y} \rangle = \langle c, \hat{x} \rangle \geq v_{\min}(P) \geq v_{\max}(D).$$

Eftersom de båda ytterleden är lika, råder det likhet överallt, vilket innebär att $\hat{y}$ är en maximipunkt och $\hat{x}$ är en minimipunkt. □

Sats 12.2.3 (Dualitetssatsen i linjär programmering)**.** *Antag att åtminstone ett av de båda duala LP-problemen*

$$\begin{aligned} &\text{(P)} && \min \ \langle c, x \rangle \\ &&& \text{då} \quad Ax - b \in V^+,\ x \in U^+ \end{aligned}$$

och

$$\begin{aligned} &\text{(D)} && \max \ \langle b, y \rangle \\ &&& \text{då} \quad c - A^T y \in U,\ y \in V \end{aligned}$$

har tillåtna punkter. Då har de båda problemen samma optimala värde.

Under förutsättning att åtminstone ett av de båda duala problemen har tillåtna punkter gäller med andra ord:

(i) $X = \emptyset \ \Leftrightarrow$ målfunktionen $\langle b, y \rangle$ är inte uppåt begränsad på Y.
(ii) $Y = \emptyset \ \Leftrightarrow$ målfunktionen $\langle c, x \rangle$ är inte nedåt begränsad på X.
(iii) Om $X \neq \emptyset$ och $Y \neq \emptyset$, så finns det en punkt $\hat{x} \in X$ och en punkt $\hat{y} \in Y$ så att $\langle b, y \rangle \leq \langle b, \hat{y} \rangle = \langle c, \hat{x} \rangle \leq \langle c, x \rangle$ för alla $x \in X$ och alla $y \in Y$.

Dualitetssatsen i linjär programmering är ett specialfall av den allmänna dualitetssatsen för konvexa problem eftersom bivillkorsfunktionerna är affina, men vi ger här ett alternativt bevis som direkt bygger på följande variant av Farkas lemma.

Lemma. *Systemet*

$$\begin{cases} \langle c, x \rangle \leq \alpha \\ \quad x \in X \end{cases} \tag{12.2}$$

har en lösning om och endast om följande två system båda saknar lösningar:

(12.3-A) $\begin{cases} \langle b, y\rangle > \alpha \\ y \in Y \end{cases}$ *och* (12.3-B) $\begin{cases} \langle b, y\rangle = 1 \\ -A^T y \in U \\ y \in V. \end{cases}$

Bevis. Systemet (12.2), dvs.

$$\begin{cases} \langle c, x\rangle \leq \alpha \\ Ax - b \in V^+ \\ x \in U^+ \end{cases}$$

är lösbart om och endast om följande homogeniserade system är lösbart:

(12.2′) $$\begin{cases} \langle c, x\rangle \leq \alpha t \\ Ax - bt \in V^+ \\ x \in U^+ \\ t \in \mathbf{R} \\ t > 0. \end{cases}$$

(Om x löser systemet (12.2), så löser $(x, 1)$ systemet (12.2′), och om (x, t) löser systemet (12.2′), så löser x/t systemet (12.2).) Vi kan skriva systemet (12.2′) på en mer kompakt form genom att införa matrisen

$$\tilde{A} = \begin{bmatrix} \alpha & -c^T \\ -b & A \end{bmatrix}$$

och vektorerna $\tilde{x} = (t, x) \in \mathbf{R} \times \mathbf{R}^n$ och $d = (-1, 0) \in \mathbf{R} \times \mathbf{R}^n$, nämligen som

(12.2″) $$\begin{cases} \tilde{A}\tilde{x} \in \mathbf{R}_+ \times V^+ \\ \tilde{x} \in \mathbf{R} \times U^+ \\ d^T\tilde{x} < 0. \end{cases}$$

Enligt sats 3.3.2 är systemet (12.2″) lösbart om och endast om följande duala system saknar lösningar:

(12.3″) $$\begin{cases} d - \tilde{A}^T\tilde{y} \in \{0\} \times U \\ \tilde{y} \in \mathbf{R}_+ \times V. \end{cases}$$

Eftersom

$$\tilde{A}^T = \begin{bmatrix} \alpha & -b^T \\ -c & A^T \end{bmatrix},$$

kommer systemet (12.3″) utskrivet med $\tilde{y} = (s, y) \in \mathbf{R} \times \mathbf{R}^m$ att få utseendet

$$\text{(12.3}'\text{)} \qquad \begin{cases} -1 - \alpha s + \langle b, y \rangle = 0 \\ cs - A^T y \in U \\ y \in V \\ s \geq 0. \end{cases}$$

Systemet (12.2) är med andra ord lösbart om och endast om systemet (12.3′) saknar lösningar, och genom att dela upp i fallen $s > 0$ och $s = 0$ ser vi att systemet (12.3′) saknar lösningar om och endast om de båda systemen

$$\begin{cases} \langle b, y/s \rangle = \alpha + 1/\mathrm{s} \\ c - A^T(y/s) \in U \\ y/s \in V \\ s > 0 \end{cases} \quad \text{och} \quad \begin{cases} \langle b, y \rangle = 1 \\ -A^T y \in U \\ y \in V \end{cases}$$

saknar lösningar, och detta är uppenbarligen fallet om och endast om systemen (12.3-A) och (12.3-B) båda saknar lösningar. □

Bevis för dualitetssatsen. Vi återgår nu till beviset för dualitetssatsen, och på grund av svag dualitet behöver vi bara visa olikheten

$$\text{(12.4)} \qquad v_{\min}(P) \leq v_{\max}(D).$$

Vi delar upp beviset för denna olikhet på tre fall.

Fall 1, $Y \neq \emptyset$ och systemet (12.3-B) saknar lösning.

Om $v_{\max}(D) = \infty$, så är olikheten (12.4) trivialt uppfylld. Antag därför att $v_{\max}(D) < \infty$. På grund av definitionen av $v_{\max}(D)$ saknar i så fall systemet (12.3-A) lösning för $\alpha = v_{\max}(D)$. Båda systemen (12.3) är med andra ord utan lösning, så det följer av lemmat att systemet (12.2) är lösbart för detta α-värde. Det finns alltså en tillåten punkt $\hat{x}$ så att $\langle c, \hat{x} \rangle \leq v_{\max}(D)$. Följaktligen är $v_{\min}(P) \leq \langle c, \hat{x} \rangle \leq v_{\max}(D)$.

Observera att vi som bonus också får ett bevis för att minimeringsproblemet har en optimal lösning $\hat{x}$.

Fall 2, $Y = \emptyset$ och systemet (12.3-B) saknar lösning.

Nu är systemet (12.3-A) olösbart för alla värden på α, så det följer av lemmat att systemet (12.2) är lösbart för all α-värden, och detta betyder att målfunktionen $\langle c, x \rangle$ inte är nedåt begränsad på X. Följaktligen är $v_{\min}(P) = -\infty = v_{\max}(D)$ i detta fall.

Fall 3, systemet (12.3-B) har en lösning

I detta fall ger lemmat att systemet (12.2) saknar lösning för samtliga värden på α, och detta medför att mängden X av tillåtna punkter är tom. Följaktligen är polyedern Y av tillåtna punkter i det duala problemet icke-tom. Välj en punkt $y_0 \in Y$, låt $\overline{y}$ vara en lösning till systemet (12.3-B) och betrakta punkterna $y^t = y_0 + t\overline{y}$. För $t > 0$ är y^t en konisk kombination av vektorer i V, så y^t ligger i V. Vidare är $c - A^T y^t = (c - A^T y_0) - tA^T\overline{y}$ en konisk kombination av vektorer i U och ligger därför i U. Detta innebär att y^t ligger i Y för $t > 0$, och eftersom

$$\langle b, y^t\rangle = \langle b, y_0\rangle + t\langle b, \overline{y}\rangle = \langle b, y_0\rangle + t \to +\infty$$

då $t \to \infty$, är $v_{\max}(D) = \infty$. Olikheten (12.4) är med andra ord triviallt uppfylld. □

Komplementaritetssatsen

Sats 12.2.4 (Komplementaritetssatsen)**.** *Antag att $\hat{x}$ är en tillåten punkt för LP-problemet (P) och $\hat{y}$ är en tillåten punkt för det duala LP-problemet (D). Då är de båda punkterna optimala för respektive problem om och endast om*

$$\langle c - A^T\hat{y}, \hat{x}\rangle = \langle A\hat{x} - b, \hat{y}\rangle = 0.$$

Bevis. Observera först att på grund av definitionerna av polyedrarna X och Y av tillåtna punkter är $\langle Ax - b, y\rangle \geq 0$ för alla punkter $x \in X$ och $y \in V$, medan $\langle c - A^T y, x\rangle \geq 0$ för alla punkter $y \in Y$ och $x \in U$.

Om $\hat{x}$ och $\hat{y}$ är optimala lösningar till respektive problem, så är därför speciellt $\langle A\hat{x} - b, \hat{y}\rangle \geq 0$ och $\langle c - A^T\hat{y}, \hat{x}\rangle \geq 0$, och på grund av dualitetssatsen är $\langle c, \hat{x}\rangle = \langle b, \hat{y}\rangle$. Det följer att

$$\langle c, \hat{x}\rangle - \langle A\hat{x} - b, \hat{y}\rangle \leq \langle c, \hat{x}\rangle = \langle b, \hat{y}\rangle \leq \langle b, \hat{y}\rangle + \langle c - A^T\hat{y}, \hat{x}\rangle = \langle c, \hat{x}\rangle - \langle A\hat{x} - b, \hat{y}\rangle.$$

Eftersom de båda ytterleden i denna olikhet är lika, råder det likhet på alla ställena, dvs. $\langle A\hat{x} - b, \hat{y}\rangle = \langle c - A^T\hat{y}, \hat{x}\rangle = 0$.

Omvänt, om $\langle c - A^T\hat{y}, \hat{x}\rangle = \langle A\hat{x} - b, \hat{y}\rangle = 0$, så är $\langle c, \hat{x}\rangle = \langle A^T\hat{y}, \hat{x}\rangle$ och $\langle b, \hat{y}\rangle = \langle A\hat{x}, \hat{y}\rangle$, och eftersom $\langle A^T\hat{y}, \hat{x}\rangle = \langle A\hat{x}, \hat{y}\rangle$, är således $\langle c, \hat{x}\rangle = \langle b, \hat{y}\rangle$. Det följer därför av optimalitetskriteriet att de båda punkterna är optimala. □

Låt oss för tydlighets skull formulera komplementaritetssatsen i det viktiga specialfall då de duala problemen har det utseende som beskrivs som fall 2 i exempel 12.2.1.

Korollarium 12.2.5. *Antag att $\hat{x}$ och $\hat{y}$ är tillåtna punkter i de duala problemen*

$$(\mathrm{P}_2) \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{då} \quad Ax \geq b,\ x \geq 0 \end{aligned}$$

resp.

$$(\mathrm{D}_2) \qquad \begin{aligned} &\max \ \langle b, y\rangle \\ &\text{då} \quad A^T y \leq c,\ y \geq 0. \end{aligned}$$

Då är $\hat{x}$ och $\hat{y}$ optimala lösningar om och endast om

$$\begin{cases} (A\hat{x})_i > b_i & \Rightarrow \quad \hat{y}_i = 0 \\ \hat{x}_j > 0 & \Rightarrow \quad (A^T\hat{y})_j = c_j \end{cases} \tag{12.5}$$

dvs. om $\hat{x}$ satisfierar den i:te olikheten i systemet $Ax \geq b$ strikt, så är motsvarande duala variabel $\hat{y}_i = 0$, och om $\hat{x}_j > 0$, så satisfierar $\hat{y}$ systemet $A^T y \leq c$ med likhet i den j:te olikheten.

Bevis. I detta fall är nämligen $\langle A\hat{x} - b, \hat{y}\rangle = \sum_{i=1}^{m}((A\hat{x})_i - b_i)\hat{y}_i$ en summa av icke-negativa termer, så $\langle A\hat{x} - b, \hat{y}\rangle = 0$ om och endast om alla termer är lika med 0, dvs. om och endast om $(A\hat{x})_i > b_i$ medför att $\hat{y}_i = 0$.

Analogt fås att $\langle c - A^T\hat{y}, \hat{x}\rangle = 0$ gäller om och endast om $\hat{x}_j > 0$ medför att $(A^T\hat{y})_j = c_j$. Korollariet är därför bara en omformulering av föregående sats i det fall då de duala problemen har den speciella formen i fall 2 av exempel 12.2.1. □

Den nyfikne läsaren undrar måhända om implikationerna i villkoret (12.5) kan ersättas med ekvivalenser. Följande triviala exempel visar att så inte är fallet.

Exempel 12.2.4. Betrakta de duala problemen

$$\begin{aligned} &\min \ x_1 + 2x_2 \\ &\text{då} \quad x_1 + 2x_2 \geq 2,\ x \geq 0 \end{aligned} \qquad \text{och} \qquad \begin{aligned} &\max \ 2y \\ &\text{då} \quad \begin{cases} y \leq 1 \\ 2y \leq 2,\ y \geq 0 \end{cases} \end{aligned}$$

med $A = c^T = \begin{bmatrix} 1 & 2 \end{bmatrix}$ och $b = [2]$. För de optimala punkterna $\hat{x} = (2, 0)$ och $\hat{y} = 1$ är $\hat{x}_2 = 0$ och $(A^T\hat{y})_2 = 2 = c_2$, vilket visar att (12.5) i det här fallet inte gäller med ekvivalens.

Minimeringsproblemet har emellertid flera optimala lösningar; punkterna på sträckan mellan $(2, 0)$ och $(0, 1)$ är optimala, och för $0 < t < 1$ uppfyller de optimala paren $\hat{x} = (2 - 2t, t)$ och $\hat{y} = 1$ villkoret (12.5) med ekvivalens. □

Den sista slutsatsen i exemplet ovan kan generaliseras. I alla duala problem med tillåtna punkter finns det ett par av optimala lösningar $\hat{x}$ och $\hat{y}$ som uppfyller (12.5) med implikationerna ersatta av ekvivalenser. Se övning 12.8.

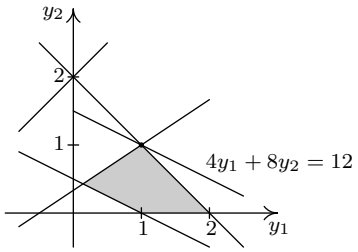
EXEMPEL 12.2.5. LP-problemet

$$\begin{array}{ll} \min & -x_1 + 2x_2 + x_3 + 2x_4 \\ \text{då} & \begin{cases} -x_1 - x_2 - 2x_3 + x_4 \geq 4 \\ -2x_1 + x_2 + 3x_3 + x_4 \geq 8 \\ x_1, x_2, x_3, x_4 \geq 0 \end{cases} \end{array}$$

löser man lätt genom att först lösa det duala problemet

$$\begin{array}{ll} \max & 4y_1 + 8y_2 \\ \text{då} & \begin{cases} -y_1 - 2y_2 \leq -1 \\ -y_1 + y_2 \leq 2 \\ -2y_1 + 3y_2 \leq 1 \\ y_1 + y_2 \leq 2 \\ y_1, y_2 \geq 0 \end{cases} \end{array}$$

grafiskt och sedan använda komplementaritetssatsen.



Figur 12.3. Grafisk lösning till maximeringsproblemet i exempel 12.2.5.

Den grafiska lösningen fås ur figur 12.3, som visar att $\hat{y} = (1, 1)$ är en optimal punkt och att värdet är 12. Eftersom $\hat{y}$ satisfierar de två första bivillkoren med sträng olikhet och $\hat{y}_1 > 0$ och $\hat{y}_2 > 0$, är den optimala lösningen $\hat{x}$ till minimeringsproblemet en lösning till systemet

$$\begin{cases} -x_1 - x_2 - 2x_3 + x_4 = 4 \\ -2x_1 + x_2 + 3x_3 + x_4 = 8 \\ x_1 = 0 \\ x_2 = 0 \\ x_1, x_2, x_3, x_4 \geq 0. \end{cases}$$

Detta system har lösningen $\hat{x} = \left(0, 0, \frac{4}{5}, \frac{28}{5}\right)$. Som man lätt verifierar är värdet i denna punkt 12, vilket det ju skall vara enligt dualitetssatsen. □

Övningar

12.1 I LP-problemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax \geq b \end{array}$$

antas A och c vara givna medan högerledsvektorn b får variera. Antag att problemet har ett ändligt värde för något högerled b. Visa att för varje högerled b är i så fall antingen värdet ändligt eller också saknas det tillåtna punkter. Visa också att problemets optimala värde är en konvex funktion av b.

12.2 Ge exempel på duala problem som båda saknar tillåtna punkter.

12.3 Använd dualitet för att visa att $(3, 0, 1)$ är en optimal lösning till LP-problemet

$$\begin{array}{ll} \min & 2x_1 + 4x_2 + 3x_3 \\ \text{då} & \begin{cases} 2x_1 + 3x_2 + 4x_3 \geq 10 \\ x_1 + 2x_2 \qquad\ \geq 3 \\ 2x_1 + 7x_2 + 2x_3 \geq 5, \ \ x \geq 0. \end{cases} \end{array}$$

12.4 Visa att kolonnspelarens och radspelarens problem i ett tvåpersoners nollsummespel (se kapitel 9.4) är duala problem.

12.5 Undersök hur den optimala lösningen till LP-problemet

$$\begin{array}{ll} \max & x_1 + x_2 \\ \text{då} & \begin{cases} tx_1 + x_2 \geq -1 \\ x_1 \qquad\ \leq 2 \\ x_1 - x_2 \geq -1 \end{cases} \end{array}$$

varierar då den reella parametern t varierar.

12.6 Dualitetssatsen följer ur Farkas lemma (korollarium 3.3.3). Visa omvänt att Farkas lemma följer ur dualitetssatsen genom att betrakta de duala problemen

$$\begin{array}{lll} \min & \langle c, x\rangle & \text{och} \\ \text{då} & Ax \geq 0 & \end{array} \qquad \begin{array}{ll} \max & \langle 0, y\rangle \\ \text{då} & A^T y = c, \ y \geq 0 \end{array}$$

12.7 Sätt $Y = \{y \in \mathbf{R}^m \mid c - A^T y \in U, \ y \in V\}$, där U och V är slutna konvexa koner, och antag att $Y \neq \emptyset$.

a) Visa att $\operatorname{recc} Y = \{y \in \mathbf{R}^m \mid -A^T y \in U, \ y \in V\}$.

b) Visa att systemet (12.3-B) lösbart om och endast om $-b$ inte ligger i dualkonen till $\operatorname{recc} Y$.

c) Använd resultatet i b) för att visa att slutsatsen i fall 3 av beviset för dualitetssatsen, dvs. att $v_{\max}(D) = \infty$ om (och endast om) systemet (12.3-B) har en lösning, följer av sats 12.1.1.

12.8 Antag att de duala problemen

$$\begin{array}{lll} \min & \langle c, x\rangle & \text{och} \\ \text{då} & Ax \geq b,\ x \geq 0 & \end{array} \qquad \begin{array}{lll} \max & \langle b, y\rangle & \\ \text{då} & A^T y \leq c,\ y \geq 0 & \end{array}$$

båda har tillåtna punkter. Visa att det finns optimala lösningar $\hat{x}$ och $\hat{y}$ i respektive problem som uppfyller

$$\begin{cases} (A\hat{x})_i > b_i & \Leftrightarrow \quad \hat{y}_i = 0 \\ \hat{x}_j > 0 & \Leftrightarrow \quad (A^T\hat{y})_j = c_j. \end{cases}$$

[Ledning: På grund av komplementaritetssatsen räcker det att visa att följande system av olikheter har en lösning: $Ax \geq b$, $x \geq 0$, $A^T y \leq c$, $y \geq 0$, $\langle b, y\rangle \geq \langle c, x\rangle$, $Ax + y > b$, $Ay - c < x$. Detta system är lösbart om och endast om följande homogena system är lösbart: $Ax - bt \geq 0$, $x \geq 0$, $-A^T y + ct \geq 0$, $y \geq 0$, $-\langle c, x\rangle + \langle b, y\rangle \leq 0$, $Ax + y - bt > 0$, $x - A^T y + ct > 0$, $t > 0$. Lösbarheten kan nu avgöras med hjälp av sats 3.3.7.]

Del III

Simplexalgoritmen

Kapitel 13

Simplexalgoritmen

För praktiskt bruk finns det något förenklat två olika slags lösningsmetoder för LP-problem. I båda genereras en följd av tillåtna punkter med successivt bättre målfunktionsvärden. *Simplex*metoderna introducerades av Dantzig i slutet av 1940-talet och i dem är de genererade punkterna extremalpunkter till polyedern av tillåtna punkter, och man erhåller dem genom att förflytta sig utefter polyederns kantlinjer. I *inrepunkt*-metoderna genereras i stället punkter i det inre av polyedern; dessa metoder härstammar från tekniker för icke-linjär programmering som utvecklades av Fiacco och McCormick på 1960-talet, men det var först efter Karmarkars nydanande analys 1984 som metoderna började användas på LP-problem.

I det här kapitlet skall vi beskriva och analysera simplexalgoritmen.

13.1 Standarform

För att lösa ett LP-problem med simplexalgoritmen måste man först formulera problemet på en speciell form. Den variant av simplexalgoritmen som vi skall studera förutsätter att problemet är ett minimeringsproblem, att alla ingående variabler är icke-negativa och att alla övriga bivillkor är formulerade som likheter.

Definition. Ett LP-problem har *standardform* om det har formen

$$\begin{array}{ll} \min & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{då} & \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\ x_1, x_2, \ldots, x_n \geq 0. \end{cases} \end{array}$$

Genom att införa koefficientmatriserna

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \quad \text{och} \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

får vi följande kompakta skrivsätt för ett LP-problem på standardform:

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax = b,\ x \geq 0. \end{array}$$

Som vi har sett i kapitel 9 kan man transformera varje LP-problem till ett ekvivalent LP-problem på standardform med hjälp av slack/surplus-variabler och genom att ersätta oinskränkta variabler med differenser av icke-negativa variabler.

Dualitet

I kapitel 12.2 gav vi en allmän definition av begreppet dualitet och visade att duala LP-problem har samma optimala värde utom i det fall då båda problemen saknar tillåtna punkter. Vi kommer i vår beskrivning av simplexalgoritmen att behöva ett specialfall av dualitet, och för att göra framställningen oberoende av resultaten i föregående kapitel upprepar vi definitionen av dualitet för detta specialfall.

Definition. LP-problemet

$$\text{(D)} \qquad \begin{array}{ll} \max & \langle b, y \rangle \\ \text{då} & A^T y \leq c \end{array}$$

kallas *dualt* till LP-problemet

$$\text{(P)} \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax = b,\ x \geq 0. \end{array}$$

Vi kommer att utnyttja följande triviala del av dualitetssatsen.

Sats 13.1.1 (Svag dualitet)**.** *Om x är en tillåten punkt i minimeringsproblemet (P) och y är en tillåten punkt i det duala maximeringsproblemet (D), dvs. om $Ax = b$, $x \geq 0$ och $A^T y \leq c$, så är*

$$\langle b, y \rangle \leq \langle c, x \rangle.$$

Bevis. Av olikheterna $A^T y \leq c$ och $x \geq 0$ följer att $\langle x, A^T y \rangle \leq \langle x, c \rangle$. Därför är

$$\langle b, y \rangle = \langle Ax, y \rangle = \langle x, A^T y \rangle \leq \langle x, c \rangle = \langle c, x \rangle. \qquad \square$$

Korollarium 13.1.2 (Optimalitetskriteriet)**.** *Om $\overline{x}$ är en tillåten punkt för minimeringsproblemet (P), $\overline{y}$ är en tillåten punkt för det duala maximeringsproblemet (D) och $\langle b, \overline{y}\rangle = \langle c, \overline{x}\rangle$, så är $\overline{x}$ och $\overline{y}$ optimala punkter i respektive problem.*

Bevis. Av förutsättningarna och sats 13.1.1, tillämpad på punkten $\overline{y}$ och en godtycklig tillåten punkt x i minimeringsproblemet, följer att

$$\langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle \leq \langle c, x\rangle$$

för alla tillåtna punkter x. Detta innebär att $\overline{x}$ är en minimipunkt. Analogt visas att $\overline{y}$ är en maximipunkt. □

13.2 Informell beskrivning av simplexalgoritmen

I det här avsnittet skall vi beskriva huvuddragen i simplexalgoritmen med hjälp av några enkla exempel. Den precisa formuleringen av algoritmen och beviset för att den fungerar ges i avsnitten 13.4 och 13.5.

Exempel 13.2.1. Vi börjar med ett helt trivialt problem, nämligen

$$\begin{array}{ll} \min & f(x) = x_3 + 2x_4 \\ \text{då} & \begin{cases} x_1 \quad + 2x_3 - x_4 = 2 \\ \quad x_2 - \ x_3 + x_4 = 3, \ x \geq 0. \end{cases} \end{array}$$

Eftersom koefficienterna i målfunktionen $f(x)$ är positiva och $x \geq 0$, är det klart att $f(x) \geq 0$ för alla tillåtna punkter x. Det finns vidare en tillåten punkt med $x_3 = x_4 = 0$, ty insättning i bivillkoren ger $x_1 = 2$ och $x_2 = 3$. Minimum är därför lika med 0 och $x = (2, 3, 0, 0)$ är problemets (unika) minimipunkt. □

Betrakta nu ett godtyckligt problem på formen

$$\begin{array}{ll} \min & c_{m+1}x_{m+1} + \cdots + c_n x_n + d \\ \text{då} & \begin{cases} x_1 + \ a_{1\,m+1}x_{m+1} + \ldots + \ a_{1n}x_n = b_1 \\ x_2 + \ a_{2\,m+1}x_{m+1} + \ldots + \ a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ x_m + a_{m\,m+1}x_{m+1} + \ldots + a_{mn}x_n = b_m, \ x \geq 0 \end{cases} \end{array} \tag{13.1}$$

där

$$b_1, b_2, \ldots, b_m \geq 0.$$

Om $c_{m+1}, c_{m+2}, \ldots, c_n \geq 0$, så fungerar resonemanget i exempel 13.2.1 och det följer att minimum är lika med d och antas för $x = (b_1, \ldots, b_m, 0, \ldots, 0)$.

Ekvationssystemet i (13.1) har en mycket speciell form, ty det är löst med avseende på variablerna $x_1, x_2, \ldots, x_m$, och dessa variabler förekommer inte i målfunktionen. Rent generellt kommer en uppsättning av variabler i ett ekvationssystem att kallas *basvariabler* om det är möjligt att lösa systemet med avseende på variablerna i uppsättningen.

EXEMPEL 13.2.2. Låt oss ändra målfunktionen i exempel 13.2.1 genom att byta tecken på koefficienten för x_3. Vårt nya problem lyder således

$$\begin{aligned} &\min \ f(x) = -x_3 + 2x_4 \\ &\text{då} \ \begin{cases} x_1 \quad + 2x_3 - x_4 = 2 \\ \quad x_2 - \ x_3 + x_4 = 3, \ x \geq 0. \end{cases} \end{aligned} \tag{13.2}$$

Punkten $(2, 3, 0, 0)$ är naturligtvis fortfarande tillåten och motsvarande funktionsvärde är 0, men vi erhåller ett mindre målfunktionsvärde genom att välja $x_3 > 0$ och behålla $x_4 = 0$. Vi måste dock se till att $x_1 \geq 0$ och $x_2 \geq 0$, så den första bivillkorsekvationen ger oss begränsningen $x_1 = 2 - 2x_3 \geq 0$, dvs. $x_3 \leq 1$.

Vi transformerar nu problemet genom att lösa ekvationssystemet i (13.2) med avseende på variablerna x_2 och x_3, dvs. genom att byta basvariabler från x_1, x_2 till x_2, x_3. Gausselimination ger

$$\begin{cases} \frac{1}{2}x_1 \quad + x_3 - \frac{1}{2}x_4 = 1 \\ \frac{1}{2}x_1 + x_2 \quad + \frac{1}{2}x_4 = 4. \end{cases}$$

Därefter elimineras basvariabeln x_3 ur målfunktionen med hjälp av den första ekvationen i det nya systemet. Vi får då

$$f(x) = \tfrac{1}{2}x_1 + \tfrac{3}{2}x_4 - 1.$$

Vårt problem har därför reducerats till ett problem på formen (13.1), nämligen

$$\begin{aligned} &\min \ \tfrac{1}{2}x_1 + \tfrac{3}{2}x_4 - 1 \\ &\text{då} \ \begin{cases} \frac{1}{2}x_1 \quad + x_3 - \frac{1}{2}x_4 = 1 \\ \frac{1}{2}x_1 + x_2 \quad + \frac{1}{2}x_4 = 4, \ x \geq 0 \end{cases} \end{aligned}$$

med x_2 och x_3 som basvariabler och med icke-negativa koefficienter för övriga variabler i målfunktionen. Funktionens minimivärde är således lika med -1 och det antas i punkten $(0, 4, 1, 0)$. □

Strategin för att lösa ett problem av typen (13.1), där någon koefficient c_{m+k} är negativ, består således i att byta ut någon av basvariablerna $x_1, x_2, \ldots, x_m$ mot x_{m+k} för att på så sätt erhålla ett nytt problem som har formen (13.1). Om de nya "c-koefficienterna" är icke-negativa är saken klar, annars upprepas proceduren. Vi illustrerar med ytterligare ett exempel.

EXEMPEL 13.2.3. Betrakta problemet

$$\begin{aligned} &\min \ f(x) = 2x_1 - x_2 + x_3 - 3x_4 + x_5 \\ &\text{då} \ \begin{cases} x_1 \qquad\quad + 2x_4 - \ x_5 = 5 \\ \quad x_2 \quad + \ x_4 + 3x_5 = 4 \\ \qquad x_3 - \ x_4 + \ x_5 = 3, \ x \geq 0. \end{cases} \end{aligned} \tag{13.3}$$

Först måste vi eliminera basvariablerna x_1, x_2, x_3 ur målfunktionen och får då

$$f(x) = -5x_4 + 5x_5 + 9. \tag{13.4}$$

Eftersom koefficienten för x_4 är negativ, skall vi eliminera x_4 ur målfunktionen och ur två av bivillkorsekvationerna på ett sådant sätt att högerleden i det transformerade ekvationssystemet förblir icke-negativa. Den tredje ekvationen i (13.3) kan inte utnyttjas för denna elimination beroende på att koefficienten för x_4 är negativ. Om vi eliminerar x_4 ur den första ekvationen med hjälp av den andra, så får den resulterande ekvationen $x_1 - 2x_2 - 7x_5 = 5 - 2 \cdot 4 = -3$ ett otillåtet högerled. Det återstår således endast att utnyttja det första av bivillkoren i (13.3) för eliminationen. Vi får då följande ekvivalenta system

$$\begin{cases} \frac{1}{2}x_1 \qquad\quad + x_4 - \frac{1}{2}x_5 = \frac{5}{2} \\ -\frac{1}{2}x_1 + x_2 \qquad\quad + \frac{7}{2}x_5 = \frac{3}{2} \\ \frac{1}{2}x_1 \quad + x_3 \quad + \frac{1}{2}x_5 = \frac{11}{2}, \ x \geq 0 \end{cases} \tag{13.5}$$

med x_2, x_3, x_4 som nya basvariabler. Att det nya högerledet blir positivt då den första ekvationen i (13.3) utnyttjas för eliminationen av x_4 beror på att kvoten mellan högerled och x_4-koefficient är mindre för den första ekvationen än för den andra ($5/2 < 4/1$).

Vi eliminerar nu x_4 ur målfunktionen (13.4) och erhåller då

$$f(x) = \tfrac{5}{2}x_1 + \tfrac{5}{2}x_5 - \tfrac{7}{2}$$

som alltså skall minimeras under bivillkoren (13.5). Uppenbarligen är minimivärdet lika med $-\frac{7}{2}$, och minimipunkten är $x = (0, \frac{3}{2}, \frac{11}{2}, \frac{5}{2}, 0)$.

För att nedbringa skrivarbetet brukar man utelämna variablerna och enbart arbeta med koefficienterna i tabellform. Vi kommer att använda oss av följande uppställning. Problemet (13.3) representeras av *simplextabellen*

$$\begin{array}{ccccc|c} 1 & 0 & 0 & 2 & -1 & 5 \\ 0 & 1 & 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & -1 & 1 & 3 \\ \hline 2 & -1 & 1 & -3 & 1 & f \end{array}$$

Den övre delen av tabellen består av ekvationssystemets totalmatris, och den undre raden representerar målfunktionen f. Det vertikala strecket svarar mot likhetstecknen i (13.3).

För att eliminera basvariablerna x_1, x_2, x_3 ur målfunktionen behöver vi bara addera -2 gånger rad 1, 1 gånger rad 2 samt -1 gånger rad 3 till målfunktionsraden i tabellen ovan. Detta ger oss den nya tabellen

$$\begin{array}{ccccc|c} 1 & 0 & 0 & \underline{2} & -1 & 5 \\ 0 & 1 & 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & -1 & 1 & 3 \\ \hline \underline{0} & \underline{0} & \underline{0} & -5 & 5 & f-9 \end{array}$$

Den undre raden i ovanstående tabell svarar mot ekvation (13.4). Observera dock att konstanttermen 9 förekommer på andra sidan om likhetstecknet jämfört med (13.4), vilket förklarar minustecknet i tabellen. Vi har vidare markerat basvariabelkolonnerna genom understrykning.

Eftersom x_4-koefficienten i målfunktionen är negativ, skall tabellen transformeras så att x_4 blir ny basvariabel. Genom att jämföra kvoterna $5/2$ och $4/1$ kommer vi fram till att den första raden skall vara *pivotrad*, dvs. tas som utgångspunkt för eliminationen. Vi har därför strukit under koefficienten 2 i första raden och fjärde kolonnen (det s. k. *pivotelementet*).

Gausselimination ger upphov till den nya tabellen:

$$\begin{array}{ccccc|c} \frac{1}{2} & 0 & 0 & 1 & -\frac{1}{2} & \frac{5}{2} \\ -\frac{1}{2} & 1 & 0 & 0 & \frac{7}{2} & \frac{3}{2} \\ \frac{1}{2} & 0 & 1 & 0 & \frac{1}{2} & \frac{11}{2} \\ \hline \frac{5}{2} & \underline{0} & \underline{0} & \underline{0} & \frac{5}{2} & f+\frac{7}{2} \end{array}$$

Eftersom den nya målfunktionens koefficienter är icke-negativa, kan vi nu läsa av minimum med *ombytt tecken* i den nedre högra rutan. Minimipunkten fås genom att sätta icke-basvariablerna x_1 och x_5 lika med 0, vilket ger $x = (0, \frac{3}{2}, \frac{11}{2}, \frac{5}{2}, 0)$. □

EXEMPEL 13.2.4. Till problemet

$$\begin{array}{ll} \min & x_1 - 2x_2 + x_3 \\ \text{då} & \begin{cases} x_1 + 2x_2 + 2x_3 + x_4 = 5 \\ x_1 + x_3 + x_5 = 2 \\ x_2 - 2x_3 + x_6 = 1, \ x \geq 0 \end{cases} \end{array}$$

hör simplextabellen

$$\begin{array}{cccccc|c} 1 & 2 & 2 & 1 & 0 & 0 & 5 \\ 1 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & \underline{1} & -2 & 0 & 0 & 1 & 1 \\ \hline 1 & -2 & 1 & \underline{0} & \underline{0} & \underline{0} & f \end{array}$$

Variablerna x_4, x_5, x_6 är basvariabler, och dessa är redan eliminerade ur målfunktionen. Eftersom målfunktionens x_2-koefficient är negativ, skall vi introducera x_2 som ny basvariabel. Vi måste välja det understrukna elementet som pivotelement, ty $1/1 < 5/2$. Med hjälp av den tredje raden transformeras tabellen till

$$\begin{array}{cccccc|c} 1 & 0 & \underline{6} & 1 & 0 & -2 & 3 \\ 1 & 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 1 & -2 & 0 & 0 & 1 & 1 \\ \hline 1 & \underline{0} & -3 & \underline{0} & \underline{0} & 2 & f+2 \end{array}$$

Den sistnämnda tabellen svarar förstås mot problemet

$$\begin{array}{ll} \min & x_1 - 3x_3 + 2x_6 - 2 \\ \text{då} & \begin{cases} x_1 + 6x_3 + x_4 - 2x_6 = 3 \\ x_1 + x_3 + x_5 = 2 \\ x_2 - 2x_3 + x_6 = 1, \ x \geq 0. \end{cases} \end{array}$$

Eftersom x_3-koefficienten i målfunktionen är negativ, måste vi upprepa proceduren. Vi skall alltså införa x_3 som basvariabel, och denna gång skall den första raden användas som pivotrad, ty $3/6 < 2/1$. Den nya tabellen får utseendet:

$$\begin{array}{cccccc|c} \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 & -\frac{1}{3} & \frac{1}{2} \\ \frac{5}{6} & 0 & 0 & -\frac{1}{6} & 1 & \frac{1}{3} & \frac{3}{2} \\ \frac{1}{3} & 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 2 \\ \hline \frac{3}{2} & \underline{0} & \underline{0} & \frac{1}{2} & \underline{0} & 1 & f+\frac{7}{2} \end{array}$$

Vi kan nu avläsa minimivärdet $-\frac{7}{2}$. Minimum antas för $x = (0, 2, \frac{1}{2}, 0, \frac{3}{2}, 0)$. □

Hittills har vi skrivit funktionssymbolen f i nedre högra rutan av våra simplextabeller. Detta har vi gjort av pedagogiska skäl för att förklara varför funktionsvärdet i rutan får ombytt tecken. Kom t. ex. ihåg att den sista raden i närmast föregående simplextabell betyder att

$$\tfrac{3}{2}x_1 + \tfrac{1}{2}x_4 + x_6 = f(x) + \tfrac{7}{2}.$$

Eftersom symbolen inte har någon annan funktion, kommer vi att utelämna den i fortsättningen.

Exempel 13.2.5. Problemet

$$\begin{array}{ll} \min & f(x) = -2x_1 + x_2 \\ \text{då} & \begin{cases} x_1 - x_2 + x_3 = 3 \\ -x_1 + x_2 + x_4 = 4, \ x \geq 0 \end{cases} \end{array}$$

ger upphov till följande simplextabeller:

$$\begin{array}{rrrr|r} \underline{1} & -1 & 1 & 0 & 3 \\ -1 & 1 & 0 & 1 & 4 \\ \hline -2 & 1 & \underline{0} & \underline{0} & 0 \end{array}$$

$$\begin{array}{rrrr|r} 1 & -1 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 & 7 \\ \hline \underline{0} & -1 & 2 & \underline{0} & 6 \end{array}$$

Eftersom x_2-koefficienten i målfunktionen är negativ, skulle vi nu egentligen eliminera x_2, men det finns ingen rad som duger som pivotrad eftersom hela x_2-kolonnen i tabellen är icke-positiv. Detta innebär att målfunktionen är nedåt obegränsad, dvs. minimum saknas. För att se detta skriver vi om den sista tabellen med variabler på formen

$$\begin{array}{ll} \min & f(x) = -x_2 + 2x_3 - 6 \\ \text{då} & \begin{cases} x_1 = x_2 - x_3 + 3 \\ x_4 = \quad -x_3 + 7. \end{cases} \end{array}$$

Genom att välja $x_2 = t$ och $x_3 = 0$ får vi en tillåten punkt

$$x^t = (3 + t, t, 0, 7)$$

för varje $t \geq 0$, och eftersom $f(x^t) = -t - 6$, är målfunktionen nedåt obegränsad. □

Exemplen 13.2.4 och 13.2.5 är typiska för problem på formen (13.1). I avsnitt 13.5 skall vi nämligen visa att man alltid kan utföra iterationerna så att man får en sluttabell liknande den i exempel 13.2.4 eller den i exempel 13.2.5, och i avsnitt 13.6 skall vi visa hur man kommer igång, dvs. hur man transformerar ett godtyckligt standardproblem så att det får formen (13.1).

13.3 Baslösningar

För att beskriva och förstå simplexalgoritmen måste man först veta hur man producerar lösningar till ett linjärt ekvationssystem. Vi utgår ifrån att Gauss eliminationsmetod är välbekant och koncentrerar oss på att beskriva hur man byter från en s. k. baslösning till en annan. Vi börjar med att gå igenom den notation som vi kommer att använda oss av i resten av det här kapitlet.

De n kolonnerna i en $m \times n$-matris A kommer att betecknas A_{*1}, A_{*2}, $\ldots$, A_{*n} så att

$$A = \left[A_{*1} \; A_{*2} \; \ldots \; A_{*n} \right].$$

Vi kommer ofta att behöva betrakta delmatriser bestående av vissa kolonner i en $m \times n$-matris A. Om $1 \leq k \leq n$ och

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$$

är en permutation av k stycken element tagna från mängden $\{1, 2, \ldots, n\}$, låter vi därför $A_{*\alpha}$ beteckna $m \times k$-matrisen som består av kolonnerna $A_{*\alpha_1}$, $A_{*\alpha_2}$, $\ldots$, $A_{*\alpha_k}$ i matrisen A, dvs.

$$A_{*\alpha} = \left[A_{*\alpha_1} \; A_{*\alpha_2} \; \ldots \; A_{*\alpha_k} \right].$$

Om

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

är en kolonnmatris med n element, låter vi på motsvarande sätt x_α beteckna kolonnmatrisen

$$\begin{bmatrix} x_{\alpha_1} \\ x_{\alpha_2} \\ \vdots \\ x_{\alpha_k} \end{bmatrix}.$$

Som vanligt skiljer vi inte på kolonnmatriser med n element och vektorer i $\mathbf{R}^n$.

Vi uppfattar permutationer $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ som ordnade mängder, och tillåter oss därför att skriva $j \in \alpha$ om j är något av talen $\alpha_1, \alpha_2, \ldots, \alpha_k$. Detta gör att vi nu kan skriva summor av typen

$$\sum_{i=1}^{k} x_{\alpha_i} A_{*\alpha_i}$$

som

$$\sum_{j\in\alpha} x_j A_{*j},$$

och på matrisform som

$$A_{*\alpha}x_\alpha.$$

Definition. Låt A vara en $m \times n$-matris av rang m. En permutation

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$$

av m stycken av talen $\{1, 2, \ldots, n\}$ kallas en *basindexmängd* till matrisen A om kolonnerna i $m \times m$-matrisen $A_{*\alpha}$ bildar en bas för $\mathbf{R}^m$.

Att kolonnerna $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ bildar en bas är ekvivalent med att delmatrisen

$$A_{*\alpha} = \left[\, A_{*\alpha_1}\ A_{*\alpha_2}\ \ldots\ A_{*\alpha_m} \,\right]$$

är inverterbar. Inversen till matrisen $A_{*\alpha}$ kommer att betecknas $A_{*\alpha}^{-1}$. Denna matris, som alltså betyder $(A_{*\alpha})^{-1}$, kommer att förekomma ofta i fortsättningen − förväxla den inte med $(A^{-1})_{*\alpha}$ som ju i allmänhet inte är definierad.

Om $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ är en basindexmängd, så är förstås varje permutation av elementen i α också en basindexmängd.

Exempel 13.3.1. Matrisen

$$\begin{bmatrix} 3 & 1 & 1 & -3 \\ 3 & -1 & 2 & -6 \end{bmatrix}$$

har följande basindexmängder: $(1,2)$, $(2,1)$, $(1,3)$, $(3,1)$, $(1,4)$, $(4,1)$ $(2,3)$, $(3,2)$, $(2,4)$ och $(4,2)$. □

Vi behöver vidare ett bekvämt sätt att redovisa resultatet av att byta ut ett element i en ordnad mängd mot något annat element. Låt därför $M = (a_1, a_2, \ldots, a_n)$ vara en godtycklig n-tipel (ordnad mängd). Den n-tipel som fås genom att byta ut objektet a_r på plats r mot ett godtyckligt objekt x kommer att betecknas $M_{\hat{r}}[x]$. Med andra ord är

$$M_{\hat{r}}[x] = (a_1, \ldots, a_{r-1}, x, a_{r+1}, \ldots, a_n).$$

Vi kan förstås uppfatta en $m \times n$-matris A som en ordnad mängd av sina kolonner. Om b är en kolonnmatris med m element och $1 \le r \le n$, skriver vi därför $A_{\hat{r}}[b]$ för matrisen

$$\left[\, A_{*1}\ \ldots\ A_{*r-1}\ b\ A_{*r+1}\ \ldots\ A_{*n} \,\right].$$

Ett annat sammanhang där vi kommer att använda ovanstående notation för byte av element, är då $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ är en permutation av m stycken element tagna från mängden $\{1, 2, \ldots, n\}$. Om $1 \leq r \leq m$, $1 \leq k \leq n$ och $k \notin \alpha$, så är $\alpha_{\hat{r}}[k]$ den nya permutationen

$$(\alpha_1, \ldots, \alpha_{r-1}, k, \alpha_{r+1}, \ldots, \alpha_m).$$

Längre fram kommer vi att behöva följande enkla resultat, där ovanstående beteckningssätt kommer till användning.

Lemma 13.3.1. *Låt E vara enhetsmatrisen av ordning m, och låt b vara en kolonnmatris med m element. Då är matrisen $E_{\hat{r}}[b]$ inverterbar om och endast om $b_r \neq 0$, och i så fall är*

$$E_{\hat{r}}[b]^{-1} = E_{\hat{r}}[c],$$

där

$$c_j = \begin{cases} -b_j/b_r & \text{för } j \neq r \\ 1/b_r & \text{för } j = r. \end{cases}$$

Bevis. Beviset lämnas som enkel övning. □

Exempel 13.3.2.

$$\begin{bmatrix} 1 & 4 & 0 \\ 0 & 3 & 0 \\ 0 & 5 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -4/3 & 0 \\ 0 & 1/3 & 0 \\ 0 & -5/3 & 1 \end{bmatrix}$$

□

Linjära ekvationssystem och baslösningar

Betrakta ett linjärt ekvationssystem

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases} \tag{13.6}$$

med koefficientmatris A av rang m och högerledsmatris b. Ett sådant system kan också ekvivalent uppfattas som en vektorekvation

$$\sum_{j=1}^{n} x_j A_{*j} = b \tag{13.6′}$$

eller som en matrisekvation

$$Ax = b. \tag{13.6''}$$

Båda alternativa synsätten är, som vi skall se, fruktbara.

Att lösa ekvationssystemet (13.6), vilket man som bekant gör med hjälp av Gausselimination, innebär att man uttrycker m stycken av systemets variabler, $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ säg, som linjärkombinationer av de övriga $n-m$ variablerna $x_{\beta_1}, x_{\beta_2}, \ldots, x_{\beta_{n-m}}$ och $b_1, b_2, \ldots, b_m$. För varje tilldelning av värden till de sistnämnda β-variablerna får man en unik uppsättning av värden på de förstnämnda α-variablerna, och speciellt får man en unik lösning genom att sätta alla β-variabler lika med 0.

Detta motiverar följande definition.

Definition. Låt $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ vara en permutation av m tal tagna från mängden $\{1, 2, \ldots, n\}$, och låt $\beta = (\beta_1, \beta_2, \ldots, \beta_{n-m})$ vara en permutation av resterande $n-m$ tal. Variablerna $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ kallas *basvariabler* och variablerna $x_{\beta_1}, x_{\beta_2}, \ldots, x_{\beta_{n-m}}$ kallas *fria variabler* i systemet (13.6), om det för varje $c = (c_1, c_2, \ldots, c_{n-m}) \in \mathbf{R}^{n-m}$ finns en unik lösning x till (13.6) med $x_\beta = c$. Speciellt kallas den unika lösning som fås genom att sätta alla fria variabler lika med 0 för en *baslösning*.

Vilka m variabler som helst kan inte väljas som basvariabler; för att undersöka vilka som duger låter vi $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ vara en permutation av m tal tagna från $\{1, 2, \ldots, n\}$ och $\beta = (\beta_1, \beta_2, \ldots, \beta_{n-m})$ vara en godtycklig permutation av resterande $n-m$ tal. Vi skriver sedan om ekvationssystemet (13.6′) på formen

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = b - \sum_{j=1}^{n-m} x_{\beta_j} A_{*\beta_j}. \tag{13.6'''}$$

Om α är en basindexmängd, dvs. om kolonnerna $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ bildar en bas för $\mathbf{R}^m$, så har tydligen ekvation (13.6‴) en entydig lösning för varje tilldelning av värden på β-variablerna, och $(x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m})$ är helt enkelt koordinaterna i denna bas för vektorn $b - \sum_{j=1}^{n-m} x_{\beta_j} A_{*\beta_j}$. För motsvarande baslösning $\overline{x}$, som definieras av att $\overline{x}_{\beta_j} = 0$ för alla j, är tydligen $(\overline{x}_{\alpha_1}, \overline{x}_{\alpha_2}, \ldots, \overline{x}_{\alpha_m})$ lika med koordinaterna för vektorn b.

Antag omvänt att varje tilldelning av värden på β-variablerna bestämmer α-variablernas värden entydigt. Då har speciellt ekvationen

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = b \tag{13.7}$$

en entydig lösning, och detta medför i sin tur att ekvationen

$$\sum_{j=1}^{m} x_{\alpha_j} A_{*\alpha_j} = 0 \tag{13.8}$$

endast kan ha den triviala lösningen $x_{\alpha_j} = 0$ för alla j; i motsatt fall skulle vi nämligen få flera lösningar till (13.7) genom att till en given lösning addera en icke-trival lösning till (13.8). Kolonnerna $A_{*\alpha_1}, A_{*\alpha_2}, \ldots, A_{*\alpha_m}$ är med andra ord linjärt oberoende, och eftersom de är m till antalet, bildar de en bas för $\mathbf{R}^m$, så α är en basindexmängd.

Vi har med andra ord bevisat följande samband.

Sats 13.3.2. *Variablerna $x_{\alpha_1}, x_{\alpha_2}, \ldots, x_{\alpha_m}$ är basvariabler i systemet* (13.6) *om och endast om α är en basindexmängd till koefficientmatrisen A.*

Vi kan uttrycka den mot basindexmängden α svarande baslösningen på matrisform. Genom att skriva matrisekvationen (13.6″) på formen

$$A_{*\alpha} x_\alpha + A_{*\beta} x_\beta = b$$

samt multiplicera från vänster med matrisen $A_{*\alpha}^{-1}$ erhåller vi

$$\begin{aligned} &x_\alpha + A_{*\alpha}^{-1} A_{*\beta} x_\beta = A_{*\alpha}^{-1} b, \qquad \text{dvs.} \\ &x_\alpha = A_{*\alpha}^{-1} b - A_{*\alpha}^{-1} A_{*\beta} x_\beta, \end{aligned}$$

vilket framställer basvariablerna som linjärkombinationer av de fria variablerna och b:s koordinater. Baslösningen fås genom att sätta $x_\beta = 0$ och ges av att

$$\overline{x}_\alpha = A_{*\alpha}^{-1} b\,, \;\; \overline{x}_\beta = 0.$$

Vi sammanfattar detta resultat i följande sats.

Sats 13.3.3. *Den mot basindexmängden α svarande baslösningen $\overline{x}$ till ekvationssystemet $Ax = b$ är bestämd av att*

$$\overline{x}_\alpha = A_{*\alpha}^{-1} b \quad \textit{och} \quad \overline{x}_k = 0 \quad \textit{för } k \notin \alpha.$$

I en baslösning är $n - m$ av variablerna satta till noll och därför högst m variabler skilda från noll. Naturligtvis kan även någon eller några av basvariablerna råka vara noll. Vi gör därför följande definition.

Definition. En baslösning $\overline{x}$ kallas *icke-degenererad* om $\overline{x}_i \neq 0$ för m stycken index i och *degenererad* om $\overline{x}_i \neq 0$ för färre än m index i.

Två basindexmängder α och α' som är permutationer av varandra ger naturligtvis upphov till samma baslösning $\overline{x}$. Antalet olika baslösningar till ett system $Ax = b$ med m ekvationer och n obekanta är därför högst lika med antalet delmängder med m element som kan väljas från mängden $\{1, 2, \ldots, n\}$, dvs. högst lika med $\binom{n}{m}$. Antalet är mindre om det finns m kolonner i matrisen A som är linjärt beroende.

Exempel 13.3.3. Ekvationssystemet

$$\begin{cases} 3x_1 + x_2 + \ x_3 - 3x_4 = 3 \\ 3x_1 - x_2 + 2x_3 - 6x_4 = 3 \end{cases}$$

har − bortsett från permutationer − följande basindexmängder: $(1,2)$, $(1,3)$, $(1,4)$, $(2,3)$ och $(2,4)$, och motsvarande baslösningar är i tur och ordning $(1,0,0,0)$, $(1,0,0,0)$, $(1,0,0,0)$, $(0,1,2,0)$ resp. $(0,1,0,-\frac{2}{3})$. Baslösningen $(1,0,0,0)$ är degenererad, medan övriga två baslösningar är icke-degenererade. □

Skälet till att vi intresserar oss för basindexmängder och baslösningar är att optimala värden i LP-problem antas i extremalpunkter, och extremalpunkterna är baslösningar. Vi har nämligen följande karakterisering av extremalpunkter.

Sats 13.3.4. *Antag att A är en $m \times n$-matris av rang m, att $b \in \mathbf{R}^m$ och att $c \in \mathbf{R}^n$. Då gäller:*

(i) $\overline{x}$ är en extremalpunkt till polyedern $X = \{x \in \mathbf{R}^n \mid Ax = b,\, x \geq 0\}$ om och endast om $\overline{x}$ är en icke-negativ baslösning till systemet $Ax = b$, dvs. om och endast om det finns en basindexmängd α till matrisen A så att $\overline{x}_\alpha = A_{*\alpha}^{-1} b \geq 0$ och $\overline{x}_k = 0$ för $k \notin \alpha$.

(ii) $\overline{y}$ är en extremalpunkt till polyedern $Y = \{y \in \mathbf{R}^m \mid A^T y \leq c\}$ om och endast om $A^T \overline{y} \leq c$ och det finns en basindexmängd α till matrisen A sådan att $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$.

Bevis. (i) Enligt sats 5.1.1 är punkten $\overline{x}$ en extremalpunkt till polyedern X om och endast om $\overline{x} \geq 0$ och de m ekvationerna i systemet $Ax = b$ tillsammans med $n - m$ stycken av ekvationerna $x_1 = 0$, $x_2 = 0$, …, $x_n = 0$ har $\overline{x}$ som entydig lösning. Om vi låter $\alpha_1, \alpha_2, \ldots, \alpha_m$ vara index för de resterande m stycken ekvationerna $x_i = 0$, så är $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ en basindexmängd och $\overline{x}$ motsvarande baslösning.

(ii) På grund av samma sats är $\overline{y}$ en extremalpunkt till polyedern Y om och endast om $\overline{y} \in Y$ och $\overline{y}$ är den entydiga lösningen till något av de kvadratiska ekvationssystem som fås genom att välja m stycken av de n ekvationerna i systemet $A^T y = c$. Om index för de valda ekvationerna är

$\alpha_1, \alpha_2, \ldots, \alpha_m$, så har det kvadratiska systemet formen $(A_{*\alpha})^T y = c_\alpha$, och detta ekvationsystem har entydig lösning $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$ om och endast om matrisen $A_{*\alpha}$ är inverterbar, dvs. om och endast om α är en basindexmängd till A. □

EXEMPEL 13.3.4. Av sats 13.3.4 och resultatet i exempel 13.3.3 följer att polyedern X av lösningar till systemet

$$\begin{cases} 3x_1 + x_2 + \ x_3 - 3x_4 = 3 \\ 3x_1 - x_2 + 2x_3 - 6x_4 = 3, \ x \geq 0 \end{cases}$$

har två extremalpunkter, nämligen $(1, 0, 0, 0)$ och $(0, 1, 2, 0)$. Den "duala" polyedern Y av lösningar till systemet

$$\begin{cases} \ \ 3y_1 + 3y_2 \leq \ \ 2 \\ \ \ y_1 - \ \ y_2 \leq \ \ 1 \\ \ \ y_1 + 2y_2 \leq \ \ 1 \\ -3y_1 - 6y_2 \leq -1 \end{cases}$$

har tre extremalpunkter, nämligen $(\frac{5}{6}, -\frac{1}{6})$, $(\frac{1}{3}, \frac{1}{3})$ och $(\frac{7}{9}, -\frac{2}{9})$, som hör ihop med basindexmängderna $(1, 2)$, $(1, 3)$ och $(2, 4)$. (De mot basindexmängderna $(1, 4)$ och $(2, 3)$ svarande lösningarna $y = (1, -\frac{1}{3})$ och $y = (1, 0)$ ligger inte i polyedern Y.) □

Byte av basindexmängd

Vi skall nu diskutera hur man genererar en svit av baslösningar genom att successivt byta ut ett element i taget i basindexmängden.

Sats 13.3.5. *Antag att $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ är en basindexmängd till ekvationssystemet $Ax = b$ med motsvarande baslösning $\overline{x}$. Låt k vara ett kolonnindex som inte tillhör basindexmängden α, och definiera kolonnmatrisen v i $\mathbf{R}^n$ genom att sätta*

$$v_\alpha = A_{*\alpha}^{-1} A_{*k}, \quad v_k = -1 \quad \textit{och} \quad v_j = 0 \quad \textit{för } j \notin \alpha \cup \{k\}.$$

(i) Då är $Av = 0$, så det följer att $\overline{x} - tv$ är en lösning till systemet $Ax = b$ för alla $t \in \mathbf{R}$.

(ii) Antag att $1 \leq r \leq m$ och definiera en ny ordnad mängd α' genom att byta ut elementet på plats r i α mot k, dvs.

$$\alpha' = \alpha_{\hat{r}}[k] = (\alpha_1, \ldots, \alpha_{r-1}, k, \alpha_{r+1}, \ldots, \alpha_m).$$

Då är α' en basindexmängd om och endast om $v_{\alpha_r} \neq 0$. I så fall är vidare

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1}$$

och för den mot α' svarande baslösningen $\overline{x}'$ gäller att

$$\overline{x}' = \overline{x} - \tau v,$$

där

$$\tau = \overline{x}_{\alpha_r} / v_{\alpha_r}.$$

(iii) De båda baslösningarna $\overline{x}$ och $\overline{x}'$ är identiska om och endast om $\tau = 0$. Om baslösningen $\overline{x}$ är icke-degenererad, så är $\overline{x} \neq \overline{x}'$.

Vi kommer att kalla kolonnvektorn v för den till basindexmängden α och indexet k associerade *sökvektorn* eftersom vi erhåller den nya baslösningen $\overline{x}'$ genom att utgå från $\overline{x}$ och söka i den riktning som ges av (minus) v.

Bevis. (i) Av v:s definition följer omedelbart att

$$Av = \sum_{j\in\alpha} v_j A_{*j} + \sum_{j\notin\alpha} v_j A_{*j} = A_{*\alpha} v_\alpha - A_{*k} = A_{*k} - A_{*k} = 0.$$

(ii) Mängden α' är en basindexmängd om och endast om matrisen $A_{*\alpha'}$ är inverterbar. Nu är

$$\begin{aligned} A_{*\alpha}^{-1} A_{*\alpha'} &= A_{*\alpha}^{-1} \left[A_{*\alpha_1} \ \dots \ A_{*\alpha_{r-1}} \ A_{*k} \ A_{*\alpha_{r+1}} \ \dots \ A_{*\alpha_m} \right] \\ &= \left[A_{*\alpha}^{-1} A_{*\alpha_1} \ \dots \ A_{*\alpha}^{-1} A_{*\alpha_{r-1}} \ A_{*\alpha}^{-1} A_{*k} \ A_{*\alpha}^{-1} A_{*\alpha_{r+1}} \ \dots \ A_{*\alpha}^{-1} A_{*\alpha_m} \right] \\ &= \left[E_{*1} \ \dots \ E_{*r-1} \ v_\alpha \ E_{*r+1} \ \dots \ E_{*m} \right] = E_{\hat{r}}[v_\alpha], \end{aligned}$$

där förstås E betecknar enhetsmatrisen av ordning m. Följaktligen är

$$A_{*\alpha'} = A_{*\alpha} E_{\hat{r}}[v_\alpha].$$

Matrisen $A_{*\alpha'}$ är således inverterbar om och endast om matrisen $E_{\hat{r}}[v_\alpha]$ är inverterbar, vilket enligt lemma 13.3.1 gäller om och endast om $v_{\alpha_r} \neq 0$. För inversen $A_{*\alpha'}^{-1}$ gäller i förekommande att

$$A_{*\alpha'}^{-1} = \left(A_{*\alpha} E_{\hat{r}}[v_\alpha] \right)^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1}.$$

Sätt $x^\tau = \overline{x} - \tau v$; enligt (i) är x^τ en lösning till $Ax = b$, så för att visa att x^τ är lika med den mot basindexmängden α' svarande baslösningen $\overline{x}'$ räcker det att visa att $x_j^\tau = 0$ för alla $j \notin \alpha'$, dvs. för $j = \alpha_r$ och för $j \notin \alpha \cup \{k\}$.

För $j = \alpha_r$ är $x_j^\tau = \overline{x}_{\alpha_r} - \tau v_{\alpha_r} = 0$ på grund av definitionen av τ, och för $j \notin \alpha \cup \{k\}$ är såväl $\overline{x}_j$ som v_j lika med 0 varför $x_j^\tau = \overline{x}_j - \tau v_j = 0$.

(iii) Det följer att $\overline{x}' = \overline{x}$ om och endast om $\tau v = 0$, och eftersom $v_k = -1$, är detta ekvivalent med att $\tau = 0$. Om baslösningen $\overline{x}$ är icke-degenererad, så är speciellt $\overline{x}_{\alpha_r} \neq 0$, vilket implicerar att $\tau \neq 0$. □

Korollarium 13.3.6. *Behåll förutsättningarna i sats 13.3.5 och antag dessutom att $\overline{x} \geq 0$, att indexmängden*

$$I_+ = \{j \in \alpha \mid v_j > 0\}$$

inte är tom, samt att indexet r valts så att $\alpha_r \in I_+$ och

$$\tau = \overline{x}_{\alpha_r}/v_{\alpha_r} = \min\{\overline{x}_j/v_j \mid j \in I_+\}.$$

Då är $\overline{x}' \geq 0$.

Bevis. För $j \notin \alpha'$ är förstås $\overline{x}'_j = 0$ så det räcker att visa att $\overline{x}'_j \geq 0$ för alla $j \in \alpha \cup \{k\}$.

Vi börjar med att konstatera att $\tau = \overline{x}_{\alpha_r}/v_{\alpha_r} \geq 0$ eftersom $\overline{x} \geq 0$. För $j = k$ är därför

$$\overline{x}'_j = \overline{x}_k - \tau v_k = 0 + \tau \geq 0,$$

och för $j \in \alpha \setminus I_+$ är

$$\overline{x}'_j = \overline{x}_j - \tau v_j \geq \overline{x}_j \geq 0$$

eftersom $v_j \leq 0$. För $j \in I_+$ är slutligen $\overline{x}_j/v_j \geq \tau$, så det följer att

$$\overline{x}'_j = \overline{x}_j - \tau v_j \geq 0.$$

Därmed är beviset klart. □

13.4 Simplexalgoritmen

Den variant av simplexalgoritmen som vi skall beskriva förutsätter att LP-problemt har standardform. Vi utgår därför från problemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax = b,\ x \geq 0 \end{array}$$

där A är en $m \times n$-matris, $b \in \mathbf{R}^m$ och $c \in \mathbf{R}^n$.

Vi förutsätter vidare att

$$\operatorname{rang} A = m = \text{antalet rader i } A.$$

Detta är naturligtvis inte någon inskränkning, ty om $\operatorname{rang} A < m$ är antingen systemet $Ax = b$ olösbart, eller också är $(m - \operatorname{rang} A)$ stycken av de ingående ekvationerna konsekvenser av de övriga, och de kan därför strykas utan att lösningsmängden ändras.

Låt oss kalla en basindexmängd α till matrisen A och motsvarande baslösning $\overline{x}$ till systemet $Ax = b$ för en *tillåten basindexmängd* resp. *tillåten baslösning* om $\overline{x}$ är en tillåten punkt i vårt standardproblem, dvs. om $\overline{x} \geq 0$.

Simplexalgoritmen utgår ifrån att vi har en tillåten basindexmängd α att starta ifrån. I avsnitt 13.6 kommer vi att visa hur man hittar en sådan mängd genom att tillämpa simplexalgoritmen på ett s. k. artificiellt problem.

Beräkna först motsvarande tillåtna baslösning $\overline{x}$, dvs.

$$\overline{x}_\alpha = A_{*\alpha}^{-1} b \geq 0,$$

och därefter talet $\lambda \in \mathbf{R}$ och kolonnmatriserna $\overline{y} \in \mathbf{R}^m$ och $z \in \mathbf{R}^n$, som definieras av att

$$\begin{aligned} \lambda &= \langle c, \overline{x} \rangle = \langle c_\alpha, \overline{x}_\alpha \rangle \\ \overline{y} &= (A_{*\alpha}^{-1})^T c_\alpha \\ z &= c - A^T \overline{y}. \end{aligned}$$

Talet λ är med andra ord målfunktionens värde i punkten $\overline{x}$.

Observera att

$$z_\alpha = c_\alpha - (A^T \overline{y})_{\alpha *} = c_\alpha - (A_{*\alpha})^T \overline{y} = c_\alpha - c_\alpha = 0,$$

så för att beräkna vektorn z behöver vi bara beräkna koordinaterna

$$z_j = c_j - (A_{*j})^T \overline{y} = c_j - \langle A_{*j}, \overline{y} \rangle$$

för alla index $j \notin \alpha$. Talen z_j brukar kallas *reducerade kostnader*.

Lemma 13.4.1. *Talet λ och vektorerna $\overline{x}$, $\overline{y}$ och z har följande egenskaper:*

(i) $\langle z, \overline{x} \rangle = 0$*, dvs. vektorerna z och $\overline{x}$ är ortogonala.*

(ii) $Ax = 0 \Rightarrow \langle c, x \rangle = \langle z, x \rangle$.

(iii) $Ax = b \Rightarrow \langle c, x \rangle = \lambda + \langle z, x \rangle$.

(iv) Om v är den till basindexmängden α och indexet $k \notin \alpha$ hörande sökvektorn, så är $\langle c, \overline{x} - tv \rangle = \lambda + t z_k$.

Bevis. (i) Eftersom $z_j = 0$ för $j \in \alpha$ och $\overline{x}_j = 0$ för $j \notin \alpha$, är

$$\langle z, \overline{x} \rangle = \sum_{j \in \alpha} z_j \overline{x}_j + \sum_{j \notin \alpha} z_j \overline{x}_j = 0 + 0 = 0.$$

(ii) Definitionen av z ger omedelbart att

$$\langle z, x \rangle = \langle c, x \rangle - \langle A^T \overline{y}, x \rangle = \langle c, x \rangle - \langle \overline{y}, Ax \rangle = \langle c, x \rangle$$

för alla x som uppfyller $Ax = 0$.

(iii) Om $Ax = b$, så är

$$\begin{aligned} \langle c, x \rangle - \langle z, x \rangle &= \langle A^T \overline{y}, x \rangle = \langle \overline{y}, Ax \rangle = \langle (A_{*\alpha}^{-1})^T c_\alpha, b \rangle = \langle c_\alpha, A_{*\alpha}^{-1} b \rangle \\ &= \langle c_\alpha, \overline{x}_\alpha \rangle = \lambda. \end{aligned}$$

(iv) Eftersom $Av = 0$, följer det av (ii) att

$$\langle c, \overline{x} - tv \rangle = \langle c, \overline{x} \rangle - t \langle c, v \rangle = \lambda - t \langle z, v \rangle = \lambda + t z_k. \qquad \square$$

Följande sats innehåller alla väsentliga ingredienser i simplexalgoritmen.

Sats 13.4.2. *Låt α, $\overline{x}$, λ, $\overline{y}$ och z vara definierade som ovan.*

(i) ***(Optimalitet)*** *Om $z \geq 0$, så är $\overline{x}$ en optimal lösning till minimeringsproblemet*

$$\begin{aligned} &\min && \langle c, x\rangle \\ &\text{då} && Ax = b,\ x \geq 0 \end{aligned}$$

och $\overline{y}$ en optimal lösning till det duala maximeringsproblemet

$$\begin{aligned} &\max && \langle b, y\rangle \\ &\text{då} && A^T y \leq c \end{aligned}$$

med λ som det optimala värdet. Om dessutom $z_j > 0$ för alla $j \notin \alpha$, så har minimeringsproblemet unik optimal lösning.

(ii) *Antag att $z \ngeq 0$, och låt k vara ett index (som inte kan tillhöra α) sådant att $z_k < 0$. Låt vidare v vara den till α och k hörande sökvektorn, dvs.*

$$v_\alpha = A_{*\alpha}^{-1} A_{*k}, \quad v_k = -1, \quad v_j = 0 \quad \text{för } j \notin \alpha \cup \{k\},$$

samt sätt $x^t = \overline{x} - tv$. Beroende på om $v \leq 0$ eller $v \not\leq 0$ gäller följande:

(ii a) ***(Obegränsad målfunktion)*** *Om $v \leq 0$, så är punkterna x^t tillåtna i minimeringsproblemet för alla $t \geq 0$ och $\langle c, x^t\rangle \to -\infty$ då $t \to \infty$. Målfunktionen i minimeringsproblemet är därför nedåt obegränsad, och det duala maximeringsproblemet saknar tillåtna punkter.*

(ii b) ***(Iterationssteget)*** *Definiera om $v \not\leq 0$ indexmängden α' och talet τ som i sats 13.3.5 (ii) med index r valt som i korollarium 13.3.6. Då är α' en tillåten basindexmängd med $\overline{x}' = \overline{x} - \tau v$ som motsvarande tillåtna baslösning, och*

$$\langle c, \overline{x}'\rangle = \langle c, \overline{x}\rangle + \tau z_k \leq \langle c, \overline{x}\rangle.$$

Om $\tau > 0$ är alltså $\langle c, \overline{x}'\rangle < \langle c, \overline{x}\rangle$.

Bevis. (i) Antag att $z \geq 0$ och att x är en godtycklig tillåten punkt i minimeringsproblemet. Då är $\langle z, x\rangle \geq 0$ (eftersom $x \geq 0$), så det följer av (iii) i lemma 13.4.1 att $\langle c, x\rangle \geq \lambda = \langle c, \overline{x}\rangle$. Punkten $\overline{x}$ är således en minimipunkt och minimivärdet är λ.

Av $z \geq 0$ följer också att $A^T\overline{y} = c - z \leq c$, dvs. $\overline{y}$ är en tillåten punkt i det duala maximeringsproblemet, och

$$\langle b, \overline{y}\rangle = \langle \overline{y}, b\rangle = \langle (A_{*\alpha}^{-1})^T c_\alpha, b\rangle = \langle c_\alpha, A_{*\alpha}^{-1} b\rangle = \langle c_\alpha, \overline{x}_\alpha\rangle = \langle c, \overline{x}\rangle,$$

så det följer av optimalitetskriteriet (korollarium 13.1.2) att $\overline{y}$ är en optimal lösning till det duala problemet.

Antag nu att $z_j > 0$ för alla $j \notin \alpha$. Om x är en tillåten punkt $\neq \overline{x}$, så är $x_{j_0} > 0$ för något index $j_0 \notin \alpha$, och det följer att $\langle z, x\rangle = \sum_{j\notin\alpha} z_j x_j \geq z_{j_0} x_{j_0} > 0$. Lemma 13.4.1 (iii) medför därför att $\langle c, x\rangle > \lambda = \langle c, \overline{x}\rangle$. Detta visar att minimipunkten är unik.

(ii a) Enligt sats 13.3.5 är x^t en lösning till ekvationen $Ax = b$ för alla reella tal t. Om $v \leq 0$, så är vidare $x^t = \overline{x} - tv \geq \overline{x} \geq 0$ för $t \geq 0$, dvs. punkterna x^t är tillåtna i minimeringsproblemet, och lemma 13.4.1 (iv) medför att

$$\lim_{t\to\infty} \langle c, x^t\rangle = \lambda + \lim_{t\to\infty} z_k t = -\infty.$$

Målfunktionen i minimeringsproblemet är således nedåt obegränsad.

Antag att det duala maximeringsproblemet har en tillåten punkt y. På grund av den svaga dualitetssatsen är då speciellt $\langle b, y\rangle \leq \langle c, x^t\rangle$ för alla $t \geq 0$, vilket strider mot att högerledet går mot $-\infty$ då $t \to \infty$. Av denna motsägelse följer att det inte finns några tillåtna punkter i det duala problemet.

(ii b) Enligt korollarium 13.3.6 är α' en tillåten baslösning med x^τ som motsvarande tillåtna baslösning, och olikheten $\langle c, \overline{x}'\rangle \leq \langle c, \overline{x}\rangle$ följer nu direkt av lemma 13.4.1 (iv) eftersom $\tau \geq 0$. □

Sats 13.4.2 ger oss nu följande algoritm för att lösa standardproblemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax = b,\ x \geq 0. \end{array}$$

Simplexalgoritmen

Givet en tillåten basindexmängd α.

1. Beräkna matrisen $A_{*\alpha}^{-1}$, motsvarande tillåtna baslösning $\overline{x}$, dvs. $\overline{x}_\alpha = A_{*\alpha}^{-1} b$ och $\overline{x}_j = 0$ för $j \notin \alpha$, och talet $\lambda = \langle c_\alpha, \overline{x}_\alpha\rangle$.

Upprepa steg 2–8 till dess att stopp inträffar.

2. Beräkna vektorn $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$ och talen $z_j = c_j - \langle A_{*j}, \overline{y}\rangle$ för $j \notin \alpha$.
3. *Stoppkriterium:* Stoppa om $z_j \geq 0$ för alla $j \notin \alpha$. Optimal lösning: $\overline{x}$. Optimalt värde: λ. Optimal dual lösning: $\overline{y}$.
4. Välj annars ett index k med $z_k < 0$, beräkna motsvarande sökvektor v, dvs. $v_\alpha = A_{*\alpha}^{-1} A_{*k}$, $v_k = -1$ och $v_j = 0$ för $j \notin \alpha \cup \{k\}$, och sätt $I_+ = \{j \in \alpha \mid v_j > 0\}$.
5. *Stoppkriterium:* Stoppa om $I_+ = \emptyset$. Optimalt värde: $-\infty$.
6. Sätt annars $\tau = \min\{\overline{x}_j / v_j \mid j \in I_+\}$ och bestäm ett index r så att $\alpha_r \in I_+$ och $\overline{x}_{\alpha_r} / v_{\alpha_r} = \tau$.

7. Sätt $\alpha' = \alpha_{\hat{r}}[k]$ och beräkna inversen $A^{-1}_{*\alpha'} = E_{\hat{r}}[v_\alpha]^{-1} A^{-1}_{*\alpha}$.
8. *Uppdatera:* $\alpha := \alpha'$, $A^{-1}_{*\alpha} := A^{-1}_{*\alpha'}$, $\overline{x} := \overline{x} - \tau v$ samt $\lambda := \lambda + \tau z_k$.

Innan man kan kalla proceduren ovan för en algoritm i betydelsen mekanisk beräkning som en maskin kan utföra, behöver man förstås precisera hur man i stegen 4 och 6 skall välja k i det fall då $z_j < 0$ för flera index j, och r om $\overline{x}_j/v_j = \tau$ för fler än ett index $j \in I_+$.

En enkel regel, som fungerar bra i de flesta sammanhangen, är att som index k välja det index j som minimerar z_j (och om det finns flera sådana index välja det minsta av dessa), och att som index r välja det minsta av alla index i för vilka $\overline{x}_{\alpha_i}/v_{\alpha_i} = \tau$. Vi skall återkomma till valet av k och r senare; för den närmaste diskussionen av algoritmen spelar det nämligen ingen roll hur man gör valet.

Vi behöver vidare en metod för att hitta en första tillåten basindexmängd att starta simplexalgoritmen ifrån. Detta problem skall vi återkomma till och lösa i avsnitt 13.6.

Antag nu att vi tillämpar simplexalgoritmen på ett LP-problem på standardform och startar från en tillåten basindexmängd. Att algoritmen levererar en optimal lösning om den stoppar under steg 3 och att målfunktionen är nedåt obegränsad om den stoppar under steg 5 följer förstås av sats 13.4.2.

Låt oss därför undersöka vad som händer om algoritmen inte stoppar. I så fall genereras en tillåten basindexmängd varje gång algoritmen kommer till steg 7, och vi får därför en oändlig följd $\alpha^1, \alpha^2, \alpha^3, \ldots$ av tillåtna basindexmängder med tillhörande tillåtna baslösningar $\overline{x}^1, \overline{x}^2, \overline{x}^3, \ldots$. Eftersom antalet olika basindexmängder är ändligt, kan inte alla basindexmängderna i den genererade följden vara olika, utan någon basindexmängd α^p måste upprepas efter ytterligare ett antal, q säg, iterationer. Detta betyder att $\alpha^p = \alpha^{p+q}$ och $\overline{x}^p = \overline{x}^{p+q}$ och medför i sin tur att följden $\alpha^p, \alpha^{p+1}, \ldots, \alpha^{p+q-1}$ upprepas periodiskt i all oändlighet. Man uttrycker detta genom att säga att algoritmen *cyklar*. På grund av (ii b) i sats 13.4.2 är

$$\langle c, \overline{x}^p \rangle \geq \langle c, \overline{x}^{p+1} \rangle \geq \cdots \geq \langle c, \overline{x}^{p+q} \rangle = \langle c, \overline{x}^p \rangle,$$

och detta medför förstås att

$$\langle c, \overline{x}^p \rangle = \langle c, \overline{x}^{p+1} \rangle = \cdots = \langle c, \overline{x}^{p+q-1} \rangle.$$

Under alla iterationerna i cykeln är således talet τ lika med 0, så baslösningarna $\overline{x}^p, \overline{x}^{p+1}, \ldots, \overline{x}^{p+q-1}$ är identiska och degenererade. Om simplexalgoritmen inte stoppar utan cyklar, så beror det alltså på att den ”fastnar” i en degenerad baslösning.

Av ovanstående resonemang följer omedelbart:

Sats 13.4.3. *Simplexalgoritmen, startad från en godtycklig tillåten basindexmängd, stoppar om samtliga tillåtna baslösningar i LP-problemet är icke-degenererade.*

Cykling kan inträffa, och vi skall ge ett exempel på detta i nästa avsnitt. Teoretiskt är detta naturligtvis besvärande, men eftersom det är mycket ovanligt att praktiska problem cyklar, saknar cykling praktisk betydelse. De små avrundningsfel som introduceras under den numeriska behandlingen av ett LP-problem har också en hälsosam effekt eftersom de kan förvandla degenererade baslösningar till icke-degenererade och därigenom har en tendens att förhindra cykling.

Det finns också en enkel regel för val av index k och r, *Blands regel*, som omöjliggör cykling och som vi skall beskriva i nästa avsnitt.

Exempel

EXEMPEL 13.4.1. Vi illustrerar nu simplexalgoritmen genom att lösa minimeringsproblemet

$$\begin{array}{ll} \min & x_1 - x_2 + x_3 \\ \text{då} & \begin{cases} -2x_1 + x_2 + \ x_3 \le 3 \\ -x_1 + x_2 - 2x_3 \le 3 \\ 2x_1 - x_2 + 2x_3 \le 1, \ x \ge 0. \end{cases} \end{array}$$

Vi börjar med att skriva problemet på standardform genom att introducera tre slackvariabler:

$$\begin{array}{ll} \min & x_1 - x_2 + x_3 \\ \text{då} & \begin{cases} -2x_1 + x_2 + \ x_3 + x_4 = 3 \\ -x_1 + x_2 - 2x_3 + x_5 = 3 \\ 2x_1 - x_2 + 2x_3 + x_6 = 1, \ x \ge 0. \end{cases} \end{array}$$

På matrisform blir detta förstås

$$\begin{array}{ll} \min & c^T x \\ \text{då} & Ax = b, \, x \ge 0 \end{array}$$

med

$$A = \begin{bmatrix} -2 & 1 & 1 & 1 & 0 & 0 \\ -1 & 1 & -2 & 0 & 1 & 0 \\ 2 & -1 & 2 & 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix} \quad \text{och}$$

$$c^T = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Vi konstaterar att vi kan starta simplexalgoritmen med

$$\alpha = (4,5,6), \quad A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \overline{x}_\alpha = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix},$$

$$\lambda = \langle c_\alpha, \overline{x}_\alpha \rangle = c_\alpha^T \overline{x}_\alpha = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix} = 0.$$

1:a iterationen:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$z_{1,2,3} = c_{1,2,3} - (A_{*1,2,3})^T \overline{y} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} - \begin{bmatrix} -2 & -1 & 2 \\ 1 & 1 & -1 \\ 1 & -2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

Eftersom $z_2 = -1 < 0$, är

$$k = 2$$

$$v_\alpha = A_{*\alpha}^{-1} A_{*k} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad v_2 = -1$$

$$I_+ = \{j \in \alpha \mid v_j > 0\} = \{4,5\}$$

$$\tau = \min\{\overline{x}_j / v_j \mid j \in I_+\} = \min\{\overline{x}_4/v_4, \overline{x}_5/v_5\} = \min\{3/1, 3/1\} = 3$$

för $\alpha_1 = 4$, dvs.

$$r = 1$$

$$\alpha' = \alpha_{\hat{r}}[k] = (4,5,6)_{\hat{1}}[2] = (2,5,6)$$

$$E_{\hat{r}}[v_\alpha]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} = \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} - 3 \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}$$

$$\lambda' = \lambda + \tau z_k = 0 + 3\,(-1) = -3.$$

Uppdatering: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ och $\lambda := \lambda'$.

2:a iterationen:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

$$z_{1,3,4} = c_{1,3,4} - (A_{*1,3,4})^T \overline{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} -2 & -1 & 2 \\ 1 & -2 & 2 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}.$$

Eftersom $z_1 = -1 < 0$, är

$$k = 1$$

$$v_\alpha = A_{*\alpha}^{-1} A_{*k} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \quad v_1 = -1$$

$$I_+ = \{j \in \alpha \mid v_j > 0\} = \{5\}$$

$$\tau = \overline{x}_5 / v_5 = 0/1 = 0 \quad \text{för } \alpha_2 = 5, \text{ dvs.}$$

$$r = 2$$

$$\alpha' = \alpha_{\hat{r}}[k] = (2,5,6)_{\hat{2}}[1] = (2,1,6)$$

$$E_{\hat{r}}[v_\alpha]^{-1} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} - 0 \begin{bmatrix} -2 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}$$

$$\lambda' = \lambda + \tau z_k = -3 + 0\,(-1) = -3.$$

Uppdatering: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ och $\lambda := \lambda'$.

3:e iterationen:

$$\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} -1 & -1 & 1 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$$

$$z_{3,4,5} = c_{3,4,5} - (A_{*3,4,5})^T \overline{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & -2 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Eftersom $z_3 = -1 < 0$, är

$$
\begin{aligned}
k &= 3 \\
v_\alpha = A_{*\alpha}^{-1} A_{*k} &= \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} -5 \\ -3 \\ 3 \end{bmatrix}, \quad v_3 = -1 \\
I_+ &= \{j \in \alpha \mid v_j > 0\} = \{6\} \\
\tau &= \overline{x}_6 / v_6 = 4/3 \quad \text{för } \alpha_3 = 6, \text{ dvs.} \\
r &= 3 \\
\alpha' &= \alpha_{\hat{r}}[k] = (2,1,6)_{\hat{3}}[3] = (2,1,3)
\end{aligned}
$$

$$
E_{\hat{r}}[v_\alpha]^{-1} = \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & -3 \\ 0 & 0 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & \frac{5}{3} \\ 0 & 1 & 1 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}
$$

$$
A_{*\alpha'}^{-1} = E_{\hat{r}}[v_\alpha]^{-1} A_{*\alpha}^{-1} = \begin{bmatrix} 1 & 0 & \frac{5}{3} \\ 0 & 1 & 1 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -1 & 2 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & 2 & \frac{5}{3} \\ 0 & 1 & 1 \\ \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix}
$$

$$
\overline{x}'_{\alpha'} = \overline{x}_{\alpha'} - \tau v_{\alpha'} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} - \frac{4}{3} \begin{bmatrix} -5 \\ -3 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{29}{3} \\ 4 \\ \frac{4}{3} \end{bmatrix}
$$

$$
\lambda' = \lambda + \tau z_k = -3 + \frac{4}{3}(-1) = -\frac{13}{3}.
$$

Uppdatering: $\alpha := \alpha'$, $A_{*\alpha}^{-1} := A_{*\alpha'}^{-1}$, $\overline{x}_\alpha := \overline{x}'_{\alpha'}$ och $\lambda := \lambda'$.

4:e iterationen:

$$
\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha = \begin{bmatrix} \frac{2}{3} & 0 & \frac{1}{3} \\ 2 & 1 & 0 \\ \frac{5}{3} & 1 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ -1 \\ -\frac{1}{3} \end{bmatrix}
$$

$$
z_{4,5,6} = c_{4,5,6} - (A_{*4,5,6})^T \overline{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{3} \\ -1 \\ -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \\ \frac{1}{3} \end{bmatrix}.
$$

Eftersom $z_{4,5,6} > 0$, är lösningen optimal. Minimum $= -\frac{13}{3}$ antas i punkten $\overline{x} = (4, \frac{29}{3}, \frac{4}{3}, 0, 0, 0)$. Det ursprungliga minimeringsproblemet har förstås samma minimivärde och antar sitt minimum för $(x_1, x_2, x_3) = (4, \frac{29}{3}, \frac{4}{3})$. □

Den variant av simplexalgoritmen som vi presenterat är utmärkt för datorberäkningar, men den är svåröverskådlig om man önskar utföra beräkningar för hand. Då är det bättre att använda den tabellform som vi utnyttjade i

avsnitt 13.2, även om detta medför att man utför onödiga beräkningar. Till LP-problemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax = b,\ x \geq 0 \end{array}$$

associerar vi följande simplextabell:

$$\begin{array}{c|c|c} A & b & E \\ \hline c^T & 0 & 0^T \end{array} \tag{13.9}$$

Kolonnen längst till höger i tabellen har vi tagit med enbart för att förklara hur tabellräkningen fungerar; den kommer att utelämnas när vi i fortsättningen räknar konkreta exempel.

Låt α vara en tillåten basindexmängd med motsvarande baslösning $\overline{x}$. Den övre delen $[\,A \mid b \mid E\,]$ av tabellen kan vi förstås uppfatta som en matris, och vi multiplicerar nu denna matris från vänster med $A_{*\alpha}^{-1}$. Detta resulterar i tabellen

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & A_{*\alpha}^{-1}b & A_{*\alpha}^{-1} \\ \hline c^T & 0 & 0^T \end{array}$$

Subtrahera nu den övre delen av tabellen multiplicerad från vänster med c_α^T från den nedre radmatrisen i tabellen. Detta ger oss tabellen

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & A_{*\alpha}^{-1}b & A_{*\alpha}^{-1} \\ \hline c^T - c_\alpha^T A_{*\alpha}^{-1}A & -c_\alpha^T A_{*\alpha}^{-1}b & -c_\alpha^T A_{*\alpha}^{-1} \end{array}$$

Vi använder nu de beteckningar som vi införde i definitionen av simplexalgoritmen, dvs. $A_{*\alpha}^{-1}b = \overline{x}_\alpha$, $c_\alpha^T A_{*\alpha}^{-1} = ((A_{*\alpha}^{-1})^T c_\alpha)^T = \overline{y}^T$, $c^T - c_\alpha^T A_{*\alpha}^{-1}A = c^T - \overline{y}^T A = z^T$ och $c_\alpha^T A_{*\alpha}^{-1}b = c_\alpha^T \overline{x}_\alpha = \langle c_\alpha, \overline{x}_\alpha\rangle = \lambda$. Tabellen ovan övergår då i tabellen

$$\begin{array}{c|c|c} A_{*\alpha}^{-1}A & \overline{x}_\alpha & A_{*\alpha}^{-1} \\ \hline z^T & -\lambda & -\overline{y}^T \end{array} \tag{13.10}$$

Observera att enhetsmatrisens kolonner ingår som kolonner i matrisen $A_{*\alpha}^{-1}A$; enhetsmatriskolonnen E_{*j} är kolonn nummer α_j i matrisen $A_{*\alpha}^{-1}A$. Vidare är $z_{\alpha_j} = 0$.

Vid praktisk räkning använder man sig förstås av Gausselimination för att komma från tabell (13.9) till tabell (13.10).

Om $z^T \geq 0$, vilket vi kan avgöra med hjälp av den nedre raden i (13.10), är baslösningen $\overline{x}$ optimal, och i tabellen kan vi också avläsa den optimala lösningen $\overline{y}$ till det duala maximeringsproblemet. (I många fall finns enhetsmatrisens kolonner redan med som kolonner i matrisen A, och då behöver man naturligtvis inte tillfoga enhetsmatrisen i högerledet av tabell (13.9) för att kunna avläsa lösningen till det duala problemet.)

Om $z^T \not\geq 0$, väljer man ett kolonnindex k med $z_k < 0$, och betraktar motsvarande kolonn $a = A_{*\alpha}^{-1} A_{*k}$ $(= v_\alpha)$ i den övre delen av tabellen.

I fallet $a \leq 0$ är minimeringsproblemet nedåt obegränsat. I motsatt fall väljer man ett index $i = r$ som minimerar kvoterna $\overline{x}_{\alpha_i}/a_i$ $(= \overline{x}_{\alpha_i}/v_{\alpha_i})$ för alla positiva a_i. Detta innebär att r är radindex för den rad i tabellen som har minst kvot $\overline{x}_{\alpha_i}/a_i$ bland alla rader med positivt a_i. Slutligen transformerar man simplextabellen genom att pivotera kring elementet på plats (r, k).

Exempel 13.4.2. Vi löser exempel 13.4.1 på nytt − denna gång genom att utföra samtliga beräkningar i tabellform. Starttabellen har formen

$$\begin{array}{rrrrrr|r} -2 & \underline{1} & 1 & 1 & 0 & 0 & 3 \\ -1 & 1 & -2 & 0 & 1 & 0 & 3 \\ 2 & -1 & 2 & 0 & 0 & 1 & 1 \\ \hline 1 & -1 & 1 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

och i det här fallet är det förstås onödigt att upprepa enhetsmatrisens kolonner i högerledet för att också kunna lösa det duala problemet.

Basindexmängden $\alpha = (4, 5, 6)$ är tillåten, och eftersom $A_{*\alpha} = E$ och $c_\alpha^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$, kan vi direkt läsa av $z^T = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 \end{bmatrix}$ och $-\lambda = 0$ i den nedre delen av tabellen. Optimalitetsvillkoret är inte uppfyllt eftersom $z_2 = -1 < 0$. Vi går därför vidare genom att sätta $k = 2$. I det här fallet är kvoterna mellan elementen i högerledet och de positiva elementen i andra kolonnen desamma och lika med 3/1 i första och andra raden. Vi kan därför välja $r = 1$ eller $r = 2$ och bestämmer oss för det mindre av de båda talen, dvs. vi sätter $r = 1$. Tabellen transformeras sedan genom pivotering kring elementet på plats $(1, 2)$. Genom att sedan fortsätta i samma stil får vi följande sekvens av tabeller:

$$\begin{array}{cccccc|c} -2 & 1 & 1 & 1 & 0 & 0 & 3 \\ \underline{1} & 0 & -3 & -1 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 1 & 4 \\ \hline -1 & \underline{0} & 2 & 1 & \underline{0} & \underline{0} & 3 \end{array}$$

$\alpha = (2, 5, 6), \quad k = 1, \quad r = 2$

$$\begin{array}{cccccc|c} 0 & 1 & -5 & -1 & 2 & 0 & 3 \\ 1 & 0 & -3 & -1 & 1 & 0 & 0 \\ 0 & 0 & \underline{3} & 1 & 0 & 1 & 4 \\ \hline \underline{0} & \underline{0} & -1 & 0 & 1 & \underline{0} & 3 \end{array}$$

$\alpha = (2, 1, 6), \quad k = 3, \quad r = 3$

$$\begin{array}{cccccc|c} 0 & 1 & 0 & \frac{2}{3} & 2 & \frac{5}{3} & \frac{29}{3} \\ 1 & 0 & 0 & 0 & 1 & 1 & 4 \\ 0 & 0 & 1 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{4}{3} \\ \hline \underline{0} & \underline{0} & \underline{0} & \frac{1}{3} & 1 & \frac{1}{3} & \frac{13}{3} \end{array}$$

$\alpha = (2, 1, 3)$

Nu är optimalitetsvillkoret uppfyllt. Minimum $-\frac{13}{3}$ antas i baslösningen $\overline{x} = (4, \frac{29}{3}, \frac{4}{3}, 0, 0, 0)$. Det duala problemets optimala lösning är $(-\frac{1}{3}, -1, -\frac{1}{3})$. □

När vi i fortsättningen använder simplexalgoritmen kommer vi att använda tabellvarianten för att redovisa våra beräkningar eftersom det är den mest överskådligliga metoden.

Optimalitetsvillkoret i steg 2 av simplexalgoritmen är ett tillräckligt villkor för optimalitet, men villkoret är inte nödvändigt. En degenererad baslösning kan vara optimal utan att optimalitetsvillkoret är uppfyllt. Här följer ett trivialt exempel på detta.

Exempel 13.4.3. Problemet

$$\begin{array}{ll} \min & -x_2 \\ \text{då} & x_1 + x_2 = 0, \ \ x \geq 0 \end{array}$$

har bara en tillåten punkt, $x = (0, 0)$, som därför är optimal. Det finns två tillåtna basindexmängder, $\alpha = (1)$ och $\alpha' = (2)$, som båda hör till den degenerade baslösningen $(0, 0)$.

I basindexmängden α är $\overline{y} = \langle 0, 1\rangle = 0$ och $z_2 = -1 - \langle 0, 1\rangle = -1 < 0$, varför optimalitetsvillkoret inte är uppfyllt. I den andra basindexmängden α'

är däremot $\overline{y} = -\langle 1, 1\rangle = -1$ och $z_2 = 0 - \langle -1, 1\rangle = 1 > 0$, och optimalitetsvillkoret är nu uppfyllt.

Motsvarande simplextabeller är

$$\begin{array}{cc|c} 1 & 1 & 0 \\ \hline \underline{0} & -1 & 0 \end{array} \qquad \text{resp.} \qquad \begin{array}{cc|c} 1 & 1 & 0 \\ \hline 1 & \underline{0} & 0 \end{array}$$

$\alpha = (1)$ $\qquad\qquad$ $\alpha = (2)$ □

Vi studerar nu ett enkelt exempel där den optimala lösningen inte är unik.

EXEMPEL 13.4.4. Till problemet

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{då} & \begin{cases} x_1 + \ x_2 - x_3 \ \ \ \ \ \ = 1 \\ \qquad 2x_2 - x_3 + x_4 = 1, \ x \geq 0 \end{cases} \end{array}$$

hör följande simplextabeller:

$$\begin{array}{cccc|c} 1 & 1 & -1 & 0 & 1 \\ 0 & 2 & -1 & 1 & 1 \\ \hline 1 & 1 & 0 & 0 & 0 \end{array}$$

$\alpha = (1, 4)$

$$\begin{array}{cccc|c} 1 & 1 & -1 & 0 & 1 \\ 0 & 2 & -1 & 1 & 1 \\ \hline \underline{0} & 0 & 1 & \underline{0} & -1 \end{array}$$

$\alpha = (1, 4)$

Optimalitetsvillkoret är uppfyllt; $\overline{x} = (1, 0, 0, 1)$ är en optimal lösning, och det optimala värdet är 1. I målfunktionsraden är emellertid koefficienten på plats 2, dvs. z_2, lika med 0. Vi kan därför utföra ytterligare en iteration av simplexalgoritmen genom att välja den andra kolonnen som pivotkolonn och den andra raden som pivotrad, dvs. $k = 2$ och $r = 2$. Detta ger oss följande nya tabell:

$$\begin{array}{cccc|c} 1 & 0 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \hline \underline{0} & \underline{0} & 1 & 0 & -1 \end{array}$$

$\alpha = (1, 2)$

Optimalitetsvillkoret är uppfyllt, nu med $\hat{x} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ som optimal lösning. Eftersom mängden av optimala lösningar är konvex, är också varje punkt på sträckan mellan $\hat{x}$ och $\overline{x}$ optimal. □

13.5 Blands anticyklingsregel

Vi börjar med ett exempel av Kuhn som visar att cykling kan förekomma i degenererade LP-problem om man inte väljer kolonnindex k och radindex r på ett speciellt sätt.

EXEMPEL 13.5.1. Betrakta problemet

$$\begin{array}{ll} \min & -2x_1 - 3x_2 + x_3 + 12x_4 \\ \text{då} & \begin{cases} -2x_1 - 9x_2 + x_3 + 9x_4 + x_5 = 0 \\ \frac{1}{3}x_1 + x_2 - \frac{1}{3}x_3 - 2x_4 + x_6 = 0 \\ 2x_1 + 3x_2 - x_3 - 12x_4 + x_7 = 2, \ x \geq 0. \end{cases} \end{array}$$

Vi använder simplexalgoritmen med regeln att varje gång i steg 4 välja det kolonnindex k som minimerar z_k och i steg 6 välja det radindex r som är minst av alla tillåtna. Vår första tabell blir

$$\begin{array}{rrrrrrr|r} -2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\ 2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\ \hline -2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

med $\alpha = (5, 6, 7)$ som tillåten basindexmängd. Enligt vår regel för val av k skall vi välja $k = 2$. För radindex r finns det här bara ett alternativ, nämligen $r = 2$. Tabellen pivoteras nu kring elementet på plats $(2, 2)$, vilket resulterar i följande tabell

$$\begin{array}{rrrrrrr|r} \underline{1} & 0 & -2 & -9 & 1 & 9 & 0 & 0 \\ \frac{1}{3} & 1 & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -6 & 0 & -3 & 1 & 2 \\ \hline -1 & \underline{0} & 0 & 6 & \underline{0} & 3 & \underline{0} & 0 \end{array}$$

med $\alpha = (5, 2, 7)$. Nu är $k = 1$, men det finns två radindex i som ger samma minsta värde för kvoterna $\overline{x}_{\alpha_i}/v_{\alpha_i}$, nämligen 1 och 2. Enligt vår regel skall vi välja $r = 1$. Pivotering kring elementet på plats $(1, 1)$ ger nästa tabell

$$\begin{array}{rrrrrrr|r} 1 & 0 & -2 & -9 & 1 & 9 & 0 & 0 \\ 0 & 1 & \frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 0 \\ 0 & 0 & 2 & 3 & -1 & -12 & 1 & 2 \\ \hline \underline{0} & \underline{0} & -2 & -3 & 1 & 12 & \underline{0} & 0 \end{array}$$

med $\alpha = (1, 2, 7)$, $k = 4$, $r = 2$.

Algoritmen fortsätter nu med följande tabeller:

$$\begin{array}{rrrrrrr|r}
1 & 9 & \underline{1} & 0 & -2 & -9 & 0 & 0 \\
0 & 1 & \frac{1}{3} & 1 & -\frac{1}{3} & -2 & 0 & 0 \\
0 & -3 & 1 & 0 & 0 & -6 & 1 & 2 \\
\hline
\underline{0} & 3 & -1 & \underline{0} & 0 & 6 & \underline{0} & 0
\end{array}$$

$$\alpha = (1, 4, 7), \;\; k = 3, \;\; r = 1$$

$$\begin{array}{rrrrrrr|r}
1 & 9 & 1 & 0 & -2 & -9 & 0 & 0 \\
-\frac{1}{3} & -2 & 0 & 1 & \frac{1}{3} & \underline{1} & 0 & 0 \\
-1 & -12 & 0 & 0 & 2 & 3 & 1 & 2 \\
\hline
\underline{1} & 12 & \underline{0} & \underline{0} & -2 & -3 & \underline{0} & 0
\end{array}$$

$$\alpha = (3, 4, 7), \;\; k = 6, \;\; r = 2$$

$$\begin{array}{rrrrrrr|r}
-2 & -9 & 1 & 9 & \underline{1} & 0 & 0 & 0 \\
-\frac{1}{3} & -2 & 0 & 1 & \frac{1}{3} & 1 & 0 & 0 \\
0 & -6 & 0 & -3 & 1 & 0 & 1 & 2 \\
\hline
0 & 6 & \underline{0} & 3 & -1 & \underline{0} & \underline{0} & 0
\end{array}$$

$$\alpha = (3, 6, 7), \;\; k = 5, \;\; r = 1$$

$$\begin{array}{rrrrrrr|r}
-2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\
\frac{1}{3} & \underline{1} & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\
2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\
\hline
-2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0
\end{array}$$

$$\alpha = (5, 6, 7)$$

Efter sex iterationer har vi således kommit tillbaka till starttabellen. Simplexalgoritmen cyklar. □

Vi introducerar nu en regel för val av index k och index r som kommer att omöjliggöra cykling.

Blands regel: *Välj index k under steg 4 av simplexalgoritmen så att*

$$k = \min\{j \mid z_j < 0\}$$

och index r under steg 6 så att

$$\alpha_r = \min\{j \in I_+ \mid \overline{x}_j / v_j = \tau\}.$$

EXEMPEL 13.5.2. Betrakta återigen minimeringsproblemet i exempel 13.5.1 och använd nu simplexalgoritmen med Blands regel. Detta ger oss följande sekvens av tabeller:

$$\begin{array}{rrrrrrr|r}
-2 & -9 & 1 & 9 & 1 & 0 & 0 & 0 \\
\underline{\frac{1}{3}} & 1 & -\frac{1}{3} & -2 & 0 & 1 & 0 & 0 \\
2 & 3 & -1 & -12 & 0 & 0 & 1 & 2 \\
\hline
-2 & -3 & 1 & 12 & \underline{0} & \underline{0} & \underline{0} & 0
\end{array}$$

$$\alpha = (5, 6, 7), \quad k = 1, \quad r = 2$$

$$\begin{array}{rrrrrrr|r}
0 & -3 & -1 & -3 & 1 & 6 & 0 & 0 \\
1 & 3 & -1 & -6 & 0 & 3 & 0 & 0 \\
0 & -3 & \underline{1} & 0 & 0 & -6 & 1 & 2 \\
\hline
\underline{0} & 3 & -1 & 0 & \underline{0} & 6 & \underline{0} & 0
\end{array}$$

$$\alpha = (5, 1, 7), \quad k = 3, \quad r = 3$$

$$\begin{array}{rrrrrrr|r}
0 & -6 & 0 & -3 & 1 & 0 & 0 & 2 \\
1 & 0 & 0 & -6 & 0 & -3 & 1 & 2 \\
0 & -3 & 1 & 0 & 0 & -6 & 1 & 2 \\
\hline
\underline{0} & 0 & \underline{0} & 0 & \underline{0} & 12 & 1 & 2
\end{array}$$

$$\alpha = (5, 1, 3)$$

Optimalitetskriteriet är nu uppfyllt, och minimivärdet -2 antas i punkten $\overline{x} = (2, 0, 2, 0, 2, 0, 0)$. □

Sats 13.5.1. *Simplexalgoritmen, startad från en godtycklig tillåten basindexmängd, stoppar för varje LP-problem om Blands regel används.*

Bevis. Vi visar satsen med ett motsägelsebevis. Antag därför att vi har ett LP-problem där simplexalgoritmen cyklar, och låt $\overline{x}$ vara den gemensamma baslösningen under iterationerna i cykeln.

Låt $\mathcal{C}$ beteckna mängden av index k för de variabler x_k som under iterationerna i cykeln övergår från att vara basvariabler till att vara fria variabler. Eftersom dessa variabler måste återkomma som basvariabler under cykeln, är $\mathcal{C}$ förstås också lika med indexmängden för de variabler x_k som under cykeln går från att vara fria variabler till att vara basvariabler. Vidare är $\overline{x}_k = 0$ för alla $k \in \mathcal{C}$.

Sätt

$$q = \max\{j \mid j \in \mathcal{C}\},$$

och låt α vara den basindexmängd som är i bruk under den iteration i cykeln då variabeln x_q övergår från basvariabel till fri variabel, och låt x_k vara den fria variabel som träder in i stället för x_q. I den efterkommande basindexmängden har med andra ord q ersatts av k. För motsvarande sökvektor v och reducerad kostnadsvektor z gäller att

$$z_k < 0 \quad \text{och} \quad v_q > 0,$$

och eftersom index k väljs enligt Blands regel, är

$$z_j \geq 0 \quad \text{för } j < k.$$

Vidare är $k < q$ på grund av definitionen av q, ty $k \in \mathcal{C}$.

Betrakta nu basindexmängden α' i en iteration när x_q återkommer som basvariabel från att ha varit fri variabel. För motsvarande reducerade kostnadsvektor z' gäller då på grund av Blands regel för valet av inträdande index, i det här fallet alltså q, att

$$z'_j \geq 0 \quad \text{för } j < q \quad \text{och} \quad z'_q < 0. \tag{13.11}$$

Speciellt är alltså $z'_k \geq 0$.

Eftersom $Av = 0$, $v_k = -1$ och $v_j = 0$ för $j \notin \alpha \cup \{k\}$, samt $z_j = 0$ för $j \in \alpha$, följer det av lemma 13.4.1 att

$$\sum_{j\in\alpha} z'_j v_j - z'_k = \langle z', v\rangle = \langle c, v\rangle = \langle z, v\rangle = \sum_{j\in\alpha} z_j v_j + z_k v_k = -z_k > 0,$$

så

$$\sum_{j\in\alpha} z'_j v_j > z'_k \geq 0.$$

Det finns följaktligen ett index $j_0 \in \alpha$ så att $z'_{j_0} v_{j_0} > 0$, och speciellt är alltså $z'_{j_0} \neq 0$, vilket innebär att j_0 inte kan tillhöra indexmängden α'. Variabeln x_{j_0} är således en basvariabel under en iteration i cykeln och en fri variabel under en annan. Detta innebär att j_0 är ett index i $\mathcal{C}$, så $j_0 \leq q$ på grund av definitionen av q. Fallet $j_0 = q$ är omöjligt eftersom $v_q > 0$ och $z'_q < 0$. Således är $j_0 < q$, och det följer därför av (13.11) att $z'_{j_0} > 0$. Detta medför i sin tur att $v_{j_0} > 0$, ty produkten $z'_{j_0} v_{j_0}$ är positiv. Indexet j_0 tillhör således mängden $I_+ = \{j \in \alpha \mid v_j > 0\}$, och eftersom $\overline{x}_{j_0}/v_{j_0} = 0 = \tau$, är

$$\min\{j \in I_+ \mid \overline{x}_j/v_j = \tau\} \leq j_0 < q.$$

Det strider därför mot Blands regel att välja q som det index som skall lämna basindexmängden α, och vi har därmed erhållit en motsägelse som visar att cykling inte kan förekomma. □

Anmärkning. För att förhindra cykling i simplexalgoritmen behöver man inte använda Blands regel hela tiden; det räcker att använda den när man kommer till en iteration där $\tau = 0$.

13.6 Simplexalgoritmen, fas 1

Simplexalgoritmen förutsätter att det finns en tillåten basindexmängd att starta ifrån. För vissa problem får man automatiskt en sådan när problemet skrivs på standardform. Så är exempelvis fallet för problem av typen

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax \le b,\ x \ge 0 \end{array}$$

där A är en $m \times n$-matris och högerledet b är icke-negativt. Genom att introducera m slackvariabler $s_{n+1}, s_{n+2}, \ldots, s_{n+m}$ samt sätta

$$s = (s_{n+1}, s_{n+2}, \ldots, s_{n+m})$$

får vi standardproblemet

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax + Es = b,\ x, s \ge 0, \end{array}$$

och här är det klart att slackvariablerna duger som basvariabler, dvs. att $\overline{x} = 0$, $\overline{s} = b$ är en tillåten baslösning med $\alpha = (n+1, n+2, \ldots, n+m)$ som motsvarande basindexmängd.

I andra fall är det inte alls uppenbart hur man skall hitta en tillåten basindexmängd att starta ifrån, men man kan alltid generera en sådan genom att använda simplexalgoritmen på ett lämpligt artificiellt problem.

Betrakta ett godtyckligt LP-problem på standardform

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax = b,\ x \ge 0, \end{array}$$

där A är en $m \times n$-matris. Vi kan utan inskränkning anta att $b \ge 0$, ty om någon koefficient b_j är negativ multiplicerar vi bara motsvarande ekvation med -1.

Vi börjar med att välja en $m \times k$-matris B så att matrisen

$$A' = \begin{bmatrix} A & B \end{bmatrix}$$

får rang m och systemet

$$A' \begin{bmatrix} x \\ y \end{bmatrix} = Ax + By = b$$

får en uppenbar tillåten basindexmängd α^0. De nya y-variablerna kallas *artificiella variabler*, och vi numrerar dem så att $y = (y_{n+1}, y_{n+2}, \ldots, y_{n+k})$.

Ett trivialt sätt att åstadkomma detta är att välja B lika med enhetsmatrisen E av ordning m, ty då är

$$\alpha^0 = (n+1, n+2, \ldots, n+m)$$

en tillåten basindexmängd med $(\overline{x}, \overline{y}) = (0, b)$ som motsvarande tillåtna baslösning. Ofta innehåller emellertid A-matrisen redan ett antal enhetskolonner, och då är det lämpligt att bara komplettera med de enhetskolonner som saknas för att A' skall innehålla enhetsmatrisen som delmatris.

Låt nu

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix}^T$$

vara $k \times 1$-matrisen bestående av k stycken ettor och betrakta följande artificiella LP-problem:

$$\begin{array}{ll} \min & \langle \mathbf{1}, y \rangle = y_{n+1} + \cdots + y_{n+k}. \\ \text{då} & Ax + By = b,\ x, y \geq 0 \end{array}$$

Detta problems optimala värde är uppenbarligen ≥ 0, och *värdet är lika med noll om och endast om det finns en tillåten punkt på formen* $(x, 0)$, *dvs. om och endast om det finns en icke-negativ lösning till systemet* $Ax = b$.

Vi löser därför det artificiella problemet med hjälp av simplexalgoritmen med α^0 som första tillåtna basindexmängd. Eftersom målfunktionen är nedåt begränsad, stoppar algoritmen (eventuellt behöver vi använda Blands tilläggsregel) efter ett ändligt antal iterationer i en tillåten basindexmängd α, där optimalitetsvillkoret är uppfyllt. Låt $(\overline{x}, \overline{y})$ beteckna motsvarande baslösning.

Vi har nu följande två möjligheter.

Fall 1. *Det artificiella problemets värde är större än noll.*

Det ursprungliga problemet saknar i detta fall tillåtna lösningar.

Fall 2. *Det artificiella problemets värde är lika med noll.*

I detta fall är nödvändigtvis $\overline{y} = 0$ och $A\overline{x} = b$.

Om $\alpha \subseteq \{1, 2, \ldots, n\}$, så är α också en tillåten basindexmängd till matrisen A, och vi kan starta simplexalgoritmen för vårt ursprungliga problem med utgångspunkt från α, $A_{*\alpha}^{-1}$ och baslösningen $\overline{x}$.

Om $\alpha \not\subseteq \{1, 2, \ldots, n\}$, sätter vi

$$\alpha' = \alpha \cap \{1, 2, \ldots, n\}.$$

Då är kolonnerna $\{A_{*k} \mid k \in \alpha'\}$ linjärt oberoende, och vi kan konstruera en indexmängd $\beta \supseteq \alpha'$ som är maximal med avseende på egenskapen att kolonnerna $\{A_{*k} \mid k \in \beta\}$ är linjärt oberoende.

Om rang $A = m$ kommer naturligtvis β att bestå av m element, och β är i så fall en basindexmängd till matrisen A. Eftersom $\overline{x}_j = 0$ för alla $j \notin \alpha'$, och därmed speciellt för alla $j \notin \beta$, är $\overline{x}$ också den till basindexmängden β hörande baslösningen till systemet $Ax = b$. Det följer att basindexmängden β är en tillåten basindexmängd för vårt ursprungliga problem. Vi kan också notera att baslösningen $\overline{x}$ är degenererad.

Om rang $A < m$, så kommer β bara att bestå av $p = \operatorname{rang} A$ stycken element, men vi kan nu stryka $m - p$ av ekvationerna i systemet $Ax = b$ utan att lösningsmängden förändras. Detta resulterar i ett nytt ekvivalent LP-problem med en koefficientmatris av rang p, och för detta problem är β en tillåten basindexmängd med $\overline{x}$ som motsvarande baslösning.

För att lösa ett typiskt LP-problem behöver man således normalt använda simplexalgoritmen två gånger. I fas 1 använder vi simplexalgoritmen för att lösa det artificiella problemet och genererar på så sätt en tillåten basindexmängd α till det ursprungliga LP-problemet. I fas 2 används simplexalgoritmen för att lösa det ursprungliga problemet med utgångspunkt från basindexmängden α.

Exempel 13.6.1. Vi illustrerar tekniken på det enkla problemet

$$\begin{array}{ll} \min & x_1 + 2x_2 + x_3 - 2x_4 \\ \text{då} & \begin{cases} x_1 + x_2 + x_3 - x_4 = 2 \\ 2x_1 + x_2 - x_3 + 2x_4 = 3 \\ x_1, x_2, x_3, x_4 \geq 0. \end{cases} \end{array}$$

Fas 1 består i att lösa det artificiella problemet

$$\begin{array}{ll} \min & y_5 + y_6 \\ \text{då} & \begin{cases} x_1 + x_2 + x_3 - x_4 + y_5 = 2 \\ 2x_1 + x_2 - x_3 + 2x_4 + y_6 = 3 \\ x_1, x_2, x_3, x_4, y_5, y_6 \geq 0. \end{cases} \end{array}$$

Räkningarna redovisas i tabellform. Den första tabellen är

$$\begin{array}{cccccc|c} 1 & 1 & 1 & -1 & 1 & 0 & 2 \\ 2 & 1 & -1 & 2 & 0 & 1 & 3 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

Vi börjar med att eliminera basvariablerna ur målfunktionen och får sedan följande följd av tabeller.

$$\begin{array}{cccccc|c} 1 & 1 & 1 & -1 & 1 & 0 & 2 \\ \underline{2} & 1 & -1 & 2 & 0 & 1 & 3 \\ \hline -3 & -2 & 0 & -1 & \underline{0} & \underline{0} & -5 \end{array}$$

$$\alpha = (5, 6), \quad k = 1, \quad r = 2$$

$$\begin{array}{cccccc|c} 0 & \frac{1}{2} & \underline{\frac{3}{2}} & -2 & 1 & -\frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & \frac{1}{2} & \frac{3}{2} \\ \hline \underline{0} & -\frac{1}{2} & -\frac{3}{2} & 2 & \underline{0} & \frac{3}{2} & -\frac{1}{2} \end{array}$$

$$\alpha = (5, 1), \quad k = 3, \quad r = 1$$

$$\begin{array}{cccccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{5}{3} \\ \hline \underline{0} & 0 & \underline{0} & 0 & 1 & 1 & 0 \end{array}$$

$$\alpha = (3, 1)$$

Ovanstående sluttabell för det artificiella problemet visar att $\alpha = (3, 1)$ är en tillåten basindexmängd i det ursprungliga problemet med $\overline{x} = (\frac{5}{3}, 0, \frac{1}{3}, 0)$ som motsvarande baslösning. Vi kan därför övergå till fas 2 med följande tabell som starttabell.

$$\begin{array}{cccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \frac{1}{3} & \frac{5}{3} \\ \hline 1 & 2 & 1 & -2 & 0 \end{array}$$

Genom att eliminera basvariablerna ur målfunktionen erhålles tabellen

$$\begin{array}{cccc|c} 0 & \frac{1}{3} & 1 & -\frac{4}{3} & \frac{1}{3} \\ 1 & \frac{2}{3} & 0 & \underline{\frac{1}{3}} & \frac{5}{3} \\ \hline \underline{0} & 1 & \underline{0} & -1 & -2 \end{array}$$

$$\alpha = (3, 1), \quad k = 4, \quad r = 2$$

vilken efter en iteration leder till följande tabell, som uppfyller optimalitetskriteriet

4	3	1	0	7
3	2	0	1	5
3	3	$\underline{0}$	$\underline{0}$	3

$\alpha = (3, 4)$

Minimivärdet är således lika med -3 och antas i punkten $\overline{x} = (0, 0, 7, 5)$. □

Eftersom arbetsvolymen växer med antalet artificiella variabler, bör man inte införa fler sådana än nödvändigt. För minimeringsproblemet

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax \leq b,\ x \geq 0 \end{array}$$

räcker det alltid med högst en artificiell variabel. Genom att introducera slackvariabler $s = (s_{n+1}, s_{n+2}, \ldots, s_{n+m})$ får man först det ekvivalenta problemet

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax + Es = b,\ x, s \geq 0 \end{array}$$

på standardform. Om $b \geq 0$ klarar man sig, som vi redan noterat, utan artificiella variabler. Låt annars i_0 vara index för den mest negativa koordinaten i högerledet b, och subtrahera ekvation nr i_0 i systemet $Ax + Es = b$ från alla övriga ekvationer med negativt högerled, samt byt slutligen tecken på ekvation nr i_0. Resultatet är ett med $Ax+Es = b$ ekvivalent ekvationssystem på formen $A'x + E's = b'$, där $b' \geq 0$ och samtliga kolonner i matrisen E' utom kolonn nr i_0 är är lika med motsvarande kolonner i enhetsmatrisen E. Fas 1 av simplexalgoritmen på problemet

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & A'x + E's = b',\ x, s \geq 0 \end{array}$$

behöver därför bara en artificiell variabel.

Existens av optimal lösning; dualitetssatsen

Simplexalgoritmen har naturligtvis sin främsta betydelse som en effektiv algoritm för att lösa LP-problem, men vi kan också använda den för att ge alternativa bevis för viktiga teoretiska resultat. Dessa är korollarier till följande sats.

Sats 13.6.1. *Varje LP-problem på standardform med tillåtna punkter har en tillåten basindexmängd där simplexalgoritmen stoppar.*

Bevis. Med Blands regel stoppar säkert fas 1 av simplexalgoritmen i en tillåten basindexmängd som kan användas som utgångspunkt för fas 2, och Blands regel garanterar att fas 2 också stoppar i en tillåten basindexmängd, där nu ett av de två stoppkriterierna i simplexalgoritmen är uppfyllt. □

Som första följdsats får vi nu ett nytt bevis för att varje LP-problem med ändligt värde har optimala lösningar (sats 12.1.1).

Korollarium 13.6.2. *Varje linjärt minimeringsproblem med tillåtna punkter och nedåt begränsad målfunktion har en optimal lösning.*

Bevis. Eftersom varje LP-problem kan ersättas med ett ekvivalent LP-problem på standardform, räcker det att betrakta sådana problem, och en tillåten basindexmängd där simplexalgoritmen stoppar måste, om målfunktionen är nedåt begränsad, vara en basindexmängd där optimalitetskriteriet är uppfyllt och där motsvarande baslösning följaktligen är en optimal lösning till LP-problemet. □

Vi får också ett algoritmiskt bevis för dualitetssatsen.

Korollarium 13.6.3 (Dualitetssatsen)**.** *Om det linjära minimeringsproblemet*

$$\begin{aligned} &\min \ \langle c, x\rangle \\ &\text{då} \quad Ax = b,\ x \geq 0 \end{aligned}$$

har tillåtna punkter, så har det samma optimala värde som det duala maximeringsproblemet

$$\begin{aligned} &\max \ \langle b, y\rangle \\ &\text{då} \quad A^T y \leq c. \end{aligned}$$

Bevis. Låt α vara den tillåtna basindexmängd där simplexalgoritmen stoppar. Om optimalitetskriteriet är uppfyllt, så följer det av sats 13.4.2 (i) att minimeringsproblemet och det duala maximeringsproblemet har samma ändliga optimala värde. Om algoritmen istället stoppar därför att målfunktionen i minimeringsproblemet är nedåt obegränsad, så saknar det duala problemet tillåtna punkter enligt sats 13.4.2 (ii a), och båda problemen har per definition värdet $-\infty$. □

Genom att skriva om allmänna minimeringsproblem på standardform kan man också härleda den allmänna formen av dualitetssatsen ur ovanstående specialfall.

13.7 Känslighetsanalys

I avsnitt 12.1 studerade vi det optimal värdets och den optimala lösningens beroende av koefficienterna i målfunktionen. I det här avsnittet skall vi med anknytning till simplexalgoritmen studera samma fråga och också hur lösningen till LP-problemet

$$\text{(P)} \qquad \begin{aligned} &\min \ \langle c, x\rangle \\ &\text{då} \quad Ax = b,\ x \geq 0 \end{aligned}$$

beror av högerledet b. I verkliga LP-problem är ofta koefficienterna i målfunktion och bivillkor inte exakt kända utan uppskattningar, och det är då förstås betydelsefult att veta hur känslig den optimala lösningen är för felaktigheter i indata. Även om indata är exakta är det intressant att veta hur den optimala lösningen påverkas av att en eller flera koefficienter ändras.

Låt α vara en basindexmängd till matrisen A, och låt $\overline{x}(b)$ beteckna motsvarande baslösning till systemet $Ax = b$, dvs.

$$\overline{x}(b)_\alpha = A_{*\alpha}^{-1}b \quad \text{och} \quad \overline{x}(b)_j = 0 \text{ för alla } j \notin \alpha.$$

Antag att LP-problemet (P) har en optimal lösning för ett visst värde på b och c och att denna lösning erhållits genom att simplexalgoritmen stoppat i basindexmängden α. För att så skall vara fallet måste dels baslösningen $\overline{x}(b)$ vara en tillåten baslösning, dvs.

$$\text{(i)} \qquad A_{*\alpha}^{-1}b \geq 0,$$

dels optimalitetsvillkoret $z \geq 0$ i simplexalgoritmen vara uppfyllt. Eftersom

$$z = c - A^T\overline{y} \quad \text{och} \quad \overline{y} = (A_{*\alpha}^{-1})^T c_\alpha,$$

är $z = c - (A_{*\alpha}^{-1}A)^T c_\alpha$, så optimalitetsvillkoret kan skrivas på formen

$$\text{(ii)} \qquad z(c) = c - (A_{*\alpha}^{-1}A)^T c_\alpha \geq 0.$$

Omvänt, för alla b och c som uppfyller villkoren (i) och (ii) är $\overline{x}(b)$ en optimal lösning till LP-problemet (P) eftersom optimalitetsvillkoret i simplexalgoritmen är uppfyllt.

Villkoret (i) är ett system av homogena linjära olikheter i variablerna b_1, $b_2, \ldots, b_m$ och definierar en polyedrisk kon B_α i $\mathbf{R}^m$, medan (ii) är ett system av homogena linjära olikheter i variablerna $c_1, c_2, \ldots, c_n$ och definierar en polyedrisk kon C_α i $\mathbf{R}^n$. Sammanfattningsvis har vi således följande resultat.

För alla $b \in B_\alpha$ och alla $c \in C_\alpha$ är $\overline{x}(b)$ en optimal lösning till LP-problemet (P).

Antag nu att vi löst problemet (P) för givna värden på b och c med $\overline{x} = \overline{x}(b)$ som optimal lösning och λ som optimalt värde. Villkoret (ii) bestämmer hur mycket vi kan ändra koefficienterna i målfunktionen utan att ändra den optimala lösningen; $\overline{x}$ är fortfarande en optimal lösning till det störda problemet

$$\begin{array}{ll} \text{min} & \langle c + \Delta c, x\rangle \\ \text{då} & Ax = b,\ x \geq 0 \end{array} \tag{P$'$}$$

om $z(c + \Delta c) = z(c) + z(\Delta c) \geq 0$, dvs. om

$$\Delta c - (A_{*\alpha}^{-1}A)^T(\Delta c)_\alpha \geq -z(c). \tag{13.12}$$

Det optimala värdet ändras förstås i så fall till $\lambda + \langle \Delta c, \overline{x}\rangle$.

Olikhet (13.12) definierar en polyeder i variablerna Δc_1, Δc_2, ..., Δc_n. Om speciellt $\Delta c_j = 0$ för alla j utom för $j = k$, dvs. om vi endast varierar c_k-koefficienten i målfunktionen, så bestämmer olikhet (13.12) ett (eventuellt obegränsat) slutet intervall $[-d_k, e_k]$ kring 0 för Δc_k.

Om vi istället ändrar koefficienterna i bivillkorens högerled och ersätter vektorn b med $b + \Delta b$, så är $\overline{x}(b + \Delta b)$ en optimal lösning till problemet

$$\begin{array}{ll} \text{min} & \langle c, x\rangle \\ \text{då} & Ax = b + \Delta b,\ x \geq 0 \end{array}$$

så länge som lösningen är tillåten, dvs. så länge som $A_{*\alpha}^{-1}(b + \Delta b) \geq 0$. Detta ger oss efter förenkling villkoret

$$A_{*\alpha}^{-1}(\Delta b) \geq -\overline{x}(b)_\alpha,$$

som är ett system av linjära olikheter som bestämmer hur Δb kan väljas. Om $\Delta b_i = 0$ för alla index utom för $i = k$, så är lösningsmängden för Δb_k ett intervall av typen $[-d_k, e_k]$.

Utskrifterna till datorprogram för simplexalgoritmen innehåller i allmänhet information om dessa intervall.

Exempel 13.7.1. En person studerar dietproblemet

$$\begin{array}{ll} \text{min} & \langle c, x\rangle \\ \text{då} & Ax \geq b,\ \ x \geq 0 \end{array}$$

i ett konkret fall med sex livsmedel och fyra krav på näringsämnen. Den aktuella prisvektorn c och högerledet b i bivillkoren ges av att $c^T = (1, 2, 3, 4, 1, 6)$ och $b^T = (10, 15, 20, 18)$. Efter att ha formulerat sitt problem löser personen det med hjälp av ett datorprogram, och datorkörningen ger följande utskrifter.

Resultatrapport

Optimalt värde: 8.52
Optimal lösning:

Livsmedel 1:	5.73
Livsmedel 2:	0.00
Livsmedel 3:	0.93
Livsmedel 4:	0.00
Livsmedel 5:	0.00
Livsmedel 6:	0.00

Känslighetsrapport

Variabel	Värde	Mål-koefficient	Tillåten minskning	Tillåten ökning
Livsmedel 1:	5.73	1.00	0.14	0.33
Livsmedel 2:	0.00	2.00	1.07	∞
Livsmedel 3:	0.93	3.00	2.00	0.50
Livsmedel 4:	0.00	4.00	3.27	∞
Livsmedel 5:	0.00	1.00	0.40	∞
Livsmedel 6:	0.00	6.00	5.40	∞

Bivillkor	Slutgiltigt värde	Skuggpris	Begränsning högerled	Tillåten minskning	Tillåten ökning
Ämne 1:	19.07	0.00	10.00	∞	9.07
Ämne 2:	31.47	0.00	15.00	∞	16.47
Ämne 3:	20.00	0.07	20.00	8.00	7.00
Ämne 4:	18.00	0.40	18.00	4.67	28.67

Av känslighetsrapporten framgår att den optimala lösningen förblir oförändrad om priset på livsmedel 1 stiger med högst 0.33 eller minskar med högst 0.14, allt annat oförändrat. En förändring av priset med z enheter inom detta intervall förändrar naturligtvis det optimala värdet med $5.73\,z$.

En prissänkning på livsmedel 4 med högst 3.27 eller en obegränsad prisökning på samma livsmedel, allt annat oförändrat, påverkar inte den optimala lösningen och inte heller det optimala värdet.

Enligt olikhet (13.12) är mängden av prisändringar som lämnar den optimala lösningen oförändrad en polyeder och således konvex. I vårt exempel blir därför den optimala lösningen oförändrad om exempelvis priset på livsmedlen 1, 2 och 3 höjs med 0.20, 1.20 resp. 0.10, ty eftersom

$$\frac{0.20}{0.33} + \frac{1.20}{\infty} + \frac{0.10}{0.50} \leq 1,$$

är $\Delta c = (0.20,\ 1.20,\ 0.10,\ 0,\ 0,\ 0)$ en konvex kombination av tillåtna prishöjningar på varje enskilt livsmedel för sig.

Av känslighetsrapporten framgår också hur den optimala lösningen påverkas av vissa förändringar i högerledet b. Om kravet på näringsinnehåll för ämne 1 ändras från 10 till 15, så förblir den optimala lösningen oförändrad eftersom bivillkoret inte är bindande och ökningen 5 är mindre än den tillåtna ökningen.

Om istället b_4 ökar med säg 20 enheter från 18 till 38, en ökning som ligger inom ramen för den tillåtna, så kommer den nya optimala lösningen att ges av samma basindexmängd som tidigare, dvs. den optimala dieten kommer fortfarande bara att bestå av livsmedlen 1 och 3, men det optimala värdet ökar med $20 \cdot 0.40$ till 16.52 eftersom skuggpriset på ämne 4 är 0.40. □

13.8 Duala simplexalgoritmen

Simplexalgoritmen startar, när den tillämpas på ett problem

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax = b,\ x \geq 0 \end{array}$$

med begränsat värde, i en tillåten basindexmängd α^0 och genererar sedan en ändlig följd $(\alpha^k, \overline{x}^k, \overline{y}^k)_{k=0}^p$ av basindexmängder α^k, motsvarande baslösningar $\overline{x}^k$ och vektorer $\overline{y}^k$ med följande egenskaper:

(i) Baslösningarna $\overline{x}^k$ är extremalpunkter till polyedern

$$X = \{x \in \mathbf{R}^n \mid Ax = b,\ x \geq 0\}$$

av tillåtna punkter.

(ii) Sträckorna $[\overline{x}^k, \overline{x}^{k+1}]$ är kantlinjer i polyedern X.

(iii) Målfunktionsvärdena $(\langle c, \overline{x}^k\rangle)_{k=0}^p$ bildar en avtagande följd.

(iv) $\langle b, \overline{y}^k\rangle = \langle c, \overline{x}^k\rangle$ för alla k.

(v) Algoritmen stoppar efter p iterationer när optimalitetsvillkoret är uppfyllt, och då är $\overline{y}^p$ en extremalpunkt till polyedern

$$Y = \{y \in \mathbf{R}^m \mid A^T y \leq c\}.$$

(vi) $\overline{x}^p$ är en optimal lösning, och $\overline{y}^p$ är en optimal lösning till det duala problemet

$$\begin{array}{ll} \max & \langle b, y\rangle \\ \text{då} & A^T y \leq c. \end{array}$$

(vii) För $0 \leq k \leq p-1$ ligger däremot vektorerna $\overline{y}^k$ utanför polyedern Y.

Istället för att leta sig fram till den optimala lösningen $\overline{x}^p$ genom att stega utefter kanter i polyedern X till dess att man kommer till en basindexmängd som också svarar mot en extremalpunkt i polyedern Y, kan man stega sig fram utefter kanter i polyedern Y. Denna observation leder till följande metod för att lösa minimeringsproblemet.

Duala simplexagoritmen

Givet en basindexmängd α sådan att $z = c - A^T\overline{y} \geq 0$, där $\overline{y} = (A_{*\alpha}^{-1})^T c_\alpha$.
Upprepa steg 1–4 till dess att stopp inträffar.

1. Beräkna den till α hörande baslösningen $\overline{x}$.
2. *Stoppkriterium:* Stoppa om $\overline{x} \geq 0$.
   Optimal lösning: $\overline{x}$. Optimal dual lösning: $\overline{y}$.
   Stoppa också om någon bivillkorsekvation är av typen $a'_{i1}x_1 + a'_{i2}x_2 + \cdots + a'_{in}x_n = b'_i$ med $b'_i > 0$ och $a'_{ij} \leq 0$ för alla j, vilket innebär att det primala problemet saknar tillåtna lösningar.
3. Generera annars en ny basindexmängd α' genom att byta ut ett index i α på ett sådant sätt att den nya reducerade kostnadsvektorn z' förblir icke-negativ och $\langle b, \overline{y}' \rangle \geq \langle b, \overline{y} \rangle$, där $\overline{y}' = (A_{*\alpha'}^{-1})^T c_{\alpha'}$
4. *Uppdatera:* $\alpha := \alpha'$, $\overline{y} := \overline{y}'$.

Vi avstår från att precisera pivoteringsreglerna för att algoritmen ska fungera utan nöjer oss med att räkna igenom ett exempel.

Exempel 13.8.1. Vi ska lösa minimeringsproblemet

$$\begin{array}{ll} \min & x_1 + 2x_2 + 3x_3 \\ \text{då} & \begin{cases} 2x_1 \quad + \quad x_3 \geq 9 \\ x_1 + 2x_2 \quad \geq 12 \\ \quad x_2 + 2x_3 \geq 15, \ x \geq 0 \end{cases} \end{array}$$

med hjälp av den duala simplexalgoritmen, och vi börjar därför med att skriva problemet på standardform:

$$\begin{array}{ll} \min & x_1 + 2x_2 + 3x_3 \\ \text{då} & \begin{cases} 2x_1 \quad + \quad x_3 - x_4 \quad = 9 \\ x_1 + 2x_2 \quad - x_5 \quad = 12 \\ \quad x_2 + 2x_3 \quad - x_6 = 15, \ x \geq 0. \end{cases} \end{array}$$

Motsvarande simplextabell ser förstås ut så här:

$$\begin{array}{cccccc|c} 2 & 0 & 1 & -1 & 0 & 0 & 9 \\ 1 & 2 & 0 & 0 & -1 & 0 & 12 \\ 0 & 1 & 2 & 0 & 0 & -1 & 15 \\ \hline 1 & 2 & 3 & \underline{0} & \underline{0} & \underline{0} & 0 \end{array}$$

Som jämförelse skriver vi också ut det duala maximeringsproblemet:

$$\begin{array}{ll} \max & 9y_1 + 12y_2 + 15y_3 \\ \text{då} & \begin{cases} 2y_1 + \ y_2 \qquad\quad \le 1 \\ \qquad 2y_2 + \ y_3 \le 2 \\ \ y_1 \qquad\quad + 2y_3 \le 3, \ y \ge 0. \end{cases} \end{array}$$

Vi kan starta den duala simplexalgoritmen med utgångspunkt från basindexmängden $\alpha = (4, 5, 6)$, och som vanligt har vi markerat baskolonnerna med understrykning. Motsvarande baslösning $\overline{x}$ är förstås inte tillåten eftersom $\overline{x}_\alpha = (-9, -12, -15)$ har negativa koordinater. Raden $[\,1\ 2\ 3\ 0\ 0\ 0\,]$ i den nedre delen av tabellen är lika med reducerade kostnadsvektorn $z^T = c^T - \overline{y}^T A$. Radmatrisen $\overline{y}^T = c_\alpha^T A_{*\alpha}^{-1} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ kan också avläsas i nedre raden; den står under matrisen $-E$. Vektorn $\overline{y}$ ligger i polyedern Y av tillåtna punkter till det duala problemet eftersom $z^T \ge 0$.

Vi kommer nu successivt att byta ut ett element i taget i basindexmängden. Som pivotrad r väljer vi genomgående den rad som svarar mot den mest negativa koordinaten i $\overline{x}_\alpha$, dvs. i första iterationen den tredje raden i ovanstående simplextabell. För att den reducerade kostnadsvektorn ska förbli icke-negativ efter pivoteringen måste vi som pivotkolonn välja en kolonn k, där matriselementet a_{rk} är positivt och kvoten z_k/a_{rk} är så liten som möjligt. I tabellen ovan är detta den tredje kolonnen, så vi pivoterar kring elementet på plats $(3, 3)$. Detta leder till följande tabell:

$$\begin{array}{rrrrrr|r} 2 & -\frac{1}{2} & 0 & -1 & 0 & \frac{1}{2} & \frac{3}{2} \\ 1 & 2 & 0 & 0 & -1 & 0 & 12 \\ 0 & \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & \frac{15}{2} \\ \hline 1 & \frac{1}{2} & \underline{0} & \underline{0} & \underline{0} & \frac{3}{2} & -\frac{45}{2} \end{array}$$

I den nya tabellen är $\alpha = (4, 5, 3)$, $\overline{x}_\alpha = (-\frac{3}{2}, -12, \frac{15}{2})$ och $\overline{y} = (0, 0, \frac{3}{2})$. Det mest negativa elementet i $\overline{x}_\alpha$ står i den andra raden, och för positiva koefficienter a'_{2k} är kvoterna z_k/a'_{2k} minst för $k = 2$. Pivotering kring elementet på plats $(2, 2)$ leder till tabellen

$$\begin{array}{rrrrrr|r} \frac{9}{4} & 0 & 0 & -1 & -\frac{1}{4} & \frac{1}{2} & \frac{9}{2} \\ \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & 0 & 6 \\ -\frac{1}{4} & 0 & 1 & 0 & \frac{1}{4} & -\frac{1}{2} & \frac{9}{2} \\ \hline \frac{3}{4} & \underline{0} & \underline{0} & \underline{0} & \frac{1}{4} & \frac{3}{2} & -\frac{51}{2} \end{array}$$

Nu är $\alpha = (4, 2, 3)$, $\overline{x}_\alpha = (-\frac{9}{2}, 6, \frac{9}{2})$ och $\overline{y} = (0, \frac{1}{4}, \frac{3}{2})$. Vi skall den här gången välja elementet i rad 1 och kolonn 1 som pivotelement, vilket leder till nästa tabell.

$$\begin{array}{cccccc|c} 1 & 0 & 0 & -\frac{4}{9} & -\frac{1}{9} & \frac{2}{9} & 2 \\ 0 & 1 & 0 & \frac{2}{9} & -\frac{4}{9} & -\frac{1}{9} & 5 \\ 0 & 0 & 1 & -\frac{1}{9} & \frac{2}{9} & -\frac{4}{9} & 5 \\ \hline \underline{0} & \underline{0} & \underline{0} & \frac{1}{3} & \frac{1}{3} & \frac{4}{3} & -27 \end{array}$$

Här är $\alpha = (1, 2, 3)$, $\overline{x}_\alpha = (2, 5, 5)$ och $\overline{y} = (\frac{1}{3}, \frac{1}{3}, \frac{4}{3})$. Eftersom $\overline{x}_\alpha \geq 0$, är optimalitetsvillkoret uppfyllt. Det optimala värdet är således lika med 27, och minimeringsproblemet antar sitt minimum i punkten $(2, 5, 5, 0, 0, 0)$, medan det duala maximeringsproblemet har $(\frac{1}{3}, \frac{1}{3}, \frac{4}{3})$ som optimal lösning. Minimipunkten i vårt ursprungliga minimeringsproblem är förstås $(2, 5, 5)$. □

13.9 Komplexitet

Hur många iterationer behövs det för att lösa ett LP-problem med simplexalgoritmen? Svaret beror naturligtvis på problemets storlek, och på vilka pivoteringsregler som används. Empiriskt har man funnit att antalet iterationer i stort sett växer linjärt med antalet rader m och sublinjärt med antalet kolonner n för realistiska tillämpningsproblem. I de flesta verkliga problemen är vidare n en liten multipel av m, vanligtvis högst $10m$. Antalet iterationer brukar därför ligga någonstans mellan m och $4m$, vilket betyder att simplexalgoritmen i allmänhet fungerar mycket bra.

Algoritmens uppförande i värsta fallet är emellertid dåligt (för alla kända pivoteringsregler). Klee och Minty har konstruerat ett exempel där antalet iterationer växer exponentiellt med problemets storlek.

Exempel 13.9.1 (Klee och Minty, 1972). I LP-problemet

$$\begin{aligned} \max\ & 2^{n-1}x_1 + 2^{n-2}x_2 + \cdots + 2x_{n-1} + x_n \\ \text{då}\ & \begin{cases} x_1 & & & & \leq 5 \\ 4x_1 + & x_2 & & & \leq 25 \\ 8x_1 + & 4x_2 + & x_3 & & \leq 125 \\ \vdots & & & & \vdots \\ 2^n x_1 + & 2^{n-1}x_2 + & \ldots + & 4x_{n-1} + x_n & \leq 5^n \end{cases} \end{aligned}$$

med n variabler och n linjära olikheter har polyedern av tillåtna punkter 2^n extremalpunkter. Om simplexalgoritmen används på det ekvivalenta standardproblemet och startas i den mot $x = 0$ svarande tillåtna baslösningen och om man i varje iterationssteg väljer kolonnen med minst reducerad

kostnad z_k som pivotkolonn, så kommer algoritmen att besöka samtliga 2^n tillåtna baslösningar innan den slutligen stoppar i den optimala lösningen $(0, 0, \ldots, 5^n)$. Antalet iterationer är därför lika med 2^n och växer således exponentiellt med n. □

En algoritm för att lösa ett problem i n variabler kallas *strikt polynomiell*, om det finns ett positivt heltal k så att antalet elementära räkneoperationer i algoritmen växer med n som högst $O(n^k)$. I många algoritmer beror antalet operationer också på indatas storlek. En algoritm kallas *polynomiell*, om antalet räkneoperationer växer som ett polynom i L, där L är antalet binära bitar som behövs för att representera all indata (dvs. i linjär programmering matriserna A, b och c).

Gausselimination är en strikt polynomiell algoritm, ty antalet aritmetiska operationer för att lösa ett linjärt ekvationssystem med n ekvationer och n obekanta är $O(n^3)$.

Klee–Mintys och andra liknande exempel visar att simplexalgoritmen inte är strikt polynomiell. Men även om värstafallet-uppförandet är dåligt, så fungerar simplexalgoritmen erfarenhetsmässigt bra. Detta styrks av probabilistiska analyser, som gjorts av Borgwardt (1987), Smale (1983), Adler och Megiddo (1985), m. fl. Ett exempel på ett resultat av denna analys är att (en variant av) simplexalgoritmen, givet en viss speciell sannolikhetsfördelning för indata, i genomsnitt konvergerar efter $O(m^2)$ iterationer, där m är antalet bivillkor.

Men det finns polynomiella algoritmer som löser LP-problem (med rationella koefficienter som indata). Leonid Khachiyan visade 1979 att den s. k. *ellipsoidalgoritmen* är en sådan algoritm. I ellipsoidmetoden reduceras ett LP-problem till problemet att hitta en lösning till ett system $Ax > b$ av strikta olikheter med begränsad lösningsmängd, och algoritmen genererar en följd av krympande ellipsoider, som alla garanterat innehåller alla lösningar till systemet. Om någon ellipsoids centrum satisfierar samtliga olikheter, så är en lösning funnen. Annars stoppar processen när en genererad ellipsoid har alltför liten volym för att innehålla samtliga lösningar, om det finns några, med slutsatsen att det inte finns några lösningar.

Ellipsoidmetoden löser LP-problem på standardform med inputstorlek L och n variabler med $O(n^4L)$ aritmetiska operationer. Det stod emellertid snart klart att ellipsoidmetoden trots detta inte kunde konkurrera med simplexalgoritmen på praktiska problem, eftersom den konvergerar långsamt för problem av måttlig storlek (och orsaken till det är förstås att den underförstådda konstanten i O-uppskattningen är stor).

En ny polynomiell algoritm upptäcktes 1984 av Narendra Karmarkar. Hans algoritm genererar en följd av punkter, som ligger i det inre av mängden

av tillåtna punkter och konvergerar mot en optimal punkt. Algoritmen utnyttjar upprepad centrering av de genererade punkterna med hjälp av en projektiv skalningstransformation. Den teoretiska komplexitetsgränsen för den ursprungliga versionen av algoritmen är också den $O(n^4L)$.

Eftersom Karmarkars algoritm visade sig vara konkurrenskraftig med simplexalgoritmen på praktiska problem, blev hans upptäckt startpunkten för ett intensivt utvecklingsarbete av alternativa s. k. inrepunktsmetoder för LP-problem och mer generella konvexa problem. Vi kommer att studera en sådan algoritm i kapitel 18.

Det är fortfarande ett öppet problem huruvida det finns någon strikt polynomiell algoritm för LP-problem.

Övningar

13.1 Skriv följande problem på standardform

a) $\min \ 2x_1 - 2x_2 + x_3$ då $\begin{cases} x_1 + x_2 - x_3 \geq 3 \\ x_1 + x_2 - x_3 \leq 2 \\ x_1, x_2, x_3 \geq 0 \end{cases}$

b) $\min \ x_1 + 2x_2$ då $\begin{cases} x_1 + x_2 \geq 1 \\ x_2 \geq -2 \\ x_1 \geq 0. \end{cases}$

13.2 Bestäm samtliga icke-negativa baslösningar till följande ekvationssystem

a) $\begin{cases} 5x_1 + 3x_2 + x_3 = 40 \\ x_1 + x_2 + x_3 = 10 \end{cases}$

b) $\begin{cases} x_1 - 2x_2 - x_3 + x_4 = 3 \\ 2x_1 + 5x_2 - 3x_3 + 2x_4 = 6. \end{cases}$

13.3 Formulera det duala problemet till

$$\min \ x_1 + x_2 + 4x_3$$
$$\text{då} \ \begin{cases} x_1 - x_3 = 1 \\ x_1 + 2x_2 + 7x_3 = 7, \ x \geq 0 \end{cases}$$

samt visa att $(1, 3, 0)$ är en optimal lösning och att $(\frac{1}{2}, \frac{1}{2})$ är en optimal lösning till det duala problemet.

13.4 Lös följande LP-problem med hjälp av simplexalgoritmen.

a) $\min \ -x_4$ då $\begin{cases} x_1 + x_4 = 1 \\ x_2 + 2x_4 = 2 \\ x_3 - x_4 = 3, \ x \geq 0 \end{cases}$

b) $\max \ 2x_1 - x_2 + x_3 - 3x_4 + x_5$ då $\begin{cases} x_1 + 2x_4 - x_5 = 15 \\ x_2 + x_4 + x_5 = 12 \\ x_3 - 2x_4 + x_5 = 9, \ x \geq 0 \end{cases}$

c) $\max\ 15x_1 + 12x_2 + 14x_3$
$$\text{då} \quad \begin{cases} 3x_1 + 2x_2 + 5x_3 \leq 6 \\ x_1 + 3x_2 + 3x_3 \leq 3 \\ 5x_3 \leq 4,\ x \geq 0 \end{cases}$$

d) $\max\ 2x_1 + x_2 + 3x_3 + x_4 + 2x_5$
$$\text{då} \quad \begin{cases} x_1 + 2x_2 + x_3 + x_5 \leq 10 \\ x_2 + x_3 + x_4 + x_5 \leq 8 \\ x_1 + x_3 + x_4 \leq 5,\ x \geq 0 \end{cases}$$

e) $\min\ x_1 - 2x_2 + x_3$
$$\text{då} \quad \begin{cases} x_1 + x_2 - 2x_3 \leq 3 \\ x_1 - x_2 + x_3 \leq 2 \\ -x_1 - x_2 + x_3 \leq 0,\ x \geq 0 \end{cases}$$

f) $\min\ x_1 - x_2 + 2x_3 - 3x_4$
$$\text{då} \quad \begin{cases} 2x_1 + 3x_2 + x_3 = 2 \\ x_1 + 3x_2 + x_3 + 5x_4 = 4,\ x \geq 0. \end{cases}$$

13.5 Genomför i detalj stegen i simplexalgoritmen på problemet

$$\min\ -x_2 + x_4$$
$$\text{då} \quad \begin{cases} x_1 + x_4 + x_5 = 1 \\ x_2 - 2x_4 - x_5 = 1 \\ x_3 + 2x_4 + x_5 = 3,\ x \geq 0. \end{cases}$$

Är den optimala lösningen unik?

13.6 Använd tekniken med artificiella variabler för att lösa LP-problemet

$$\max\ x_1 + 2x_2 + 3x_3 - x_4$$
$$\text{då} \quad \begin{cases} x_1 + 2x_2 + 3x_3 = 15 \\ 2x_1 + x_2 + 5x_3 = 20 \\ x_1 + 2x_2 + x_3 + x_4 = 10,\ x \geq 0. \end{cases}$$

13.7 Visa med hjälp av simplexalgoritmen att följande system av likheter och olikheter är lösbara.

a) $\begin{cases} 3x_1 + x_2 + 2x_3 + x_4 + x_5 = 2 \\ 2x_1 - x_2 + x_3 + x_4 + 4x_5 = 3,\ x \geq 0 \end{cases}$

b) $\begin{cases} x_1 - x_2 + 2x_3 + x_4 \geq 6 \\ -2x_1 + x_2 - 2x_3 + 7x_4 \geq 1 \\ x_1 - x_2 + x_3 - 3x_4 \geq -1,\ x \geq 0. \end{cases}$

13.8 Lös LP-problemet

$$\min\ x_1 + 2x_2 + 3x_3$$
$$\text{då} \quad \begin{cases} 2x_1 + x_3 \geq 3 \\ x_1 + 2x_2 \geq 4 \\ x_2 + 2x_3 \geq 5,\ x \geq 0. \end{cases}$$

13.9 Skriv följande problem på standardform samt lös det därefter med hjälp av simplexalgoritmen

$$\begin{array}{ll}\min & 8x_1 - x_2 \\ \text{då} & \begin{cases} 3x_1 + \ x_2 \geq \ 1 \\ \ x_1 - \ x_2 \leq \ 2 \\ \ x_1 + 2x_2 = 20, \ x \geq 0. \end{cases}\end{array}$$

13.10 Lös följande LP-problem med hjälp av den duala simplexalgoritmen.

a) $\begin{array}{ll}\min & 2x_1 + x_2 + 3x_3 \\ \text{då} & \begin{cases} \ x_1 + x_2 + \ x_3 \geq 2 \\ 2x_1 - x_2 \qquad \geq 1 \\ \qquad x_2 + 2x_3 \geq 2, \ x \geq 0 \end{cases}\end{array}$

b) $\begin{array}{ll}\min & x_1 + 2x_2 \\ \text{då} & \begin{cases} \ x_1 \qquad - 2x_3 \geq -5 \\ -2x_1 + 3x_2 - \ x_3 \geq -4 \\ -2x_1 + 5x_2 - \ x_3 \geq \ \ 2, \ x \geq 0 \end{cases}\end{array}$

c) $\begin{array}{ll}\min & 3x_1 + 2x_2 + 4x_3 \\ \text{då} & \begin{cases} 4x_1 \qquad + 2x_3 \geq 5 \\ \ x_1 + 3x_2 + 2x_3 \geq 4, \ x \geq 0. \end{cases}\end{array}$

13.11 Visa att för alla b_1, b_2 med $b_2 \geq b_1 \geq 0$ är $\overline{x} = \big(b_1, \frac{1}{2}(b_2 - b_1), 0\big)$ en optimal lösning till problemet

$$\begin{array}{ll}\min & x_1 + x_2 + 4x_3 \\ \text{då} & \begin{cases} x_1 \qquad - \ x_3 = b_1 \\ x_1 + 2x_2 + 7x_3 = b_2, \ x \geq 0. \end{cases}\end{array}$$

13.12 Undersök hur den optimala lösningen till LP-problemet

$$\begin{array}{ll}\max & 2x_1 + tx_2 \\ \text{då} & \begin{cases} \ x_1 + x_2 \leq 5 \\ 2x_1 + x_2 \leq 7, \ x \geq 0 \end{cases}\end{array}$$

varierar då den reella parametern t varierar.

13.13 En skofabrikant tillverkar två skomodeller A och B. Begränsad tillgång på läder gör att tillverkat antal par x_A och x_B av de två modellerna måste uppfylla olikheterna

$$x_A \leq 1000, \quad 4x_A + 3x_B \leq 4100, \quad 3x_A + 5x_B \leq 5000.$$

Försäljningspriset för A och B är 500 resp. 350 kr per par. Det kostar 200 kr att tillverka ett par skor av modell B, medan tillverkningskostnaden för skor av modell A är osäker på grund av att maskinen krånglat och därför

endast kan uppskattas till att ligga mellan 300 och 410 kr per par. Visa att fabrikanten trots detta kan bestämma hur många par av skor han skall tillverka av varje modell för att maximera sin vinst.

13.14 Dietkonstnären Hugo vill tillgodose sitt dagliga behov av vitaminerna P, Q och R genom att endast leva på mjölk och bröd. Hans dagsbehov av vitaminerna är 6, 12 resp. 4 mg. En liter mjölk kostar 7.50 kr och innehåller 2 mg P, 2 mg Q och ingenting av R; en limpa bröd kostar 20 kr och innehåller 1 mg P, 4 mg Q och 4 mg R. Vitaminerna är inte giftiga, så en eventuell överdosering gör inget. Hugo vill komma undan så billigt som möjligt. Vilken daglig matsedel skall han välja? Antag att mjölkpriset börjar stiga. Hur högt kan det bli utan att Hugo ser någon anledning att byta matsedel?

13.15 Visa i lemma 13.4.1 att om z_k är en reducerad kostnad och v är motsvarande sökvektor, så är z_k lika med riktningsderivatan med avseende på riktningen $-v$ av målfunktionen $\langle c, x\rangle$.

13.16 I den här övningen skall vi skissera en metod för att förhindra cykling i simplexalgoritmen. Betrakta problemet

$$\text{(P)} \qquad \begin{aligned} &\min && \langle c, x\rangle \\ &\text{då} && Ax = b,\ x \geq 0 \end{aligned}$$

och låt α vara en godtycklig tillåten basindexmängd med motsvarande baslösning $\overline{x}$. Definiera för varje positivt tal ϵ en ny vektor $\overline{x}(\epsilon) \in \mathbf{R}^n$ genom att sätta

$$\overline{x}(\epsilon)_\alpha = \overline{x}_\alpha + (\epsilon, \epsilon^2, \ldots, \epsilon^m) \quad \text{och} \quad \overline{x}(\epsilon)_j = 0 \text{ för alla } j \notin \alpha,$$

samt en ny vektor $b(\epsilon) \in \mathbf{R}^m$ genom att sätta

$$b(\epsilon) = A\overline{x}(\epsilon).$$

Då är uppenbarligen $\overline{x}(\epsilon)$ en icke-negativ baslösning till systemet $Ax = b(\epsilon)$ med α som motsvarande basindexmängd. Koordinaterna i vektorn $b(\epsilon)$ är vidare polynom av grad m i variabeln ϵ.

a) Visa att för alla utom ändligt många tal $\epsilon > 0$ är samtliga baslösningar till systemet $Ax = b(\epsilon)$ icke-degenererade. Följaktligen finns det ett tal $\epsilon_0 > 0$ så att samtliga baslösningar är icke-degenererade om $0 < \epsilon < \epsilon_0$.

b) Visa att om $0 < \epsilon < \epsilon_0$ och β är en tillåten basindexmängd för problemet

$$\text{(P}_\epsilon\text{)} \qquad \begin{aligned} &\min && \langle c, x\rangle \\ &\text{då} && Ax = b(\epsilon),\ x \geq 0 \end{aligned}$$

så är β också en tillåten basindexmängd för problemet (P).

c) Om $\epsilon > 0$ är tillräckligt litet kommer därför simplexalgoritmen tillämpad på problemet (P$_\epsilon$) att stoppa med en tillåten basindexmängd β, som även är tillåten för problemet (P). Visa att β i så fall också uppfyller stoppvillkoren för problemet (P).

Man kan således undvika cykling i degenererade LP-problem med följande metod: Stör högerledet genom att bilda $\overline{x}(\epsilon)$ och kolonnmatrisen $b(\epsilon)$, där ϵ är ett litet tal. Använd simplexalgoritmen på det störda problemet. Algoritmen stoppar i någon basindexmängd β. Motsvarande ostörda problem stoppar i samma basindexmängd.

13.17 Antag att $\mathcal{A}$ är en polynomiell algoritm för att lösa system av linjära olikheter. För lösbara olikheter $Cx \geq b$ producerar algoritmen en lösning $\overline{x}$, vilket anges med utskriften $\mathcal{A}(C, b) = \overline{x}$, och för olösbara olikheter resulterar algoritmen i utskriften $\mathcal{A}(C, b) = \emptyset$. Konstruera med hjälp av algoritmen $\mathcal{A}$ en polynomiell algoritm för att lösa godtyckliga LP-problem

$$\begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax \geq b,\ x \geq 0. \end{array}$$

13.18 Genomför räkningarna i simplexalgoritmen för Klee och Mintys exempel då $n = 3$.

Del IV

Inrepunktsmetoder

Kapitel 14

Descentmetoder

De vanligaste numeriska algoritmerna för att minimera differentierbara funktioner av flera variabler är s.k. *descentalgoritmer*. Dessa genererar med utgångspunkt från en lämpligt vald startpunkt iterativt en följd av punkter med avtagande funktionsvärden, och processen avbryts när man erhållit ett funktionsvärde som enligt något kriterium kan anses approximera minimivärdet tillräckligt bra. Det finns dock ingen algoritm som fungerar för helt godtyckliga funktioner; för att garantera konvergens mot minimivärdet eller att algoritmen ska upptäcka att inget sådant finns, behövs det speciella antaganden om funktionerna som ska minimeras. Konvexitet är ett sådant antagande, och för konvexa funktioner kan man också i många fall bestämma konvergenshastigheten.

I det här kapitlet beskrivs descentmetoder i allmänna ordalag, och vi exemplifierar med den allra enklaste descentmetoden, brantaste lutningsmetoden.

14.1 Allmänna principer

Vi skall studera optimeringsproblemet

$$\text{(P)} \qquad \min f(x)$$

där funktionen f är differentierbar och definierad på en öppen delmängd X av $\mathbf{R}^n$. Vi förutsätter att problemet har en lösning, dvs. att det finns en optimal punkt $\hat{x} \in X$ och sätter $f_{\min} = f(\hat{x})$. Ett antagande som garanterar existensen av en (unik) optimal lösning är enligt korollarium 8.1.7 att funktionen f är starkt konvex och har någon sluten icke-tom subnivåmängd.

Syftet är att utifrån en vald *startpunkt* $x_0 \in X$ iterativt generera en följd av punkter $x_1, x_2, x_3, \ldots$ i X med avtagande funktionsvärden och med

egenskapen att $f(x_k) \to f_{\min}$ då $k \to \infty$, och för att i iterationssteg k, givet punkten x_k, konstruera nästa punkt x_{k+1} väljer vi (om $f'(x_k) \neq 0$) först en vektor v_k så att funktionen $\phi_k(t) = f(x_k + tv_k)$ är strängt avtagande då $t = 0$. Sedan gör vi en *linjesökning* utefter halvlinjen $x_k + tv_k$, $t > 0$, och bestämmer med hjälp av specificerade regler punkten $x_{k+1} = x_k + h_k v_k$ så att $f(x_{k+1}) < f(x_k)$. Vektorn v_k kallas *sökriktningen*, och det positiva talet h_k kallas *steglängden*. Algoritmen avbryts när differensen $f(x_k) - f_{\min}$ är mindre än någon i förväg vald toleransnivå.

Schematiskt kan vi beskriva en typisk descentalgoritm så här:

Descentalgoritm

Givet en startpunkt $x \in X$.
Upprepa
1. Välj (om $f'(x) \neq 0$) en sökriktning v och en steglängd $h > 0$ så att $f(x + hv) < f(x)$.
2. *Uppdatera:* $x := x + hv$.

till dess att stoppkriteriet är uppfyllt.

Olika strategier för att välja sökriktning, olika sätt att utföra linjesökningen samt olika stoppkriterier ger förstås upphov till olika algoritmer.

Sökriktningen

Tillåtna sökriktningar i iterationssteg k är vektorer v_k som uppfyller

$$\langle f'(x_k), v_k \rangle < 0,$$

ty detta garanterar att funktionen $\phi_k(t) = f(x_k + tv_k)$ är avtagande i punkten $t = 0$ eftersom $\phi_k'(0) = \langle f'(x_k), v_k \rangle$. Vi kommer att studera två sätt att välja sökriktning.

I *brantaste lutningsmetoden* väljer man $v_k = -f'(x_k)$, vilket är ett tillåtet val eftersom $\langle f'(x_k), v_k \rangle = -\|f'(x_k)\|^2 < 0$. Lokalt ger detta val snabbast minskning av funktionsvärdet.

I *Newtons metod* väljer man istället $v_k = -f''(x_k)^{-1} f'(x_k)$, förutsatt att andraderivatan $f''(x_k)$ existerar och är positivt definit. I så fall är nämligen $\langle f'(x_k), v_k \rangle = -\langle f'(x_k), f''(x_k)^{-1} f'(x_k) \rangle < 0$.

Linjesökningen

Givet sökrikningen v_k finns det flera tänkbara strategier för att välja steglängden h_k.

1. *Exakt linjesökning.* Steglängden h_k bestäms genom minimering av envariabelfunktionen $t \mapsto f(x_k + tv_k)$. Denna metod används för teoretiska undersökningar av algoritmers uppförande men nästan aldrig i praktiken på grund av beräkningskostnaderna för att utföra den endimensionella minimeringen.

2. Steglängdsföljden $(h_k)_{k=1}^{\infty}$ är given *apriori,* t. ex. som $h_k = h$ eller som $h_k = h/\sqrt{k+1}$ för någon positiv konstant h. Detta är en enkel regel som ofta används inom konvex optimering.

3. Steglängden h_k i punkten x_k definieras som $h_k = \rho(x_k)$ för någon given funktion ρ. Denna teknik används i analysen av Newtons metod för självkonkordanta funktioner.

4. *Armijos regel.* För givna parametrar $\alpha, \beta \in]0,1[$ sätter man

$$h_k = \beta^m,$$

där m är det minsta icke-negativa heltalet för vilket punkten $x_k + \beta^m v_k$ ligger i f:s definitionsmängd X och olikheten

$$f(x_k + \beta^m v_k) \leq f(x_k) + \alpha\beta^m \langle f'(x_k), v_k \rangle \tag{14.1}$$

är uppfylld.

Det finns säkert ett sådant m, eftersom $\beta^n \to 0$ då $n \to \infty$ och

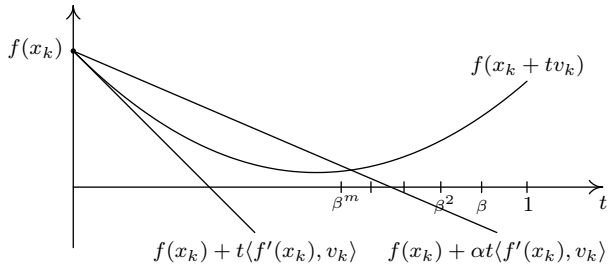
$$\lim_{t \to 0} \frac{f(x_k + tv_k) - f(x_k)}{t} = \langle f'(x_k), v_k \rangle < \alpha \langle f'(x_k), v_k \rangle.$$

Talet m bestäms med hjälp av enkel "backtracking": Starta med $m = 0$ och undersök om $x_k + \beta^m v_k \in X$ och olikheten (14.1) gäller. Om så inte är fallet, öka m med 1 och upprepa till dess att detta gäller. Figur 14.1 illustrerar processen.

För konvexa funktioner är funktionsvärdesminskningen per steglängd i iterationssteg k, dvs. kvoten $(f(x_k) - f(x_{k+1}))/h_k$, mindre än eller lika med $-\langle f'(x_k), v_k \rangle$ för varje val av steglängd h_k. Med steglängd h_k vald enligt Armijos regel blir samma kvot också $\geq -\alpha\langle f'(x_k), v_k \rangle$. Minskningen per steglängd är med Armijos regel med andra ord minst α av vad den maximalt kan vara. Ju mindre α, desto mindre minskning i varje iterationssteg. I praktiska algoritmer brukar typiska α-värden ligga mellan mellan 0.01 och 0.3.

Parametern β är avgörande för hur många backtrackingsteg som behövs. Ju större β, desto fler backtrackingsteg, dvs. desto finare linjesökning. Parametern β väljs ofta mellan 0.1 och 0.8.

Armijos regel förekommer i olika varianter och används i ett flertal praktiska algoritmer.



Figur 14.1. Armijos regel: Som steglängd väljs $h_k = \beta^m$, där m är det minsta icke-negativa heltalet med egenskapen att $f(x_k + \beta^m v_k) \le f(x_k) + \alpha\beta^m\langle f'(x_k), v_k\rangle$.

Stoppkriterier

Eftersom det optimala värdet i allmänhet inte är känt i förväg, kan stoppkriteriet inte formuleras direkt i termer av $f_{\min}$. Intuitivt förefaller det rimligt att x ligger nära minimipunkten $\hat{x}$ om derivatan $f'(x)$ är jämförelsevis liten, och nästa sats visar att så är fallet under lämpliga förutsättningar på målfunktionen.

Sats 14.1.1. *Låt $f\colon X \to \mathbf{R}$ vara en differentierbar, μ-starkt konvex funktion, och antag att funktionen har ett minimum i punkten $\hat{x} \in X$. För alla $x \in X$ är då*

(i) $$f(x) - f(\hat{x}) \le \frac{1}{2\mu}\|f'(x)\|^2 \qquad \textit{och}$$

(ii) $$\|x - \hat{x}\| \le \frac{1}{\mu}\|f'(x)\|.$$

Bevis. På grund av konvexitetsantagandet är

(14.2) $$f(y) \ge f(x) + \langle f'(x), y - x\rangle + \tfrac{1}{2}\mu\|y - x\|^2$$

för alla $x, y \in X$. Högra sidan av olikheten (14.2) är en konvex kvadratisk funktion i variabeln y som minimeras för $y = x - \mu^{-1}f'(x)$, och minimivärdet är lika med $f(x) - \frac{1}{2}\mu^{-1}\|f'(x)\|^2$. För alla $y \in X$ är därför

$$f(y) \ge f(x) - \tfrac{1}{2}\mu^{-1}\|f'(x)\|^2.$$

Vi får nu olikheten (i) genom att som y välja minimipunkten $\hat{x}$.

Byt nu y mot x och x mot $\hat{x}$ i olikheten (14.2); den reduceras då, beroende på att $f'(\hat{x}) = 0$, till olikheten

$$f(x) \ge f(\hat{x}) + \tfrac{1}{2}\mu\|x - \hat{x}\|^2,$$

som i kombination med (i) ger (ii). □

Vi återvänder nu till descentalgoritmen och vår diskussion av stoppkriteriet. Sätt

$$S = \{x \in X \mid f(x) \leq f(x_0)\},$$

där x_0 är den valda startpunkten, och antag att subnivåmängden S är konvex och att målfunktionen f är μ-starkt konvex på S. Alla av descentalgoritmen genererade punkterna $x_1, x_2, x_3, \ldots$ kommer förstås att ligga i S eftersom funktionsvärdena är avtagande, och det följer därför av sats 14.1.1 att $f(x_k) < f_{\min} + \epsilon$ om $\|f'(x_k)\| < (2\mu\epsilon)^{1/2}$.

Som stoppkriterium kan vi således använda villkor av typen

$$\|f'(x_k)\| \leq \eta,$$

vilket garanterar att $f(x_k) - f_{\min} \leq \eta^2/2\mu$ och att $\|x_k - \hat{x}\| \leq \eta/\mu$. Ett problem i sammanhanget är att konvexitetskonstanten μ sällan är känd.

Konvergenshastighet

Låt oss säga att en konvergent följd $x_0, x_1, x_2, \ldots$ av punkter med gränsvärde $\hat{x}$ konvergerar *minst linjärt* om det finns en konstant $c < 1$ sådan att

$$\|x_{k+1} - \hat{x}\| \leq c\|x_k - \hat{x}\| \tag{14.3}$$

för alla k, och att konvergensen är *minst kvadratisk* om det finns en konstant C sådan att

$$\|x_{k+1} - \hat{x}\| \leq C\|x_k - \hat{x}\|^2 \tag{14.4}$$

för alla k. Vi säger vidare att konvergensen *inte är bättre än linjär* resp. *kvadratisk* om

$$\varliminf_{k\to\infty} \frac{\|x_{k+1} - \hat{x}\|}{\|x_k - \hat{x}\|^\alpha} > 0$$

för $\alpha = 1$ resp. $\alpha = 2$.

Observera att olikheten (14.3) medför att följden $(x_k)_0^\infty$ konvergerar mot $\hat{x}$, ty det följer genom induktion att

$$\|x_k - \hat{x}\| \leq c^k\|x_0 - \hat{x}\|$$

för alla k.

På motsvarande sätt medför olikheten (14.4) att följden $(x_k)_0^\infty$ konvergerar mot $\hat{x}$ om startpunkten x_0 uppfyller villkoret $\|x_0-\hat{x}\| < C^{-1}$, ty induktion ger i det här fallet att

$$\|x_k - \hat{x}\| \le C^{-1}\big(C\|x_0 - \hat{x}\|\big)^{2^k}$$

för alla k.

En iterativ metod, som när den appliceras på funktioner i en given funktionsklass genererar konvergenta följder, säges *konvergera linjärt* resp. *kvadratiskt* för funktionsklassen ifråga, om de genererade följderna är minst linjärt resp. minst kvadratiskt konvergenta och det finns någon följd som inte är bättre än linjärt resp. kvadratiskt konvergent.

14.2 Brantaste lutningsmetoden

I det här avsnittet ska vi analysera brantaste lutningsalgoritmen då steglängden är konstant. Den iterativa formuleringen av den aktuella varianten av algoritmen ser ut så här:

Brantaste lutningsalgoritmen med konstant steglängd
Givet en startpunkt x och en steglängd h.
Upprepa
1. Beräkna sökriktningen $v = -f'(x)$.
2. *Uppdatera:* $x := x + hv$.

till dess att stoppkriteriet är uppfyllt.

För starkt konvexa funktioner med Lipschitzkontinuerlig derivata och minimipunkt konvergerar algoritmen linjärt mot minimipunkten förutsatt att steglängden är tillräckligt liten och startpunkten väljs tillräckligt nära minimipunkten. Detta är huvudbudskapet i följande sats (och exempel 14.2.1).

Sats 14.2.1. *Antag att f är en funktion med lokal minimipunkt $\hat{x}$ och att det finns en öppen omgivning U av $\hat{x}$ där funktionens restriktion $f|_U$ är μ-starkt konvex och differentierbar med Lipschitzkontinuerlig derivata och Lipschitzkonstant L. Då konvergerar brantaste lutningsalgoritmen med konstant steglängd h minst linjärt mot $\hat{x}$, om steglängden är tillräckligt liten och startpunkten x_0 ligger tillräckligt nära $\hat{x}$.*

Mer precist gäller: Om bollen med centrum i $\hat{x}$ och radie $\|x_0 - \hat{x}\|$ ligger i U och $h \le \mu/L^2$, så ligger de av algoritmen genererade punkterna $(x_k)_0^\infty$ i U och

$$\|x_{k+1} - \hat{x}\| \le c\|x_k - \hat{x}\|,$$

för alla k, där $c = \sqrt{1 - h\mu}$.

Bevis. Antag induktivt att de av brantaste lutningsalgoritmen genererade punkterna $x_0, x_1, \ldots, x_k$ ligger i U och att $\|x_k - \hat{x}\| \leq \|x_0 - \hat{x}\|$. Eftersom restriktionen $f|_U$ är μ-starkt konvex och $f'(\hat{x}) = 0$, är

$$\langle f'(x_k), x_k - \hat{x}\rangle = \langle f'(x_k) - f'(\hat{x}), x_k - \hat{x}\rangle \geq \mu\|x_k - \hat{x}\|^2$$

enligt sats 7.3.1, och Lipschitzkontinuiteten hos derivatan ger oss olikheten

$$\|f'(x_k)\| = \|f'(x_k) - f'(\hat{x})\| \leq L\|x_k - \hat{x}\|.$$

Genom att kombinera de två olikheterna erhålls olikheten

$$\begin{aligned}\langle f'(x_k), x_k - \hat{x}\rangle &\geq \mu\|x_k - \hat{x}\|^2 = \frac{\mu}{2}\|x_k - \hat{x}\|^2 + \frac{\mu}{2}\|x_k - \hat{x}\|^2 \\ &\geq \frac{\mu}{2}\|x_k - \hat{x}\|^2 + \frac{\mu}{2L^2}\|f'(x_k)\|^2.\end{aligned}$$

För nästa punkt $x_{k+1} = x_k - hf'(x_k)$ erhålls därför olikheten

$$\begin{aligned}\|x_{k+1} - \hat{x}\|^2 &= \|x_k - hf'(x_k) - \hat{x}\|^2 = \|(x_k - \hat{x}) - hf'(x_k)\|^2 \\ &= \|x_k - \hat{x}\|^2 - 2h\langle f'(x_k), x_k - \hat{x}\rangle + h^2\|f'(x_k)\|^2 \\ &\leq \|x_k - \hat{x}\|^2 - h\mu\|x_k - \hat{x}\|^2 - h\frac{\mu}{L^2}\|f'(x_k)\|^2 + h^2\|f'(x_k)\|^2 \\ &= (1 - h\mu)\|x_k - \hat{x}\|^2 + h\big(h - \frac{\mu}{L^2}\big)\|f'(x_k)\|^2.\end{aligned}$$

För $h \leq \mu/L^2$ är således $\|x_{k+1} - \hat{x}\|^2 \leq (1 - h\mu)\|x_k - \hat{x}\|^2$, vilket visar att olikheten i satsen gäller med $c = \sqrt{1 - h\mu} < 1$, och att punkten x_{k+1} också uppfyller induktionsförutsättningen eftersom den ligger närmare $\hat{x}$ än vad x_k gör. Brantaste lutningsalgoritmen konvergerar därför minst linjärt för funktionen f under angivna förutsättningar på h och x_0. □

För μ-starkt konvexa funktioner med Lipschitzkontinuerlig derivata som är definierade på hela $\mathbf{R}^n$ kan vi erhålla något skarpare resultat.

Sats 14.2.2. *Antag att $f \in \mathcal{S}_{\mu,L}(\mathbf{R}^n)$. Då konvergerar brantaste lutningsalgoritmen med konstant steglängd h och godtycklig startpunkt x_0 minst linjärt mot funktionens minimipunkt $\hat{x}$ om*

$$0 < h \leq \frac{2}{\mu + L}.$$

För den av algoritmen genererade punktföljden $(x_k)_0^\infty$ gäller mer precist att

$$\|x_k - \hat{x}\| \leq \Big(1 - \frac{2h\mu L}{\mu + L}\Big)^{k/2}\|x_0 - \hat{x}\|. \tag{14.5}$$

För $h = \dfrac{2}{\mu + L}$ *är*

$$\|x_k - \hat{x}\| \le \Big(\frac{Q-1}{Q+1}\Big)^k \|x_0 - \hat{x}\| \qquad \textit{och} \tag{14.6}$$

$$f(x_k) - f_{\min} \le \frac{L}{2}\Big(\frac{Q-1}{Q+1}\Big)^{2k} \|x_0 - \hat{x}\|^2, \tag{14.7}$$

där $Q = L/\mu$ *är funktionsklassens konditionstal.*

Bevis. Enligt korollarium 8.1.7 har funktionen en unik minimipunkt $\hat{x}$, och som i beviset för föregående sats är

$$\|x_{k+1} - \hat{x}\|^2 = \|x_k - \hat{x}\|^2 - 2h\langle f'(x_k), x_k - \hat{x}\rangle + h^2\|f'(x_k)\|^2.$$

Eftersom $f'(\hat{x}) = 0$, följer det nu av sats 7.4.4 (med $x = \hat{x}$ och $v = x_k - \hat{x}$) att

$$\langle f'(x_k), x_k - \hat{x}\rangle \ge \frac{\mu L}{\mu + L}\|x_k - \hat{x}\|^2 + \frac{1}{\mu + L}\|f'(x_k)\|^2,$$

vilket insatt i likheten ovan efter förenkling ger

$$\|x_{k+1} - \hat{x}\|^2 \le \Big(1 - \frac{2h\mu L}{\mu + L}\Big)\|x_k - \hat{x}\|^2 + h\Big(h - \frac{2}{\mu + L}\Big)\|f'(x_k)\|^2.$$

För $h \le 2/(\mu + L)$ är därför

$$\|x_{k+1} - \hat{x}\| \le \Big(1 - \frac{2h\mu L}{\mu + L}\Big)^{1/2}\|x_k - \hat{x}\|,$$

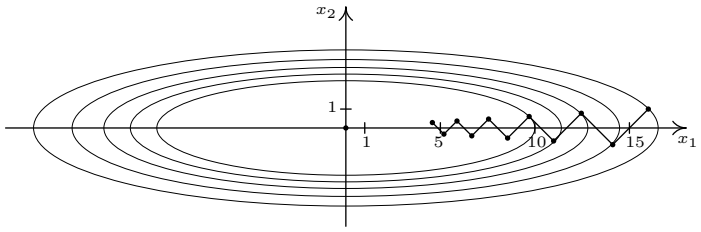
vilket efter iteration leder till olikheten (14.5).

Insättning av $h = 2(\mu + L)^{-1}$ i (14.5) ger olikheten (14.6), och den sista olikheten (14.7) följer av (14.6) och sats 1.1.2 eftersom $f'(\hat{x}) = 0$. □

Konvergenshastigheten i satserna 14.2.1 och 14.2.2 beror av konditionstalet $Q \ge 1$. Ju mindre Q desto snabbare konvergens. I praktiken känner man naturligtvis sällan konstanterna μ och L och därmed heller inte Q, så satserna 14.2.1 och 14.2.2 är av kvalitativ karaktär och kan sällan användas för att förutse hur många iterationer som behövs för att uppnå en viss precision.

Följande exempel visar att resultatet i sats 14.2.2 är skarpt i den meningen att olikheten (14.6) inte kan skärpas.

Exempel 14.2.1. Betrakta funktionen $f(x) = \frac{1}{2}(\mu x_1^2 + L x_2^2)$, där $0 < \mu \le L$. Funktionen tillhör klassen $\mathcal{S}_{\mu,L}(\mathbf{R}^2)$, $f'(x) = (\mu x_1, L x_2)$, och minimipunkten är förstås $\hat{x} = (0, 0)$.



Figur 14.2. I figuren visas några nivåkurvor till funktionen $f(x) = \frac{1}{2}(x_1^2 + 16x_2^2)$ samt hur brantaste lutningsalgoritmen fortskrider då $x^{(0)} = (16, 1)$ valts som startpunkt. Funktionens konditionstal är $Q = 16$, så konvergensen mot minimipunkten $(0, 0)$ är relativt långsam − i varje iterationssteg förbättras avståndet från $x^{(k)}$ till origo med faktorn 15/17.

Låt oss nu utföra brantaste lutningsalgoritmen med konstant steglängd $h = 2(\mu + L)^{-1}$ och startpunkt (L, μ). Med $\alpha = \frac{Q-1}{Q+1}$ får vi i de två första iterationerna

$$\begin{aligned} x^{(0)} &= (L, \mu) \\ f'(x^{(0)}) &= (\mu L, \mu L) \\ x^{(1)} &= x^{(0)} - hf'(x^{(0)}) = \alpha(L, -\mu) \\ f'(x^{(1)}) &= \alpha(\mu L, -\mu L) \\ x^{(2)} &= x^{(1)} - hf'(x^{(1)}) = \alpha^2(L, \mu) \end{aligned}$$

och det följer nu med induktion att

$$x^{(k)} = \alpha^k(L, (-1)^k\mu).$$

Följaktligen är

$$\|x^{(k)} - \hat{x}\| = \alpha^k\sqrt{L^2 + \mu^2} = \alpha^k\|x^{(0)} - \hat{x}\|,$$

så olikheten (14.6) gäller i detta fall med likhet.

Det är slutligen värt att notera att $2(\mu+L)^{-1}$ är precis den steglängd som erhålls om man använder sig av exakt linjesökning i varje iterationssteg. □

Brantaste lutningsalgoritmen är inte invariant under affina koordinatbyten. Konvergenshastigheten kan således förbättras om man först gör ett koordinatbyte som minskar konditionstalet.

EXEMPEL 14.2.2. Vi fortsätter med funktionen $f(x) = \frac{1}{2}(\mu x_1^2 + Lx_2^2)$ från

föregående exempel. Gör variabelbytet $y_1 = \sqrt{\mu}\, x_1$, $y_2 = \sqrt{L}\, x_2$ och sätt

$$g(y) = f(x) = \tfrac{1}{2}(y_1^2 + y_2^2).$$

Funktionen g har konditionstalet $Q = 1$, så brantaste lutningsalgoritmen startad i en godtycklig punkt $y^{(0)}$ ger efter ett iterationssteg $y^{(1)} = (0, 0)$, dvs. minimipunkten. □

Brantaste lutningsalgoritmen konvergerar alltför långsamt för att vara praktiskt användbar i realistiska problem. I nästa kapitel skall vi därför i detalj studera en mer effektiv metod för optimering, Newtons metod.

Övningar

14.1 Använd brantaste lutningsalgoritmen på problemet

$$\min x_1^2 + 2x_2^2.$$

Starta i punkten $(1, 1)$ och gör tre iterationer.

14.2 Låt f vara funktionen $f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$ med $X = \{x \in \mathbf{R}^2 \mid x_1 > 1\}$ som definitionsmängd och sätt $x^{(0)} = (2, 2)$.

a) Visa att subnivåmängden $\{x \in X \mid f(x) \leq f(x^{(0)})\}$ inte är sluten.

b) Uppenbarligen är $f_{\min} = \inf f(x) = \frac{1}{2}$, men visa att brantaste lutningsmetoden med start i punkten $x^{(0)}$ och med linjesökning enligt Armijos regel med parametervärden $\alpha \leq \frac{1}{2}$ och $\beta < 1$, genererar en följd $x^{(k)} = (a_k, a_k)$, $k = 0, 1, 2, \ldots$, av punkter som konvergerar mot punkten $(1, 1)$. Funktionsvärdena $f(x^{(k)})$ konvergerar följaktligen inte mot $f_{\min}$ utan mot 1.
[Ledning: Visa att $a_{k+1} - 1 \leq (1 - \beta)(a_k - 1)$ för alla k.]

14.3 Antag att brantaste lutningsalgoritmen med konstant steglängd konvergerar mot punkten $\hat{x}$ när den appliceras på den kontinuerligt differentierbara funktionen f. Visa att $\hat{x}$ är en stationär punkt till funktionen, dvs. $f'(\hat{x}) = 0$.

Kapitel 15

Newtons metod

Den bärande idén i Newtons metod består i att i varje iterationssteg approximera funktionen som ska minimeras med funktionens Taylorpolynom av grad två i den aktuella punkten, och att som descentriktning välja en vektor som minimerar detta polynom. Som sökriktning i punkten x till funktionen f väljs följaktligen en vektor som minimerar Taylorpolynomet

$$P(v) = f(x) + Df(x)[v] + \tfrac{1}{2}D^2f(x)[v,v] = f(x) + \langle f'(x), v\rangle + \tfrac{1}{2}\langle v, f''(x)v\rangle,$$

och eftersom $P'(v) = f'(x) + f''(x)v$, fås den minimerande sökvektorn som lösning till ekvationen

$$f''(x)v = -f'(x).$$

Om andraderivatan $f''(x)$ är positivt definit, så har ekvationen en entydig lösning och lösningsvektorn v minimerar Taylorpolynomet.

Iterationsstegen i Newtons metod är förstås mer arbetskrävande än iterationsstegen i brantaste lutningsmetoden eftersom man för att bestämma sökriktningarna behöver beräkna andraderivatan och lösa kvadratiska ekvationssystem. Detta kompenseras emellertid, som vi ska se, mer än väl av att konvergensen mot minimivärdet är mycket snabbare.

15.1 Newtonriktning och Newtondekrement

Eftersom sökrikningarna i Newtons metod erhålls genom minimering av andragradspolynom, börjar vi med att undersöka när sådana polynom har ett minimum. Ett nödvändigt villkor för att ett andragradspolynom skall vara nedåt begränsat är att polynomet är konvext, så vi kan redan från början inskränka oss till att undersöka konvexa andragradspolynom.

Sats 15.1.1. *Ett andragradspolynom*

$$P(v) = \tfrac{1}{2}\langle v, Av\rangle + \langle b, v\rangle + c$$

i n variabler, där A är en positivt semidefinit symmetrisk operator, är nedåt begränsat på $\mathbf{R}^n$ om och endast om ekvationen

$$Av = -b \tag{15.1}$$

har en lösning.

Om polynomet är nedåt begränsat, så har det ett minimum, och vektorn $\hat{v}$ är en minimipunkt om och endast om $A\hat{v} = -b$.

Om $\hat{v}$ är en minimipunkt till polynomet P, så är vidare

$$P(v) - P(\hat{v}) = \tfrac{1}{2}\langle v - \hat{v}, A(v - \hat{v})\rangle \tag{15.2}$$

för alla $v \in \mathbf{R}^n$.

Om $\hat{v}_1$ och $\hat{v}_2$ är minimipunkter till polynomet, så är $\langle \hat{v}_1, A\hat{v}_1\rangle = \langle \hat{v}_2, A\hat{v}_2\rangle$.

Anmärkning. Ett annat sätt att uttrycka att ekvationen (15.1) är lösbar är förstås att säga att vektorn $-b$, och därmed också vektorn b, ligger i bildrummet till operatorn A. Men bildrummet till en operator på ett ändligdimensionellt rum är lika med det ortogonala komplementet till operatorns nollrum. Ekvationen (15.1) är följaktligen lösbar om och endast om

$$Av = 0 \Rightarrow \langle b, v\rangle = 0.$$

Bevis. Antag först att ekvation (15.1) saknar lösning. Då finns det enligt anmärkningen ovan en vektor v sådan att $Av = 0$ och $\langle b, v\rangle \neq 0$. Det följer att

$$P(tv) = \tfrac{1}{2}\langle v, Av\rangle t^2 + \langle b, v\rangle t + c = \langle b, v\rangle t + c$$

för alla $t \in \mathbf{R}$, och eftersom koefficienten för t är nollskild innebär detta att polynomet P inte är nedåt begränsat.

Antag nu att $A\hat{v} = -b$. Då är

$$\begin{aligned} P(v) - P(\hat{v}) &= \tfrac{1}{2}(\langle v, Av\rangle - \langle \hat{v}, A\hat{v}\rangle) + \langle b, v\rangle - \langle b, \hat{v}\rangle \\ &= \tfrac{1}{2}(\langle v, Av\rangle - \langle \hat{v}, A\hat{v}\rangle) - \langle A\hat{v}, v\rangle + \langle A\hat{v}, \hat{v}\rangle \\ &= \tfrac{1}{2}(\langle v, Av\rangle + \langle \hat{v}, A\hat{v}\rangle - \langle A\hat{v}, v\rangle - \langle \hat{v}, Av\rangle) \\ &= \tfrac{1}{2}\langle v - \hat{v}, A(v - \hat{v})\rangle \geq 0 \end{aligned}$$

för alla $v \in \mathbf{R}^n$. Detta visar att polynomet P är nedåt begränsat, att $\hat{v}$ är en minimipunkt och att likheten (15.2) gäller.

Eftersom varje positivt semidefinit symmetrisk operator A har en unik positivt semidefinit symmetrisk kvadratrot $A^{1/2}$, kan vi skriva likheten (15.2) på formen

$$P(v) = P(\hat{v}) + \tfrac{1}{2}\langle A^{1/2}(v-\hat{v}), A^{1/2}(v-\hat{v})\rangle = P(\hat{v}) + \tfrac{1}{2}\|A^{1/2}(v-\hat{v})\|^2.$$

Om också v är en minimipunkt till P, så är därför $A^{1/2}(v-\hat{v}) = 0$, och detta medför förstås att $A(v-\hat{v}) = A^{1/2}(A^{1/2}(v-\hat{v})) = 0$, dvs. att $Av = A\hat{v} = -b$. Varje minimipunkt til P fås således som lösning till ekvation (15.1).

Om slutligen $\hat{v}_1$ och $\hat{v}_2$ är två minimipunkter till polynomet så är $A\hat{v}_1 = A\hat{v}_2$ $(= -b)$, och det följer att $\langle \hat{v}_1, A\hat{v}_1\rangle = \langle \hat{v}_1, A\hat{v}_2\rangle = \langle A\hat{v}_1, \hat{v}_2\rangle = \langle A\hat{v}_2, \hat{v}_2\rangle = \langle \hat{v}_2, A\hat{v}_2\rangle$. □

Sats 15.1.1 reducerar problemet att minimera ett konvext kvadratiskt polynom i n variabler till problemet att lösa ett kvadratiskt linjärt ekvationssystem i n variabler, vilket är ett i sammanhanget trivialt numeriskt problem som kan utföras med $O(n^3)$ aritmetiska operationer.

Vi är nu redo att definiera ingredienserna i Newtons metod.

Definition. Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar funktion med en öppen delmängd X av $\mathbf{R}^n$ som definitionsmängd, och antag att $x \in X$ är en punkt där andraderivatan $f''(x)$ är positivt semidefinit.

Med en *Newtonriktning* Δx_{nt} till funktionen f i punkten x menas en lösning v till ekvationen

$$f''(x)v = -f'(x).$$

Anmärkning. Det följer av anmärkningen efter sats 15.1.1 att det finns en Newtonriktning i punkten x om och endast om implikationen

$$f''(x)v = 0 \Rightarrow \langle f'(x), v\rangle = 0$$

gäller. Att Newtonriktning saknas är följaktligen ekvivalent med att det finns en vektor w sådan att $f''(x)w = 0$ och $\langle f'(x), w\rangle = 1$.

Newtonriktningen Δx_{nt} är naturligtvis entydigt bestämd som

$$\Delta x_{\text{nt}} = -f''(x)^{-1}f'(x),$$

om andraderivatan $f''(x)$ är icke-singulär, dvs. positivt definit.

Enligt sats 15.1.1 är Newtonriktningen Δx_{nt}, när den existerar, en minimerande vektor till Taylorpolynomet

$$P(v) = f(x) + \langle f'(x), v\rangle + \tfrac{1}{2}\langle v, f''(x)v\rangle$$

och för differensen $P(0) - P(\Delta x_{\text{nt}})$ gäller att

$$P(0) - P(\Delta x_{\text{nt}}) = \tfrac{1}{2}\langle 0 - \Delta x_{\text{nt}}, f''(x)(0 - \Delta x_{\text{nt}})\rangle = \tfrac{1}{2}\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}}\rangle.$$

Om vi använder oss av Taylorapproximationen $f(x+v) \approx P(v)$, så är alltså

$$f(x) - f(x + \Delta x_{\text{nt}}) \approx P(0) - P(\Delta x_{\text{nt}}) = \tfrac{1}{2}\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}}\rangle,$$

vilket visar att $\frac{1}{2}\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}}\rangle$ (för små Δx_{nt}) är en approximation till den förbättring, dvs. minskning, av funktionsvärdet som vi får genom att ersätta $f(x)$ med $f(x+\Delta x_{\text{nt}})$. För att kunna beskriva denna minskning litet enklare gör vi följande definition.

Definition. *Newtondekrementet* $\lambda(f,x)$ till funktionen f i punkten x är en kvantitet som definieras av att

$$\lambda(f,x) = \sqrt{\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}}\rangle}$$

i punkter x där funktionen har en Newtonriktning Δx_{nt}, och

$$\lambda(f,x) = +\infty$$

i punkter x där funktionen saknar Newtonriktning.

Observera att definitionen av Newtondekrement är oberoende av vilken Newtonriktning Δx_{nt} som väljs i punkten x i de fall då Newtonriktningen inte är unik. Detta följer direkt av den sista utsagan i sats 15.1.1.

I termer av Newtondekrementet är alltså

$$f(x) - f(x + \Delta x_{\text{nt}}) \approx \tfrac{1}{2}\lambda(f,x)^2$$

för små värden på Δx_{nt}.

Per definition är $f''(x)\Delta x_{\text{nt}} = -f'(x)$, så det följer att Newtondekrementet, när det är ändligt, också kan beräknas med formeln

$$\lambda(f,x) = \sqrt{-\langle \Delta x_{\text{nt}}, f'(x)\rangle}.$$

I punkter x där andraderivatan är positivt definit blir alltså speciellt

$$\lambda(f,x) = \sqrt{\langle f''(x)^{-1}f'(x), f'(x)\rangle}.$$

Exempel 15.1.1. Den konvexa envariabelfunktionen

$$f(x) = -\ln x, \;\; x > 0$$

har Newtondekrement

$$\lambda(f,x) = \sqrt{\langle x^2(-x^{-1}), -x^{-1}\rangle} = \sqrt{(-x)\cdot(-x^{-1})} = 1$$

överallt. □

I punkter med en Newtonriktning kan vi också, genom att använda oss av att $f''(x)$ har en positivt semidefinit symmetrisk kvadratrot, uttrycka Newtondekrementet med hjälp av den euklidiska normen $\|\cdot\|$ på följande sätt:

$$\lambda(f,x) = \sqrt{\langle f''(x)^{1/2}\Delta x_{\mathrm{nt}}, f''(x)^{1/2}\Delta x_{\mathrm{nt}}\rangle} = \|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|.$$

Den förbättring i funktionsvärde som erhålles genom att ta ett steg i Newtonriktningen Δx_{nt} är alltså proportionell mot $\|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|^2$ och inte mot $\|\Delta x_{\mathrm{nt}}\|^2$, vilket motiverar att vi introducerar följande seminorm.

Definition. Antag att f är en två gånger differentierbar funktion i en omgivning av punkten x och att andraderivatan $f''(x)$ är positivt semidefinit. Med den till funktionen f och punkten x hörande *lokala seminormen* menas funktionen $\|\cdot\|_x \colon \mathbf{R}^n \to \mathbf{R}_+$ som definieras av att

$$\|v\|_x = \sqrt{\langle v, f''(x)v\rangle} = \|f''(x)^{1/2}v\|.$$

Man verifierar omedelbart att $\|\cdot\|_x$ verkligen är en seminorm på $\mathbf{R}^n$, och eftersom

$$\{v \in \mathbf{R}^n \mid \|v\|_x = 0\} = \mathcal{N}(f''(x)),$$

där $\mathcal{N}(f''(x))$ är nollrummet till $f''(x)$, är seminormen $\|\cdot\|_x$ en norm om och endast om den positivt semidefinita andraderivatan $f''(x)$ är icke-singulär, dvs. positivt definit.

Uttryckt med hjälp av den lokala normen är alltå Newtondekrementet

$$\lambda(f,x) = \|\Delta x_{\mathrm{nt}}\|_x$$

i de punkter x där det finns en Newtonriktning.

Exempel 15.1.2. Låt oss undersöka Newtondekrementet $\lambda(f,x)$ för konvext kvadratiska polynom f, dvs. funktioner på formen

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c,$$

där A är en positivt semidefinit operator. Nu är $f'(x) = Ax + b$, $f''(x) = A$ och $\|v\|_x = \sqrt{\langle v, Av\rangle}$, så seminormen $\|\cdot\|_x$ är densamma i alla punkter x.

Om Δx_{nt} är en Newtonriktning till f i punkten x, så är per definition

$$A\Delta x_{\mathrm{nt}} = -(Ax + b),$$

och det följer att $A(x + \Delta x_{\mathrm{nt}}) = -b$. Enligt sats 15.1.1 är därför funktionen f nedåt begränsad.

Om f inte är nedåt begränsad, saknas det följaktligen Newtonriktning i varje punkt x, vilket per definition innebär att $\lambda(f,x) = +\infty$ för alla x.

Antag omvänt att funktionen f är nedåt begränsad. Då finns det en vektor v_0 sådan att $Av_0 = -b$, och det följer att

$$f''(x)(v_0 - x) = Av_0 - Ax = -b - Ax = -f'(x).$$

För varje x är med andra ord vektorn $v_0 - x$ en Newtonriktning till f i punkten x, så Newtondekrementet $\lambda(f,x)$ är ändligt i alla punkter och

$$\lambda(f,x) = \|v_0 - x\|_x.$$

Om funktionen är nedåt begränsad utan att vara konstant, så är nödvändigtvis $A \neq 0$, och vi kan välja en vektor w sådan att $\|w\|_x = \sqrt{\langle w, Aw\rangle} = 1$. Sätt nu $x_k = kw + v_0$, där k är ett positivt tal; då är

$$\lambda(f,x_k) = \|v_0 - x_k\|_{x_k} = k\|w\|_{x_k} = k,$$

och vi drar speciellt slutsatsen att $\sup_{x\in\mathbf{R}^n} \lambda(f,x) = +\infty$.

För konstanta funktioner f, fallet $A = 0$, $b = 0$, är $\|v\|_x = 0$ för alla x och v, och följakligen $\lambda(f,x) = 0$ för alla x.

Sammanfattningsvis har vi kommit fram till följande resultat:

Nedåt obegränsade konvex-kvadratiska funktioner (vilket inkluderar alla icke-konstanta affina funktioner) har oändligt Newtondekrement i alla punkter. Nedåt begränsade konvex-kvadratiska funktioner f har ändligt Newtondekrement i alla punkter, men $\sup_x \lambda(f,x) = \infty$, såvida inte funktionen är konstant. □

Vi ska ge en alternativ karakterisering av Newtondekrementet och börjar med en användbar olikhet.

Sats 15.1.2. *Om $\lambda(f,x) < \infty$ så är*

$$|\langle f'(x), v\rangle| \leq \lambda(f,x)\|v\|_x$$

för alla $v \in \mathbf{R}^n$.

Bevis. Att $\lambda(f,x)$ är ändligt betyder att det finns en Newtonriktning Δx_{nt} till f i punkten x. Per definition är alltså $f''(x)\Delta x_{\mathrm{nt}} = -f'(x)$, och det följer nu med hjälp av Cauchy–Schwarz olikhet att

$$\begin{aligned}|\langle f'(x), v\rangle| &= |\langle f''(x)\Delta x_{\mathrm{nt}}, v\rangle| = |\langle f''(x)^{1/2}\Delta x_{\mathrm{nt}}, f''(x)^{1/2}v\rangle| \\ &\leq \|f''(x)^{1/2}\Delta x_{\mathrm{nt}}\|\|f''(x)^{1/2}v\| = \lambda(f,x)\|v\|_x.\end{aligned}$$

□

Sats 15.1.3. *Antag som tidigare att x är en punkt där andraderivatan $f''(x)$ är positivt semidefinit. Då är*

$$\lambda(f,x) = \sup_{\|v\|_x \leq 1} \langle f'(x), v\rangle.$$

Bevis. Antag först att $\lambda(f,x) < \infty$. Det följer då omedelbart av olikheten i sats 15.1.2 att

$$\langle f'(x), v\rangle \leq \lambda(f,x)$$

för alla vektorer v med $\|v\|_x \leq 1$. I fallet $\lambda(f,x) = 0$ gäller olikheten ovan med likhet för $v = 0$, så antag att $\lambda(f,x) > 0$. För vektorn $v = -\lambda(f,x)^{-1}\Delta x_{\text{nt}}$ gäller då att $\|v\|_x = 1$ medan $\langle f'(x), v\rangle = -\lambda(f,x)^{-1}\langle f'(x), \Delta x_{\text{nt}}\rangle = \lambda(f,x)$. Detta visar att likheten i satsen gäller om Newtondekrementet $\lambda(f,x)$ är ett ändligt tal.

Antag härnäst att $\lambda(f,x) = +\infty$, dvs. att x är en punkt där det inte finns någon Newtonriktning. Då finns det enligt anmärkningen efter definitionen av Newtonriktning en vektor w sådan att $f''(x)w = 0$ och $\langle f'(x), w\rangle = 1$. För alla positiva tal t är då $\|tw\|_x = t\|w\|_x = t\sqrt{\langle w, f''(x)w\rangle} = 0 \leq 1$ och $\langle f'(x), tw\rangle = t$, varav följer att $\sup_{\|v\|_x \leq 1}\langle f'(x), v\rangle = +\infty = \lambda(f,x)$. □

Ibland behöver vi jämföra $\|\Delta x_{\text{nt}}\|$, $\|f'(x)\|$ och $\lambda(f,x)$ med varandra, och det gör vi med följande sats.

Sats 15.1.4. *Låt $\lambda_{\min}$ och $\lambda_{\max}$ beteckna det minsta resp. det största egenvärdet till den positivt semidefinita andraderivatan $f''(x)$, och antag att Newtondekrementet $\lambda(f,x)$ är ändligt. Då är*

$$\lambda_{\min}^{1/2}\|\Delta x_{\text{nt}}\| \leq \lambda(f,x) \leq \lambda_{\max}^{1/2}\|\Delta x_{\text{nt}}\|$$

och

$$\lambda_{\min}^{1/2}\lambda(f,x) \leq \|f'(x)\| \leq \lambda_{\max}^{1/2}\lambda(f,x).$$

Bevis. Om A är en godtycklig positivt semidefinit operator med minsta och största egenvärde $\mu_{\min}$ resp. $\mu_{\max}$, så är

$$\mu_{\min}\|v\| \leq \|Av\| \leq \mu_{\max}\|v\|$$

för alla vektorer v.

Satsens båda olikheterna följer nu direkt ur denna generella olikhet genom att välja $A = f''(x)^{1/2}$ och $v = \Delta x_{\text{nt}}$ resp. $v = f''(x)^{1/2}\Delta x_{\text{nt}}$, eftersom $\|f''(x)^{1/2}\Delta x_{\text{nt}}\| = \lambda(f,x)$,

$$\|f''(x)^{1/2}(f''(x)^{1/2}\Delta x_{\text{nt}})\| = \|f''(x)\Delta x_{\text{nt}}\| = \|f'(x)\|,$$

och $\lambda_{\min}^{1/2}$ och $\lambda_{\max}^{1/2}$ är det minsta resp. det största egenvärdet till operatorn $f''(x)^{1/2}$. □

Sats 15.1.4 är ett lokalt resultat, men om funktionen f är μ-starkt konvex, så är $\lambda_{\min} \geq \mu$, och om andraderivatans norm $\|f''(x)\|$ begränsas av någon konstant M, så är $\lambda_{\max} = \|f''(x)\| \leq M$ för alla x i funktionens definitionsmängd. Vi får därför följande korollarium till sats 15.1.4.

Korollarium 15.1.5. *Om $f\colon X \to \mathbf{R}$ är en två gånger differentierbar μ-starkt konvex funktion, så är*

$$\mu^{1/2}\|\Delta x_{\mathrm{nt}}\| \leq \lambda(f,x) \leq \mu^{-1/2}\|f'(x)\|$$

för alla $x \in X$. Om dessutom $\|f''(x)\| \leq M$, så är

$$M^{-1/2}\|f'(x)\| \leq \lambda(f,x) \leq M^{1/2}\|\Delta x_{\mathrm{nt}}\|.$$

För starkt konvexa funktioner med begränsad andraderivata kan vi uppskatta avståndet från en godtycklig punkt x till minimipunkten med hjälp av Newtondekrementet i punkten x. Vi har nämligen följande resultat.

Sats 15.1.6. *Om funktionen $f\colon X \to \mathbf{R}$ är μ-starkt konvex, $\|f''(x)\| \leq M$ för alla $x \in X$ och funktionen har minimum i punkten $\hat{x}$, så är*

$$f(x) - f(\hat{x}) \leq \frac{M}{2\mu}\lambda(f,x)^2$$

och

$$\|x - \hat{x}\| \leq \frac{\sqrt{M}}{\mu}\lambda(f,x)$$

för alla $x \in X$.

Bevis. Satsen följer av sats 14.1.1 och uppskattningen $\|f'(x)\| \leq M^{1/2}\lambda(f,x)$ i korollarium 15.1.5. □

Newtondekrementet är invariant under surjektiva affina koordinattransformationer. Mer generellt har vi följande resultat.

Sats 15.1.7. *Låt $A\colon \mathbf{R}^m \to \mathbf{R}^n$ vara en affin avbildning, låt f vara en två gånger kontinuerligt differentierbar funktion som är definierad på en delmängd av $\mathbf{R}^n$, och sätt $g = f \circ A$. Låt vidare y vara en punkt i g:s definitionsmängd, och antag att andraderivatan f'' är positivt semidefinit i punkten $x = Ay$. Då är andraderivatan $g''(y)$ positivt semidefinit och för de två funktionernas Newtondekrement gäller olikheten*

$$\lambda(g,y) \leq \lambda(f,x).$$

Om den affina avbildningen A är surjektiv, så är vidare $\lambda(g,y) = \lambda(f,x)$.

Bevis. Den affina avbildningen kan skrivas på formen $Ay = Cy + b$, där C är en linjär avbildning och b är en vektor, och kedjeregeln ger oss sambanden

$$\langle g'(y), w\rangle = \langle f'(x), Cw\rangle \quad \text{och} \quad \langle w, g''(y)w\rangle = \langle Cw, f''(x)Cw\rangle$$

för godtyckliga vektorer w i $\mathbf{R}^m$. Av den sistnämnda identiteten följer speciellt att andraderivatan $g''(y)$ är positivt semidefinit om andraderivatan $f''(x)$ är det, samt sambandet

$$\|w\|_y = \|Cw\|_x$$

för de lokala seminormerna i punkterna y och x. På grund av sats 15.1.3 är därför

$$\lambda(g, y) = \sup_{\|w\|_y \le 1} \langle g'(y), w\rangle = \sup_{\|Cw\|_x \le 1} \langle f'(x), Cw\rangle \le \sup_{\|v\|_x \le 1} \langle f'(x), v\rangle = \lambda(f, x).$$

Om den affina avbildningen är surjektiv, så är C en surjektiv linjär avbildning, varför $v = Cw$ genomlöper hela $\mathbf{R}^n$ då w genomlöper $\mathbf{R}^m$. I detta fall övergår därför den enda olikheten i ovanstående kedja av likheter och olikheter i likhet, vilket innebär att $\lambda(g, y) = \lambda(f, x)$. □

15.2 Newtons metod

Newtons metod för minimera en två gånger differentierbar funktion f är en descentmetod, där sökriktningen i varje iteration ges av Newtonriktningen Δx_{nt} i den aktuella punkten. Stoppkriteriet formuleras med hjälp av Newtondekrementet; algoritmen avbryts när dekrementet är tillräckligt litet. I korta drag ser därför algoritmen ut så här:

Newtons metod

Givet startpunkt $x \in \operatorname{dom} f$ och tolerans $\epsilon > 0$.
Upprepa

1. Beräkna Newtonriktning Δx_{nt} och Newtondekrement $\lambda(f, x)$ i punkten x.
2. *Stoppkriterium:* Avbryt om $\lambda(f, x)^2 \le 2\epsilon$.
3. Bestäm annars en steglängd $h > 0$.
4. *Uppdatera:* $x := x + h\Delta x_{\text{nt}}$.

I den *rena* Newtonmetoden sätts $h = 1$ i varje iteration, medan s. k. *dämpade* Newtonmetoder utnyttjar variabel steglängd, som beräknas genom linjesökning med (någon variant av) Armijos regel eller på annat sätt.

Stoppkriteriet motiveras av att $\frac{1}{2}\lambda(f,x)^2$ approximerar funktionsvärdesminskningen $f(x)-f(x+\Delta x_{\mathrm{nt}})$, och om denna minskningen är liten så lönar det sig inte att fortsätta.

Newtons metod fungerar i allmänhet väl för funktioner som är konvexa i en omgivning av den optimala punkten, men den bryter förstås ihop om den träffar på en punkt där andraderivatan är singulär och Newtonriktning saknas. Vi skall visa att den rena metoden under lämpliga förutsättningar på målfunktionen f konvergerar mot minimipunkten om startpunkten ligger tillräckligt nära minimipunkten. För att erhålla konvergens för godtyckliga startpunkter behöver man använda sig av metoder med dämpning.

Exempel 15.2.1. För nedåt begränsade konvexa kvadratiska polynom

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$$

ger Newtons rena metod en optimal lösning efter endast en iteration, oberoende av valet av startpunkt x. Under den första iterationen är nämligen $f'(x) = Ax + b$, $f''(x) = A$ och $A\Delta x_{\mathrm{nt}} = -(Ax + b)$, så det uppdaterade x-värdet $x^+ = x + \Delta x_{\mathrm{nt}}$ satisfierar ekvationen

$$f'(x^+) = Ax^+ + b = Ax + A\Delta x_{\mathrm{nt}} + b = 0,$$

vilket betyder att x^+ är en minimipunkt. □

Invarians under koordinatbyte

Till skillnad från brantaste lutningsmetoden är Newtonmetoden invariant under affina koordinatbyten.

Sats 15.2.1. *Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar funktion med positivt definit andraderivata, och antag att Newtons rena algoritm med x_0 som startpunkt genererar följden $(x_k)_0^\infty$. Låt vidare $A\colon Y \to X$ vara en affin koordinatavbildning, dvs. restriktionen till Y av en bijektiv affin avbildning, och sätt $g = f \circ A$. Då genererar Newtons rena algoritm tillämpad på funktionen g med $y_0 = A^{-1}x_0$ som startpunkt en följd $(y_k)_0^\infty$ med egenskapen att $Ay_k = x_k$ för alla k.*

De två följderna har vidare i varje iterationssteg samma Newtondekrement, så stoppvillkoret uppfylls samtidigt av de båda följderna.

Bevis. Påståendet om Newtondekrementen följer av sats 15.1.7. Sambandet mellan följderna följer med induktion om vi visar att $Ay = x$ medför att $A(y+\Delta y_{\mathrm{nt}}) = x+\Delta x_{\mathrm{nt}}$, där $\Delta x_{\mathrm{nt}} = -f''(x)^{-1}f'(x)$ och $\Delta y_{\mathrm{nt}} = -g''(y)^{-1}g'(y)$ är respektive funktioners entydigt bestämda Newtonriktningar i punkterna x respektive y.

Den affina avbildningen A kan skrivas på formen $Ay = Cy + b$, där C är en inverterbar linjär avbildning och b är en vektor, och kedjeregeln ger för $x = Ay$ att $g'(y) = C^T f'(x)$ och $g''(y) = C^T f''(x)C$. Det följer att

$$\begin{aligned} C\Delta y_{\text{nt}} &= -Cg''(y)^{-1}g'(y) = -CC^{-1}f''(x)^{-1}(C^T)^{-1}C^T f'(x) \\ &= -f''(x)^{-1}f'(x) = \Delta x_{\text{nt}}, \end{aligned}$$

så

$$A(y+\Delta y_{\text{nt}}) = C(y+\Delta y_{\text{nt}})+b = Cy+b+C\Delta y_{\text{nt}} = Ay+\Delta x_{\text{nt}} = x+\Delta x_{\text{nt}}. \ \square$$

Lokal konvergens

Vi skall nu studera Newtonmetodens konvergensegenskaper och börjar med den rena metoden.

Sats 15.2.2. *Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar, μ-starkt konvex funktion med minimum i punkten $\hat{x}$, och antag att funktionens andraderivata f'' är Lipschitzkontinuerlig med Lipschitzkonstant L. Låt x vara en punkt i X och sätt*

$$x^+ = x + \Delta x_{\text{nt}},$$

där Δx_{nt} är Newtonriktningen i punkten. Då är

$$\|x^+ - \hat{x}\| \le \frac{L}{2\mu}\|x - \hat{x}\|^2.$$

Om punkten x^+ ligger i X är vidare

$$\|f'(x^+)\| \le \frac{L}{2\mu^2}\|f'(x)\|^2.$$

Bevis. Det minsta egenvärdet till andraderivatan $f''(x)$ är enligt sats 7.3.2 större än eller lika med μ. Andraderivatan är följaktligen inverterbar och $f''(x)^{-1}$:s största egenvärde är mindre än eller lika med μ^{-1}. Härav följer att

$$\|f''(x)^{-1}\| \le \mu^{-1}. \tag{15.3}$$

För att uppskatta normen av $x^+ - \hat{x}$ gör vi först omskrivningen

$$\begin{aligned} x^+ - \hat{x} &= x + \Delta x_{\text{nt}} - \hat{x} = x - \hat{x} - f''(x)^{-1}f'(x) \\ &= f''(x)^{-1}\big(f''(x)(x-\hat{x}) - f'(x)\big) = -f''(x)^{-1}w \end{aligned} \tag{15.4}$$

med

$$w = f'(x) - f''(x)(x - \hat{x}).$$

För $0 \leq t \leq 1$ definierar vi sedan vektorn

$$w(t) = f'(\hat{x} + t(x - \hat{x})) - tf''(x)(x - \hat{x});$$

då är $w = w(1) - w(0)$, beroende på att $f'(\hat{x}) = 0$. Kedjeregeln ger att

$$w'(t) = \big(f''(\hat{x} + t(x - \hat{x})) - f''(x)\big)(x - \hat{x}),$$

och genom att utnytta att andraderivatan är Lipschitzkontinuerlig erhålls uppskattningen

$$\begin{aligned}\|w'(t)\| &\leq \|f''(\hat{x} + t(x - \hat{x})) - f''(x)\| \, \|x - \hat{x}\| \\ &\leq L\|\hat{x} + t(x - \hat{x}) - x\|\|x - \hat{x}\| = L(1 - t)\|x - \hat{x}\|^2.\end{aligned}$$

Integrering av olikheten ovan ger

$$\begin{aligned}\|w\| = \Big\|\int_0^1 w'(t)\,dt\Big\| &\leq \int_0^1 \|w'(t)\|\,dt \leq L\|x - \hat{x}\|^2 \int_0^1 (1 - t)\,dt \\ &= \frac{L}{2}\|x - \hat{x}\|^2.\end{aligned} \tag{15.5}$$

Likheten (15.4) tillsammans med olikheterna (15.3) och (15.5) leder nu till uppskattningen

$$\|x^+ - \hat{x}\| = \|f''(x)^{-1}w\| \leq \|f''(x)^{-1}\|\|w\| \leq \frac{L}{2\mu}\|x - \hat{x}\|^2,$$

som är det första påståendet i satsen.

För att bevisa det andra påståendet antar vi att x^+ liggger i X och betraktar för $0 \leq t \leq 1$ vektorerna

$$v(t) = f'(x + t\Delta x_{\text{nt}}) - tf''(x)\Delta x_{\text{nt}},$$

och noterar att

$$v(1) - v(0) = f'(x^+) - f''(x)\Delta x_{\text{nt}} - f'(x) = f'(x^+) + f'(x) - f'(x) = f'(x^+).$$

Eftersom $v'(t) = \big(f''(x + t\Delta x_{\text{nt}}) - f''(x)\big)\Delta x_{\text{nt}}$, följer det av Lipschitzkontinuiteten att

$$\|v'(t)\| \leq \|f''(x + t\Delta x_{\text{nt}}) - f''(x)\| \, \|\Delta x_{\text{nt}}\| \leq L\|\Delta x_{\text{nt}}\|^2 t,$$

och genom integrering av denna olikhet erhålls den sökta uppskattningen

$$\|f'(x^+)\| = \Big\|\int_0^1 v'(t)\,dt\Big\| \leq \int_0^1 \|v'(t)\|\,dt \leq \frac{L}{2}\|\Delta x_{\text{nt}}\|^2 \leq \frac{L}{2\mu^2}\|f'(x)\|^2,$$

där den sista olikheten följer av korollarium 15.1.5. □

En konsekvens av föregående sats är att den rena Newtonmetoden konvergerar kvadratiskt mot minimipunkten för funktioner med positivt definit andraderivata som inte varierar alltför snabbt i en omgivning av minimipunkten, förutsatt att startpunkten valts tillräckligt nära minimipunkten. Mer precist gäller:

Sats 15.2.3. *Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar, μ-starkt konvex funktion med minimum i punkten $\hat{x}$, och antag att funktionens andraderivata f'' är Lipschitzkontinuerlig med Lipschitzkonstant L. Antag vidare att $0 < r \leq 2\mu/L$ och att den öppna bollen $B(\hat{x}; r)$ ligger i X.*

Då genererar Newtons rena metod med startpunkt $x_0 \in B(\hat{x}; r)$ en följd av punkter $(x_k)_0^\infty$ sådan att

$$\|x_k - \hat{x}\| \leq \frac{2\mu}{L}\Big(\frac{L}{2\mu}\|x_0 - \hat{x}\|\Big)^{2^k}$$

för alla k, och följden konvergerar därför mot minimipunkten $\hat{x}$, då $k \to \infty$.

Konvergensen är således mycket snabb. Exempelvis blir

$$\|x_k - \hat{x}\| \leq \frac{2\mu}{L}\,2^{-2^k}$$

om startpunkten väljs så att $\|x_0 - \hat{x}\| \leq \mu/L$, och då är $\|x_k - \hat{x}\| \leq 10^{-19}\mu/L$ redan för $k = 6$.

Bevis. Med beteckningarna i sats 15.2.2 är $x_{k+1} = x_k^+$, och om punkten x_k ligger i bollen $B(\hat{x}; r)$ är därför enligt samma sats

$$\|x_{k+1} - \hat{x}\| \leq \frac{L}{2\mu}\|x_k - \hat{x}\|^2. \tag{15.6}$$

Speciellt är alltså $\|x_{k+1} - \hat{x}\| < Lr^2/2\mu \leq r$, dvs. punkten x_{k+1} ligger också i bollen $B(\hat{x}; r)$. Det följer därför med induktion att alla punkterna i följden $(x_k)_0^\infty$ ligger i $B(\hat{x}; r)$, och olikheten i satsens fås nu genom upprepad användning av olikheten (15.6). □

Global konvergens

Under lämpliga förutsättningar på målfunktionen konvergerar Newtons dämpade metod för en godtyckligt vald startpunkt. Dämpningen behövs bara under en inledande fas; när algoritmen producerat en punkt med tillräckligt liten gradient kan steglängden fortsättningsvis väljas lika med 1, och konvergensen sker då med kvadratisk hastighet.

Följande sats beskriver ett konvergensresultat för starkt konvexa funktioner med Lipschitzkontinuerlig andraderivata.

Sats 15.2.4. *Låt $f\colon X \to \mathbf{R}$ vara en två gånger differentierbar, starkt konvex funktion med Lipschitzkontinuerlig andraderivata. Låt vidare x_0 vara en punkt i X och antag att subnivåmängden $S = \{x \in X \mid f(x) \leq f(x_0)\}$ är sluten.*

Då har funktionen en unik minimipunkt $\hat{x}$, och Newtons dämpade algoritm, med x_0 som startpunkt och med linjesökning enligt Armijos regel med parametrar $0 < \alpha < \frac{1}{2}$ och $0 < \beta < 1$, genererar en följd $(x_k)_0^\infty$ i S som konvergerar mot minimipunkten.

Algoritmen övergår efter en inledande fas med dämpning i en kvadratiskt konvergent fas, där steglängden hela tiden är 1.

Bevis. Att f har en unik minimipunkt följer av korollarium 8.1.7.

Antag att f är μ-starkt konvex och låt L vara andraderivatans Lipschitzkonstant. Subnivåmängden S är kompakt eftersom den är begränsad enligt sats 8.1.6, så avståndet från S till randen av den öppna mängden X är positivt. Fixera ett positivt tal r som är mindre än detta avstånd och som också uppfyller olikheten

$$r \leq \mu/L.$$

För $x \in S$ sätter vi nu

$$x^+ = x + h\Delta x_{\mathrm{nt}},$$

där h är steglängden enligt Armijos regel. Speciellt är alltså $x_{k+1} = x_k^+$ för alla k.

Kärnan i beviset består i att visa att det finns två positiva konstanter γ och $\eta \leq \mu r$, sådana att följande två implikationer gäller för alla $x \in S$:

(i) $\quad \|f'(x)\| \geq \eta \;\Rightarrow\; f(x^+) - f(x) \leq -\gamma;$

(ii) $\quad \|f'(x)\| < \eta \;\Rightarrow\; h = 1 \;\&\; \|f'(x^+)\| < \eta.$

Antag att vi visat (i) och (ii). Om $\|f'(x_k)\| \geq \eta$ för $0 \leq k < m$, så är på grund av egenskapen (i)

$$f_{\min} - f(x_0) \leq f(x_m) - f(x_0) = \sum_{k=0}^{m-1} (f(x_k^+) - f(x_k)) \leq -m\gamma.$$

Detta kan inte gälla för alla m, så därför finns det ett minsta heltal k_0 så att $\|f'(x_{k_0})\| < \eta$, och detta heltal måste uppfylla olikheten

$$k_0 \leq \big(f(x_0) - f_{\min}\big)/\gamma.$$

Det följer nu med induktion av (ii) att steglängden h är lika med 1 för alla $k \geq k_0$. Från och med iteration nummer k_0 övergår således den dämpade algoritmen i den rena algoritmen. På grund av sats 14.1.1 är

$$\|x_{k_0} - \hat{x}\| \leq \mu^{-1}\|f'(x_{k_0})\| < \mu^{-1}\eta \leq r \leq \mu L^{-1},$$

så det följer av sats 15.2.3 att följden $(x_k)_0^\infty$ konvergerar mot $\hat{x}$. Mer precist gäller uppskattningen

$$\|x_{k+k_0} - \hat{x}\| \le \frac{2\mu}{L}\Big(\frac{L}{2\mu}\|x_{k_0} - \hat{x}\|\Big)^{2^k} \le \frac{2\mu}{L}\,2^{-2^k}$$

för $k \ge 0$.

Det återstår därför bara att bevisa existensen av talen η och γ med egenskaperna (i) och (ii). Sätt för den skull

$$S_r = S + \overline{B}(x; r);$$

mängden S_r är en konvex och kompakt delmängd av X, och de kontinuerliga funktionerna f' och f'' är därför begränsade på S_r, dvs. det finns konstanter K och M så att

$$\|f'(x)\| \le K \quad \text{och} \quad \|f''(x)\| \le M$$

för alla $x \in S_r$. Det följer av sats 7.4.1 att derivatan f' är Lipschitzkontinuerlig med Lipschitzkonstant M på S_r, dvs.

$$\|f'(y) - f'(x)\| \le M\|y - x\|$$

för $x, y \in S_r$.

Vi definierar nu våra tal η och γ genom att sätta

$$\eta = \min\Big\{\frac{3(1-2\alpha)\mu^2}{L}, \mu r\Big\} \text{ och } \gamma = \frac{\alpha\beta c\mu}{M}\eta^2, \text{ där } c = \min\Big\{\frac{1}{M}, \frac{r}{K}\Big\},$$

och vi ska först uppskatta steglängden h i en given punkt $x \in S$. Eftersom

$$\|\Delta x_{\mathrm{nt}}\| \le \mu^{-1}\|f'(x)\| \le \mu^{-1}K,$$

ligger punkten $x + t\Delta x_{\mathrm{nt}}$ i S_r och därmed också i X om $0 \le t \le r\mu K^{-1}$. Funktionen

$$g(t) = f(x + t\Delta x_{\mathrm{nt}})$$

är därför definierad för dessa t-värden, och eftersom funktionen f är μ-starkt konvex och derivatan är Lipschitzkontinuerlig med konstant M på S_r, följer det av sats 1.1.2 och korollarium 15.1.5 att

$$\begin{aligned} f(x + t\Delta x_{\mathrm{nt}}) &\le f(x) + t\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \tfrac{1}{2}M\|\Delta x_{\mathrm{nt}}\|^2 t^2 \\ &\le f(x) + t\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \tfrac{1}{2}M\mu^{-1}\lambda(f,x)^2 t^2 \\ &= f(x) + t\big(1 - \tfrac{1}{2}M\mu^{-1}t\big)\langle f'(x), \Delta x_{\mathrm{nt}}\rangle. \end{aligned}$$

För talet $\hat{t} = c\mu$ gäller dels att det ligger i intervallet $[0, r\mu K^{-1}]$, dels att $\hat{t} \le \mu M^{-1}$. Följaktligen är

$$1 - \tfrac{1}{2}M\mu^{-1}\hat{t} \ge \tfrac{1}{2} \ge \alpha,$$

vilket insatt i olikheten ovan ger att

$$f(x+\hat{t}\Delta x_{\mathrm{nt}}) \le f(x) + \alpha\hat{t}\,\langle f'(x), \Delta x_{\mathrm{nt}}\rangle.$$

Enligt Armijos regel är därför $\hat{t} < \beta^{m-1}$, så steglängden $h = \beta^m$ uppfyller olikheten

$$h \ge \beta\hat{t} = \beta c\mu,$$

och för punkten $x^+ = x + h\Delta x_{\mathrm{nt}}$ får vi uppskattningen

$$\begin{aligned} f(x^+) - f(x) &\le \alpha h\langle f'(x), \Delta x_{\mathrm{nt}}\rangle = -\alpha h\,\lambda(f,x)^2 \\ &\le -\alpha\beta c\mu\,\lambda(f,x)^2 \le -\alpha\beta c\mu M^{-1}\|f'(x)\|^2 = -\gamma\eta^{-2}\|f'(x)\|^2. \end{aligned}$$

Detta medför att implikationen (i) gäller.

För att visa den återstående implikationen (ii) återvänder vi till funktionen $g(t) = f(x + t\Delta x_{\mathrm{nt}})$ och antar att $\|f'(x)\| < \eta$. Eftersom

$$\|\Delta x_{\mathrm{nt}}\| \le \mu^{-1}\|f'(x)\| < \mu^{-1}\eta \le r,$$

är funktionen g säkert definierad för $0 \le t \le 1$. Vidare är

$$g'(t) = \langle f'(x + t\Delta x_{\mathrm{nt}}), \Delta x_{\mathrm{nt}}\rangle \text{ och } g''(t) = \langle \Delta x_{\mathrm{nt}}, f''(x + t\Delta x_{\mathrm{nt}})\Delta x_{\mathrm{nt}}\rangle.$$

På grund av Lipschitzkontinuitetsantagandet är

$$\begin{aligned} |g''(t) - g''(0)| &= |\langle \Delta x_{\mathrm{nt}}, f''(x + t\Delta x_{\mathrm{nt}})\Delta x_{\mathrm{nt}}\rangle - \langle \Delta x_{\mathrm{nt}}, f''(x)\Delta x_{\mathrm{nt}}\rangle| \\ &\le \|f''(x + t\Delta x_{\mathrm{nt}}) - f''(x)\|\,\|\Delta x_{\mathrm{nt}}\|^2 \le tL\|\Delta x_{\mathrm{nt}}\|^3, \end{aligned}$$

så det följer, eftersom $g''(0) = \lambda(f,x)^2$ och $\|\Delta x_{\mathrm{nt}}\| \le \mu^{-1/2}\lambda(f,x)$, att

$$g''(t) \le \lambda(f,x)^2 + tL\|\Delta x_{\mathrm{nt}}\|^3 \le \lambda(f,x)^2 + tL\mu^{-3/2}\,\lambda(f,x)^3.$$

Genom att integrera denna olikhet över intervallet $[0,t]$ erhålles olikheten

$$g'(t) - g'(0) \le t\lambda(f,x)^2 + \tfrac{1}{2}t^2L\mu^{-3/2}\lambda(f,x)^3.$$

Nu är $g'(0) = \langle f'(x), \Delta x_{\mathrm{nt}}\rangle = -\lambda(f,x)^2$, så det följer att

$$g'(t) \le -\lambda(f,x)^2 + t\lambda(f,x)^2 + \tfrac{1}{2}t^2L\mu^{-3/2}\lambda(f,x)^3,$$

och ytterligare en integration leder till olikheten

$$g(t) - g(0) \le -t\lambda(f,x)^2 + \tfrac{1}{2}t^2\lambda(f,x)^2 + \tfrac{1}{6}t^3L\mu^{-3/2}\lambda(f,x)^3.$$

Insättning av $t = 1$ ger slutligen

$$\begin{aligned} f(x+\Delta x_{\mathrm{nt}}) &\le f(x) - \tfrac{1}{2}\lambda(f,x)^2 + \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)^3 \\ &= f(x) - \lambda(f,x)^2\big(\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)\big) \\ &= f(x) + \langle f'(x), \Delta x_{\mathrm{nt}}\rangle\big(\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x)\big). \end{aligned} \tag{15.7}$$

Av antagandet $\|f'(x)\| < \eta$ följer att

$$\lambda(f,x) \le \mu^{-1/2}\|f'(x)\| < \mu^{-1/2}\eta \le \mu^{-1/2}\cdot 3(1-2\alpha)\mu^2 L^{-1} = 3(1-2\alpha)\mu^{3/2}L^{-1}.$$

Följaktligen är

$$\tfrac{1}{2} - \tfrac{1}{6}L\mu^{-3/2}\lambda(f,x) > \alpha,$$

och insättning av detta i olikheten (15.7) resulterar i olikheten

$$f(x+\Delta x_{\mathrm{nt}}) \le f(x) + \alpha\langle f'(x), \Delta x_{\mathrm{nt}}\rangle,$$

som visar att steglängden h är lika med 1.

Iterationssteget från x till $x^+ = x + h\Delta x_{\mathrm{nt}}$ sker således enligt den rena Newtonmetoden. På grund av olikheten

$$\|x - \hat{x}\| \le \mu^{-1}\|f'(x)\| < \mu^{-1}\eta \le r,$$

som gäller enligt sats 14.1.1, ligger vidare punkten x i bollen $B(\hat{x}; r)$, så det följer av den lokala konvergenssatsen 15.2.2 att

$$\|f'(x^+)\| \le \frac{L}{2\mu^2}\|f'(x)\|^2, \tag{15.8}$$

och eftersom $\eta \le \mu r \le \mu^2/L$, är därför

$$\|f'(x^+)\| < \frac{L}{2\mu^2}\eta^2 \le \frac{\eta}{2} < \eta.$$

Därmed är beviset fullständigt. □

Genom att iterera olikheten (15.8) erhåller man i själva verket uppskattningen

$$\|f'(x_k)\| \le \frac{2\mu^2}{L}\Big(\frac{L}{2\mu^2}\|f'(x_{k_0})\|\Big)^{2^{k-k_0}} < \frac{2\mu^2}{L}2^{-2^{k-k_0}}$$

för $k \ge k_0$, och det följer därför av sats 14.1.1 att

$$f(x_k) - f_{\min} < \frac{2\mu^3}{L^2}2^{-2^{k-k_0+1}}$$

för $k \ge k_0$. Genom att kombinera denna uppskattning med den tidigare erhållna begränsningen på k_0 erhålles en övre gräns på antalet iterationer som krävs för att uppskatta minimivärdet $f_{\min}$ med en given precision − om

$$k > \frac{f(x_0) - f_{\min}}{\gamma} + \log_2\log_2\frac{2\mu^3}{L^2\epsilon},$$

så är säkert $f(x_k) - f_{\min} < \epsilon$. Uppskattningen har dock inte något praktiskt värde, ty de ingående konstanterna γ, μ och L är aldrig kända i konkreta fall.

En annan skönhetsfläck är att konstanterna, till skillnad från själva algoritmen, är koordinatberoende. För självkonkordanta funktioner är det emellertid möjligt att genomföra konvergensanalysen i Newtons algoritm utan okända konstanter, och vi ska göra detta i kapitel 16.5.

15.3 Bivillkor i form av likheter

Det behövs bara små modifikationer i Newtons algoritm för att algoritmen också skall fungera för konvexa optimeringsproblem med bivillkor i form av affina likheter.

Betrakta ett konvext problem av typen

$$\text{(P)} \qquad \begin{aligned} &\min && f(x) \\ &\text{då} && Ax = b \end{aligned}$$

där $f\colon \Omega \to \mathbf{R}$ är en två gånger kontinuerligt diffferentierbar konvex funktion, $\Omega \subseteq \mathbf{R}^n$, och A är en $m \times n$-matris.

Problemets Lagrangefunktion är

$$L(x, y) = f(x) + (Ax - b)^T y = f(x) + x^T A^T y - b^T y,$$

så på grund av Karush–Kuhn–Tuckers sats (sats 11.2.1) är $\hat{x} \in \Omega$ en optimal lösning om och endast om det finns en vektor $\hat{y} \in \mathbf{R}^m$ så att

$$\begin{cases} f'(\hat{x}) + A^T\hat{y} = 0 \\ \quad A\hat{x} \qquad\quad = b. \end{cases} \tag{15.9}$$

Minimeringsproblemet (P) är därför ekvivalent med problemet att lösa ekvationssystemet (15.9).

Exempel 15.3.1. För konvexa kvadratiska funktioner

$$f(x) = \tfrac{1}{2}\langle x, Px\rangle + \langle q, x\rangle + r$$

får ekvationssystemet (15.9) formen

$$\begin{cases} P\hat{x} + A^T\hat{y} = -q \\ A\hat{x} \qquad\quad = \quad b, \end{cases}$$

och detta är ett kvadratiskt linjärt ekvationssystem med symmetrisk koefficientmatris av ordning $m + n$. Ekvationssystemet har entydig lösning om $\operatorname{rang} A = m$ och $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$. Se övning 15.4. Speciellt har det alltså entydig lösning om P är positivt definit och $\operatorname{rang} A = m$. □

Vi återvänder nu till det generella konvexa minimeringsproblemet (P) och låter X beteckna mängden av tillåtna punkter, dvs.

$$X = \{x \in \Omega \mid Ax = b\}.$$

För optimeringsproblem utan bivillkor är descentriktningen Δx_{nt} i punkten x en vektor som minimerar Taylorpolynomet av grad två till funktionen $f(x+v)$. Minimeringen sker över alla vektorer v i $\mathbf{R}^n$, och som ny punkt x^+ med mindre funktionsvärde än $f(x)$ väljs $x^+ = x + h\Delta x_{\text{nt}}$ för lämplig steglängd h. I problem med bivillkor måste den nya punkten x^+ förstås också vara en tillåten punkt, och för detta krävs att $A\Delta x_{\text{nt}} = 0$. Detta betyder att minimeringen av Taylorpolynomet bara kan ske över vektorer v som satisfierar villkoret $Av = 0$.

För optimeringsproblem med bivillkor i form av likheter leds vi därför till följande modifierade version av den tidigare definitionen av Newtonriktning och Newtondekrement.

Definition. I konvexa optimeringsproblem (P) med bivillkor kallas en vektor Δx_{nt} för en *Newtonriktning* till målfunktionen f i punkten $x \in X$ om det finns en vektor $w \in \mathbf{R}^m$ så att

$$\begin{cases} f''(x)\Delta x_{\text{nt}} + A^T w = -f'(x) \\ A\Delta x_{\text{nt}} = 0. \end{cases} \tag{15.10}$$

Kvantiteten

$$\lambda(f,x) = \sqrt{\langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}} \rangle}$$

kallas *Newtondekrementet*.

Det följer av exempel 15.3.1 att Newtonriktningen Δx_{nt} (om den existerar) är en optimal lösning till minimeringsproblemet

$$\begin{array}{ll} \min & f(x) + \langle f'(x), v \rangle + \frac{1}{2}\langle v, f''(x)v \rangle \\ \text{då} & Av = 0. \end{array}$$

För lösningar $(\Delta x_{\text{nt}}, w)$ till systemet (15.10) är vidare

$$\begin{aligned} -\langle f'(x), \Delta x_{\text{nt}} \rangle &= \langle f''(x)\Delta x_{\text{nt}} + A^T w, \Delta x_{\text{nt}} \rangle \\ &= \langle f''(x)\Delta x_{\text{nt}}, \Delta x_{\text{nt}} \rangle + \langle w, A\Delta x_{\text{nt}} \rangle \\ &= \langle f''(x)\Delta x_{\text{nt}}, \Delta x_{\text{nt}} \rangle + \langle w, 0 \rangle = \langle \Delta x_{\text{nt}}, f''(x)\Delta x_{\text{nt}} \rangle, \end{aligned}$$

så det följer att

$$\lambda(f,x) = \sqrt{-\langle f'(x), \Delta x_{\text{nt}} \rangle},$$

precis som i fallet utan bivillkor.

Att Δx_{nt} är en descentriktning följer av räkningen

$$\frac{d}{dt} f(x + t\Delta x_{\text{nt}})\big|_{t=0} = \langle f'(x), \Delta x_{\text{nt}} \rangle = -\lambda(f, x)^2 \leq 0,$$

som visar att målfunktionen f avtar i Newtonriktningen.

Om $P(v)$ är Taylorpolynomet av grad två till funktionen $f(x + v)$, så är vidare

$$\begin{aligned} f(x) - f(x + \Delta x_{\text{nt}}) &\approx P(0) - P(\Delta x_{\text{nt}}) \\ &= -\langle f'(x), \Delta x_{\text{nt}} \rangle - \tfrac{1}{2} \langle \Delta x_{\text{nt}}, f''(x) \Delta x_{\text{nt}} \rangle = \tfrac{1}{2} \lambda(f, x)^2, \end{aligned}$$

också det som i det bivillkorslösa fallet.

Med vår modifierade definition av Newtonriktning kan vi nu kopiera Newtons metod ordagrant för konvexa minimeringsproblem av typen

$$\begin{array}{ll} \min & f(x) \\ \text{då} & Ax = b. \end{array}$$

Algoritmen ser ut så här:

Newtons metod

Givet startpunkt $x \in \Omega$ som satisfierar bivillkoret $Ax = b$, och tolerans $\epsilon > 0$

Upprepa

1. Beräkna Newtonriktningen Δx_{nt} i punkten x genom att lösa ekvationssystemet (15.10) och beräkna Newtondekrementet $\lambda(f, x)$.
2. *Stoppkriterium:* Avbryt om $\lambda(f, x)^2 \leq 2\epsilon$.
3. Bestäm annars en steglängd $h > 0$.
4. *Uppdatera:* $x := x + h\Delta x_{\text{nt}}$.

Elimination av bivillkoren

Ett alternativt sätt att angripa optimeringsproblemet

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & Ax = b, \end{array}$$

med $x \in \Omega$ som implicit villkor och $r = \operatorname{rang} A$, är att lösa ekvationssystemet $Ax = b$ och uttrycka r variabler som linjärkombinationer av de övriga $n - r$ variablerna. De förstnämnda variablerna kan sedan elimineras ur målfunktionen, och man erhåller på så sätt ett optimeringsproblem i $n - r$ variabler utan explicita bivillkor som kan attackeras med Newtons metod.

Vi skall beskriva detta angreppssätt litet mer detaljerat och jämföra med metoden ovan.

Antag att mängden X av tillåtna punkter inte är tom, välj en punkt $a \in X$ och fixera en affin parametrisering

$$x = \xi(z), \quad z \in \tilde{\Omega}$$

av X med $\xi(0) = a$. Eftersom $\{x \in \mathbf{R}^n \mid Ax = b\} = a + \mathcal{N}(A)$, kan parametriseringen skrivas på formen

$$\xi(z) = a + Cz$$

där $C\colon \mathbf{R}^p \to \mathbf{R}^n$ är en injektiv linjär avbildning, vars värdemängd $\mathcal{V}(C)$ sammanfaller med avbildningen A:s nollmängd $\mathcal{N}(A)$, och $p = n - \operatorname{rang} A$. Definitionsmängden $\tilde{\Omega} = \{z \in \mathbf{R}^p \mid a + Cz \in \Omega\}$ är en öppen, konvex delmängd av $\mathbf{R}^p$.

Ett praktiskt sätt att konstruera parametriseringen är förstås att lösa ekvationssystemet $Ax = b$ med Gausselimination.

Definiera slutligen funktionen $\tilde{f}\colon \tilde{\Omega} \to \mathbf{R}$ genom att sätta $\tilde{f}(z) = f(\xi(z))$. Då är problemet (P) är ekvivalent med det konvexa optimeringsproblemet

$$(\tilde{\mathrm{P}}) \qquad \min \tilde{f}(z)$$

som saknar explicita bivillkor.

Låt nu Δx_{nt} vara en Newtonriktning till funktionen f i punkten x, dvs. en vektor som satisfierar systemet (15.10) för lämpligt vald vektor w. Vi skall visa att Δx_{nt} motsvaras av en Newtonriktning Δz_{nt} till funktionen $\tilde{f}$ i punkten $z = \xi^{-1}(x)$ och att $\Delta x_{\mathrm{nt}} = C\Delta z_{\mathrm{nt}}$.

Eftersom $A\Delta x_{\mathrm{nt}} = 0$ och $\mathcal{N}(A) = \mathcal{V}(C)$, finns det en unik vektor v så att $\Delta x_{\mathrm{nt}} = Cv$. Enligt kedjeregeln är $\tilde{f}'(z) = C^T f'(x)$ och $\tilde{f}''(z) = C^T f''(x)C$, så det följer av den första ekvationen i systemet (15.10) att

$$\begin{aligned}\tilde{f}''(z)v = C^T f''(x)Cv = C^T f''(x)\Delta x_{\mathrm{nt}} &= -C^T f'(x) - C^T A^T w \\ &= -\tilde{f}'(z) - C^T A^T w.\end{aligned}$$

Om S är en godtycklig linjär avbildning, så är $\mathcal{N}(S) = \mathcal{V}(S^T)^\perp$, och genom att tillämpa detta resultat från den linjära algebran på avbildningarna C^T och A, samt utnyttja att $\mathcal{V}(C) = \mathcal{N}(A)$, erhåller vi likheterna

$$\mathcal{N}(C^T) = \mathcal{V}(C)^\perp = \mathcal{N}(A)^\perp = \mathcal{V}(A^T)^{\perp\perp} = \mathcal{V}(A^T),$$

som medför att $C^T A^T w = 0$. Följaktligen är

$$\tilde{f}''(z)v = -\tilde{f}'(z),$$

och vektorn v är således en Newtonriktning till funktionen $\tilde{f}$ i punkten z. Vektorn $\Delta z_{\text{nt}} = v$ är således den eftersökta Newtonriktningen.

Iterationssteget $z \to z^+ = z + h\Delta z_{\text{nt}}$ i Newtons metod för problemet ($\tilde{\text{P}}$) leder från punkten $z = \xi^{-1}(x)$ i $\tilde{\Omega}$ till en punkt z^+ vars bild i X är

$$\begin{aligned}\xi(z^+) &= \xi(z + h\Delta z_{\text{nt}}) = a + C(z + h\Delta z_{\text{nt}}) = a + Cz + hC(\Delta z_{\text{nt}}) \\ &= \xi(z) + h\Delta x_{\text{nt}} = x + h\Delta x_{\text{nt}},\end{aligned}$$

vilket är precis den punkt som erhålles när Newtons metod används direkt på problemet (P) med bivillkor.

Vi noterar också att Newtondekrementen är lika i motsvarande punkter eftersom

$$\begin{aligned}\lambda(\tilde{f}, z)^2 &= -\langle \tilde{f}'(z), \Delta z_{\text{nt}} \rangle = -\langle C^T f'(x), \Delta z_{\text{nt}} \rangle = -\langle f'(x), C\Delta z_{\text{nt}} \rangle \\ &= -\langle f'(x), \Delta x_{\text{nt}} \rangle = \lambda(f, x)^2.\end{aligned}$$

Sammanfattningsvis har vi alltså visat följande resultat.

Sats 15.3.1. *Låt $(x_k)_0^\infty$ vara en följd av punkter som erhålls när Newtons metod används på problemet* (P) *med bivillkor. Newtons metod tillämpad på det problem* ($\tilde{\text{P}}$) *som erhålls genom att eliminera bivillkoren och med $\xi^{-1}(x_0)$ som startpunkt, genererar då en följd $(z_k)_0^\infty$ med egenskapen att $x_k = \xi(z_k)$ för alla k.*

Konvergensanalys

Det krävs ingen ny konvergensanalys för den modifierade varianten av Newtons metod, utan vi kan på grund av sats 15.3.1 direkt tillämpa resultaten i sats 15.2.4. Om restriktionen av funktionen $f\colon \Omega \to \mathbf{R}$ till mängden X av tillåtna punkter är starkt konvex och andraderivatan är Lipschitzkontinuerlig, så är nämligen också funktionen $\tilde{f}\colon \tilde{\Omega} \to \mathbf{R}$ starkt konvex och andraderivatan $\tilde{f}''$ Lipschitzkontinuerlig. (Se övning 15.5.) Förutsatt att x_0 är en tillåten startpunkt och subnivåmängden $\{x \in X \mid f(x) \le f(x_0)\}$ är sluten, kommer därför Newtons dämpade algoritm tillämpad på problemet (P) att konvergera mot problemets minimipunkt. Tillräckligt nära minimipunkten kommer vidare steglängden h vara lika med 1 och konvergensen vara kvadratisk.

Övningar

15.1 Bestäm Newtonriktningen, Newtondekrementet och den lokala normen i en godtycklig punkt $x > 0$ för funktionen $f(x) = x \ln x - x$.

15.2 Låt f vara funktionen $f(x_1, x_2) = -\ln x_1 - \ln x_2 - \ln(4 - x_1 - x_2)$ med $X = \{x \in \mathbf{R}^2 \mid x_1 > 0, x_2 > 0, x_1 + x_2 < 4\}$ som definitionsmängd. Bestäm Newtonriktningen, Newtondekrementet och den lokala normen i punkten x för

a) $x = (1, 1)$ b) $x = (1, 2)$.

15.3 Bestäm en Newtonriktning, Newtondekrementet och den lokala normen till funktionen $f(x_1, x_2) = \mathrm{e}^{x_1+x_2} + x_1 + x_2$ i en godtycklig punkt $x \in \mathbf{R}^2$.

15.4 Antag att P är en symmetrisk positivt semidefinit $n \times n$-matris och att A är en godtycklig $m \times n$-matris. Visa att matrisen

$$M = \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}$$

är inverterbar om och endast om $\operatorname{rang} A = m$ och $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$.

15.5 Antag att funktionen $f\colon \Omega \to \mathbf{R}$ är två gånger differentierbar och konvex, låt $x = \xi(z) = a + Cz$ vara en affin parametrisering av mängden

$$X = \{x \in \Omega \mid Ax = b\},$$

och definiera funktionen $\tilde{f}$ genom att sätta $\tilde{f}(z) = f(\xi(z))$, precis som i avsnitt 15.3. Låt vidare σ beteckna det minsta egenvärdet till den symmetriska avbildningen C^TC.

a) Visa att funktionen $\tilde{f}$ är $\mu\sigma$-starkt konvex om restriktionen av f till X är μ-starkt konvex.

b) Antag att matrisen A har full rang och att det finns konstanter K och M så att $Ax = b$ medför att

$$\left\| \begin{bmatrix} f''(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\| \leq K \quad \text{och} \quad \|f''(x)\| \leq M.$$

Visa att i så fall är funktionen $\tilde{f}$ μ-starkt konvex för $\mu = \sigma K^{-2} M^{-1}$.

Kapitel 16

Självkonkordanta funktioner

Självkonkordanta funktioner introducerades av Nesterov och Nemirovski i slutet av 1980-talet som en produkt av deras analys av konvergenshastigheten i Newtons metod. De klassiska konvergensresultaten för två gånger kontinuerligt deriverbara funktioner förutsätter att andraderivatan är Lipschitzkontinuerlig, och de i bevisen erhållna resultaten för konvergenshastigheten beror bl.a. av Lipschitzkonstanten. En uppenbar svaghet hos resultaten är att Lipschitzkonstantens värde, till skillnad från Newtons metod, inte är invariant under affina koordinattransformationer.

Att en funktion $f\colon X \to \mathbf{R}$, där X är en öppen konvex delmängd av $\mathbf{R}^n$, har en Lipschitzkontinuerlig andraderivata med Lipschitzkonstant L betyder att

$$\|f''(y) - f''(x)\| \leq L\|y - x\|$$

för alla $x, y \in X$, och för restriktionen $\phi_{x,v}(t) = f(x + tv)$ av funktionen f till en linje genom x med riktningsvektor v innebär detta att

$$|\phi''_{x,v}(t) - \phi''_{x,v}(0)| = |\langle v, (f''(x+tv) - f''(x))v\rangle| \leq L\|x+tv-x\|\|v\|^2 = L|t|\|v\|^3.$$

Om funktionen f är tre gånger differentierbar, är följaktligen

$$|\phi'''_{x,v}(0)| = \lim_{t\to 0}\Big|\frac{\phi''_{x,v}(t) - \phi''_{x,v}(0)}{t}\Big| \leq L\|v\|^3.$$

Men

$$\phi'''_{x,v}(0) = \sum_{i,j,k=1}^{n} \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} v_i v_j v_k = D^3 f(x)[v, v, v],$$

så ett nödvändigt villkor för att en tre gånger differentierbar funktion f skall ha en Lipschitzkontinuerlig andraderivata med Lipschitzkonstant L är att

$$|D^3 f(x)[v, v, v]| \leq L\|v\|^3 \tag{16.1}$$

för alla $x \in X$ och alla $v \in \mathbf{R}^n$, och man kan lätt visa att detta också är ett tillräckligt villkor.

Att Lipschitzkonstantens värde inte är affint invariant beror på att det inte finns någon naturlig koppling mellan den euklidiska normen $\|\cdot\|$ och funktionen f. Analysen av en funktions uppförande förenklas om man istället använder en norm som är anpassad till nivåytornas form, och för funktioner med positivt semidefinit andraderivata $f''(x)$ ges en sådan (semi)norm av den lokala seminormen $\|\cdot\|_x$ som vi introducerade i det förra kapitlet och som definieras av att $\|v\|_x = \sqrt{\langle v, f''(x)v\rangle}$. Nesterov–Nemirovskis genidrag var att i olikheten (16.1) byta ut $\|v\|$ mot just $\|v\|_x$, och för den klass av konvexa funktioner som erhölls på detta sätt kunde de beskriva konvergenshastigheten i Newtons metod på ett affint oberoende sätt och med hjälp av absoluta konstanter.

16.1 Självkonkordanta funktioner

Vi är nu redo för Nesterov–Nemirovskis definition av självkonkordanta funktioner och att studera dessa funktioners grundläggande egenskaper.

Definition. Låt $f\colon X \to \mathbf{R}$ vara en tre gånger kontinuerligt differentierbar funktion, vars definitionsmängd X är en öppen konvex delmängd av $\mathbf{R}^n$. Funktionen kallas *självkonkordant* om den är konvex, och olikheten

$$\big|D^3 f(x)[v,v,v]\big| \le 2\big(D^2 f(x)[v,v]\big)^{3/2} \tag{16.2}$$

gäller för alla $x \in X$ och alla $v \in \mathbf{R}^n$.

Eftersom $D^2 f(x)[v,v] = \|v\|_x^2$, där $\|\cdot\|_x$ är den av funktionen f definierade lokala seminormen i punkten x, kan den definierande olikheten (16.2) också skrivas på formen

$$\big|D^3 f(x)[v,v,v]\big| \le 2\|v\|_x^3,$$

och det är denna kortare variant som vi kommer att föredra, när vi arbetar med en enda funktion f.

Anmärkning 1. Det är inget speciellt med konstanten 2 i olikheten (16.2). Om funktionen f istället satisfierar olikheten med konstanten 2 bytt mot K, så är den genom skalning erhållna funktionen $F = \frac{1}{4}K^2 f$ självkonkordant. Valet av 2 som konstant underlättar emellertid formuleringarna av ett antal resultat.

Anmärkning 2. För funktioner f definierade på delmängder av reella axeln är $D^2 f(x)[v,v] = f''(x)v^2$ och $D^3 f(x)[v,v,v] = f'''(x)v^3$ för alla $v \in \mathbf{R}$, så en

konvex funktion $f\colon X \to \mathbf{R}$ av en variabel är självkonkordant om och endast om

$$|f'''(x)| \le 2f''(x)^{3/2}$$

för alla $x \in X$.

Anmärkning 3. Uttryckt med hjälp av restriktionen $\phi_{x,v}$ av funktionen f till den räta linjen genom x med riktningen v kan olikheten

$$\big|D^3 f(x+tv)[v,v,v]\big| \le 2\big(D^2 f(x+tv)[v,v]\big)^{3/2}$$

ekvivalent skrivas $|\phi'''_{x,v}(t)| \le 2\phi''_{x,v}(t)^{3/2}$. En tre gånger kontinuerligt deriverbar konvex funktion av flera variabler är därför självkonkordant om och endast om alla dess restriktioner till linjer är självkonkordanta.

Exempel 16.1.1. Den konvexa funktionen $f(x) = -\ln x$, $x > 0$, är självkonkordant, ty olikheten (16.2) gäller med likhet eftersom $f''(x) = x^{-2}$ och $f'''(x) = -2x^{-3}$. □

Exempel 16.1.2. Konvexa kvadratiska funktioner $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$ är självkonkordanta eftersom $D^3 f(x)[v,v,v] = 0$ för alla x och v.

Speciellt är alltså affina funktioner och funktionen $x \mapsto \|x\|^2$, där $\|\cdot\|$ är den euklidiska normen, självkonkordanta. □

Uttrycket

$$D^3 f(x)[u,v,w] = \sum_{i,k,k=1}^{n} \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} u_i v_j w_k$$

är en symmetrisk trilinjär form i variablerna u, v, w om funktionen f är tre gånger kontinuerligt differentierbar i en omgivning av punkten x. För självkonkordanta funktioner gäller följande generalisering av olikheten (16.2) i definitionen av självkonkordans.

Sats 16.1.1. *Antag att $f\colon X \to \mathbf{R}$ är en självkonkordant funktion. För alla $x \in X$ och alla vektorer u, v, w i $\mathbf{R}^n$ är*

$$\big|D^3 f(x)[u,v,w]\big| \le 2\|u\|_x \|v\|_x \|w\|_x.$$

Bevis. Beviset bygger på en generell sats om normen av symmetriska trilinjära former, som bevisas i ett appendix till det här kapitlet.

Antag först att andraderivatan $f''(x)$ är positivt definit. Då är $\|\cdot\|_x$ en norm med $\langle u, v\rangle_x = \langle u, f''(x)v\rangle$ som motsvarande skalärprodukt. Vi kan därför tillämpa appendixets sats 1 på den trilinjära formen $D^3 f(x)[u,v,w]$

med $\|\cdot\|_x$ som underliggande norm, och det följer att

$$\sup_{u,v,w\neq 0} \frac{\left|D^3 f(x)[u,v,w]\right|}{\|u\|_x\|v\|_x\|w\|_x} = \sup_{v\neq 0} \frac{\left|D^3 f(x)[v,v,v]\right|}{\|v\|_x^3} \leq 2,$$

vilket är ekvivalent med påståendet i satsen.

För att också klara av punkter x där andraderivatan $f''(x)$ är singulär betraktar vi för $\epsilon > 0$ skalärprodukten

$$\langle u, v\rangle_{x,\epsilon} = \langle u, f''(x)v\rangle + \epsilon\langle u, v\rangle,$$

där $\langle\cdot\,,\cdot\rangle$ är standardskalärprodukten, och motsvarande norm

$$\|v\|_{x,\epsilon} = \sqrt{\langle v, v\rangle_{x,\epsilon}} = \sqrt{\|v\|_x^2 + \epsilon\|v\|^2}.$$

Uppenbarligen är $\|v\|_x \leq \|v\|_{x,\epsilon}$ för alla vektorer v, och eftersom funktionen f är självkonkordant är därför också $|D^3 f(x)[v,v,v]| \leq 2\|v\|_{x,\epsilon}^3$ för alla v. Det följer därför av sats 1 i appendixet att

$$\begin{aligned}|D^3 f(x)[u,v,w]| &\leq 2\|u\|_{x,\epsilon}\|v\|_{x,\epsilon}\|w\|_{x,\epsilon} \\ &= 2\sqrt{(\|u\|_x^2 + \epsilon\|u\|^2)(\|v\|_x^2 + \epsilon\|v\|^2)(\|w\|_x^2 + \epsilon\|u\|w^2)},\end{aligned}$$

och vi erhåller nu olikheten i satsen genom att låta $\epsilon \to 0$. □

Sats 16.1.2. *Om funktionen $f\colon X \to \mathbf{R}$ är självkonkordant, så har andraderivatan $f''(x)$ samma nollrum $\mathcal{N}(f''(x))$ i alla punkter $x \in X$*

Bevis. Vi påminner om att $\mathcal{N}(f''(x)) = \{v \mid \|v\|_x = 0\}$.

Låt x och y vara två punkter i X. Av symmetriskäl räcker det förstås att visa inklusionen $\mathcal{N}(f''(x)) \subseteq \mathcal{N}(f''(y))$.

Antag därför att $v \in \mathcal{N}(f''(x))$ och sätt $x^t = x + t(y - x)$. Eftersom X är en öppen konvex mängd, finns det säkert ett tal $a > 1$ så att punkten x^t ligger i X för $0 \leq t \leq a$, och vi definierar nu en funktion $g\colon [0, a] \to \mathbf{R}$ genom att sätta

$$g(t) = D^2 f(x^t)[v,v] = \|v\|_{x^t}^2.$$

Då är $g(0) = \|v\|_x^2 = 0$ och $g(t) \geq 0$ för $0 \leq t \leq a$, och eftersom

$$g'(t) = D^3 f(x^t)[v,v,y-x],$$

följer det av sats 16.1.1 att

$$|g'(t)| \leq 2\|v\|_{x^t}^2\|y - x\|_{x^t} = 2g(t)\|y - x\|_{x^t}.$$

Men seminormen

$$\|y - x\|_{x^t} = \sqrt{D^2 f(x^t)[y-x, y-x]}$$

beror kontinuerligt av t och är därför uppåt begränsad av någon konstant C på intervallet $[0, a]$. Följaktligen är

$$|g'(t)| \leq 2Cg(t)$$

för $0 \leq t \leq a$, och det följer därför av sats 2 i detta kapitels appendix att $g(t) = 0$ för alla t; speciellt är alltså $g(1) = \|v\|_y^2 = 0$, vilket visar att $v \in \mathcal{N}(f''(y))$. Därmed är inklusionen $\mathcal{N}(f''(x)) \subseteq \mathcal{N}(f''(y))$ bevisad □

Eftersom $f''(x)$ är icke-singulär om och endast om $\mathcal{N}(f''(x)) = \{0\}$, är följande korollarium bara ett specialfall av sats 16.1.2.

Korollarium 16.1.3. *Andraderivatan till en självkonkordant funktion är antingen icke-singulär i alla punkter eller singulär i alla punkter.*

Vi kommer att kalla en självkonkordant funktion *icke-degenererad* om dess andraderivata är positivt definit i alla punkter, och enligt ovanstående korollarium är så fallet om andraderivatan är positivt definit i en enda punkt.

En icke-degenererad självkonkordant funktion är alltså speciellt strikt konvex.

Självkonkordansbevarande operationer

Sats 16.1.4. *Om funktionen f är självkonkordant och $\alpha \geq 1$, så är funktionen αf också självkonkordant.*

Bevis. För $\alpha \geq 1$ är $\alpha \leq \alpha^{3/2}$. Följaktligen är

$$\begin{aligned}\big|D^3(\alpha f)(x)[v,v,v]\big| &= \alpha\big|D^3 f(x)[v,v,v]\big| \leq 2\alpha\big(D^2 f(x)[v,v]\big)^{3/2} \\ &\leq 2\big(\alpha D^2 f(x)[v,v]\big)^{3/2} = 2\big(D^2(\alpha f)(x)[v,v]\big)^{3/2}. \quad \square\end{aligned}$$

Sats 16.1.5. *Om funktionerna f och g är självkonkordanta, så är summan $f + g$ självkonkordant på sin definitionsmängd* $\operatorname{dom} f \cap \operatorname{dom} g$.

Bevis. Genom att utnyttja den elementära olikheten

$$a^{3/2} + b^{3/2} \leq (a+b)^{3/2},$$

som gäller för alla icke-negativa tal a, b (och som enkelt visas genom kvadrering) och triangelolikheten, erhålles

$$\begin{aligned}\big|D^3(f+g)(x)[v,v,v]\big| &= \big|D^3 f(x)[v,v,v] + D^3 g(x)[v,v,v]\big| \\ &\leq 2\big(D^2 f(x)[v,v]\big)^{3/2} + 2\big(D^2 g(x)[v,v]\big)^{3/2} \\ &\leq 2\big(D^2 f(x)[v,v] + D^2 g(x)[v,v]\big)^{3/2} \\ &= 2\big(D^2(f+g)(x)[v,v]\big)^{3/2},\end{aligned}$$

vilket är den sökta olikheten. □

Sats 16.1.6. *Om funktionen $f\colon X \to \mathbf{R}$ är självkonkordant, där X är en öppen konvex delmängd av $\mathbf{R}^n$, och A är en affin avbildning från $\mathbf{R}^m$ till $\mathbf{R}^n$, så är den sammansatta funktionen $g = f \circ A$ självkonkordant på sin definitionsmängd $A^{-1}(X)$.*

Bevis. Den affina avbildningen A kan skrivas på formen $Ay = Cy + b$, där C är en linjär avbildning och b är en vektor. Låt nu y vara en punkt i $A^{-1}(X)$ och u vara en vektor i $\mathbf{R}^m$, och sätt $x = Ay$ och $v = Cu$. Enligt kedjeregeln är

$$\begin{aligned} D^2g(y)[u,u] &= D^2f(Ay)[Cu,Cu] = D^2f(x)[v,v] \qquad \text{och} \\ D^3g(y)[u,u,u] &= D^3f(Ay)[Cu,Cu,Cu] = D^3f(x)[v,v,v], \end{aligned}$$

och det följer att

$$\begin{aligned} \big|D^3g(y)[u,u,u]\big| &= \big|D^3f(x)[v,v,v]\big| \le 2\big(D^2f(x)[v,v]\big)^{3/2} \\ &= 2\big(D^2g(y)[u,u]\big)^{3/2}. \end{aligned}$$

□

Exempel 16.1.3. Det följer av exempel 16.1.1 och sats 16.1.6 att funktionen $f(x) = -\ln(b - \langle c, x\rangle)$ är självkonkordant i det öppna halvrummet

$$\{x \in \mathbf{R}^n \mid \langle c, x\rangle < b\}.$$

□

Exempel 16.1.4. Antag att polyedern

$$X = \bigcap_{j=1}^{p} \{x \in \mathbf{R}^n \mid \langle c_j, x\rangle \le b_j\}$$

har ett icke-tomt inre. Då är funktionen $f(x) = -\sum_{j=1}^{p} \ln(b_j - \langle c_j, x\rangle)$ självkonkordant med det inre av X som definitionsmängd. □

16.2 Slutna självkonkordanta funktioner

I avsnitt 6.7 studerades godtyckliga konvexa funktioners recessiva delrum. För slutna självkonkordanta funktioner har det recessiva delrummet följande egenskaper.

Sats 16.2.1. *Antag att $f\colon X \to \mathbf{R}$ är en sluten självkonkordant funktion. Funktionens recessiva delrum V_f är då lika med nollrummet $\mathcal{N}(f''(x))$ till andraderivatan $f''(x)$ i en godtycklig punkt $x \in X$. Vidare gäller:*

(i) $X = X + V_f$.

(ii) *För alla vektorer $v \in V_f$ är $f(x+v) = f(x) + Df(x)[v]$.*
(iii) *Om $\lambda(f,x) < \infty$, så är $f(x+v) = f(x)$ för alla $v \in V_f$.*

Bevis. Påståendena (i) och (ii) gäller för det recessiva delrummet till en godtycklig differentierbar konvex funktion enligt sats 6.7.1, så vi behöver bara bevisa de övriga påståendena. Låt för den skull x vara en godtycklig punkt i X och v en godtycklig vektor i $\mathbf{R}^n$, och betrakta restriktionen $\phi_{x,v}(t) = f(x+tv)$ av f till linjen genom x med riktning v. Restriktionens definitionsmängd är ett öppet intervall $I =]\alpha, \beta[$ kring 0.

Antag nu först att $v \in V_f$. För alla $t \in I$ är då

$$\phi_{x,v}(t) = f(x) + tDf(x)[v]$$

på grund av egenskapen (ii), och härav följer speciellt att

$$\|v\|_x^2 = D^2 f(x)[v,v] = \phi''_{x,v}(0) = 0,$$

dvs. att vektorn v ligger i nollrummet till andraderivatan $f''(x)$. Detta bevisar inklusionen $V_f \subseteq \mathcal{N}(f''(x))$. Observera att denna inklusion gäller för godtyckliga två gånger differentierbara konvexa funktioner, alltså utan antagande om självkonkordans och slutenhet.

För att visa den omvända inklusionen $\mathcal{N}(f''(x)) \subseteq V_f$, antar vi istället att v är en vektor i $\mathcal{N}(f''(x))$. Eftersom $\mathcal{N}(f''(x+tv)) = \mathcal{N}(f''(x))$ för alla $t \in I$ på grund av sats 16.1.2, är nu

$$\phi''_{x,v}(t) = D^2 f(x+tv)[v,v] = \|v\|_{x+tv}^2 = 0$$

för alla $t \in I$, och det följer att

$$\phi_{x,v}(t) = \phi_{x,v}(0) + \phi'_{x,v}(0)t = f(x) + Df(x)[v]\, t.$$

Om $\beta < \infty$, så är $x + \beta v$ en randpunkt till X och $\lim_{t\to\beta} \phi_{x,v}(t) < \infty$. Detta strider emellertid enligt korollarium 8.2.2 mot att funktionen f är sluten. Alltså är $\beta = \infty$, och på motsvarande sätt är $\alpha = -\infty$. Detta betyder att $I =]-\infty, \infty[$, och därför ligger säkert talet 1 i I. Vi drar slutsatsen att punkten $x+v$ ligger i mängden X och att $f(x+v) = \phi_{x,v}(1) = f(x)+Df(x)[v]$ för alla $x \in X$ och alla $v \in \mathcal{N}(f''(x))$, och sats 6.7.1 ger oss nu inklusionen $\mathcal{N}(f''(x)) \subseteq V_f$. Detta visar att $V_f = \mathcal{N}(f''(x))$.

Antag slutligen att $\lambda(f,x) < \infty$. Då finns det per definition en Newtonriktning i punkten x, och detta medför enligt anmärkningen efter definitionen av Newtonriktning att implikationen

$$f''(x)v = 0 \Rightarrow Df(x)[v] = 0$$

gäller. Eftersom $V_f = \mathcal{N}(f''(x))$ följer det nu av påstående (ii) att $f(x+v) = f(x)$ för alla $v \in V_f$. □

Problemet att minimera en degenererad sluten självkonkordant funktion $f\colon X \to \mathbf{R}$ med ändligt Newtondekrement $\lambda(f,x)$ i alla punkter $x \in X$ kan nu reduceras till problemet att minimera en icke-degenererad sluten självkonkordant funktion på följande vis.

Antag att definitionsmängden X är en delmängd av $\mathbf{R}^n$ och låt V_f beteckna f:s recessiva delrum. Sätt $m = \dim V_f^\perp$ och låt $A\colon \mathbf{R}^m \to \mathbf{R}^n$ vara en godtycklig injektiv linjär avbildning med $V_f^\perp$ som bildmängd, samt sätt $X_0 = A^{-1}(X)$. Då är X_0 en öppen delmängd av $\mathbf{R}^m$, och vi får en funktion $g\colon X_0 \to \mathbf{R}$ genom att definiera $g(y) = f(Ay)$ för $y \in X_0$.

Funktionen g är självkonkordant på grund av sats 16.1.6. Etersom en punkt (y,t) tillhör epigrafen till g om och endast om (Ay,t) tillhör epigrafen till f, följer det vidare att g också är en sluten funktion.

Antag att $v \in \mathcal{N}(g''(y))$. Eftersom $g''(y) = A^T f''(Ay)A$, är

$$\langle Av, f''(Ay)Av\rangle = \langle v, A^T f''(Ay)Av\rangle = \langle v, g''(y)v\rangle = 0,$$

vilket innebär att Av är en vektor i $\mathcal{N}(f''(Ay))$, dvs. i det recessiva delrummet V_f. Men eftersom Av per definition också ligger i $V_f^\perp$ och $V_f \cap V_f^\perp = \{0\}$, följer det att $Av = 0$, och eftersom A är en injektiv avbildning är $v = 0$. Detta visar att $\mathcal{N}(g''(y)) = \{0\}$, så funktionen g är icke-degenererad.

Varje vektor $x \in X$ har en entydig uppdelning på formen $x = x_1 + x_2$ med $x_1 \in V_f^\perp$ och $x_2 \in V_f$, och x_1 $(= x - x_2)$ ligger enligt sats 16.2.1 i X. Följaktligen finns det en unik punkt $y \in X_0$ så att $Ay = x_1$. Enligt samma sats är därför $g(y) = f(Ay) = f(x_1) = f(x)$.

Funktionerna f och g har således samma värdemängder, och $\hat{y}$ är en minimipunkt till funktionen g om och endast om $A\hat{y}$, och därmed också alla punkter på formen $A\hat{y} + v$ med $v \in V_f$, är en minimipunkt till f.

Vi noterar också för framtida bruk att enligt sats 15.1.7 är

$$\lambda(g,y) \le \lambda(f,Ay) = \lambda(f,Ay+v).$$

för alla $y \in X_0$ och $v \in V_f$. (I föreliggande fall råder faktiskt likhet mellan Newtondekrementen, vilket vi lämnar som övning att visa.)

Korollarium 16.2.2. *Om definitionsmängden X till en sluten självkonkordant funktion $f\colon X \to \mathbf{R}$ inte innehåller någon linje, så är funktionen icke-degenererad.*

Bevis. Enligt sats 16.2.1 är $X = X + V_f$. Så om f är degenererad, innehåller X alla linjer genom punkter i X med riktningar som ges av de nollskilda vektorerna i V_f. Om X inte innehåller några linjer, måste därför funktionen f vara icke-degenererad. □

Korollarium 16.2.3. *En sluten självkonkordant funktion är icke-degenererad om och endast om den är strikt konvex.*

Bevis. Andraderivatan $f''(x)$ till en icke-degenererad självkonkordant funktion f är positivt definit för alla x i definitionsmängden, och detta medför att f är strikt konvex.

Om en sluten självkonkordant funktion f är degenererad, så är det recessiva delrummet V_f icke-trivialt, och enligt sats 16.2.1 är restriktionen $\phi_{x,v}(t) = f(x+tv)$ av f till en linje med en riktning som ges av en nollskild vektor $v \in V_f$ affin, vilket innebär att f inte kan vara strikt konvex. □

16.3 Grundläggande olikheter för den lokala seminormen

Grafen till en konvex funktion f ligger ovanför sina tangentplan, och om funktionen dessutom är μ-starkt konvex, så är det vertikala avståndet mellan punkten $(y, f(y))$ på grafen och tangentplanet genom $(x, f(x))$ större än eller lika med $\frac{1}{2}\mu\|y-x\|^2$. Om funktionen är självkonkordant, begränsas samma avstånd också nedåt, men nu av ett uttryck som är en funktion av den lokala normen $\|y-x\|_x$. Den aktuella funktionen ρ definieras i följande lemma, som också beskriver de egenskaper hos ρ som vi kommer att behöva.

Lemma 16.3.1. *Låt* $\rho\colon]-\infty, 1[\to \mathbf{R}$ *vara funktionen*

$$\rho(t) = -t - \ln(1-t).$$

(i) Funktionen ρ *är konvex, strängt avtagande i intervallet* $]-\infty, 0]$ *och strängt växande i intervallet* $[0, 1[$*, och* $\rho(0) = 0$*.*

(ii) För $0 \le t < 1$ *är*

$$\rho(t) \le \frac{t^2}{2(1-t)},$$

och speciellt är alltså $\rho(t) \le t^2$ *om* $0 \le t \le \frac{1}{2}$*.*

(iii) Om $s < 1$ *och* $t < 1$*, så är* $\rho(s) + \rho(t) \ge -st$*.*

(iv) Om $s \ge 0$*,* $0 \le t < 1$ *och* $\rho(-s) \le \rho(t)$*, så är* $s \le \dfrac{t}{1-t}$*.*

Bevis. (i) följer genom derivering, och (ii) genom Taylorutveckling, ty i intervallet $0 \le t < 1$ är

$$\rho(t) = \tfrac{1}{2}t^2 + \tfrac{1}{3}t^3 + \tfrac{1}{4}t^4 + \cdots \le \tfrac{1}{2}t^2(1 + t + t^2 + \cdots) = \tfrac{1}{2}t^2(1-t)^{-1}.$$

Påstående (iii) följer av den elementära olikheten $x - \ln(1+x) \ge 0$, som medför att

$$\begin{aligned} st + \rho(s) + \rho(t) &= st - s - t - \ln(1-s) - \ln(1-t) \\ &= st - s - t - \ln(1 + st - s - t) \ge 0. \end{aligned}$$

Eftersom ρ är strängt avtagande i intervallet $]-\infty, 0]$, följer påstående (iv) om vi visar att $\rho(-s) \geq \rho(t)$ för $s = t/(1-t)$. Vi sätter för den skull

$$g(t) = \rho\big(-\frac{t}{1-t}\big) - \rho(t)$$

för $0 \leq t < 1$, och får efter förenkling

$$g(t) = t - 1 + (1-t)^{-1} + 2\ln(1-t).$$

Eftersom $g(0) = 0$ och

$$g'(t) = 1 + (1-t)^{-2} - 2(1-t)^{-1} = t^2(1-t)^{-2} \geq 0,$$

är $g(t) \geq 0$ för alla $t \in [0, 1[$, och därmed är påstående (iv) bevisat. □

Nästa sats används för att uppskatta differenser av typen $\|w\|_y - \|w\|_x$, $Df(y)[w] - Df(x)[w]$ och $f(y) - f(x) - Df(x)[y-x]$ i termer av $\|w\|_x$, $\|y-x\|_x$ och funktionen ρ.

Sats 16.3.2. *Låt $f\colon X \to \mathbf{R}$ vara en sluten självkonkordant funktion, och antag att x är en punkt i X och att $\|y-x\|_x < 1$. Då ligger punkten y också i X, och följande olikheter gäller för vektorn $v = y - x$ och för godtyckliga vektorer w:*

$$\frac{\|v\|_x}{1+\|v\|_x} \leq \|v\|_y \leq \frac{\|v\|_x}{1-\|v\|_x} \tag{16.3}$$

$$\frac{\|v\|_x^2}{1+\|v\|_x} \leq Df(y)[v] - Df(x)[v] \leq \frac{\|v\|_x^2}{1-\|v\|_x} \tag{16.4}$$

$$\rho(-\|v\|_x) \leq f(y) - f(x) - Df(x)[v] \leq \rho(\|v\|_x) \tag{16.5}$$

$$(1-\|v\|_x)\|w\|_x \leq \|w\|_y \leq \frac{\|w\|_x}{1-\|v\|_x} \tag{16.6}$$

$$Df(y)[w] - Df(x)[w] \leq D^2f(x)[v,w] + \frac{\|v\|_x^2\|w\|_x}{1-\|v\|_x} \leq \frac{\|v\|_x\|w\|_x}{1-\|v\|_x}. \tag{16.7}$$

De vänstra delarna av de tre olikheterna (16.3)*,* (16.4) *och* (16.5) *gäller dessutom med $v = y - x$ för alla $y \in X$.*

Bevis. Vi sparar beviset för att y ligger i X till sist och börjar med att visa att olikheterna (16.3–16.7) gäller under tilläggsantagandet $y \in X$.

I. Vi startar med olikheten (16.6). Om $\|w\|_x = 0$, så är $\|w\|_z = 0$ för alla $z \in X$ enligt sats 16.1.2, och olikheten gäller följaktligen i detta fall. Antag därför att w är en godtycklig vektor med $\|w\|_x \neq 0$, sätt $x^t = x + t(y-x)$ och definiera funktionen ψ genom att sätta

$$\psi(t) = \|w\|_{x^t}^{-1} = \big(D^2f(x^t)[w,w]\big)^{-1/2}.$$

Funktionen är definierad i ett öppet intervall som innehåller intervallet $[0,1]$, $\psi(0) = \|w\|_x^{-1}$ och $\psi(1) = \|w\|_y^{-1}$. Det följer vidare med hjälp av sats 16.1.1 att

$$
\begin{aligned}
(16.8) \qquad |\psi'(t)| &= \frac{1}{2}\big|\big(D^2 f(x^t)[w,w]\big)^{-3/2} D^3 f(x^t)[w,w,v]\big| \\
&= \frac{1}{2}\|w\|_{x^t}^{-3}\big|D^3 f(x^t)[w,w,v]\big| \le \frac{1}{2}\|w\|_{x^t}^{-3} \cdot 2\|w\|_{x^t}^2 \|v\|_{x^t} \\
&= \|w\|_{x^t}^{-1}\|v\|_{x^t} = \psi(t)\|v\|_{x^t}.
\end{aligned}
$$

Om $\|v\|_x = 0$, så är $\|v\|_z = 0$ för alla $z \in X$, och följaktligen $\psi'(t) = 0$ för $0 \le t \le 1$. Detta medför att $\psi(1) = \psi(0)$, dvs. att $\|w\|_y = \|w\|_x$. Olikheterna (16.3) och (16.6) gäller således i fallet $\|v\|_x = 0$.

Antag därför fortsättningsvis att $\|v\|_x \neq 0$, och välj först $w = v$ i definitionen av funktionen ψ. Olikheten (16.8) innebär i detta specialfall att $|\psi'(t)| \le 1$ för $t \in [0,1]$, och på grund av medelvärdessatsen är därför $\psi(0) - 1 \le \psi(1) \le \psi(0) + 1$. Den högra delen av denna olikhet innebär att $\|v\|_y^{-1} \le \|v\|_x^{-1} + 1$, vilket efter omstuvning ger den vänstra delen av olikheten (16.3). Observera också att detta gäller även om $\|v\|_x \ge 1$. Den vänstra delen av samma olikhet ger på motsvarande sätt upphov till den högra delen av olikheten (16.3), nu under förutsättning att $\|v\|_x < 1$.

För att visa olikheten (16.6) återvänder vi till funktionen ψ för allmänt w. Eftersom $\|tv\|_x = t\|v\|_x < 1$ för $0 \le t \le 1$, följer det av den redan bevisade olikheten (16.3) (med $x^t = x + tv$ istället för y) att

$$
\|v\|_{x^t} = \frac{1}{t}\|tv\|_{x^t} \le \frac{1}{t} \cdot \frac{\|tv\|_x}{1 - \|tv\|_x} = \frac{\|v\|_x}{1 - t\|v\|_x}.
$$

Insättning av denna uppskattning i (16.8) ger oss följande olikhet för derivatan av funktionen $\ln \psi(t)$:

$$
|(\ln \psi(t))'| = \frac{|\psi'(t)|}{\psi(t)} = \|v\|_{x^t} \le \frac{\|v\|_x}{1 - t\|v\|_x}.
$$

Integrera nu denna olikhet över intervallet $[0,1]$; detta resulterar i uppskattningen

$$
\begin{aligned}
\big|\ln \frac{\|w\|_y}{\|w\|_x}\big| = \big|\ln \frac{\psi(0)}{\psi(1)}\big| &= \big|\ln \psi(1) - \ln \psi(0)\big| = \Big|\int_0^1 (\ln \psi(t))'\, dt\Big| \\
&\le \int_0^1 \frac{\|v\|_x}{1 - t\|v\|_x}\, dt = -\ln(1 - \|v\|_x),
\end{aligned}
$$

som efter exponentiering ger

$$
1 - \|v\|_x \le \frac{\|w\|_y}{\|w\|_x} \le (1 - \|v\|_x)^{-1},
$$

vilket är olikheten (16.6).

II. För att visa olikheten (16.4) sätter vi

$$\phi(t) = Df(x^t)[v],$$

där som tidigare $x^t = x + t(y - x)$. Då är

$$\phi'(t) = D^2 f(x^t)[v, v] = \|v\|_{x^t}^2 = t^{-2}\|tv\|_{x^t}^2,$$

så genom att använda oss av olikheten (16.3) får vi för $0 \leq t \leq 1$ olikheten

$$\frac{\|v\|_x^2}{(1 + t\|v\|_x)^2} = \frac{1}{t^2}\frac{\|tv\|_x^2}{(1 + \|tv\|_x)^2} \leq \phi'(t) \leq \frac{1}{t^2}\frac{\|tv\|_x^2}{(1 - \|tv\|_x)^2} = \frac{\|v\|_x^2}{(1 - t\|v\|_x)^2}.$$

Den vänstra delen av olikheten gäller med $v = y - x$ för alla $y \in X$, och den högra delen gäller om $\|v\|_x < 1$, och genom att integrera olikheten över intervallet [0,1] erhåller vi olikheten (16.4).

III. För att bevisa olikheten (16.5) startar vi med funktionen

$$\Phi(t) = f(x^t) - Df(x)[v]\, t,$$

och noterar att

$$\Phi(1) - \Phi(0) = f(y) - f(x) - Df(x)[v]$$

och att

$$\Phi'(t) = Df(x^t)[v] - Df(x)[v].$$

Genom att i olikheten (16.4) byta y mot x^t får vi olikheten

$$\frac{t\|v\|_x^2}{1 + t\|v\|_x} \leq \Phi'(t) \leq \frac{t\|v\|_x^2}{1 - t\|v\|_x},$$

där den högra delen av olikheten bara gäller om $\|v\|_x < 1$. Genom att integrera ovanstående olikhet över intervallet $[0, 1]$ erhåller vi slutligen

$$\rho(-\|v\|_x) = \int_0^1 \frac{t\|v\|_x^2}{1 + t\|v\|_x}\, dt \leq \Phi(1) - \Phi(0) \leq \int_0^1 \frac{t\|v\|_x^2}{1 - t\|v\|_x}\, dt = \rho(\|v\|_x),$$

dvs. olikheten (16.5).

IV. Beviset för olikheten (16.7) är analogt med beviset för olikheten (16.4), men vi definierar den här gången funktionen ϕ genom att sätta

$$\phi(t) = Df(x^t)[w].$$

Nu är $\phi'(t) = D^2 f(x^t)[w,v]$ och $\phi''(t) = D^3 f(x^t)[w,v,v]$, så det följer av sats 16.1.1 och olikheten (16.6) att

$$|\phi''(t)| \le 2\|w\|_{x^t}\|v\|_{x^t}^2 \le 2\frac{\|w\|_x\|v\|_x^2}{(1-t\|v\|_x)^3}.$$

Genom att integrera denna olikhet över intervallet $[0,s]$, där $s \le 1$, får vi uppskattningen

$$\begin{aligned}\phi'(s) - \phi'(0) &\le \int_0^s |\phi''(t)|\,dt \le 2\|w\|_x \int_0^s \frac{\|v\|_x^2\,dt}{(1-t\|v\|_x)^3} \\ &= \|w\|_x\Big[\frac{\|v\|_x}{(1-s\|v\|_x)^2} - \|v\|_x\Big],\end{aligned}$$

och ytterligare en integration över intervallet $[0,1]$ resulterar i olikheten

$$\phi(1) - \phi(0) - \phi'(0) \le \int_0^1 (\phi'(s) - \phi'(0))\,ds \le \frac{\|w\|_x\|v\|_x^2}{1-\|v\|_x},$$

som är den vänstra delen av olikheten (16.7).

På grund av Cauchy–Schwarz olikhet är

$$\begin{aligned}D^2 f(x)[v,w] &= \langle v, f''(x)w\rangle = \langle f''(x)^{1/2}v, f''(x)^{1/2}w\rangle \\ &\le \|f''(x)^{1/2}v\|\|f''(x)^{1/2}w\| = \|v\|_x\|w\|_x,\end{aligned}$$

och vi får nu omedelbart den högra delen av olikheten (16.7) genom att ersätta $D^2 f(x)[v,w]$ med majoranten $\|v\|_x\|w\|_x$.

V. Det återstår nu endast att visa att villkoret $\|y-x\|_x < 1$ medför att punkten y ligger i X.

Antag därför motsatsen, dvs. att det finns en punkt y utanför X med $\|y-x\|_x < 1$. Då skär sträckan $[x,y]$ randen till X i en punkt $x + \bar{t}v$, där $\bar{t}$ är ett tal i intervallet $]0,1]$. Funktionen ρ är växande i intervallet $[0,1[$, så $\rho(t\|v\|_x) \le \rho(\|v\|_x)$ om $0 \le t < \bar{t}$. Det följer därför av olikheten (16.5) att

$$f(x+tv) \le f(x) + tDf(x)[v] + \rho(t\|v\|_x) \le f(x) + |Df(x)[v]| + \rho(\|v\|_x) < +\infty$$

för alla t i intervallet $[0,\bar{t}[$. Detta strider emellertid mot att funktionen f är sluten, ty eftersom $x + \bar{t}v$ är en randpunkt, är $\lim_{t\to\bar{t}} f(x+tv) = +\infty$. Punkten y ligger följaktligen i X. □

16.4 Minimering

Det här avsnittet handlar om minimering av självkonkordanta funktioner, och de erhållna resultaten bygger till stor del på följande sats, som också är en grundbult för Newtons algoritm, som vi kommer att studera i nästa avsnitt.

Sats 16.4.1. *Låt $f\colon X \to \mathbf{R}$ vara en sluten självkonkordant funktion, antag att x är en punkt i X med ändligt Newtondekrement $\lambda = \lambda(f,x)$ och låt Δx_{nt} vara en Newtonriktning i punkten x. Då ligger punkten*

$$x^+ = x + (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$$

i X och

$$f(x^+) \le f(x) - \rho(-\lambda).$$

Anmärkning. Om f har en minimipunkt $\hat{x}$, så är alltså speciellt

$$f(\hat{x}) \le f(x) - \rho(-\lambda)$$

för alla $x \in X$ med ändligt Newtondekrement λ.

Bevis. Vektorn $v = (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$ har lokal seminorm

$$\|v\|_x = (1+\lambda)^{-1}\|\Delta x_{\mathrm{nt}}\|_x = \lambda(1+\lambda)^{-1} < 1,$$

så det följer av sats 16.3.2 att punkten $x^+ = x + v$ ligger i X och att

$$\begin{aligned} f(x^+) &\le f(x) + Df(x)[v] + \rho(\|v\|_x) = f(x) + \frac{1}{1+\lambda}\langle f'(x), \Delta x_{\mathrm{nt}}\rangle + \rho(\frac{\lambda}{1+\lambda}) \\ &= f(x) - \frac{\lambda^2}{1+\lambda} - \frac{\lambda}{1+\lambda} - \ln\frac{1}{1+\lambda} = f(x) - \lambda + \ln(1+\lambda) \\ &= f(x) - \rho(-\lambda). \end{aligned}$$

□

Sats 16.4.2. *Om en sluten självkonkordant funktion $f\colon X \to \mathbf{R}$ är nedåt begränsad, så är Newtondekrementet $\lambda(f,x)$ ändligt i varje punkt $x \in X$ och $\inf_{x\in X}\lambda(f,x) = 0$.*

Bevis. Låt v vara en godtycklig vektor i det recessiva delrummet $V_f = \mathcal{N}(f''(x))$. Enligt sats 16.2.1 är

$$f(x+tv) = f(x) + t\langle f'(x), v\rangle$$

för alla $t \in \mathbf{R}$ vilket, eftersom f förutsatts vara nedåt begränsad, medför att $\langle f'(x), v\rangle = 0$. Detta bevisar implikationen

$$f''(x)v = 0 \Rightarrow \langle f'(x), v\rangle = 0,$$

som innebär att det finns en Newtonriktning i punkten x och att följaktligen $\lambda(f,x) < \infty$.

Om det finns ett positivt tal δ så att $\lambda(f,x) \geq \delta$ för alla $x \in X$, så ger upprepad användning av sats 16.4.1 med start i en godtycklig punkt $x_0 \in X$ upphov till en följd $(x_k)_0^\infty$ av punkter i X, definierade av att $x_{k+1} = x_k^+$, sådana att

$$f(x_k) \leq f(x_0) - k\rho(-\delta)$$

för alla k, och eftersom $\rho(-\delta) > 0$, strider detta mot antagandet att funktionen f är nedåt begränsad. Följaktligen är $\inf_{x\in X} \lambda(f,x) = 0$. □

Sats 16.4.3. *Låt $f\colon X \to \mathbf{R}^n$ vara en icke-degenererad sluten självkonkordant funktion och antag att $\lambda(f,x_0) < 1$ för någon punkt $x_0 \in X$. Då är funktionens alla subnivåmängder kompakta.*

Bevis. Subnivåmängderna är slutna eftersom funktionen är sluten, och för att visa att de också är begränsade räcker det på grund av sats 6.8.3 att visa att subnivåmängden $S = \{x \in X \mid f(x) \leq f(x_0)\}$ är begränsad.

Så låt x vara en godtycklig punkt i S, och sätt $r = \|x - x_0\|_{x_0}$ och $\lambda_0 = \lambda(f,x_0)$. Enligt sats 16.3.2 är

$$f(x) \geq f(x_0) + Df(x_0)[x - x_0] + \rho(-r),$$

och sats 15.1.2 ger att

$$Df(x_0)[x - x_0] = \langle f'(x_0), x - x_0\rangle \geq -\lambda(f,x_0)\|x - x_0\|_{x_0} = -\lambda_0 r.$$

Genom att kombinera de båda olikheterna erhålls olikheten

$$f(x_0) \geq f(x) \geq f(x_0) - \lambda_0 r + \rho(-r),$$

som medför att

$$r - \ln(1+r) = \rho(-r) \leq \lambda_0 r,$$

dvs. att

$$(1-\lambda_0)r \leq \ln(1+r).$$

Vi kan nu dra slutsatsen att $r \leq r_0$, där r_0 är den unika positiva roten till ekvationen $(1-\lambda_0)r = \ln(1+r)$. Subnivåmängden S är således inkluderad i ellipsoiden $\{x \in \mathbf{R}^n \mid \|x - x_0\|_{x_0} \leq r_0\}$ och är därmed begränsad. □

Sats 16.4.4. *Låt $f\colon X \to \mathbf{R}$ vara en sluten självkonkordant funktion, och antag att $\lambda(f,x_0) < 1$ för någon punkt $x_0 \in X$. Då antar funktionen ett minimum.*

Bevis. Om funktionen f är icke-degenererad, så är funktionens subnivåmängder kompakta enligt föregående sats, och funktionen antar därför ett minimum på subnivåmängden $\{x \in X \mid f(x) \leq f(x_0)\}$. Detta minimum är förstås också ett globalt minimum till f. Minimipunkten är vidare unik eftersom en icke-degenererad självkonkordant funktion är strikt konvex.

Om funktionen f är degenererad, så finns det enligt diskussionen efter sats 16.2.1 en icke-degenererad självkonkordant funktion $g\colon X_0 \to \mathbf{R}$ med samma värdemängd som f. Sambandet mellan de båda funktionerna har formen $g(y) = f(Ay + v)$, där A är en injektiv linjär avbildning och v är en godtycklig vektor i det recessiva delrummet V_f. Till x_0 hör vidare en punkt $y_0 \in X_0$ sådan att $Ay_0 + v = x_0$ för något $v \in V_f$, och för den gäller att $\lambda(g, y_0) \leq \lambda(f, x_0) < 1$. Enligt den redan bevisade delen av satsen har därför funktionen g en unik minimipunkt $\hat{y}$, och detta medför att samtliga punkter i mängden $A\hat{y} + V_f$ är minimipunkter till f. □

Sats 16.4.5. *Varje sluten självkonkordant funktion $f\colon X \to \mathbf{R}$ som är nedåt begränsad, har en minimipunkt.*

Bevis. Det följer av sats 16.4.2 att det finns en punkt $x_0 \in X$ sådan att $\lambda(f, x_0) < 1$, så satsen är ett korollarium till sats 16.4.4. □

Nästa sats beskriver hur väl en given punkt approximerar minimipunkten till en sluten självkonkordant funktion.

Sats 16.4.6. *Låt $f\colon X \to \mathbf{R}$ vara en sluten självkonkordant funktion med minimipunkt $\hat{x}$. Om $x \in X$ är en godtycklig punkt med Newtondekrement $\lambda = \lambda(f, x) < 1$, så är*

$$\rho(-\lambda) \leq f(x) - f(\hat{x}) \leq \rho(\lambda) \tag{16.9}$$

$$\frac{\lambda}{1+\lambda} \leq \|x - \hat{x}\|_x \leq \frac{\lambda}{1-\lambda} \tag{16.10}$$

$$\|x - \hat{x}\|_{\hat{x}} \leq \frac{\lambda}{1-\lambda}. \tag{16.11}$$

Anmärkning. För $t \leq \frac{1}{2}$ är $\rho(t) \leq t^2$. Om $\lambda(f, x) \leq \frac{1}{2}$, så är därför

$$f(x) - f_{\min} \leq \lambda(f, x)^2$$

på grund av olikheten (16.9).

Bevis. Sätt för att förenkla beteckningarna $v = x - \hat{x}$ och $r = \|v\|_x$.

Den vänstra delen av olikheten (16.9) följer direkt av anmärkningen efter sats 16.4.1. För att visa den högra delen av samma olikhet erinrar vi om olikheten

$$\langle f'(x), v\rangle \leq \lambda(f, x)\|v\|_x = \lambda r, \tag{16.12}$$

och kombinerar den med den vänstra delen av olikheten (16.5) i sats 16.3.2 och olikheten (iii) i lemma 16.3.1, vilket leder till följande kedja av olikheter:

$$\begin{aligned} f(\hat{x}) = f(x-v) &\geq f(x) + \langle f'(x), -v\rangle + \rho(-\|-v\|_x) \\ &= f(x) - \langle f'(x), v\rangle + \rho(-r) \\ &\geq f(x) - \lambda r + \rho(-r) \geq f(x) - \rho(\lambda). \end{aligned}$$

Därmed är olikheten (16.9) fullständigt bevisad.

Eftersom $x - v = \hat{x}$ och $f'(\hat{x}) = 0$, följer det av olikheten (16.12) och den vänstra delen av olikheten (16.4) att

$$\lambda r \geq \langle f'(x), v\rangle = \langle f'(x-v), -v\rangle - \langle f'(x), -v\rangle \geq \frac{\|-v\|_x^2}{1 + \|-v\|_x} = \frac{r^2}{1+r},$$

och genom att lösa olikheten ovan med avseende på r erhåller vi den högra delen av olikheten (16.10).

Den vänstra delen av samma olikhet gäller uppenbarligen om $r \geq 1$. Antag därför att $r < 1$. På grund av olikheten (16.7) är nu

$$\langle f'(x), w\rangle = \langle f'(x-v), -w\rangle - \langle f'(x), -w\rangle \leq \frac{\|-v\|_x \|-w\|_x}{1 - \|-v\|_x} = \frac{r}{1-r}\|w\|_x,$$

så det följer att

$$\lambda = \sup_{\|w\|_x \leq 1} \langle f'(x), w\rangle \leq \frac{r}{1-r},$$

vilket ger den vänstra delen av olikheten (16.10).

För att visa den återstående olikheten (16.11) använder vi den vänstra delen av olikheten (16.5) med y bytt mot x och x bytt mot $\hat{x}$, vilket ger oss olikheten

$$\rho(-\|x - \hat{x}\|_{\hat{x}}) \leq f(x) - f(\hat{x}).$$

Enligt den redan visade olikheten (16.9) är vidare $f(x) - f(\hat{x}) \leq \rho(\lambda)$, så det följer att $\rho(-\|x - \hat{x}\|_{\hat{x}}) \leq \rho(\lambda)$, och enligt lemma 16.3.1 betyder detta att $\|x - \hat{x}\|_{\hat{x}} \leq \dfrac{\lambda}{1-\lambda}$. □

Sats 16.4.7. *Antag att $f\colon X \to \mathbf{R}$ är en sluten självkonkordant funktion, att $X \subseteq \mathbf{R}^n$, och att*

$$\nu = \sup\{\lambda(f, x) \mid x \in X\} < 1.$$

Då är X lika med hela rummet $\mathbf{R}^n$, och funktionen f är konstant.

Bevis. Det följer av sats 16.4.4 att funktionen har en minimipunkt $\hat{x}$ och av olikheten (16.9) i sats 16.4.6 att

$$\rho(-\nu) \leq f(x) - f(\hat{x}) \leq \rho(\nu)$$

för alla $x \in X$, så funktionen är begränsad på X. Eftersom f är en sluten funktion, kan X följaktligen inte ha någon randpunkt utan sammanfaller med hela $\mathbf{R}^n$.

Om v är en godtycklig vektor i $\mathbf{R}^n$, så får vi vidare genom att tillämpa olikheten (16.11) med $x = \hat{x} + tv$ att

$$t\|v\|_{\hat{x}} = \|x - \hat{x}\|_{\hat{x}} \leq \frac{\lambda(f,x)}{1-\lambda(f,x)} \leq \frac{\nu}{1-\nu}$$

för alla $t > 0$, vilket medför att $\|v\|_{\hat{x}} = 0$. Funktionens recessiva delrum V_f är således lika med $\mathbf{R}^n$, och då är funktionen konstant enligt sats 16.2.1. □

16.5 Newtons metod för självkonkordanta målfunktioner

I det här avsnittet skall vi visa att Newtons metod konvergerar när målfunktionen $f\colon X \to \mathbf{R}$ är sluten, självkonkordant och nedåt begränsad. Vi skall också ge en uppskattning av antalet iterationer som behövs för att erhålla minimivärdet med en given precision ϵ – en uppskattning som bara beror av ϵ och differensen mellan funktionsvärdet i startpunkten och minimivärdet. Algoritmen startar med en dämpad fas som inte kräver någon linjesökning eftersom steglängden i punkten x kan väljas lika med $1/(1+\lambda(f,x))$, och övergår sedan i en ren fas med kvadratisk konvergens när Newtondekrementet blivit tillräckligt litet.

Den dämpade fasen

Under den dämpade fasen genererar man punkterna x_k i Newtons algoritm rekursivt genom att för $k \geq 0$ sätta

$$x_{k+1} = x_k + \frac{1}{1+\lambda_k} v_k,$$

där $\lambda_k = \lambda(f, x_k)$ är Newtondekrementet i punkten x_k och v_k är en Newtonriktning i samma punkt, dvs.

$$f''(x_k)v_k = -f'(x_k).$$

Enligt sats 16.4.1 kommer alla genererade punkter x_k att ligga i X, givet att startpunkten x_0 är en punkt i X, och

$$f(x_{k+1}) - f(x_k) \leq -\rho(-\lambda_k).$$

Om $\delta > 0$ och $\lambda_k \geq \delta$, så är $\rho(-\lambda_k) \geq \rho(-\delta)$ eftersom funktionen $\rho(t)$ är avtagande för $t < 0$. Om x_N är den första punkten i följden för vilken $\lambda_N = \lambda(f, x_N) < \delta$, är därför

$$\begin{aligned} f_{\min} - f(x_0) \leq f(x_N) - f(x_0) &= \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x_k)) \\ &\leq -\sum_{k=0}^{N-1} \rho(-\lambda_k) \leq -\sum_{k=0}^{N-1} \rho(-\delta) = -N\rho(-\delta), \end{aligned}$$

varav följer att $N \leq (f(x_0) - f_{\min})/\rho(-\delta)$. Vi har därmed visat följande sats.

Sats 16.5.1. *Låt $f\colon X \to \mathbf{R}$ vara en sluten, självkonkordant och nedåt begränsad funktion. Med Newtons dämpade metod och steglängd $1/(1+\lambda(f,x))$ behövs det högst*

$$\left\lfloor \frac{f(x_0) - f_{\min}}{\rho(-\delta)} \right\rfloor$$

iterationer för att från en godtycklig startpunkt x_0 i X generera en punkt $x \in X$ med Newtondekrement $\lambda(f,x) < \delta$.

Lokal konvergens

Vi övergår nu till att studera Newtons rena metod för startpunkter som ligger tillräcklig nära en minimipunkt $\hat{x}$. För motsvarande analys av Newtons metod med dämpning hänvisas till övning 16.6.

Sats 16.5.2. *Låt $f\colon X \to \mathbf{R}$ vara en sluten självkonkordant funktion, och antag att $x \in X$ är en punkt med Newtondekrement $\lambda(f,x) < 1$. Låt Δx_{nt} vara en Newtonriktning i punkten och sätt*

$$x^+ = x + \Delta x_{\mathrm{nt}}.$$

Då ligger x^+ i X och

$$\lambda(f, x^+) \leq \Big(\frac{\lambda(f,x)}{1 - \lambda(f,x)}\Big)^2.$$

Bevis. Att x^+ ligger i X följer av sats 16.3.2 eftersom $\|\Delta x_{\mathrm{nt}}\|_x = \lambda(f,x) < 1$. För att visa olikheten för $\lambda(f,x^+)$ använder vi först olikhet (16.7) i samma sats med $v = x^+ - x = \Delta x_{\mathrm{nt}}$ och får då

$$\begin{aligned}\langle f'(x^+), w\rangle &\le \langle f'(x), w\rangle + \langle f''(x)\Delta x_{\mathrm{nt}}, w\rangle + \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)} \\ &= \langle f'(x), w\rangle + \langle -f'(x), w\rangle + \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)} = \frac{\lambda(f,x)^2\|w\|_x}{1-\lambda(f,x)}.\end{aligned}$$

På grund av olikheten (16.6) i sats 16.3.2 är vidare

$$\|w\|_x \le \frac{\|w\|_{x^+}}{1-\lambda(f,x)},$$

så det följer att

$$\langle f'(x^+), w\rangle \le \frac{\lambda(f,x)^2\|w\|_{x^+}}{(1-\lambda(f,x))^2},$$

och detta medför att

$$\lambda(f,x^+) = \sup_{\|w\|_{x^+}\le 1} \langle f'(x^+), w\rangle \le \frac{\lambda(f,x)^2}{(1-\lambda(f,x))^2}.$$

Därmed är satsen fullständigt bevisad. □

För Newtons rena algoritm kan vi nu visa följande konvergensresultat.

Sats 16.5.3. *Antag att funktionen $f\colon X \to \mathbf{R}$ är sluten och självkonkordant och att x_0 är en punkt i X med Newtondekrement*

$$\lambda(f,x_0) \le \delta < \overline{\lambda} = \tfrac{1}{2}(3-\sqrt{5}) = 0.381966\ldots.$$

Definiera följden $(x_k)_0^\infty$ rekursivt genom att sätta

$$x_{k+1} = x_k + v_k,$$

där v_k är en Newtonriktning i punkten x_k.

Då konvergerar följden $(f(x_k))_0^\infty$ mot funktionens minimivärde $f_{\min}$, och om $\epsilon > 0$ så är

$$f(x_k) - f_{\min} < \epsilon$$

för $k > A + \log_2(\log_2 B/\epsilon)$, där A och B är konstanter som bara beror av δ.

Om funktionen f är icke-degenererad, så konvergerar vidare följden $(x_k)_0^\infty$ mot funktionens unika minimipunkt.

Bevis. Det kritiska talet $\overline{\lambda}$ är rot till ekvationen $(1-\lambda)^2 = \lambda$ och för $0 \le \lambda < \overline{\lambda}$ är $\lambda < (1-\lambda)^2$.

Sätt $K(\lambda) = (1-\lambda)^{-2}$; i intervallet $[0, \overline{\lambda}[$ är funktionen K växande och $K(\lambda)\lambda < 1$. Det följer därför av sats 16.5.2 att för alla punkter $x \in X$ med $\lambda(f,x) \le \delta < \overline{\lambda}$ är

$$\lambda(f,x^+) \le K(\lambda(f,x))\,\lambda(f,x)^2 \le K(\delta)\lambda(f,x)^2 \le K(\delta)\delta\lambda(f,x) \le \lambda(f,x) \le \delta.$$

Sätt nu $\lambda_k = \lambda(f,x_k)$. Det följer då av ovanstående olikhet med hjälp av induktion att $\lambda_k \le \delta$ och att

$$\lambda_{k+1} \le K(\delta)\lambda_k^2$$

för alla k, och den sistnämnda olikheten medför i sin tur att

$$\lambda_k \le K(\delta)^{-1}\big(K(\delta)\lambda_0\big)^{2^k} \le (1-\delta)^2\big(K(\delta)\delta\big)^{2^k}.$$

Följaktligen går λ_k mot 0 då $k \to \infty$, ty $K(\delta)\delta < 1$, och eftersom

$$f(x_k) - f_{\min} \le \lambda_k^2$$

om $\lambda_k \le \frac{1}{2}$ enligt sats 16.4.6 och anmärkningen efter densamma, är

$$\lim_{k\to\infty} f(x_k) = f_{\min}.$$

För att bevisa den återstående feluppskattningen kan vi utan inskränkning anta att $\epsilon < \delta^2$, ty för $\epsilon > \delta^2$ är redan

$$f(x_0) - f_{\min} \le \lambda(f,x_0)^2 \le \delta^2 < \epsilon.$$

Sätt

$$A = -\log_2\big(-2\log_2(K(\delta)\delta)\big) \quad \text{och} \quad B = (1-\delta)^4;$$

då är $0 < B \le 1$, och eftersom $B/\epsilon \ge (1-\delta)^4/\delta^2 = (K(\delta)\delta)^{-2} > 1$, är $\log_2(\log_2 B/\epsilon)$ ett väldefinierat tal. För $k > A + \log_2(\log_2 B/\epsilon)$ är slutligen

$$\lambda_k^2 \le (1-\delta)^4\big(K(\delta)\delta\big)^{2^{k+1}} < \epsilon,$$

och följaktligen också $f(x_k) - f_{\min} < \epsilon$.

Om funktionen f är icke-degenererad, så har den en unik minimipunkt $\hat{x}$, och det följer nu av olikheten (16.11) i sats 16.4.6 att

$$\|x_k - \hat{x}\|_{\hat{x}} \le \frac{\lambda_k}{1-\lambda_k} \to 0, \quad \text{då } k \to \infty,$$

och eftersom $\|\cdot\|_{\hat{x}}$ är en norm, betyder detta att $x_k \to \hat{x}$. □

För $\delta = 1/3$ blir konstanterna $A = 0.268\ldots$ och $B = 16/81$, vilket för $\epsilon = 10^{-30}$ ger $A + \log_2(\log_2 B/\epsilon) = 6.87$. Om startpunkten x_0 i Newtons algoritm uppfyller $\lambda(f,x_0) < 1/3$, behövs det således högst 7 iterationer för att erhålla en punkt med ett funktionsvärde som approximerar minimivärdet med ett fel som understiger 10^{-30}.

Newtons metod för självkonkordanta funktioner

Genom att kombinera Newtons dämpade metod med $1/(1 + \lambda(f, x))$ som dämpningsfaktor och Newtons rena metod får vi följande variant av Newtons metod.

Newtons metod

Givet ett positivt tal $\delta < \frac{1}{2}(3 - \sqrt{5})$, startpunkt $x_0 \in X$ och tolerans $\epsilon > 0$.

1. *Initiera:* $x := x_0$.
2. Beräkna Newtondekrementet $\lambda = \lambda(f, x)$.
3. Gå till punkt 8 om $\lambda < \delta$.
4. Beräkna annars en Newtonriktning Δx_{nt} i punkten x.
5. *Uppdatera:* $x := x + (1 + \lambda)^{-1}\Delta x_{\text{nt}}$.
6. Gå tillbaka till punkt 2.
7. Beräkna Newtondekrementet $\lambda = \lambda(f, x)$.
8. *Stoppkriterium:* Avbryt om $\lambda < \sqrt{\epsilon}$. x är approximativ optimal punkt.
9. Beräkna annars en Newtonriktning Δx_{nt} i punkten x.
10. *Uppdatera:* $x := x + \Delta x_{\text{nt}}$.
11. Gå tillbaka till punkt 7.

Den dämpade fasen av algoritmen, dvs. stegen 2–6 pågår för en nedåt begränsad, sluten självkonkordant funktion enligt sats 16.5.1 under högst

$$\lfloor (f(x_0) - f_{\min})/\rho(-\delta) \rfloor$$

iterationer, och den rena fasen 7–11 tar enligt sats 16.5.3 slut efter högst $\lceil A + \log_2(\log_2 B/\epsilon) \rceil$ iterationer. Vi har därför följande resultat.

Sats 16.5.4. *Om funktionen f är nedåt begränsad, sluten och självkonkordant, så stoppar ovanstående Newtons metod i en punkt x där $f(x) < f_{\min} + \epsilon$ efter högst*

$$\lfloor (f(x_0) - f_{\min}))/\rho(-\delta) \rfloor + \lceil A + \log_2(\log_2 B/\epsilon) \rceil$$

iterationer, där A och B är konstanterna i sats 16.5.3.

För $\delta = 1/3$ är speciellt $1/\rho(-\delta) = 21.905$, och den andra termen kan för $\epsilon \geq 10^{-30}$ ersättas av talet 7. Det krävs alltså högst $\lfloor 22(f(x_0) - f_{\min}) \rfloor + 7$ iterationer för att erhålla en approximation av minimivärdet som uppfyller alla praktiska behov med råge.

Övningar

16.1 Visa att funktionen $f(x) = x \ln x - \ln x$ är självkonkordant på $\mathbf{R}_{++}$.

16.2 Antag att $X_i \subseteq \mathbf{R}^{n_i}$ och att funktionerna $f_i \colon X_i \to \mathbf{R}$ är självkonkordanta för $i = 1, 2, \ldots, m$. Sätt $X = X_1 \times X_2 \times \cdots \times X_m$ och definiera funktionen $f \colon X \to \mathbf{R}$ genom att för $x = (x_1, x_2, \ldots, x_m) \in X$ sätta

$$f(x_1, x_2, \ldots, x_m) = f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

Visa att funktionen f är självkonkordant.

16.3 Antag att $f \colon \mathbf{R}_{++} \to \mathbf{R}$ är en tre gånger kontinuerligt deriverbar konvex funktion och att

$$|f'''(x)| \leq 3\frac{f''(x)}{x} \qquad \text{för alla } x.$$

a) Visa att funktionen

$$g(x) = -\ln(-f(x)) - \ln x$$

med $\{x \in \mathbf{R}_{++} \mid f(x) < 0\}$ som definitionsmängd är självkonkordant.

[Ledning: Utnyttja att $3a^2b + 3a^2c + 2b^3 + 2c^3 \leq 2(a^2 + b^2 + c^2)^{3/2}$ om $a, b, c \geq 0$.]

b) Visa att funktionen

$$F(x, y) = -\ln(y - f(x)) - \ln x$$

är självkonkordant på mängden $\{(x, y) \in \mathbf{R}^2 \mid x > 0,\ \ y > f(x)\}$.

16.4 Visa att funktionen f uppfyller villkoren i föregående övning om

a) $f(x) = -\ln x$ b) $f(x) = x \ln x$ c) $f(x) = -x^p$, där $0 < p \leq 1$.

16.5 Inför för $x = (x_1, x_2, \ldots, x_n)$ i $\mathbf{R}^n$ förkortningen $x' = (x_1, x_2, \ldots, x_{n-1})$, och låt $\|\cdot\|$ beteckna den euklidiska normen i $\mathbf{R}^{n-1}$. Sätt

$$X = \{x \in \mathbf{R}^n \mid \|x'\| < x_n\},$$

och låt $f \colon X \to \mathbf{R}$ vara funktionen

$$f(x) = -\ln(x_n^2 - \|x'\|^2).$$

Visa att för alla vektorer $v \in \mathbf{R}^n$ gäller identiteten

$$D^2 f(x)[v, v] = \frac{1}{2}\big(Df(x)[v]\big)^2 + 2\frac{(x_n^2 - \|x'\|^2)(\|x'\|^2\|v'\|^2 - \langle x', v'\rangle^2) + (v_n\|x'\|^2 - x_n\langle x', v'\rangle)^2}{(x_n^2 - \|x'\|^2)^2\|x'\|^2},$$

och använd den för att dra slutsatsen att funktionen f är konvex och att $\lambda(f, x) = 2$ för alla $x \in X$.

16.6 *Konvergens för Newtons dämpade metod.*
Antag att funktionen $f \colon X \to \mathbf{R}$ är sluten och självkonkordant, och sätt om $x \in X$ är en punkt med ändligt Newtondekrement

$$x^+ = x + \frac{1}{1 + \lambda(f, x)}\Delta x_{\text{nt}},$$

där Δx_{nt} är en Newtonriktning i punkten x.

a) Enligt sats 16.3.2 ligger punkten x^+ i X. Visa att

$$\lambda(f, x^+) \leq 2\lambda(f, x)^2,$$

och att följaktligen $\lambda(f, x^+) \leq \lambda(f, x)$ om $\lambda(f, x) \leq \frac{1}{2}$.

b) Antag att x_0 är en punkt i X med Newtondekrement $\lambda(f, x_0) \leq \frac{1}{4}$, och definiera följden $(x_k)_0^\infty$ rekursivt genom att sätta $x_{k+1} = x_k^+$. Visa att

$$f(x_k) - f_{\min} \leq \tfrac{1}{4} \cdot \left(\tfrac{1}{2}\right)^{2^{k+1}},$$

och att följaktligen $f(x_k)$ konvergerar kvadratiskt mot $f_{\min}$.

Appendix

Vi börjar med ett resultat om trilinjära former, som behövdes i beviset för den fundamentala olikheten $\left|D^3 f(x)[u, v, w]\right| \leq 2\|u\|_x\|v\|_x\|w\|_x$ för självkonkordanta funktioner.

Fixera en godtycklig skalärprodukt $\langle \cdot\,, \cdot \rangle$ på $\mathbf{R}^n$ och låt $\|\cdot\|$ vara motsvarande norm, dvs. $\|v\| = \langle v, v\rangle^{1/2}$. Om $\phi(u, v, w)$ är en symmetrisk trilinjär form på $\mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^n$, definierar vi dess norm $\|\phi\|$ genom att sätta

$$\|\phi\| = \sup_{u,v,w \neq 0} \frac{|\phi(u, v, w)|}{\|u\|\|v\|\|w\|}.$$

Täljaren och nämnaren i uttrycket för $\|\phi\|$ är homogena av samma grad 3, så därför är också

$$\|\phi\| = \sup_{(u,v,w) \in S^3} |\phi(u, v, w)|,$$

där S betecknar enhetssfären i $\mathbf{R}^n$ med avseende på normen $\|\cdot\|$, dvs.

$$S = \{u \in \mathbf{R}^n \mid \|u\| = 1\}.$$

Av normdefinitionen följer vidare att

$$|\phi(u, v, w)| \leq \|\phi\|\|u\|\|v\|\|w\|$$

för alla vektorer u, v, w i $\mathbf{R}^n$.

Eftersom trilinjära former är kontinuerliga och enhetssfären är kompakt, antas supremum $\|\phi\|$ i någon punkt $(u, v, w) \in S^3$, och vi skall visa att supremum i själva verket antas i någon punkt där $u = v = w$. Detta är innebörden av följande sats.

Sats 1. *Antag att $\phi(u, v, w)$ är en symmetrisk trilinjär form. Då är*

$$\|\phi\| = \sup_{u,v,w \neq 0} \frac{|\phi(u, v, w)|}{\|u\|\|v\|\|w\|} = \sup_{v \neq 0} \frac{|\phi(v, v, v)|}{\|v\|^3}.$$

Anmärkning. Satsen är ett specialfall av motsvarande resultat för symmetriska m-multilinjära former, men vi behöver bara fallet $m = 3$. Det allmänna fallet visas med induktion.

Bevis. Sätt

$$\|\phi\|' = \sup_{v\neq 0} \frac{|\phi(v,v,v)|}{\|v\|^3} = \sup_{\|v\|=1} |\phi(v,v,v)|.$$

Vårt påstående är att $\|\phi\| = \|\phi\|'$. Uppenbarligen är $\|\phi\|' \leq \|\phi\|$, så vi behöver bara bevisa den omvända olikheten $\|\phi\| \leq \|\phi\|'$.

För beviset av denna olikhet behöver vi motsvarande resultat för symmetriska bilinjära former $\psi(u,v)$. Till en symmetrisk bilinjär form hör en symmetrisk linjär operator (matris) A sådan att $\psi(u,v) = \langle Au, v\rangle$, och om $e_1, e_2, \ldots, e_n$ är en ON-bas av egenvektorer till A och $\lambda_1, \lambda_2, \ldots, \lambda_n$ är motsvarande egenvärden med λ_1 som det till beloppet största egenvärdet, och vektorerna $u, v \in S$ har koordinaterna $u_1, u_2, \ldots, u_n$ resp. $v_1, v_2, \ldots, v_n$ med avseende på ON-basen ifråga, så är

$$\begin{aligned}|\psi(u,v)| = |\sum_{i=1}^n \lambda_i u_i v_i| &\leq \sum_{i=1}^n |\lambda_i||u_i||v_i| \leq |\lambda_1| \sum_{i=1}^n |u_i||v_i| \\ &\leq |\lambda_1| \Big(\sum_{i=1}^n u_i^2\Big)^{1/2} \Big(\sum_{i=1}^n v_i^2\Big)^{1/2} = |\lambda_1| = |\psi(e_1,e_1)|,\end{aligned}$$

vilket bevisar att $\sup_{(u,v)\in S^2} |\psi(u,v)| = \sup_{v\in S} |\psi(v,v)|$.

Vi återvänder nu till den trilinjära formen $\phi(u,v,w)$. Låt $(\hat{u},\hat{v},\hat{w})$ vara en punkt i S^3 där supremum som definierar $\|\phi\|$ antas, dvs.

$$\|\phi\| = \phi(\hat{u},\hat{v},\hat{w}),$$

och betrakta funktionen

$$\psi(u,v) = \phi(u,v,\hat{w});$$

detta är en symmetrisk bilinjär form på $\mathbf{R}^n \times \mathbf{R}^n$ och

$$\sup_{(u,v)\in S^2} |\psi(u,v)| = \|\phi\|.$$

Men för den symmetriska bilinjära formen ψ gäller enligt beviset ovan att

$$\sup_{(u,v)\in S^2} |\psi(u,v)| = \sup_{v\in S} |\psi(v,v)|.$$

Vi drar därför slutsatsen att vi utan inskränkning kan anta att $\hat{u} = \hat{v}$.

Vi har med andra ord visat att mängden

$$A = \{(v,w) \in S^2 \mid |\phi(v,v,w)| = \|\phi\|\}$$

inte är tom. Mängden A är en sluten delmängd av S^2 och följaktligen existerar talet

$$\alpha = \max\{\langle v, w\rangle \mid (v, w) \in A\},$$

och uppenbarligen är $0 \leq \alpha \leq 1$.

På grund av trilineariteten är

$$\phi(u+v, u+v, w) - \phi(u-v, u-v, w) = 4\phi(u, v, w).$$

För vektorer u, v, w i S, dvs. av norm 1, följer det därför att

$$\begin{aligned}
4|\phi(u,v,w)| &\leq |\phi(u+v,u+v,w)| + |\phi(u-v,u-v,w)| \\
&\leq |\phi(u+v,u+v,w)| + \|\phi\|\|u-v\|^2\|w\| \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + \|\phi\|(\|u+v\|^2 + \|u-v\|^2) \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + \|\phi\|(2\|u\|^2 + 2\|v\|^2) \\
&= |\phi(u+v,u+v,w)| - \|\phi\|\|u+v\|^2 + 4\|\phi\|.
\end{aligned}$$

Välj nu $(\overline{v}, \overline{w}) \in A$ så att $\langle \overline{v}, \overline{w}\rangle = \alpha$; då är på grund av ovanstående olikhet

$$\begin{aligned}
4\|\phi\| &= 4|\phi(\overline{v},\overline{v},\overline{w})| = 4|\phi(\overline{v},\overline{w},\overline{v})| \\
&\leq |\phi(\overline{v}+\overline{w},\overline{v}+\overline{w},\overline{v})| - \|\phi\|\|\overline{v}+\overline{w}\|^2 + 4\|\phi\|
\end{aligned}$$

varav följer att

$$|\phi(\overline{v}+\overline{w},\overline{v}+\overline{w},\overline{v})| \geq \|\phi\|\|\overline{v}+\overline{w}\|^2.$$

Notera att $\|\overline{v}+\overline{w}\|^2 = \|\overline{v}\|^2 + \|\overline{w}\|^2 + 2\langle \overline{v}, \overline{w}\rangle = 2 + 2\alpha > 0$, så vi kan därför bilda vektorn $\overline{z} = (\overline{v}+\overline{w})/\|\overline{v}+\overline{w}\|$ och då övergår ovanstående olikhet i

$$|\phi(\overline{z},\overline{z},\overline{v})| \geq \|\phi\|,$$

vilket medför att

$$|\phi(\overline{z},\overline{z},\overline{v})| = \|\phi\| \tag{16.13}$$

eftersom $\overline{z}$ och $\overline{v}$ är vektorer i S. Paret $(\overline{z}, \overline{v})$ ligger således i A, så det följer av definitionen av α att

$$\alpha \geq \langle \overline{z}, \overline{v}\rangle = \frac{\langle \overline{v}, \overline{v}\rangle + \langle \overline{w}, \overline{v}\rangle}{\|\overline{v}+\overline{w}\|} = \frac{1+\alpha}{\sqrt{2+2\alpha}} = \sqrt{\frac{1+\alpha}{2}},$$

och denna olikhet för α medför att $\alpha \geq 1$. Följaktligen är $\alpha = 1$, och

$$\langle \overline{z}, \overline{v}\rangle = 1 = \|\overline{z}\|\|\overline{v}\|.$$

Likheten i Cauchy–Schwarz olikhet medför att $\overline{z} = \overline{v}$. Insättning av detta i likheten (16.13) ger oss olikheten

$$\|\phi\|' \geq \phi(\overline{v},\overline{v},\overline{v}) = \|\phi\|,$$

och därmed är satsen bevisad. □

Det andra resultatet i detta appendix är följande entydighetssats för funktioner som uppfyller en speciell differentialolikhet.

Sats 2. *Antag att funktionen $y(t)$ är kontinuerligt deriverbar i intervallet $I = [0, b[$, att $y(t) \geq 0$, $y(0) = 0$ och $y'(t) \leq Cy(t)^\alpha$ för några givna konstanter $C > 0$ och $\alpha \geq 1$. Då är $y(t) = 0$ i intervallet I.*

Bevis. Sätt $a = \sup\{x \in I \mid y(t) = 0 \text{ för } 0 \leq t \leq x\}$. Vi skall visa att $a = b$ genom att visa att antagandet $a < b$ leder till en motsägelse.

På grund av kontinuiteten är $y(a) = 0$. Fixera en punkt $c \in]a, b[$ och sätt $M = \max\{y(t) \mid a \leq t \leq c\}$. Välj sedan en punkt d så att $a < d < c$ och $d - a \leq \frac{1}{2}C^{-1}M^{1-\alpha}$. Maximum av funktionen $y(t)$ på intervallet $[a, d]$ antas i någon punkt e, och på grund av supremumdefinitionen av a är $y(e) > 0$. Naturligtvis är också $y(e) \leq M$, så det följer att

$$\begin{aligned} y(e) = y(e) - y(a) &= \int_a^e y'(t)\,dt \leq C\int_a^e y(t)^\alpha\,dt \\ &\leq C(d-a)y(e)^\alpha \leq C(d-a)M^{\alpha-1}y(e) \leq \frac{1}{2}y(e), \end{aligned}$$

vilket är en motsägelse. □

Kapitel 17

Den vägföljande metoden

I det här kapitlet skall vi beskriva en metod för att lösa optimeringsproblemet

$$\begin{array}{ll} \min & f(x) \\ \text{då} & x \in X \end{array}$$

när X är en sluten delmängd av $\mathbf{R}^n$ med icke-tomt inre och f är en kontinuerlig funktion som är differentierbar i det inre av X. Vi förutsätter genomgående att $X = \mathrm{cl}(\mathrm{int}\, X)$. Ganska snart kommer vi att inskränka oss till konvexa problem, dvs. förutsätta att X och f är konvexa, och i det fallet är förstås automatiskt $X = \mathrm{cl}(\mathrm{int}\, X)$ för alla mängder med icke-tomt inre.

Descentmetoderna förutsätter att funktionen f är differentierbar i en omgivning av den optimala punkten, och om denna ligger på randen till X får vi följaktligen problem. Ett sätt att angripa problemet är att välja en funktion $F\colon \mathrm{int}\, X \to \mathbf{R}$ med egenskapen att $F(x) \to +\infty$ då x går mot randen till X och en parameter $\mu > 0$, samt att minimera funktionen $f(x) + \mu F(x)$ över $\mathrm{int}\, X$. Denna funktions minimipunkt $\hat{x}(\mu)$ ligger garanterat i det inre av X, och eftersom $f(x) + \mu F(x) \to f(x)$ då $\mu \to 0$, kan vi hoppas på att funktionsvärdet $f(\hat{x}(\mu))$ skall ligga nära f:s minimivärde om parametern μ är tillräckligt liten. Funktionen F fungerar som en barriär som hindrar den approximerande minimipunkten från att ligga på randen.

Funktionen $\mu^{-1} f(x) + F(x)$ har naturligtvis samma minimipunkt $\hat{x}(\mu)$ som $f(x) + \mu F(x)$, och av tekniska skäl fungerar det bättre att ha parametern framför målfunktionen f än framför barriärfunktionen F. I fortsättningen kommer vi därför istället, med nya beteckningar, att undersöka vad som händer med minimipunkten $\hat{x}(t)$ till funktionen $F_t(x) = tf(x) + F(x)$, när parametern t går mot $+\infty$.

17.1 Barriärer och den centrala vägen

Barriärer

Vi börjar med den formella definitionen av begreppet barriär.

Definition. Låt X vara en sluten konvex mängd med icke-tomt inre. Med en *barriär* till X menas en differentierbar funktion $F\colon \operatorname{int} X \to \mathbf{R}$, som har egenskapen att $\lim_{k\to\infty} F(x_k) = +\infty$ för alla följder $(x_k)_1^\infty$ som konvergerar mot någon randpunkt till X.

Om barriärfunktionen F har en unik minimipunkt, så kallas minimipunkten mängden X:s *analytiska centrum* (med avseende på barriären F).

Anmärkning. En konvex funktion med öppen definitionsmängd går mot oändligheten vid randen om och endast om den är sluten. En konvex differentierbar funktion $F\colon \operatorname{int} X \to \mathbf{R}$ är därför en barriär till X om och endast om den är sluten.

Om mängden X är kompakt och barriären F är strikt konvex, så har F en unik minimipunkt i $\operatorname{int} X$. Varje kompakt konvex mängd med icke-tomt inre har med andra ord ett analytiskt centrum med avseende på varje given strikt konvex barriär.

Givet en funktion $f\colon X \to \mathbf{R}$, som vi önskar minimera, och en barriär $F\colon \operatorname{int} X \to \mathbf{R}$ sätter vi

$$F_t(x) = tf(x) + F(x)$$

för $t \geq 0$ och $x \in \operatorname{int} X$. För fixt $t \geq 0$ är alltså F_t en funktion $\operatorname{int} X \to \mathbf{R}$, och $F_0 = F$. Följande sats utgör grundbulten för barriärbaserade inrepunktsmetoder för minimering.

Sats 17.1.1. *Antag att funktionen $f\colon X \to \mathbf{R}$ är kontinuerlig. Låt F vara en nedåt begränsad barriär till mängden X, och antag att funktionen F_t för varje $t > 0$ har en minimipunkt $\hat{x}(t) \in \operatorname{int} X$. Då är*

$$\lim_{t\to+\infty} f(\hat{x}(t)) = \inf_{x\in X} f(x).$$

Bevis. Sätt $v_{\min} = \inf_{x\in X} f(x)$ och $M = \inf_{x\in\operatorname{int} X} F(x)$. (Vi utesluter inte möjligheten att $v_{\min} = -\infty$, men M är förstås ett ändligt tal.)

Välj, givet talet $\eta > v_{\min}$, en punkt $x^* \in \operatorname{int} X$ så att $f(x^*) < \eta$. Då är

$$\begin{aligned} v_{\min} &\leq f(\hat{x}(t)) \leq f(\hat{x}(t)) + t^{-1}(F(\hat{x}(t)) - M) = t^{-1}\big(F_t(\hat{x}(t)) - M\big) \\ &\leq t^{-1}\big(F_t(x^*) - M\big) = f(x^*) + t^{-1}(F(x^*) - M). \end{aligned}$$

Eftersom högra ledet av denna olikhet går mot $f(x^*)$ då $t \to +\infty$, följer det att $v_{\min} \leq f(\hat{x}(t)) < \eta$ för alla tillräckligt stora tal t, vilket bevisar satsen. □

Barriärmetoden förutsätter förstås att man har en lämplig barriär till den givna mängden. För mängder av typen

$$X = \{x \in \Omega \mid g_i(x) \leq 0,\ \ i = 1, 2, \ldots, m\}$$

kommer vi att använda oss av den *logaritmiska barriärfunktionen*

$$F(x) = -\sum_{i=1}^{m} \ln(-g_i(x)). \tag{17.1}$$

Notera att barriärfunktionen F är konvex om samtliga funktioner $g_i \colon \Omega \to \mathbf{R}$ är konvexa. I det fallet är förstås också mängden X konvex, och den har ett icke-tomt inre om Slaters villkor är uppfyllt, dvs. om det finns en punkt $\overline{x} \in \Omega$ så att $g_i(\overline{x}) < 0$ för alla i.

Andra exempel på barriärer är den exponentiella barriärfunktionen

$$F(x) = \sum_{i=1}^{m} \mathrm{e}^{-1/g_i(x)}$$

och potensbarriärerna

$$F(x) = \sum_{i=1}^{m} (-g_i(x))^{-p},$$

där p är ett positivt tal.

Centrala vägen

Definition. Låt F vara en barriär till mängden X och antag att funktionerna F_t har unika minimipunkter $\hat{x}(t) \in \operatorname{int} X$ för alla $t \geq 0$. Kurvan $\{\hat{x}(t) \mid t \geq 0\}$ kallas då den *centrala vägen* för problemet $\min_{x \in X} f(x)$.

Notera att $\hat{x}(0)$ är mängdens analytiska centrum med avseende på barriären F. Den centrala vägen startar alltså i mängdens analytiska centrum.

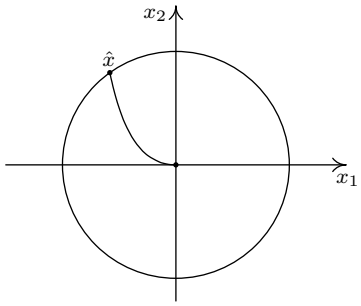
Eftersom gradienten är noll i en optimal punkt, är

$$tf'(\hat{x}(t)) + F'(\hat{x}(t)) = 0 \tag{17.2}$$

för alla punkter på den centrala vägen. För konvexa målfunktioner f och barriärfunktioner F gäller också omvändningen, dvs. $\hat{x}(t)$ ligger på den centrala vägen om och endast om ekvation (17.2) gäller.

För den till $X = \{x \in \Omega \mid g_i(x) \leq 0,\ \ i = 1, 2, \ldots, m\}$ hörande logaritmiska barriärfunktionen är speciellt

$$F'(x) = -\sum_{i=1}^{m} \frac{1}{g_i(x)} g_i'(x),$$



Figur 17.1. I figuren visas den centrala vägen för problemet att minimera funktionen $f(x) = x_1 e^{x_1+x_2}$ över $X = \{x \in \mathbf{R}^2 \mid x_1^2 + x_2^2 \leq 1\}$ med barriärfunktion $F(x) = (1 - x_1^2 - x_2^2)^{-1}$. Minimum antas i punkten $\hat{x} = (-0.5825, 0.8128)$.

så ekvationen (17.2) för den centrala vägen får i detta fall för $t > 0$ formen

$$f'(\hat{x}(t)) - \frac{1}{t}\sum_{i=1}^{m} \frac{1}{g_i(\hat{x}(t))} g_i'(\hat{x}(t)) = 0. \tag{17.3}$$

Låt oss nu betrakta ett konvext optimeringsproblem på formen

$$\text{(P)} \qquad \begin{array}{ll} \min & f(x) \\ \text{då} & g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \end{array}$$

Vi antar att Slaters villkor är uppfyllt och att problemet har en optimal lösning $\hat{x}$. Till problemet (P) hör en Lagrangefunktion

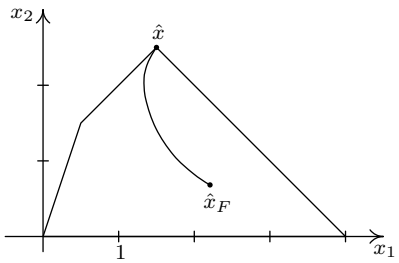
$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

och en dual funktion $\phi\colon \mathbf{R}_+^m \to \mathbf{R}$. Definiera $\hat{\lambda} \in \mathbf{R}_+^m$ genom att sätta

$$\hat{\lambda}_i = -\frac{1}{t g_i(\hat{x}(t))}.$$

På grund av ekvation (17.3) är då $L_x'(\hat{x}(t), \hat{\lambda}) = 0$, och eftersom Lagrangefunktionen är konvex i variabeln x, medför detta att funktionen $L(\cdot\,, \hat{\lambda})$ har ett minimum i punkten $\hat{x}(t)$. För den till problemet (P) duala funktionen ϕ gäller därför att

$$\phi(\hat{\lambda}) = L(\hat{x}(t), \hat{\lambda}) = f(\hat{x}(t)) - m/t.$$



Figur 17.2. Centrala vägen för LP-problemet $\min_{x \in X} f(x)$ med $f(x) = 2x_1 - 3x_2$, $X = \{x \in \mathbf{R}^2 \mid x_2 \geq 0, x_2 \leq 3x_1, x_2 \leq x_1 + 1, x_1 + x_2 \leq 4\}$ och logaritmisk barriär. Optimal punkt $\hat{x} = (1.5, 2.5)$. Punkten $\hat{x}_F$ är barriärfunktionens minimipunkt.

På grund av svag dualitet är vidare $\phi(\hat{\lambda}) \leq f(\hat{x})$, så det följer att

$$f(\hat{x}(t)) - m/t \leq f(\hat{x}).$$

Vi har därmed visat följande approximationssats, som för konvexa problem med logaritmisk barriär ger mer precis information än sats 17.1.1.

Sats 17.1.2. *För punkterna $\hat{x}(t)$ på den centrala vägen till det konvexa minimeringsproblemet (P) med logaritmisk barriär och optimal lösning $\hat{x}$ gäller att*

$$f(\hat{x}(t)) - f(\hat{x}) \leq \frac{m}{t}.$$

Notera att uppskattningen i satsen beror av antalet bivillkor m men inte av dimensionen n.

17.2 Vägföljande metoder

En strategi för att bestämma det optimala värdet till det konvexa optimeringsproblemet

$$\begin{array}{lll} \text{(P)} & \min & f(x) \\ & \text{då} & g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \end{array}$$

för två gånger kontinuerligt deriverbara mål- och bivillkorsfunktioner och med ett fel som är mindre än eller lika med ϵ, skulle i ljuset av sats 17.1.2 kunna vara att använda en logaritmisk barriär F, sätta $t = m/\epsilon$ och lösa optimeringsproblemet $\min F_t(x)$ med hjälp av Newtons metod. Strategin fungerar för små problem och med måttliga krav på noggrannhet, men man erhåller bättre resultat om man löser problemen $\min F_t(x)$ för en växande svit av t-värden till dess att $t \geq m/\epsilon$.

En enkel version av barriärmetoden eller den *vägföljande metoden*, som den också kallas, ser därför ut så här:

Vägföljande metoden

Givet startpunkter $x = x_0 \in \operatorname{int} X$ och $t = t_0 > 0$, uppdateringsparameter $\alpha > 1$ och toleransnivå $\epsilon > 0$.

Upprepa

1. Beräkna $\hat{x}(t)$ genom att minimera funktionen $F_t = tf + F$ med x som startpunkt.
2. *Uppdatera:* $x := \hat{x}(t)$.
3. *Stoppkriterium:* Avbryt om $m/t \leq \epsilon$.
4. *Öka* t*:* $t := \alpha t$.

Steget 1 kallas en *yttre iteration* eller ett *centreringssteg* eftersom det handlar om att hitta en punkt på den centrala vägen. För att minimera funktionen F_t används Newtons metod, och iterationerna i Newtons metod för att beräkna $\hat{x}(t)$ med x som startpunkt kallas *inre iterationer*.

Det är inte nödvändigt att i varje yttre iteration beräkna $\hat{x}(t)$ exakt; den centrala vägen fyller ingen annan funktion än att leda till den optimala punkten $\hat{x}$, och bra approximationer till punkter på den centrala vägen kommer också att ge upphov till en följd av punkter som konvergerar mot $\hat{x}$.

Metodens totala beräkningskostnad beror uppenbarligen av antalet inre iterationer innan stoppkriteriet utlöses och av antalet algebraiska operationer i varje Newtonsteg.

Uppdateringsparametern α

Parametern α (och startvärdet t_0) bestämmer antalet yttre iterationer som behövs för att uppnå stoppkriteriet $t \geq m/\epsilon$. För α nära 1 krävs det många yttre iterationer, men å andra sidan kräver varje yttre iteration få inre iterationer eftersom minimipunkten $x = \hat{x}(t)$ till funktionen F_t då är en mycket bra startpunkt i Newtons algoritm för problemet att minimera funktionen $F_{\alpha t}$.

För stora α-värden gäller det motsatta; det behövs få yttre iterationer, men varje yttre iterationssteg kräver nu fler Newtonsteg eftersom startpunkten $\hat{x}(t)$ ligger längre från minimipunkten $\hat{x}(\alpha t)$.

Erfarenhetsmässigt visar det sig emellertid att de två ovannämnda effekterna tenderar att balansera varandra, och totala antalet Newtonsteg är i stort sett konstant inom ett brett intervall för α. Värden mellan 10 och 20 brukar rekommenderas för praktiskt bruk.

Startvärdet t_0

Valet av startvärde t_0 är också betydelsefullt. Om t_0 är litet krävs det många yttre iterationer innan stoppkriteriet uppfylls. Om t_0 är mycket stort kräver å andra sidan den första yttre iterationen många inre iterationer för att åstadkomma en tillräckligt god approximation till punkten $\hat{x}(t_0)$ på den centrala vägen. Eftersom $f(\hat{x}(t_0)) - f(\hat{x}) \approx m/t_0$, kan det vara rimligt att välja t_0 så att m/t_0 är av ungefär samma storleksordning som $f(x_0) - f(\hat{x})$. Problemet är förstås att det optimala värdet $f(\hat{x})$ inte är känt apriori, så man får i så fall använda sig av någon lämplig uppskattning av detta. Om man exempelvis känner en tillåten punkt λ i det duala problemet med ϕ som dual funktion, så kan man använda $\phi(\lambda)$ som approximation till $f(\hat{x})$ och använda $t_0 = m/(f(x_0) - \phi(\lambda))$ som begynnelsevärde.

Startpunkten x_0

Startpunkten x_0 måste ligga i det inre av X, dvs. satisfiera samtliga bivillkor med strikt olikhet. Om en sådan punkt inte är känd i förväg, kan man använda barriärmetoden på ett artificiellt problem för att beräkna en sådan punkt eller för att konstatera att det ursprungliga problemet saknar tillåtna punkter. Proceduren kallas *fas 1* av den vägföljande metoden och fungerar så här.

Betrakta olikheterna

$$g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \tag{17.4}$$

och antag att funktionerna $g_i \colon \Omega \to \mathbf{R}$ är konvexa och två gånger kontinuerligt differentierbara. För att bestämma en punkt som uppfyller samtliga olikheter strikt eller för att avgöra att det inte finns någon sådan punkt bildar vi optimeringsproblemet

$$\begin{aligned} &\min \ s \\ &\text{då} \quad g_i(x) \leq s, \quad i = 1, 2, \ldots, m \end{aligned} \tag{17.5}$$

i variablerna x och s. Detta problem har uppenbarligen strikt tillåtna punkter – vi kan välja $x_0 \in \Omega$ godtyckligt och sedan $s_0 > \max_i g_i(x_0)$ och får på så sätt en punkt $(x_0, s_0) \in \Omega \times \mathbf{R}$ som satisfierar bivillkoren med strikt olikhet. Funktionerna $(x, s) \mapsto g_i(x) - s$ är förstås konvexa. Vi kan därför använda den vägföljande metoden på problemet (17.5) för att lösa det, och beroende på tecknet på problemets optimala värde $v_{\min}$ får vi tre fall.

1. Om $v_{\min} < 0$ har systemet (17.4) strikt tillåtna punkter. För varje tillåten punkt (x, s) till problemet (17.5) med $s < 0$ gäller att $g_i(x) < 0$ för alla i, och detta betyder att vi inte behöver lösa optimeringsproblemet (17.5) med någon större noggrannhet, utan algoritmen kan avbrytas så snart den genererat en punkt (x, s) med $s < 0$.

2. Om $v_{\min} > 0$ saknar problemet tillåtna punkter. Vi behöver inte heller nu lösa problemet med stor noggrannhet, utan kan avbryta så snart vi hittat en tillåten punkt för det duala problemet med positivt värde hos den duala funktionen eftersom detta implicerar att $v_{\min} > 0$.

3. Om $v_{\min} = 0$ och minimum antas i en punkt $(\hat{x}, \hat{s})$, där alltså $\hat{s} = 0$, är systemet (17.4) lösbart men inte strikt lösbart. Om $v_{\min} = 0$, men minimum inte antas, saknar systemet (17.4) lösningar.

   I praktiken är det förstås omöjligt att exakt avgöra att $v_{\min} = 0$, utan algoritmen slutar med slutsatsen att $|v_{\min}| < \epsilon$ för något litet positivt tal ϵ, och vi kan då endast vara säkra på att systemet $g_i(x) < -\epsilon$ saknar lösningar och att systemet $g_i(x) \leq \epsilon$ har lösningar.

Konvergensanalys

Vid ingången av yttre iteration nummer k är $t = \alpha^{k-1} t_0$. Stoppkriteriet i vägföljande metoden kommer att utlösas första gången som $m/(\alpha^{k-1} t_0) \leq \epsilon$, dvs. då $k - 1 \geq (\log(m/(\epsilon t_0))/\log \alpha$. Antalet yttre iterationer är med andra ord lika med

$$\left\lceil \frac{\log(m/(\epsilon t_0)}{\log \alpha} \right\rceil + 1$$

(för $\epsilon \leq m/t_0$).

Den vägföljande metoden fungerar därför, förutsatt att minimeringsproblemen

$$\begin{array}{ll} \min & tf(x) + F(x) \\ \text{då} & x \in \operatorname{int} X \end{array} \tag{17.6}$$

kan lösas med hjälp av Newtons metod för $t \geq t_0$, och detta gäller exempelvis om funktionen F_t uppfyller förutsättningarna i sats 15.2.4, dvs. om funktionen är starkt konvex, har en Lipschitzkontinuerlig derivata och subnivåmängden svarande mot startpunkten är sluten.

En fråga som återstår att lösa är om problemen (17.6) blir svårare och svårare, dvs. kräver fler inre iterationer, när t växer. Praktisk erfarenhet visar att så inte är fallet – i flertalet problem verkar antalet Newtonsteg vara i stort sett konstant när t växer. För problem med självkonkordanta mål- och barriärfunktioner kan man ge en exakt uppskattning av det totala antalet iterationer som behövs för att lösa optimeringsproblemet (P) med given precision, och detta kommer att vara temat i kapitel 18.

Kapitel 18

Vägföljande metoden med självkonkordant barriär

18.1 Självkonkordanta barriärer

Definition. Låt X vara en sluten konvex delmängd av $\mathbf{R}^n$ med icke-tomt inre $\operatorname{int} X$, och låt ν vara ett icke-negativt tal. En funktion $f\colon \operatorname{int} X \to \mathbf{R}$ kallas en *självkonkordant barriär till X med parameter ν*, eller kortare en *ν-självkonkordant barriär*, om funktionen är sluten, självkonkordant och icke-konstant och Newtondekrementet uppfyller olikheten

$$\lambda(f,x) \le \nu^{1/2} \tag{18.1}$$

för alla $x \in \operatorname{int} X$.

Det följer direkt av satserna 15.1.2 och 15.1.3 att olikheten (18.1) är ekvivalent med att

$$|\langle f'(x), v\rangle| \le \nu^{1/2}\|v\|_x$$

för alla vektorer $v \in \mathbf{R}^n$, vilket med andra beteckningar är detsamma som att

$$\big(Df(x)[v]\big)^2 \le \nu\, D^2 f(x)[v,v]$$

för alla $v \in \mathbf{R}^n$.

En sluten självkonkordant funktion $f\colon \Omega \to \mathbf{R}$ med $\sup_{x\in\Omega} \lambda(f,x) < 1$ är enligt sats 16.4.7 konstant, och definitionsmängden Ω är lika med hela $\mathbf{R}^n$. Det följer därför att parametern ν i en självkonkordant barriär måste vara större än eller lika med 1.

EXEMPEL 18.1.1. Funktionen $f(x) = -\ln x$ är en 1-självkonkordant barriär till intervallet $[0, \infty[$, ty funktionen är sluten och självkonkordant, och $\lambda(f, x) = 1$ för alla $x > 0$. □

EXEMPEL 18.1.2. Konvexa kvadratiska funktioner

$$f(x) = \tfrac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$$

är självkonkordanta på $\mathbf{R}^n$, men de fungerar inte som självkonkordanta barriärer, ty $\sup \lambda(f, x) = \infty$ för icke-konstanta konvexa kvadratiska funktioner f enligt exempel 15.1.2. □

Vi kommer att visa längre fram att endast delmängder av halvrum kan ha självkonkordanta barriärer, så det finns ingen självkonkordant barriär för hela $\mathbf{R}^n$.

EXEMPEL 18.1.3. Låt $g(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$ vara en icke-konstant, konvex, kvadratisk funktion, och sätt

$$f(x) = -\ln(-g(x)).$$

Då är f en 1-självkonkordant barriär till mängden $X = \{x \in \mathbf{R}^n \mid g(x) \le 0\}$.

Bevis. Låt v vara en godtycklig vektor i $\mathbf{R}^n$ och sätt för $x \in \operatorname{int} X$

$$\alpha = -\frac{1}{g(x)} Dg(x)[v] \quad \text{och} \quad \beta = -\frac{1}{g(x)} D^2 g(x)[v, v] = -\frac{1}{g(x)}\langle v, Av\rangle,$$

och notera att $\beta \ge 0$ samt att $D^3 g(x)[v, v, v] = 0$. Det följer därför av deriveringsreglerna att

$$\begin{aligned}
Df(x)[v] &= -\frac{1}{g(x)} Dg(x)[v] = \alpha, \\
D^2 f(x)[v, v] &= \frac{1}{g(x)^2}\big(Dg(x)[v]\big)^2 - \frac{1}{g(x)} D^2 g(x)[v, v] = \alpha^2 + \beta \ge 0, \\
D^3 f(x)[v, v, v] &= -\frac{2}{g(x)^3}\big(Dg(x)[v]\big)^3 + \frac{3}{g(x)^2} D^2 g(x)[v, v] Dg(x)[v] \\
&\quad - \frac{1}{g(x)} D^3 g(x)[v, v, v] = 2\alpha^3 + 3\alpha\beta.
\end{aligned}$$

Funktionen f är konvex eftersom andraderivatan är positivt semidefinit, och den är sluten eftersom $f(x) \to +\infty$ då $g(x) \to 0$. För alla $\alpha \in \mathbf{R}$ och $\beta \in \mathbf{R}_+$ är $|2\alpha^3 + 3\alpha\beta| \le 2(\alpha^2 + \beta)^{3/2}$, vilket enkelt visas genom kvadrering, och detta betyder att $\big|D^3 f(x)[v, v, v]\big| \le 2\big(D^2 f(x)[v, v]\big)^{3/2}$. Funktionen f är således självkonkordant. Vidare är förstås $\alpha^2 \le \alpha^2 + \beta$, vilket innebär att $(Df(x)[v])^2 \le D^2 f(x)[v, v]$. Funktionen f är med andra ord 1-självkonkordant. □

Av självkonkordanta barriärer kan man bilda nya med hjälp av följande tre satser.

Sats 18.1.1. *Antag att f är en ν-självkonkordant barriär till mängden X och att $\alpha \geq 1$. Då är funktionen αf en $\alpha\nu$-självkonkordant barriär till X.*

Bevis. Beviset lämnas som enkel övning □

Sats 18.1.2. *Om f är en μ-självkonkordant barriär till mängden X och g är en ν-självkonkordant barriär till mängden Y, så är summan $f + g$ en självkonkordant barriär med parameter $\mu + \nu$ till snittet $X \cap Y$. Och för varje konstant c är $f + c$ en μ-självkonkordant barriär till X.*

Bevis. Summan $f + g$ är en sluten konvex funktion, och den är självkonkordant på mängden $\operatorname{int}(X \cap Y)$ enligt sats 16.1.5. För att visa att summan är en självkonkordant barriär med parameter $(\mu + \nu)$ låter vi v vara en godtycklig vektor i $\mathbf{R}^n$ och sätter $a = D^2 f(x)[v, v]$ och $b = D^2 g(x)[v, v]$. Då är alltså per definition

$$\big(Df(x)[v]\big)^2 \leq \mu a \quad \text{och} \quad \big(Dg(x)[v]\big)^2 \leq \nu b,$$

och med hjälp av olikheten $2\sqrt{\mu\nu ab} \leq \nu a + \mu b$ mellan geometriskt och aritmetiskt medelvärde erhålls olikheten

$$\begin{aligned}\big(D(f+g)(x)[v]\big)^2 &= \big(Df(x)[v]\big)^2 + \big(Dg(x)[v]\big)^2 + 2Df(x)[v] \cdot Dg(x)[v] \\ &\leq \mu a + \nu b + 2\sqrt{\mu a \nu b} \leq \mu a + \nu b + \nu a + \mu b \\ &= (\mu + \nu)(a + b) = (\mu + \nu)\, D^2(f+g)(x)[v, v],\end{aligned}$$

som innebär att $\lambda(f + g, x) \leq (\mu + \nu)^{1/2}$.

Påståendet rörande summan $f + c$ är trivialt, ty för konstanter c är $\lambda(f, x) = \lambda(f + c, x)$. □

Sats 18.1.3. *Antag att $A\colon \mathbf{R}^m \to \mathbf{R}^n$ är en affin avbildning och att f är en ν-självkonkordant barriär till delmängden X av $\mathbf{R}^n$. Då är sammansättningen $g = f \circ A$ en ν-självkonkordant barriär till den inversa bilden $A^{-1}(X)$.*

Bevis. Beviset lämnas som övning. □

Exempel 18.1.4. Det följer av exempel 18.1.1 och satserna 18.1.2 och 18.1.3 att funktionen

$$f(x) = -\sum_{i=1}^{m} \ln(b_i - \langle a_i, x\rangle)$$

är en m-självkonkordant barriär till polyedern

$$X = \{x \in \mathbf{R}^n \mid \langle a_i, x\rangle \leq b_i, \;\; i = 1, 2, \ldots, m\}.$$ □

Att det bara är slutna konvexa delmängder av halvrum som kan ha självkonkordanta barriärer följer nu av följande sats.

Sats 18.1.4. *Om f är en ν-självkonkordant barriär till mängden X, så är*

$$\langle f'(x), y - x\rangle \leq \nu$$

för alla $x \in \operatorname{int} X$ och alla $y \in X$.

Om $x_0 \in \operatorname{int} X$ är en punkt med $c = f'(x_0) \neq 0$, så är således X en delmängd till det slutna halvrummet $\{y \in \mathbf{R}^n \mid \langle c, y\rangle \leq \nu + \langle c, x_0\rangle\}$.

Bevis. Sätt $x^t = x + t(y - x)$ och $\phi(t) = f(x^t)$. Funktionen ϕ är säkert definierad på det öppna intervallet $]\alpha, 1[$ för något negativt tal α eftersom x är en inre punkt. Vidare är

$$\phi'(t) = Df(x^t)[y - x],$$

och speciellt är alltså $\phi'(0) = Df(x)[y - x] = \langle f'(x), y - x\rangle$. Vi skall visa att $\phi'(0) \leq \nu$.

Om $\phi'(0) \leq 0$, är saken klar, så antag utan inskränkning att $\phi'(0) > 0$. På grund av ν-självkonkordans är

$$\phi''(t) = D^2 f(x^t)[y - x, y - x] \geq \nu^{-1}\big(Df(x^t)[y - x]\big)^2 = \nu^{-1}\phi'(t)^2 \geq 0.$$

Derivatan ϕ' är således växande, vilket implicerar att $\phi'(t) \geq \phi'(0) > 0$ för $t \geq 0$. Vidare är

$$\frac{d}{dt}\Big(-\frac{1}{\phi'(t)}\Big) = \frac{\phi''(t)}{\phi'(t)^2} \geq \frac{1}{\nu}$$

för alla t i intervallet $[0, 1[$, så genom att integrera den sistnämnda olikheten över intervallet $[0, \beta]$, där $\beta < 1$, erhålls olikheten

$$\frac{1}{\phi'(0)} > \frac{1}{\phi'(0)} - \frac{1}{\phi'(\beta)} = \int_0^\beta \frac{d}{dt}\Big(-\frac{1}{\phi'(t)}\Big)\, dt \geq \frac{\beta}{\nu}.$$

Följaktligen är $\phi'(0) < \nu/\beta$ för alla $\beta < 1$, vilket medför att $\phi'(0) \leq \nu$. □

Sats 18.1.5. *Antag att funktionen f är en ν-självkonkordant barriär till mängden X. Om $x \in \operatorname{int} X$, $y \in X$ och $\langle f'(x), y - x\rangle \geq 0$, så är*

$$\|y - x\|_x \leq \nu + 2\sqrt{\nu}.$$

Anmärkning. Om $x \in \operatorname{int} X$ är en minimipunkt, så är $\langle f'(x), y - x\rangle = 0$ för alla $y \in X$ eftersom $f'(x) = 0$. Följaktligen är $\|y - x\|_x \leq \nu + 2\sqrt{\nu}$ för alla $y \in X$ om x är en minimipunkt.

Bevis. Sätt $r = \|y - x\|_x$. Om $r \leq \sqrt{\nu}$ har vi ingenting att visa, så antag att $r > \sqrt{\nu}$, och betrakta för $\alpha = \sqrt{v}/r$ punkten $z = x + \alpha(y - x)$, som ligger i det inre av X eftersom $\alpha < 1$. Genom att utnyttja föregående sats med z istället för x, förutsättningen $\langle f'(x), y - x\rangle \geq 0$, sats 16.3.2 samt att $z - x = \alpha(y - x)$ och $y - z = (1 - \alpha)(y - x)$, får vi följande kedja av olikheter och likheter:

$$\begin{aligned}\nu &\geq \langle f'(z), y - z\rangle = (1 - \alpha)\langle f'(z), y - x\rangle \geq (1 - \alpha)\langle f'(z) - f'(x), y - x\rangle \\ &= \frac{1 - \alpha}{\alpha}\langle f'(z) - f'(x), z - x\rangle \geq \frac{1 - \alpha}{\alpha} \cdot \frac{\|z - x\|_x^2}{1 + \|z - x\|_x} \\ &= \frac{(1 - \alpha)\alpha\|y - x\|_x^2}{1 + \alpha\|y - x\|_x} = \frac{r\sqrt{\nu} - \nu}{1 + \sqrt{\nu}}.\end{aligned}$$

Olikheten mellan ytterleden ger efter förenkling $r \leq \nu + 2\sqrt{\nu}$, vilket är den sökta olikheten. □

Givet en självkonkordant funktion $f\colon X \to \mathbf{R}$ med motsvarande lokala seminorm $\|\cdot\|_x$ sätter vi

$$\mathcal{E}(x; r) = \{y \in \mathbf{R}^n \mid \|y - x\|_x \leq r\}.$$

Om funktionen är icke-degenererad, är seminormen en norm i varje punkt $x \in \operatorname{int} X$, och mängden $\mathcal{E}(x; r)$ är en sluten ellipsoid i $\mathbf{R}^n$ med axelriktningar som bestäms av egenvektorerna till andraderivatan $f''(x)$.

För icke-degenererade självkonkordanta barriärer har vi nu följande korollarium till sats 18.1.5.

Sats 18.1.6. *Antag att f är en icke-degenererad ν-självkonkordant barriär till den slutna konvexa mängden X. Då antar f ett minimum om och endast om mängden X är begränsad. Minimipunkten $\hat{x}_f \in \operatorname{int} X$ är i så fall unik, och*

$$\mathcal{E}(\hat{x}_f; 1) \subseteq X \subseteq \mathcal{E}(\hat{x}_f; \nu + 2\sqrt{\nu}).$$

Anmärkning. Slutna självkonkordanta funktioner, vars definitionsmängd inte innehåller någon linje, är icke-degenererade, så vi behöver därför inte explicit ange att en självkonkordant barriär till en kompakt mängd skall vara icke-degenererad.

Bevis. Slutna konvexa funktioner har slutna subnivåmängder, så om X är en begränsad mängd, är varje subnivåmängd $\{x \in \operatorname{int} X \mid f(x) \leq \alpha\}$ både sluten och begränsad, dvs. kompakt, och detta medför att f har ett minimum. Och en icke-degenererad konvex funktions minimipunkt är nödvändigtvis unik.

Antag omvänt att f har en minimipunkt $\hat{x}_f$. Enligt anmärkningen efter sats 18.1.5 är då $\|y-\hat{x}_f\|_{\hat{x}_f} \le \nu+2\sqrt{\nu}$ för alla $y \in X$, vilket är den högra inklusionen i sats 18.1.6, och inklusionen medför förstås att X är en begränsad mängd.

Den återstående vänstra inklusionen följer i sin tur av sats 16.3.2, som medför att den öppna ellipsoiden $\{y \in \mathbf{R}^n \mid \|y-x\|_x < 1\}$ är en delmängd av $\operatorname{int} X$ för varje val av $x \in \operatorname{int} X$. Den öppna ellipsoidens tillslutning $\mathcal{E}(x;1)$ är därför en delmängd av X, och det speciella valet $x = \hat{x}_f$ ger den vänstra inklusionen. □

För självkonkordanta barriärer till en mängd X behöver vi kunna jämföra motsvarande lokala normer $\|v\|_x$ och $\|v\|_y$ av en vektor v i godtyckliga inre punkter x och y, och för att åstadkomma detta behöver vi ett mått på avståndet från y till x relativt avståndet från y till randen av X utmed halvlinjen från y genom x. Följande definition ger oss det relevanta måttet.

Definition. Låt X vara en konvex sluten delmängd av $\mathbf{R}^n$ med icke-tomt inre. För $y \in \operatorname{int} X$ definierar vi en funktion $\pi_y \colon \mathbf{R}^n \to \mathbf{R}_+$ genom att sätta

$$\pi_y(x) = \inf\{t > 0 \mid y + t^{-1}(x-y) \in X\}.$$

Uppenbarligen är $\pi_y(y) = 0$. För att bestämma $\pi_y(x)$ om $x \neq y$ betraktar vi halvlinjen från y genom x; om halvlinjen skär randen till X i en punkt z, så är $\pi_y(x) = \|x-y\|/\|z-y\|$ (med avseende på godtyckliga normer), och om hela halvlinjen ligger i X, är $\pi_y(x) = 0$. Vi noterar att $\pi_y(x) < 1$ för inre punkter x, att $\pi_y(x) = 1$ för randpunkter x och att $\pi_y(x) > 1$ för punkter utanför X.

Funktionen π_y kan också uttryckas i termer av den i avsnitt 6.10 definierade Minkowskifunktionalen; som läsaren lätt kan verifiera är nämligen

$$\pi_y(x) = \phi_{-y+X}(x-y),$$

där ϕ_{-y+X} är Minskowskifunktionalen till mängden $-y+X$.

Vi kommer att behöva följande enkla uppskattning av $\pi_y(x)$.

Sats 18.1.7. *Låt X vara en kompakt konvex mängd, låt x och y vara punkter i det inre av X, och antag att*

$$B(x,r) \subseteq X \subseteq \overline{B}(0;R),$$

där bollarna är givna med avseende på en godtycklig norm $\|\cdot\|$. Då är

$$\pi_y(x) \le \frac{2R}{2R+r}.$$

Bevis. Olikheten är trivialt sann om $x = y$, så antag att $x \neq y$. Strålen från y genom x skär då randen till X i en punkt z och $\|z-y\| = \|z-x\|+\|x-y\|$. Vidare är $\|z - x\| \geq r$ och $\|x - y\| \leq 2R$, så det följer att

$$\pi_y(x) = \frac{\|x-y\|}{\|z-y\|} = \Big(1 + \frac{\|z-x\|}{\|x-y\|}\Big)^{-1} \leq \Big(1 + \frac{r}{2R}\Big)^{-1} = \frac{2R}{2R+r}. \qquad \square$$

Riktningsderivatan $\langle f'(x), v\rangle$ till en ν-självkonkordant barriärfunktion f begränsas per definiton av $\sqrt{\nu}\|v\|_x$. Vår nästa sats visar att samma riktningsderivata också begränsas av en konstant gånger $\|v\|_y$ för en godtycklig punkt y i funktionens definitionsmängd. Vidare jämförs de båda lokala normerna $\|v\|_x$ och $\|v\|_y$.

Sats 18.1.8. *Låt f vara en ν-självkonkordant barriär till X, och låt x och y vara två punkter i det inre av X. För alla vektorer v är*

$$|\langle f'(x), v\rangle| \leq \frac{\nu}{1-\pi_y(x)}\,\|v\|_y \tag{18.2}$$

och

$$\|v\|_x \leq \frac{\nu + 2\sqrt{\nu}}{1-\pi_y(x)}\|v\|_y. \tag{18.3}$$

Bevis. De båda olikheterna gäller om $y = x$ eftersom

$$|\langle f'(x), v\rangle| \leq \sqrt{\nu}\|v\|_x \leq \nu\|v\|_x$$

och $\pi_x(x) = 0$. De gäller också om $\|v\|_y = 0$, dvs. om vektorn v ligger i f:s recessiva delrum, ty då är $\|v\|_x = 0$ och $\langle f'(x), v\rangle = 0$. Antag därför fortsättningsvis att $y \neq x$ och att $\|v\|_y \neq 0$.

Betrakta först fallet $\|v\|_y = 1$, och låt s vara ett godtyckligt tal $> \nu+2\sqrt{\nu}$. På grund av satserna 16.3.2 och 18.1.5 gäller då följande påståenden:

(i) De båda punkterna $y \pm v$ ligger i X.

(ii) Minst en av de båda punkterna $x \pm \dfrac{s}{\|v\|_x}v$ ligger utanför X.

Enligt definitionen av $\pi_y(x)$ finns det vidare en vektor $z \in X$ sådan att

$$x = y + \pi_y(x)(z - y),$$

och eftersom

$$x \pm (1-\pi_y(x))v = \pi_y(x)z + (1-\pi_y(x))(y \pm v),$$

följer det av konvexiteten att

(iii) De båda punkterna $x \pm (1-\pi_y(x))v$ ligger i X.

Det följer nu av (iii) och sats 18.1.4 att

$$\langle f'(x), \pm v\rangle = \frac{1}{1-\pi_y(x)}\langle f'(x), x \pm (1-\pi_y(x))v - x\rangle \le \frac{\nu}{1-\pi_y(x)},$$

vilket innebär att

$$|\langle f'(x), v\rangle| \le \frac{\nu}{1-\pi_y(x)}.$$

Detta visar olikheten (18.2) för vektorer v med $\|v\|_y = 1$, och om v är en godtycklig vektor med $\|v\|_y \neq 0$, så får vi olikheten i satsen genom att i olikheten ovan byta v mot $v/\|v\|_y$.

Genom att kombinera de båda påståendena (ii) och (iii) drar vi slutsatsen att

$$1 - \pi_y(x) < \frac{s}{\|v\|_x},$$

dvs. att

$$\|v\|_x < \frac{s}{1-\pi_y(x)} = \frac{s}{1-\pi_y(x)}\|v\|_y,$$

och eftersom detta gäller för alla $s > \nu + 2\sqrt{\nu}$, följer det att

$$\|v\|_x \le \frac{\nu + 2\sqrt{\nu}}{1-\pi_y(x)}\|v\|_y.$$

Detta visar olikheten (18.3) i fallet $\|v\|_y = 1$, och eftersom olikheten är homogen, gäller den generellt. □

Definition. Låt $\|\cdot\|_x$ vara den till en två gånger differentierbar konvex funktion $f\colon X \to \mathbf{R}$, där $X \subseteq \mathbf{R}^n$, hörande lokala seminormen i punkten x. Med motsvarande *duala lokala norm* menas funktionen $\|\cdot\|_x^*\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ som definieras av att

$$\|v\|_x^* = \sup_{\|w\|_x \le 1} \langle v, w\rangle.$$

Man verifierar omedelbart att den duala normen är subadditiv och homogen, dvs. att $\|v+w\|_x^* \le \|v\|_x^* + \|w\|_x^*$ och $\|\lambda v\|_x^* = |\lambda|\|v\|_x^*$ för alla $v, w \in \mathbf{R}^n$ och alla reella tal λ, men $\|\cdot\|_x^*$ är bara en äkta norm på hela $\mathbf{R}^n$ i punkter x där andraderivatan $f''(x)$ är positivt definit, ty $\|v\|_x^* = \infty$ om v är en nollskild vektor i nollrummet $\mathcal{N}(f''(x))$ eftersom $\|tv\|_x = 0$ för alla $t \in \mathbf{R}$ och $\langle v, tv\rangle = t\|v\|^2 \to \infty$ då $t \to \infty$. Däremot är $\|\cdot\|_x^*$ alltid en äkta norm på delrummet $\mathcal{N}(f''(x))^\perp$. Se övning 18.2.

För Newtondekrementet $\lambda(f, x)$ gäller enligt sats 15.1.3 att

$$\lambda(f, x) = \|f'(x)\|_x^*.$$

Följande ”Cauchy–Schwarz olikhet” gäller för den lokala seminormen.

Sats 18.1.9. *Om* $\|v\|_x^* < \infty$ *så är*

$$|\langle v, w\rangle| \le \|v\|_x^* \|w\|_x$$

för alla vektorer w.

Bevis. Om $\|w\|_x \neq 0$ så är $\pm w/\|w\|_x$ två vektorer med lokal seminorm 1, varför det följer av den duala normens definition att

$$\pm\frac{1}{\|w\|_x}\langle v, w\rangle = \langle v, \pm w/\|w\|_x\rangle \le \|v\|_x^*,$$

och olikheten i satsen följer nu efter multiplikation med $\|w\|_x$.

Om istället $\|w\|_x = 0$, så är $\|tw\|_x = 0$ för alla reella tal t, och det följer av supremumdefinitionen att $t\langle v, w\rangle = \langle v, tw\rangle \le \|v\|_x^* < \infty$ för alla t, vilket medför att $\langle v, w\rangle = 0$. Så olikheten gäller även i detta fall. □

Vi kommer att behöva uppskatta $\|v\|_x^*$ på olika sätt, och vår första uppskattning är i termer av bredden hos mängden X i olika riktningar, vilket motiverar följande definition.

Definition. Till varje icke-tom delmängd X av $\mathbf{R}^n$ associerar vi en funktion $\mathrm{Var}_X\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ genom att sätta

$$\mathrm{Var}_X(v) = \sup_{x\in X}\langle v, x\rangle - \inf_{x\in X}\langle v, x\rangle.$$

För begränsade mängder X är uppenbarligen $\mathrm{Var}_X(v)$ ett ändligt tal för varje vektor v, och för enhetsvektorer v anger $\mathrm{Var}_X(v)$ bredden hos mängden X i riktningen v.

Nästa sats visar hur man kan uppskatta den duala lokala normen $\|\cdot\|_x^*$ med hjälp av Var_X.

Sats 18.1.10. *Antag att* $f\colon X \to \mathbf{R}$ *är en sluten självkonkordant funktion med en öppen konvex delmängd* X *av* $\mathbf{R}^n$ *som definitionsmängd, och låt* $\|\cdot\|_x^*$ *vara den till funktionen* f *associerade duala lokala normen i punkten* $x \in X$. *För varje* $v \in \mathbf{R}^n$ *är då*

$$\|v\|_x^* \le \mathrm{Var}_X(v).$$

Bevis. Det följer av sats 16.3.2 att punkten y ligger i $\mathrm{cl}\, X$ om $x \in X$ och $\|y - x\|_x \le 1$. Följaktligen är

$$\begin{aligned}\|v\|_x^* &= \sup_{\|w\|_x\le 1}\langle v, w\rangle = \sup_{\|y-x\|_x\le 1}\langle v, y - x\rangle \le \sup_{y\in \mathrm{cl}\, X}\langle v, y - x\rangle = \sup_{y\in X}\langle v, y - x\rangle \\ &= \sup_{y\in X}\langle v, y\rangle - \langle v, x\rangle \le \sup_{y\in X}\langle v, y\rangle - \inf_{y\in X}\langle v, y\rangle = \mathrm{Var}_X(v).\end{aligned}$$

□

Vi har tidigare definierat en sluten konvex mängd X:s analytiska centrum med avseende på en given barriär som barriärens unika minimipunkt, förutsatt att det existerar en sådan. Enligt sats 18.1.6 har varje kompakt, konvex mängd med icke-tomt inre ett analytiskt centrum med avseende på varje given ν-självkonkordant barriär. Vi kan nu erhålla en övre begränsning på den duala lokala normen $\|\cdot\|_x^*$ i en godtycklig punkt x i termer av parametern ν och den duala lokala normen i det analytiska centret.

Sats 18.1.11. *Låt X vara en kompakt konvex mängd, och låt $\hat{x}_f$ vara mängdens analytiska centrum med avseende på en ν-självkonkordant barriär f. För alla vektorer $v \in \mathbf{R}^n$ och alla $x \in \operatorname{int} X$ är då*

$$\|v\|_x^* \leq (\nu + 2\sqrt{\nu})\|v\|_{\hat{x}_f}^*.$$

Bevis. Sätt $B_1 = \mathcal{E}(x; 1)$ och $B_2 = \mathcal{E}(\hat{x}_f; \nu + 2\sqrt{\nu})$. Satserna 16.3.2 och 18.1.6 ger oss inklusionerna $B_1 \subseteq X \subseteq B_2$, så det följer med hjälp av den duala lokala normens definition att

$$\begin{aligned}\|v\|_x^* &= \sup_{\|w\|_x \leq 1} \langle v, w\rangle = \sup_{y \in B_1} \langle v, y - x\rangle \leq \sup_{y \in B_2} \langle v, y - x\rangle \\ &= \langle v, \hat{x}_f - x\rangle + \sup_{y \in B_2} \langle v, y - \hat{x}_f\rangle = \langle v, \hat{x}_f - x\rangle + \sup_{\|w\|_{\hat{x}_f} \leq \nu + 2\sqrt{\nu}} \langle v, w\rangle \\ &= \langle v, \hat{x}_f - x\rangle + (\nu + 2\sqrt{\nu})\|v\|_{\hat{x}_f}^*.\end{aligned}$$

Eftersom $\|-v\|_x^* = \|v\|_x^*$, kan vi nu utan inskränkning anta att $\langle v, \hat{x}_f - x\rangle \leq 0$, och då följer olikheten i satsen. □

18.2 Vägföljande metoden

Standardform

Låt oss säga att ett konvext optimeringsproblem har *standardform* om det är presenterat på formen

$$\begin{array}{ll}\min & \langle c, x\rangle \\ \text{då} & x \in X\end{array}$$

där X är en kompakt konvex mängd med icke-tomt inre och X har försetts med en ν-självkonkordant barriärfunktion F.

Anmärkning. Man kan visa att varje kompakt konvex mängd X har en barriärfunktion, men för att barriärfunktionen skall vara användbar i konkreta optimeringsproblem behöver den vara explicit given så att man på ett effektivt sätt kan beräkna dess partiella första- och andraderivator.

Antagandet att mängden X är begränsad är naturligtvis inte speciellt inskränkande för problem med ändliga optimala värden, ty vi kan alltid modifiera sådan problem genom att addera artificiella, mycket stora begränsningar på variablerna.

Vi påminner också om att ett godtyckligt konvext problem kan transformeras till ett ekvivalent konvext problem med linjär målfunktion med hjälp av epigrafen. (Se kapitel 9.3.)

EXEMPEL 18.2.1. Varje LP-problem med ändliga värden kan efter lämpliga transformationer skrivas på standardform. Genom att först identifiera det affina höljet till polyedern av tillåtna punkter med $\mathbf{R}^n$ för lämpligt n, kan vi utan inskränkning anta att polyedern har ett icke-tomt inre, och genom att sedan vid behov tillfoga stora begränsningar på variablerna kan vi också anta att vår polyder X av tillåtna punkter är kompakt. Om vi sedan skriver polyedern på formen

$$X = \{x \in \mathbf{R}^n \mid \langle c_i, x\rangle \leq b_i,\ i = 1, 2, \ldots, m\}, \tag{18.4}$$

så är funktionen

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - \langle c_i, x\rangle)$$

en m-självkonkordant barriär till X. □

EXEMPEL 18.2.2. Konvexa kvadratiska problem, dvs. problem av typen

$$\begin{array}{ll} \min & g(x) \\ \text{då} & x \in X \end{array}$$

där g är en konvex kvadratisk funktion och X är en begränsad polyeder i $\mathbf{R}^n$ med icke-tomt inre, får på epigrafform och med artificiell begränsning M på den nya variabeln s, formen

$$\begin{array}{ll} \min & s \\ \text{då} & (x, s) \in Y \end{array}$$

där $Y = \{(x, s) \in \mathbf{R}^n \times \mathbf{R} \mid x \in X,\ g(x) \leq s \leq M\}$ är en kompakt konvex mängd med icke-tomt inre. Om polyedern X ges som ett snitt av halvrum som i (18.4), så ser vi med hjälp av resultatet i exempel 18.1.3 att funktionen

$$F(x, s) = -\sum_{i=1}^{m} \ln(b_i - \langle c_i, x\rangle) - \ln(s - g(x)) - \ln(M - s)$$

är en $(m + 2)$-självkonkordant barriär till Y. □

Centrala vägen

Vi skall nu studera den vägföljande metoden för standardproblemet

$$\text{(SP)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & x \in X \end{array}$$

där alltså X är en kompakt konvex delmängd av $\mathbf{R}^n$ med icke-tomt inre och F är en ν-självkonkordant barriär till X. Problemets ändliga minimivärde betecknas $v_{\min}$.

För $t \geq 0$ sätter vi

$$F_t(x) = t\langle c, x\rangle + F(x).$$

Funktionerna $F_t\colon \operatorname{int} X \to \mathbf{R}$ är slutna och självkonkordanta, och eftersom mängden X är kompakt, har varje funktion F_t en unik minimipunkt $\hat{x}(t)$. Den centrala vägen $\{\hat{x}(t) \mid t \geq 0\}$ är med andra ord väldefinierad, och punkterna på den uppfyller ekvationen

$$tc + F'(\hat{x}(t)) = 0. \tag{18.5}$$

Startpunkten $\hat{x}(0)$ är per definition mängdens analytiska centrum $\hat{x}_F$ med avseende på den givna barriären F.

Vi använder Newtons metod för att bestämma minimipunkten $\hat{x}(t)$ och behöver därför i en godtycklig punkt $x \in \operatorname{int} X$ kunna beräkna Newtonsteget och Newtondekrementet med avseende på funktionen F_t.

Eftersom $F_t''(x) = F''(x)$, är den lokala normen $\|v\|_x$ av en vektor med avseende på funktionen F_t densamma för alla $t \geq 0$, nämligen

$$\|v\|_x = \sqrt{\langle v, F''(x)v\rangle}.$$

Däremot beror Newtonsteg och Newtondekrement av t; Newtonsteget i punkten x är lika med $-F''(x)^{-1}F_t'(x)$ för funktionen F_t, och för dekrementet gäller att

$$\lambda(F_t, x) = \sqrt{\langle F_t'(x), F''(x)^{-1}F_t'(x)\rangle} = \|F''(x)^{-1}F_t'(x)\|_x.$$

Stoppkriteriet i den vägföljande metoden kommer att formuleras med hjälp av följande sats.

Sats 18.2.1. *(i) För punkterna $\hat{x}(t)$ på den centrala vägen till optimeringsproblemet (SP) är*

$$\langle c, \hat{x}(t)\rangle - v_{\min} \leq \frac{\nu}{t}.$$

(ii) För $t > 0$ och alla punkter $x \in \operatorname{int} X$ som uppfyller villkoret

$$\lambda(F_t, x) \leq \kappa < 1$$

är vidare

$$\langle c, x\rangle - v_{\min} \le \frac{\nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}}{t}.$$

Bevis. (i) På grund av ekvation (18.5) är $c = -t^{-1}F'(\hat{x}(t))$. Det följer därför av sats 18.1.4 att

$$\langle c, \hat{x}(t)\rangle - \langle c, y\rangle = \frac{1}{t}\langle F'(\hat{x}(t)), y - \hat{x}(t)\rangle \le \frac{\nu}{t}$$

för alla $y \in X$, och genom att som y välja en optimal punkt till problemet (SP) erhåller vi olikheten i (i).

(ii) Eftersom $\langle c, x\rangle - v_{\min} = (\langle c, x\rangle - \langle c, \hat{x}(t)\rangle) + (\langle c, \hat{x}(t)\rangle - v_{\min})$, räcker det på grund av den redan bevisade olikheten i (i) att visa att

$$\langle c, x\rangle - \langle c, \hat{x}(t)\rangle \le \frac{\kappa}{1-\kappa} \cdot \frac{\sqrt{\nu}}{t} \tag{18.6}$$

om $x \in \operatorname{int} X$ och $\lambda(F_t, x) \le \kappa < 1$. Men det följer av sats 16.4.6 att

$$\|x - \hat{x}(t)\|_{\hat{x}(t)} \le \frac{\lambda(F_t, x)}{1 - \lambda(F_t, x)} \le \frac{\kappa}{1-\kappa},$$

så genom att utnyttja att $tc = -F'(\hat{x}(t))$ och att F är ν-självkonkordant, får vi följande kedja av likheter och olikheter:

$$\begin{aligned} t(\langle c, x\rangle - \langle c, \hat{x}(t)\rangle) &= -\langle F'(\hat{x}(t)), x - \hat{x}(t)\rangle \le \|F'(\hat{x}(t))\|^*_{\hat{x}(t)}\|x - \hat{x}(t)\|_{\hat{x}(t)} \\ &= \lambda(F, \hat{x}(t))\|x - \hat{x}(t)\|_{\hat{x}(t)} \le \sqrt{\nu}\frac{\kappa}{1-\kappa}. \end{aligned}$$

Därmed är olikheten (18.6) bevisad och beviset klart. □

Algoritmen

Den vägföljande algoritmen för att lösa standardproblemet

$$\text{(SP)} \qquad \begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & x \in X \end{array}$$

fungerar nu i korthet på följande sätt.

Vi startar med ett parametervärde $t_0 > 0$ och en punkt $x_0 \in \operatorname{int} X$, som ligger tillräckligt nära punkten $\hat{x}(t_0)$ på den centrala vägen. "Tillräckligt nära" uttrycks med hjälp av Newtondekrementet $\lambda(F_{t_0}, x_0)$, som skall vara tillräckligt litet.

Sedan uppdaterar vi parametern t genom att sätta $t_1 = \alpha t_0$ för lämpligt $\alpha > 1$ och minimerar funktionen F_{t_1} med Newtons dämpade metod och med

x_0 som startpunkt. Iterationerna i Newtons metod avbryts när man erhållit en punkt x_1, som ligger tillräckligt nära minimipunkten $\hat{x}(t_1)$ till F_{t_1}.

Därefter upprepas proceduren med $t_2 = \alpha t_1$ som ny parameter och med x_1 som startpunkt i Newtons metod för minimering av funktionen F_{t_2}, etc. Som resultat erhålls en följd t_0, x_0, t_1, x_1, t_2, x_2, ..., och proceduren avbryts när t_k blivit tillräckligt stort med x_k som approximativ optimal punkt.

Av den skissartade beskrivningen framgår att vi behöver två parameter i algoritmen, en parameter α för att beskriva uppdateringen av t, och en parameter κ för att definiera stoppkriteriet i Newtons metod. Vi skall uppskatta totala antalet inre iterationer, och uppskattningen blir enklast och tydligast om man skriver uppdateringsparametern α på formen $\alpha = 1 + \gamma/\sqrt{\nu}$.

Nedanstående precisa formulering av den vägföljande algoritmen innehåller därför parametrarna γ och κ. Tillägget "fas 2" beror på att det behövs ytterligare en fas för att generera tillåtna startvärden x_0 och t_0.

Vägföljande algoritmen, fas 2

Givet en uppdateringsparameter $\gamma > 0$, en omgivningsparameter $0 < \kappa < 1$, en toleransnivå $\epsilon > 0$, en startpunkt $x_0 \in \operatorname{int} X$ samt ett startvärde $t_0 > 0$ som uppfyller $\lambda(F_{t_0}, x_0) \leq \kappa$.

1. *Initiera:* $x := x_0$ och $t := t_0$.
2. *Stoppkriterium:* Stoppa om $\epsilon t \geq \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}$.
3. *Öka annars* t: $t := (1 + \gamma/\sqrt{\nu})t$.
4. *Uppdatera* x *genom att använda Newtons dämpade metod på funktionen* F_t *med det aktuella* x*-värdet som startpunkt:*
   (i) Beräkna Newtondekrementet $\lambda = \lambda(F_t, x)$.
   (ii) Avbryt Newtons metod om $\lambda \leq \kappa$ och gå till punkt 2.
   (iii) Beräkna annars Newtonsteget $\Delta x_{\mathrm{nt}} = -F''(x)^{-1}F_t'(x)$.
   (iv) *Uppdatera:* $x := x + (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$, och gå till (i).

Vi kan nu visa följande konvergensresultat.

Sats 18.2.2. *Antag att ovanstående vägföljande algoritm används på standardproblemet (SP) med ν-självkonkordant barriär F. Då stoppar algoritmen med en punkt $x \in \operatorname{int} X$ som uppfyller*

$$\langle c, x\rangle - v_{\min} \leq \epsilon.$$

För varje yttre iteration är antalet inre iterationer i Newtons algoritm begränsat av en konstant K, och det totala antalet inre iterationer i den vägföljande algoritmen begränsas av

$$C\sqrt{\nu}\,\ln\Big(\frac{\nu}{t_0\epsilon} + 1\Big)$$

där konstanterna K och C bara beror av κ och γ.

Bevis. Låt oss börja med att undersöka hur den inre loopen 4 i algoritmen fungerar.

Varje gång algoritmen passerar förbi rad 2 så gör den det med en punkt $x \in \operatorname{int} X$, som hör till ett t-värde med Newtondekrement $\lambda(F_t, x) \le \kappa$. Under steg 4 minimeras sedan funktionen F_s, där $s = (1 + \gamma/\sqrt{\nu})t$, med hjälp av Newtons dämpade metod med $y_0 = x$ som startpunkt. Punkterna y_0, y_1, y_2, ... som genereras av metoden ligger i $\operatorname{int} X$ på grund av sats 16.3.2, och stoppvillkoret $\lambda(F_s, y_k) \le \kappa$ medför enligt sats 16.5.1 att algoritmen terminerar efter högst $\big\lfloor \big(F_s(x) - F_s(\hat{x}(s))\big)/\rho(-\kappa) \big\rfloor$ iterationer, där ρ är funktionen

$$\rho(u) = -u - \ln(1-u).$$

Vi skall visa att det finns en konstant K, som bara beror av parametrarna κ och γ, så att

$$\left\lfloor \frac{F_s(x) - F_s(\hat{x}(s))}{\rho(-\kappa)} \right\rfloor \le K,$$

och för den skull behöver vi uppskatta differensen $F_s(x) - F_s(\hat{x}(s))$, vilket vi gör i nästa lemma.

Lemma 18.2.3. *Antag att $\lambda(F_t, x) \le \kappa < 1$. För alla $s > 0$ är då*

$$F_s(x) - F_s(\hat{x}(s)) \le \rho(\kappa) + \frac{\kappa\sqrt{\nu}}{1-\kappa} \cdot \Big|\frac{s}{t} - 1\Big| + \nu\,\rho(1 - s/t).$$

Bevis för lemmat. Vi börjar med omskrivningen

$$F_s(x) - F_s(\hat{x}(s)) = \big(F_s(x) - F_s(\hat{x}(t))\big) + \big(F_s(\hat{x}(t)) - F_s(\hat{x}(s))\big). \tag{18.7}$$

Genom att utnyttja att $tc = -F'(\hat{x}(t))$ och använda olikheten

$$|\langle F'(\hat{x}(t)), v\rangle| \le \lambda(F, \hat{x}(t))\|v\|_{\hat{x}(t)} \le \sqrt{\nu}\|v\|_{\hat{x}(t)}$$

får vi följande uppskattning av den första differensen i högerledet av (18.7):

$$\begin{aligned} F_s(x) - F_s(\hat{x}(t)) &= F_t(x) - F_t(\hat{x}(t)) + (s-t)\langle c, x - \hat{x}(t)\rangle \\ &= F_t(x) - F_t(\hat{x}(t)) - (s/t - 1)\langle F'(\hat{x}(t)), x - \hat{x}(t)\rangle \\ &\le F_t(x) - F_t(\hat{x}(t)) + |s/t - 1|\,\sqrt{\nu}\,\|x - \hat{x}(t)\|_{\hat{x}(t)}. \end{aligned} \tag{18.8}$$

Enligt sats 16.4.6 är

$$F_t(x) - F_t(\hat{x}(t)) \le \rho(\lambda(F_t, x)) \le \rho(\kappa)$$

och

$$\|x - \hat{x}(t)\|_{\hat{x}(t)} \le \frac{\lambda(F_t, x)}{1 - \lambda(F_t, x)} \le \frac{\kappa}{1-\kappa}.$$

Det följer därför av olikheten (18.8) att

$$F_s(x) - F_s(\hat{x}(t)) \leq \rho(\kappa) + \Big|\frac{s}{t} - 1\Big| \cdot \frac{\kappa\sqrt{\nu}}{1-\kappa}. \tag{18.9}$$

Det återstår nu att uppskatta den andra differensen

$$\begin{aligned} \phi(s) &= F_s(\hat{x}(t)) - F_s(\hat{x}(s)) \\ &= s\langle c, \hat{x}(t)\rangle - s\langle c, \hat{x}(s)\rangle + F(\hat{x}(t)) - F(\hat{x}(s)) \end{aligned} \tag{18.10}$$

i högerledet av (18.7).

Eftersom funktionen F är två gånger kontinuerligt differentierbar och andraderivatan $F''(x)$ är icke-singulär överallt, följer det av ekvationen

$$sc + F'(\hat{x}(s)) = 0$$

för den centrala vägen och implicita funktionssatsen att funktionen $\hat{x}(s)$ är en kontinuerligt differentierbar funktion. Implicit derivering ger

$$c + F''(\hat{x}(s))\hat{x}'(s) = 0,$$

vilket betyder att

$$\hat{x}'(s) = -F''(\hat{x}(s))^{-1}c.$$

Det följer nu av (18.10) att differensen $\phi(s)$ är kontinuerligt deriverbar med derivata

$$\begin{aligned} \phi'(s) &= \langle c, \hat{x}(t)\rangle - \langle c, \hat{x}(s)\rangle - s\langle c, \hat{x}'(s)\rangle - \langle F'(\hat{x}(s), \hat{x}'(s)\rangle \\ &= \langle c, \hat{x}(t) - \hat{x}(s)\rangle - s\langle c, \hat{x}'(s)\rangle + s\langle c, \hat{x}'(s)\rangle \\ &= \langle c, \hat{x}(t) - \hat{x}(s)\rangle, \end{aligned}$$

och ytterligare en derivering ger

$$\begin{aligned} \phi''(s) &= -\langle c, \hat{x}'(s)\rangle = \langle c, F''(\hat{x}(s))^{-1}c\rangle \\ &= \langle s^{-1}F'(\hat{x}(s)), s^{-1}F''(\hat{x}(s))^{-1}F'(\hat{x}(s))\rangle \\ &= s^{-2}\langle F'(\hat{x}(s)), F''(\hat{x}(s))^{-1}F'(\hat{x}(s))\rangle = s^{-2}\lambda(F, \hat{x}(s))^2 \leq \nu s^{-2}. \end{aligned}$$

Notera nu att $\phi(t) = \phi'(t) = 0$. Genom att integrera olikheten för $\phi''(s)$ över intervallet $[t, u]$ erhåller vi därför för $u \geq t$ uppskattningen

$$\phi'(u) = \phi'(u) - \phi'(t) \leq \int_t^u \nu s^{-2}\, ds = \nu(t^{-1} - u^{-1}),$$

och ytterligare en integration över intervallet $[t, s]$ leder för $s \geq t$ till olikheten

$$F_s(\hat{x}(t)) - F_s(\hat{x}(s)) = \phi(s) = \int_t^s \phi'(u)\,du \leq \nu \int_t^s (t^{-1} - u^{-1})\,du \\ = \nu\big(\frac{s}{t} - 1 - \ln\frac{s}{t}\big) = \nu\,\rho(1 - s/t). \tag{18.11}$$

Samma slutsats erhålls också om $s < t$ genom att först integrera olikheten för $\phi''(s)$ över intervallet $[u, t]$, och sedan den resulterande olikheten för $\phi'(u)$ över intervallet $[s, t]$.

Olikheten i lemmat följer nu slutligen av ekvation (18.7) och uppskattningarna (18.9) och (18.11). □

Fortsättning av beviset för sats 18.2.2. Genom att använde lemmats uppskattning av differensen $F_s(x) - F_s(\hat{x}(s))$ för $s = (1 + \gamma/\sqrt{\nu})t$ erhåller man olikheten

$$\left\lfloor \frac{F_s(x) - F_s(\hat{x}(s))}{\rho(-\kappa)} \right\rfloor \leq \left\lfloor \frac{\rho(\kappa) + \gamma\kappa(1-\kappa)^{-1} + \nu\,\rho(-\gamma\nu^{-1/2})}{\rho(-\kappa)} \right\rfloor.$$

För $u < 0$ är $\rho(u) = -u - \ln(1-u) \leq \frac{1}{2}u^2$, och därför är

$$\nu\,\rho(-\gamma\nu^{-1/2}) \leq \tfrac{1}{2}\gamma^2.$$

Antalet inre iterationer i varje yttre iteration majoreras därför av konstanten

$$K = \left\lfloor \frac{\rho(\kappa) + \gamma\kappa(1-\kappa)^{-1} + \frac{1}{2}\gamma^2}{\rho(-\kappa)} \right\rfloor,$$

som bara beror av konstanterna κ och γ. Exempelvis är $K = 5$ om $\kappa = 0.4$ och $\gamma = 0.32$.

Vi övergår nu till att betrakta antalet yttre iterationer. Sätt

$$\beta(\kappa) = \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu}.$$

Antag att stoppvillkoret $\epsilon t \geq \beta(\kappa)$ utlöses under iteration nummer k då alltså $t = (1 + \gamma/\sqrt{\nu})^k t_0$. För den av algoritmen levererade punkten x gäller på grund av sats 18.2.1 att

$$\langle c, x\rangle - v_{\min} \leq \epsilon,$$

så x approximerar minimipunkten med föreskriven precision.

Eftersom k är det minsta heltalet för vilket $(1 + \gamma/\sqrt{\nu})^k \geq \beta(\kappa)/t_0\epsilon$, är vidare

$$k = \left\lceil \frac{\ln(\beta(\kappa)/t_0\epsilon)}{\ln(1 + \gamma/\sqrt{\nu})} \right\rceil.$$

I intervallet $0 \le x \le 1$ är $\ln(1+\gamma x) \ge x\ln(1+\gamma)$, beroende på att funktionen $\ln(1+\gamma x)$ är konkav. Följaktligen är

$$\ln(1+\gamma/\sqrt{\nu}) \ge \frac{\ln(1+\gamma)}{\sqrt{\nu}}.$$

Vidare är $\beta(\kappa) = \nu + \kappa(1-\kappa)^{-1}\sqrt{\nu} \le \nu + \kappa(1-\kappa)^{-1}\nu = (1-\kappa)^{-1}\nu$. Detta ger oss uppskattningen

$$k \le \left\lceil \frac{\sqrt{\nu}\ln((1-\kappa)^{-1}\nu/t_0\epsilon)}{\ln(1+\gamma)} \right\rceil \le K'\sqrt{\nu}\ln\Big(\frac{\nu}{t_0\epsilon}+1\Big)$$

för antalet yttre iterationer med en konstant K' som bara beror av κ och γ, och genom att multiplicera med konstanten K erhåller vi motsvarande uppskattning för det totala antalet iterationer. □

Fas 1

För att kunna använda den vägföljande algoritmen behöver man ett $t_0 > 0$ och en punkt $x_0 \in \operatorname{int} X$ med Newtondekrement $\lambda(F_{t_0}, x_0) \le \kappa$ att starta ifrån. Eftersom centrala vägen börjar i mängdens analytiska centrum $\hat{x}_F$ och $\lambda(F, \hat{x}_F) = 0$, kan man förvänta sig att (x_0, t_0) duger som startpar bara x_0 ligger tillräckligt nära $\hat{x}_F$ och t_0 är ett tillräckligt litet positivt tal. Vi skall visa att man kan generera ett sådant par genom att lösa ett artificiellt problem, givet att man känner någon punkt $\overline{x} \in \operatorname{int} X$.

Vi definierar för den skull för $0 \le t \le 1$ funktionerna $G_t\colon \operatorname{int} X \to \mathbf{R}$ genom att för $x \in \operatorname{int} X$ sätta

$$G_t(x) = -t\langle F'(\overline{x}), x\rangle + F(x).$$

Funktionerna G_t är slutna och självkonkordanta och har unika minimipunkter, som vi betecknar $\overline{x}(t)$.

Notera att $G_0 = F$, så $\overline{x}(0) = \hat{x}_F$. Eftersom

$$G_t'(x) = -tF'(\overline{x}) + F'(x),$$

är vidare $G_1'(\overline{x}) = 0$. Detta betyder att $\overline{x}$ är funktionen G_1:s minimipunkt och att följaktligen $\overline{x}(1) = \overline{x}$. Kurvan $\{\overline{x}(t) \mid 0 \le t \le 1\}$ börjar med andra ord i analytiska centret $\hat{x}_F$ och slutar i den givna punkten $\overline{x}$. Genom att använda den vägföljande metoden, men med den skillnaden att vi följer kurvan *baklänges*, kommer vi därför att erhålla ett lämpligt startpar för fas 2 av algoritmen.

Vi använder Newtons dämpade metod för att minimera G_t och noterar att $G_t'' = F''$ för alla t, så den lokala normen med avseende på funktionen G_t överensstämmer med den lokala normen med avseende på funktionen F, och vi kan därför otvetydigt använda beteckningen $\|\cdot\|_x$ för den lokala normen i punkten x.

Algoritmen för att bestämma ett startpar (x_0, t_0) ser nu ut så här.

Vägföljande algoritmen, fas 1

Givet $\overline{x} \in \operatorname{int} X$ och parametrarna $0 < \gamma < \frac{1}{2}\sqrt{\nu}$ och $0 < \kappa < 1$.

1. *Initiera:* $x := \overline{x}$ och $t := 1$.
2. *Stoppkriterium:* Stoppa om $\lambda(F, x) < \frac{3}{4}\kappa$, och sätt $x_0 = x$.
3. *Minska* t*:* $t := (1 - \gamma/\sqrt{\nu})t$.
4. *Uppdatera* x *genom att använda Newtons dämpade metod på funktionen* G_t *med det aktuella* x*-värdet som startpunkt:*
   (i) Beräkna $\lambda = \lambda(G_t, x)$.
   (ii) Avbryt Newtons metod om $\lambda \leq \kappa/2$, och gå till punkt 2.
   (iii) Beräkna annars Newtonsteget $\Delta x_{\mathrm{nt}} = -F''(x)^{-1} G_t'(x)$.
   (iv) *Uppdatera:* $x := x + (1 + \lambda)^{-1} \Delta x_{\mathrm{nt}}$, och gå sedan till (i).

När algoritmen stoppat med en punkt x_0 definierar vi t_0 genom att sätta

$$t_0 = \max\{t \mid \lambda(F_t, x_0) \leq \kappa\}.$$

Antalet inre iterationer i fas 1 ges av följande sats.

Sats 18.2.4. *Fas 1 av den vägföljande algoritmen stoppar med en punkt* x_0 *i* $\operatorname{int} X$ *efter högst*

$$C\sqrt{\nu} \ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big)$$

inre iterationer, där konstanten C *bara beror av parametrarna* κ *och* γ*, och för talet* t_0 *gäller att* $\lambda(F_{t_0}, x_0) \leq \kappa$ *och*

$$t_0 \geq \frac{\kappa}{4 \operatorname{Var}_X(c)}.$$

Bevis. Vi börjar med att uppskatta antalet inre iterationer i varje yttre iteration; detta antal begränsas av kvoten

$$\frac{G_s(x) - G_s(\overline{x}(s))}{\rho(-\kappa/2)},$$

där $s = (1 - \gamma/\sqrt{\nu})t$, och lemma 18.2.3 ger oss majoranten

$$\rho(\kappa/2) + \frac{\kappa\sqrt{\nu}}{2 - \kappa} \cdot \frac{\gamma}{\sqrt{\nu}} + \nu\, \rho(\gamma/\sqrt{\nu})$$

till bråkets täljare. Här är $\nu\rho(\gamma/\sqrt{\nu}) \le \gamma^2$ på grund av lemma 16.3.1. Antalet inre iterationer i varje yttre iteration begränsas således av konstanten

$$\frac{\rho(\kappa/2) + \kappa(2-\kappa)^{-1}\gamma + \gamma^2}{\rho(-\kappa/2)}.$$

Vi betraktar nu de yttre iterationerna. Eftersom $F' = G_t' + tF'(\overline{x})$, är

$$\begin{aligned}(18.12)\quad \lambda(F,x) &= \|F'(x)\|_x^* = \|G_t'(x) + tF'(\overline{x})\|_x^* \le \|G_t'(x)\|_x^* + t\|F'(\overline{x})\|_x^* \\ &= \lambda(G_t,x) + t\|F'(\overline{x})\|_x^*.\end{aligned}$$

Vidare följer det av sats 18.1.11 att

$$\|F'(\overline{x})\|_x^* \le (\nu + 2\sqrt{\nu})\|F'(\overline{x})\|_{\hat{x}_F}^* \le 3\nu\|F'(\overline{x})\|_{\hat{x}_F}^*,$$

och av sats 18.1.8 att

$$\|F'(\overline{x})\|_{\hat{x}_F}^* = \sup_{\|v\|_{\hat{x}_F} \le 1} \langle F'(\overline{x}), v\rangle \le \frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})},$$

så

$$(18.13)\qquad \|F'(\overline{x})\|_x^* \le \frac{3\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x})}.$$

Under yttre iteration nummer k är $t = (1 - \gamma/\sqrt{\nu})^k$, och när Newtons metod stoppar, uppfyller punkten x villkoret $\lambda(G_t, x) \le \kappa/2$. Stoppvillkoret $\lambda(F,x) < \frac{3}{4}\kappa$ i punkt 2 av algoritmen är därför på grund av olikheten (18.12) och uppskattningen (18.13) säkert uppfyllt om

$$\frac{1}{2}\kappa + \frac{3\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x})}(1 - \gamma/\sqrt{\nu})^k \le \frac{3}{4}\kappa,$$

dvs. om

$$k\ln(1 - \gamma/\sqrt{\nu}) < -\ln\Big(\frac{12\kappa^{-1}\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x}}\Big),$$

och genom att utnyttja olikheten $\ln(1-x) \le -x$, som gäller för $0 < x < 1$, ser vi att stoppvillkoret är uppfyllt för

$$k > \frac{\sqrt{\nu}}{\gamma}\ln\Big(\frac{12\kappa^{-1}\nu^2}{1 - \pi_{\hat{x}_F}(\overline{x}}\Big).$$

Antalet yttre iterationer är därför mindre än

$$K\sqrt{\nu}\ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big),$$

där konstanten K bara beror av κ och γ. Eftersom antalet inre iterationer i varje yttre iteration begränsas av en konstant, som bara beror av κ och γ, får vi därmed satsens uppskattning för det totala antalet inre iterationer.

Det följer av definitionen av slutvärdet t_0 att $\kappa = \lambda(F_{t_0}, x_0)$, och vi får därför med hjälp av sats 18.1.10 följande olikheter

$$\begin{aligned}\kappa = \lambda(F_{t_0}, x_0) &= \|F'_{t_0}(x_0)\|^*_{x_0} = \|t_0 c + F'(x_0)\|^*_{x_0} \le t_0\|c\|^*_{x_0} + \|F'(x_0)\|^*_{x_0} \\ &= t_0\|c\|^*_{x_0} + \lambda(F, x_0) \le t_0 \operatorname{Var}_X(c) + \frac{3}{4}\kappa,\end{aligned}$$

som innebär att

$$t_0 \ge \frac{\kappa}{4 \operatorname{Var}_X c}.$$ □

Genom att kombinera de båda faserna av algoritmen får vi nu följande komplexitetsresultat.

Sats 18.2.5. *För att lösa ett standardproblem (SP) med ν-självkonkordant barriär, toleransnivå $\epsilon > 0$ och startpunkt $\overline{x} \in \operatorname{int} X$ behövs det högst*

$$C\sqrt{\nu}\ln(\nu\Phi/\epsilon + 1)$$

Newtonsteg, där

$$\Phi = \frac{\operatorname{Var}_X(c)}{1 - \pi_{\hat{x}_F}(\overline{x})}$$

och konstanten C bara beror av parametrarna γ och κ.

Bevis. Fas 1 levererar en startpunkt x_0 och ett begynnelsevärde t_0 för fas 2, som uppfyller $t_0 \ge \kappa/(4 \operatorname{Var}_X(c))$. Antalet inre iterationer i fas 2 majoreras därför av

$$O(1)\sqrt{\nu}\ln\Big(\frac{4\nu \operatorname{Var}_X(c)}{\kappa\epsilon} + 1\Big) = O(1)\sqrt{\nu}\ln\Big(\frac{\nu \operatorname{Var}_X(c)}{\epsilon} + 1\Big).$$

Totala antalet inre iterationer i de båda faserna är därför

$$\begin{aligned}&O(1)\sqrt{\nu}\ln\Big(\frac{\nu}{1 - \pi_{\hat{x}_F}(\overline{x})} + 1\Big) + O(1)\sqrt{\nu}\ln\Big(\frac{\nu \operatorname{Var}_X(c)}{\epsilon} + 1\Big) \\ &\qquad = O(1)\sqrt{\nu}\ln(\nu\Phi/\epsilon + 1).\end{aligned}$$ □

Anmärkning. Algoritmerna i det här avsnittet ger fina teoretiska komplexitetsresultat, men de är inte lämpliga för praktiskt bruk. Den huvudsakliga svagheten är den låga uppdateringsfaktorn $(1 + O(1)\nu^{-1/2})$ av straffparametern t, som leder till att totala antalet Newtonsteg blir proportionellt mot $\sqrt{\nu}$.

För ett LP-problem med $n = 1000$ variabler och $m = 10000$ olikheter skulle man behöva lösa hundratals linjära ekvationssystem med 1000 variabler, vilket kräver ojämförligt mycket mer tid än vad som behövs med simplexalgoritmen. Under flertalet yttre iterationer kan man emellertid i praktiken öka straffparametern mycket snabbare än vad som behövs för den teoretiska värstafallet-analysen, utan att för den skull behöva öka antalet Newtonsteg för att bibehålla närheten till den centrala vägen. Det finns bra praktiska implementeringar av algoritmen som använder sig av olika dynamiska strategier för att kontrollera straffparametern t och som resulterar i att det bara behövs ett måttligt totalt antal Newtonsteg, och detta oberoende av problemets storlek.

18.3 LP-problem

Vi skall nu tillämpa algoritmen i föregående avsnitt på LP-problem. Betrakta för den skull ett LP-problem på formen

$$\begin{array}{ll} \min & \langle c, x\rangle \\ \text{då} & Ax \leq b \end{array} \tag{18.14}$$

där $A = [a_{ij}]$ är en $m \times n$-matris. Vi antar att polyedern

$$X = \{x \in \mathbf{R}^n \mid Ax \leq b\}$$

av tillåtna punkter är begränsad och har ett icke-tomt inre. Begränsningsantagandet medför att $m > n$.

Den i:te raden i matrisen A betecknas a_i så att $a_i = [a_{i1}\ a_{i2}\ \ldots\ a_{in}]$. Matrismultiplikationen $a_i x$ är således väldefinierad.

Som barriär till området X använder vi den m-självkonkordanta funktionen

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i x).$$

Med godtycklig startpunkt $\overline{x} \in \operatorname{int} X$ ger den vägföljande algoritmen en ϵ-lösning, dvs. en punkt med ett målfunktionsvärde som approximerar det optimala värdet med ett fel som understiger ϵ, efter högst

$$O(1)\sqrt{m}\ln(m\Phi/\epsilon + 1)$$

inre iterationer, där

$$\Phi = \frac{\operatorname{Var}_X(c)}{1 - \pi_{\hat{x}_F}(\overline{x})}.$$

Vi skall nu uppskatta hur många aritmetiska operationer (additioner, subtraktioner, multiplikationer och divisioner) som behövs för att erhålla denna

ϵ-lösning. För varje inre iteration i Newtonalgoritmen beräknar man först barriärfunktionens gradient och hessian i den aktuella punkten x, dvs.

$$F'(x) = \sum_{i=1}^{m} \frac{a_i^T}{b_i - a_i x} \quad \text{och} \quad F''(x) = \sum_{i=1}^{m} \frac{a_i^T a_i}{(b_i - a_i x)^2}.$$

Beräkningarna kan utföras med $O(mn^2)$ aritmetiska operationer (additioner, subtraktioner, multiplikationer och divisioner).

Newtonriktningen Δx_{nt} i punkten x erhålls sedan under fas 2 som lösning till det kvadratiska ekvationssystemet

$$F''(x)\Delta x_{\text{nt}} = -(tc + F'(x)),$$

och med Gausselimination får man lösningen med hjälp av $O(n^3)$ aritmetiska operationer. Slutligen krävs det ytterligare $O(n)$ aritmetiska operationer, inklusive en kvadratrotsberäkning, för att beräkna Newtondekrementet $\lambda = \lambda(F_t, x)$ och den nya punkten $x^+ = x + (1 + \lambda)^{-1}\Delta x_{\text{nt}}$. Motsvarande kalkyl av antalet operationer gäller också för fas 1.

Av ovanstående beräkningar är gradient- och hessianberäkningen den mest kostsamma beroende på att $m > n$. Totala antalet aritmetiska operationer i varje iteration är därför $O(mn^2)$, och genom att multiplicera med antalet inre iterationer kan den totala aritmetiska "kostnaden" för den vägföljande algoritmen uppskattas till högst $O(m^{3/2}n^2)\ln(m\Phi/\epsilon + 1)$ operationer.

Den erhållna approximativa minimipunkten $\hat{x}(\epsilon)$ är en inre punkt i polyedern X, men minimum antas förstås på randen i en extremalpunkt till X. Med utgångspunkt från $\hat{x}(\epsilon)$ kan man med hjälp av en enkel procedur, som kallas *rening* och beskrivs nedan, efter ytterligare högst $O(mn^2)$ aritmetiska operationer hitta en extremalpunkt $\hat{x}$ till X med ett målfunktionsvärde som inte överstiger värdet i $\hat{x}(\epsilon)$. Detta innebär att vi har följande resultat.

Sats 18.3.1. *För LP-problemet* (18.14) *behövs det högst*

$$O(m^{3/2}n^2)\ln(m\Phi/\epsilon + 1)$$

aritmetiska operationer för att hitta en extremalpunkt $\hat{x}$ till polyedern X av tillåtna punkter med ett målfunktionsvärde som approximerar minimivärdet med ett fel som understiger ϵ.

Rening

Hur man hittar en extremalpunkt med ett lägre funktionsvärde än värdet i en given inre punkt beskrivs i beviset för följande sats.

Sats 18.3.2. *Låt*

$$\begin{aligned} &\min && \langle c, x\rangle \\ &\text{då} && Ax \leq b \end{aligned}$$

vara ett LP-problem med n variabler och m bivillkor, och antag att polyedern X av tillåtna punkter är linjefri samt att målfunktionen är nedåt begränsad på X. För varje punkt i X kan man då med högst $O(mn^2)$ aritmetiska operationer generera en extremalpunkt till X med ett målfunktionsvärde som inte överstiger värdet i den givna punkten.

Bevis. Idén är enkel; man följer en halvlinje från den givna punkten $x^{(0)}$ med icke-växande målfunktionsvärden till dess att man träffar på punkt $x^{(1)}$ i en fasad F_1 till polyedern X. Sedan följer man en halvlinje i fasaden F_1 med icke-växande funktionsvärden till dess att man träffar på en punkt $x^{(2)}$ i skärningen $F_1 \cap F_2$ av två fasader, etc. Efter n steg har man en punkt $x^{(n)}$ i skärningen av n (oberoende) fasader, dvs. en extremalpunkt, med ett funktionsvärde som är mindre än eller lika med värdet i startpunkten.

För att uppskatta antalet aritmetiska operationer studerar vi ovanstående procedur lite mer i detalj.

Vi startar genom att sätta $v^{(1)} = \mathbf{e}_1$ om $c_1 < 0$, $v^{(1)} = -\mathbf{e}_1$ om $c_1 > 0$ och $v^{(1)} = \pm\mathbf{e}_1$ om $c_1 = 0$, där tecknet i det sistnämnda fallet ska väljas så att halvlinjen $x^{(0)} + tv^{(1)}$, $t \geq 0$, skär polyederns rand, vilket är möjligt eftersom polyedern antas vara linjefri. I de två förstnämnda fallen skär halvlinjen också polyederns rand beroende på att $\langle c, x^{(0)} + tv^{(1)}\rangle = \langle c, x^{(0)}\rangle - t|c_1|$ går mot $-\infty$ då t går mot ∞ och målfunktionen antas vara nedåt begränsad på X. Skärningspunkten $x^{(1)} = x^{(0)} + t_1 v^{(1)}$ mellan halvlinjen och randen till X kan beräknas med $O(mn)$ aritmetiska operationer, eftersom vi bara behöver beräkna vektorerna $b - Ax^{(0)}$ och $Av^{(1)}$ och kvoter mellan deras koordinater för att hitta det icke-negativa parametervärdet t_1.

Efter eventuell omnumrering kan vi anta att punkten $x^{(1)}$ ligger i hyperplanet $a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$. Vi eliminerar nu variabeln x_1 ur bivillkor och målfunktion, vilket resulterar i ett system av typen

$$\begin{cases} x_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n = b'_1 \\ A' \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix} \leq b' \end{cases} \tag{18.15}$$

där A' är en $(m-1) \times (n-1)$-matris, och en målfunktion

$$c'_2 x_2 + \cdots + c'_n x_n + d'$$

som är den ursprungliga målfunktionens restriktion till den aktuella fasaden. Antalet operationer för att utföra eliminationerna är också $O(mn)$.

Efter totalt $O(mn)$ operationer har vi således hittat en punkt $x^{(1)}$ i en fasad F_1 till X med målfunktionsvärde $\langle c, x^{(1)}\rangle = \langle c, x^{(0)}\rangle - t_1|c_1| \leq \langle c, x^{(0)}\rangle$, samt bestämt fasadens ekvation (18.15) och målfunktionens restriktion till fasaden. Vi har nu ett problem av lägre dimension $n-1$ och med $m-1$ bivillkor.

Vi fortsätter nu genom att välja en vektor $v^{(2)}$ som är parallell med fasaden F_1 och längs vilken målfunktionen inte ökar. Detta sker genom att sätta $v_2^{(2)} = \pm 1$, $v_3^{(2)} = \cdots = v_n^{(2)} = 0$ (och $v_1^{(2)} = -a'_{12}v_2^{(2)}$), där tecknet för $v_2^{(2)}$ ska väljas så att målfunktionen är icke-växande utefter halvlinjen $x^{(1)} + tv^{(2)}$, $t \geq 0$, och halvlinjen skär den relativa randen till F_1. Detta innebär att $v_2^{(2)} = 1$ om $c'_2 < 0$ och $v_2^{(2)} = -1$ om $c'_2 > 0$, medan tecknet i fallet $c'_2 = 0$ bestäms av kravet att halvlinjen ska skära randen.

Vi avgör sedan var halvlinjen $x^{(1)} + tv^{(2)}$, $t \geq 0$, skär den relativa randen till F_1, dvs. något av de återstående hyperplanen. Om detta hyperplan är planet $a'_{21}x_2 + \cdots + a'_{2n}x_n = b'_2$, säg, så eliminerar vi variabeln x_2 ur de återstående bivillkoren och målfunktionen. Allt detta kan göras med högst $O(mn)$ operationer och resulterar i en punkt $x^{(2)}$ i snittet av två fasader och med ett målfunktionsvärde $\langle c, x^{(2)}\rangle = \langle c, x^{(1)}\rangle - t_2|c'_2| \leq \langle c, x^{(1)}\rangle$.

Efter totalt n iterationssteg, som kräver högst $nO(mn) = O(mn^2)$ aritmetiska operationer, har vi erhållit en extremalpunkt $\hat{x} = x^{(n)}$ med ett målfunktionsvärde som inte överstiger värdet i punkten $x^{(0)}$. Extremalpunktens koordinater får vi genom att lösa ett triangulärt ekvationssystem, vilket bara kräver $O(n^2)$ operationer. Totalt har det åtgått $O(mn^2)$ operationer. □

Exempel 18.3.1. Vi exemplifierar reningsalgoritmen genom att för problemet

$$\begin{array}{ll} \min & -2x_1 + x_2 + 3x_3 \\ \text{då} & \begin{cases} -x_1 + 2x_2 + x_3 \leq 4 \\ -x_1 + x_2 + x_3 \leq 2 \\ x_1 - 2x_2 \leq 1 \\ x_1 - x_2 - 2x_3 \leq 1 \end{cases} \end{array}$$

med $x^{(0)} = (1, 1, 1)$ som utgångspunkt bestämma en extremalpunkt med ett mindre målfunktionsvärde än $\langle c, x^{(0)}\rangle = 2$.

Eftersom $c_1 = -2 < 0$, börjar vi med att välja $v^{(1)} = (1, 0, 0)$ och bestämmer var halvlinjen $x = x^{(0)} + tv^{(1)} = (1 + t, 1, 1)$, $t \geq 0$, skär randen. Insättning i de fyra bivillkoren visar att $t = 2$ ger likhet i den tredje olikheten, samt att övriga olikheter är uppfyllda för detta t-värde. Punkten $x^{(1)} = (3, 1, 1)$ ligger med andra ord i den fasad som fås genom att skära polyedern X med stödhyperplanet $x_1 - 2x_2 = 1$. Vi eliminerar därför x_1 ur målfunktion och bivillkor med hjälp av detta hyperplans ekvation och betraktar restriktionen av målfunktionen till nämnda fasad, dvs. funktionen

$f(x) = -3x_2 + 3x_3 - 2$ då

$$\begin{cases} x_1 - 2x_2 = 1 \\ x_3 \leq 5 \\ -x_2 + x_3 \leq 3 \\ x_2 - 2x_3 \leq 0 \end{cases}$$

Den nya koefficienten för x_2 i målfunktionen är negativ, så vi följer halvlinjen $x_2 = 1 + t$, $x_3 = 1$, $t \geq 0$, i stödhyperplanet $x_1 - 2x_2 = 1$ till dess att den skär ett nytt stödhyperplan, vilket inträffar för $t = 1$, då den skär hyperplanet $x_2 - 2x_3 = 0$ i punkten $x^{(2)} = (5, 2, 1)$. Elimination av x_2 resulterar i målfunktionen $f(x) = -3x_3 - 2$ och systemet

$$\begin{cases} x_1 - 2x_2 = 1 \\ x_2 - 2x_3 = 0 \\ x_3 \leq 5 \\ -x_3 \leq 3 \end{cases}$$

Ny halvlinje i fasaden $F_1 \cap F_2$ är $x_3 = 1 + t$, som skär det tredje stödhyperplanet $x_3 = 5$ då $t = 4$, dvs. i punkten med x_3-koordinat 5. Återsubstitution ger $x^{(3)} = (21, 10, 5)$, som är en extremalpunkt med målfunktionsvärde -17. □

18.4 Komplexitet

I det här avsnittet skall vi studera hur många aritmetiska operationer som behövs för att producera en lösning till ett LP-problem med rationella koefficienter, där vi med *lösning* menar problemets optimala värde och, förutsatt att detta är ändligt, en optimal punkt. Alla kända uppskattningar av antalet operationer beror, förutom av antalet variabler och bivillkor, också av storleken på problemets koefficienter, och ett lämpligt mått på problemets storlek är antalet binära bitar som behövs för att representera samtliga koefficienter.

Definition. För vektorer $x = (x_1, x_2, \ldots, x_n)$ i $\mathbf{R}^n$ sätter vi

$$\ell(x) = \sum_{j=1}^{n} \lceil \log_2(|x_j| + 1) \rceil$$

och kallar heltalet $\ell(x)$ för vektorns *inputlängd*.

Antalet siffror (bitar) i den binära utvecklingen av ett positivt heltal z är lika med $\lceil \log_2(|z|+1) \rceil$. Den binära utvecklingen av ett negativt heltal z kräver ytterligare en bit för att ta hand om talets tecken, och det gör också representationen av $z = 0$. Antalet binära bitar för att representera en godtycklig vektor x i $\mathbf{R}^n$ med heltalskoordinater är därför högst lika med $\ell(x) + n$.

En vektors norm kan uppskattas med hjälp av inputlängden, och vi kommer att behöva följande enkla skattning i fallen $p = 1$ och $p = 2$.

Lemma 18.4.1. *För alla $x \in \mathbf{R}^n$ och $p \geq 1$ är $\|x\|_p \leq 2^{\ell(x)}$.*

Bevis. Resultatet följer av de triviala olikheterna $\sum_{j=1}^n a_j \leq \prod_{j=1}^n (a_j + 1)$, $a^p + 1 \leq (a+1)^p$ och $\log_2(a+1) \leq \lceil \log_2(a+1) \rceil$, som gäller för icke-negativa tal a, a_j, och ger att

$$\|x\|_p^p = \sum_{j=1}^n |x_j|^p \leq \prod_{j=1}^n (|x_j|^p + 1) \leq \prod_{j=1}^n (|x_j| + 1)^p \leq 2^{p\,\ell(x)}. \qquad \square$$

Vi skall nu studera LP-problem av typen

$$\text{(LP)} \qquad \begin{array}{ll} \min & \langle c, x \rangle \\ \text{då} & Ax \leq b \end{array}$$

där samtliga koefficienter i $m \times n$-matrisen $A = [a_{ij}]$ och i vektorerna b och c är heltal. Varje LP-problem med rationella koefficienter kan uppenbarligen ersättas med ett ekvivalent problem på denna form efter multiplikation med lämpliga minsta gemensamma nämnare. Polyedern av tillåtna punkter kommer att betecknas X så att

$$X = \{x \in \mathbf{R}^n \mid Ax \leq b\}.$$

Definition. Till problemet (LP) associerar vi de två talen

$$\ell(X) = \ell(A) + \ell(b) \qquad \text{och} \qquad L = \ell(X) + \ell(c) + m + n,$$

där $\ell(A)$ betecknar inputlängden hos matrisen A, uppfattad som vektor i $\mathbf{R}^{mn}$. Vi kallar heltalet $\ell(X)$ för *polyedern X:s inputlängd* och heltalet L för *LP-problemets inputlängd*[†].

Huvudresultatet i det här avsnittet är följande sats, som medför att det finns en lösningsalgoritm som är polynomiell i LP-problemets inputlängd.

[†]Det behövs $\ell(X) + mn + m$ binära bitar för att representera samtliga koefficienter i polyedern X och $L + mn$ binära bitar för att representera samtliga koefficienter i LP-problemet, så det vore mer logiskt att kalla dessa tal för polyederns resp. LP-problemets inputlängder. De fortsatta kalkylerna blir emellertid enklare med vår konvention.

Sats 18.4.2. *Det finns en algoritm som löser LP-problemet* (LP) *med högst* $O((m+n)^{7/2}L)$ *aritmetiska operationer.*

Bevis. **I.** Vi börjar med att konstatera att vi utan inskränkning kan anta att polyedern X av tillåtna punkter är linjefri. Vi kan nämligen vid behov ersätta problemet (LP) med det ekvivalenta och garanterat linjefria LP-problemet

$$\begin{array}{ll} \min & \langle c, x^+\rangle - \langle c, x^-\rangle \\ \text{då} & \begin{cases} Ax^+ - Ax^- \le b \\ -x^+ \le 0 \\ -x^- \le 0. \end{cases} \end{array}$$

Detta LP-problem i $n' = 2n$ variabler och med $m' = m + 2n$ bivillkor har inputlängd

$$\begin{aligned} L' &= 2\ell(A) + 2n + \ell(b) + 2\ell(c) + m' + n' \\ &\le 2(\ell(A) + \ell(b) + \ell(c) + m + n) + 4n = 2L + 4n \le 6L, \end{aligned}$$

så en algoritm som löser detta problem med $O((m'+n')^{7/2}L')$ operationer löser problemet (LP) med $O((m+n)^{7/2}L)$ operationer eftersom $m' + n' \le 4(m+n)$ och $L' \le 6L$.

Vi antar därför fortsättningsvis hela tiden att X *är en linjefri polyeder,* och en konsekvens av detta antagande för en icke-tom polyeder X är att $m \ge n$ och att X har minst en extremalpunkt.

Påståendet i satsen är vidare trivialt sant för LP-problem med endast en variabel, så vi kan vidare utan inskränkning anta att $m \ge n \ge 2$. Slutligen kan vi naturligtvis anta att alla raderna i matrisen A är nollskilda, ty om rad nr k är identiskt noll, så kan motsvarande bivillkor antingen strykas (om $b_k \ge 0$) eller också är polyedern av tillåtna punkter tom (om $b_k < 0$). I fortsättningen kan vi således använda oss av olikheterna

$$\ell(X) \ge \ell(A) \ge m \ge n \ge 2 \text{ och } L \ge \ell(X) + m + n \ge \ell(X) + 4.$$

II. Under ovanstående antaganden kommer vi att bevisa satsen genom att visa:

1. Med $O(m^{7/2}L)$ operationer kan man avgöra om problemets värde är $+\infty$, $-\infty$ eller ändligt, dvs. om det finns några tillåtna punkter och om målfunktionen är nedåt begränsad.
2. Givet att problemets värde är ändligt kan man sedan med $O(m^{3/2}n^2L)$ operationer bestämma en optimal lösning.

Eftersom beviset för punkt 1 utnyttjar att man bestämmer den optimala lösningen till ett lämpligt hjälpproblem (med ändligt värde), börjar vi med att visa punkt 2.

III. Som första byggsten behöver vi ett lemma som ger oss information om extremalpunkterna till polyedern X i termer av polyederns inputlängd.

Lemma 18.4.3. *(i) Låt $\hat{x}$ vara en extremalpunkt till polyedern X. För alla nollskilda koordinater $\hat{x}_j$ gäller då att*

$$2^{-\ell(X)} \leq |\hat{x}_j| \leq 2^{\ell(X)}.$$

Alla extremalpunkterna ligger således i kuben $\{x \in \mathbf{R}^n \mid \|x\|_\infty \leq 2^{\ell(X)}\}$.

(ii) Om $\hat{x}$ och $\tilde{x}$ är två extremalpunkter till X och $\langle c, \hat{x}\rangle \neq \langle c, \tilde{x}\rangle$, så är

$$|\langle c, \hat{x}\rangle - \langle c, \tilde{x}\rangle| \geq 4^{-\ell(X)}.$$

Bevis. För att bevisa påståendena i lemmat börjar vi med att erinra om *Hadamards olikhet*, som har följande lydelse: *Om $C = [c_{ij}]$ är en $k \times k$-matris med kolonner C_{*1}, C_{*2}, ..., C_{*k}, så är*

$$|\det C| \leq \prod_{j=1}^{k} \|C_{*j}\|_2 = \prod_{j=1}^{k} \Big(\sum_{i=1}^{k} c_{ij}^2\Big)^{1/2}.$$

Olikheten är geometriskt uppenbar – vänsterledet $|\det C|$ är volymen av den av matriskolonnerna uppspända hyperparallellepipeden medan högerledet är volymen av ett hyperrätblock med lika långa kanter som parallellepipedens.

Genom att kombinera Hadamards olikhet med lemma 18.4.1 erhåller vi olikheten

$$|\det C| \leq \prod_{j=1}^{k} 2^{\ell(C_{*j})} = 2^{\ell(C)}.$$

Om C är en kvadratisk delmatris till matrisen $\begin{bmatrix} A & b \end{bmatrix}$, så är uppenbarligen $\ell(C) \leq \ell(A) + \ell(b) = \ell(X)$, och det följer därför av olikheten ovan att

$$|\det C| \leq 2^{\ell(X)}. \tag{18.16}$$

Låt nu $\hat{x}$ vara en extremalpunkt till polyedern X. Då finns det enligt sats 5.1.1 en mängd $\{i_1, i_2, \ldots, i_n\}$ av radindex så att extremalpunkten $\hat{x}$ fås som unik lösning till ekvationssystemet

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = i_1, i_2, \ldots, i_n.$$

Enligt Cramers regel kan systemets lösning skrivas på formen

$$\hat{x}_j = \frac{\Delta_j}{\Delta},$$

där Δ är koefficientmatrisens determinant och Δ_j är den determinant som fås genom att i Δ byta ut kolonn nummer j mot ekvationssystemets högerled. Determinanterna Δ och Δ_j är heltal, och på grund av olikheten (18.16) är deras belopp högst lika med $2^{\ell(X)}$. För alla nollskilda koordinater $\hat{x}_j$, dvs. alla j med $\Delta_j \neq 0$, får vi därför uppskattningarna

$$|\hat{x}_j| = |\Delta_j|/|\Delta| \leq 2^{\ell(X)}/1 = 2^{\ell(X)} \text{ och } |\hat{x}_j| = |\Delta_j|/|\Delta| \geq 1/2^{\ell(X)} = 2^{-\ell(X)},$$

vilket är lemmats påstående (i).

(ii) Målfunktionens värde i extremalpunkten $\hat{x}$ är

$$\langle c, \hat{x}\rangle = \Big(\sum_{j=1}^{n} c_j\Delta_j\Big)/\Delta = T/\Delta,$$

där täljaren T är ett heltal. Om $\tilde{x}$ är en annan extremalpunkt, så är förstås analogt $\langle c, \tilde{x}\rangle = T'/\Delta'$ för något heltal T' och någon determinant Δ' med $|\Delta'| \leq 2^{\ell(X)}$. Det följer att differensen

$$\langle c, \tilde{x}\rangle - \langle c, \hat{x}\rangle = (T\Delta' - T'\Delta)/\Delta\Delta'$$

antingen är lika med noll eller, om täljaren är skild från noll, är ett tal som till beloppet är $\geq 1/|\Delta\Delta'| \geq 4^{-\ell(X)}$. □

IV. Som lösningsmetod skall vi använda den vägföljande metoden, men den förutsätter att polyedern av tillåtna punkter är begränsad och att det finns en inre punkt att starta fas 1 ifrån. För att komma runt denna svårighet betraktar vi följande hjälpproblem i $n+1$ stycken variabler och $m+2$ linjära bivillkor:

$$(\mathrm{LP}_M) \qquad \begin{array}{ll} \min & \langle c, x\rangle + Mx_{n+1} \\ \text{då} & \begin{cases} Ax + (b - \mathbf{1})x_{n+1} \leq b \\ x_{n+1} \leq 2 \\ -x_{n+1} \leq 0. \end{cases} \end{array}$$

Här är M ett positivt heltal, $\mathbf{1}$ betecknar vektorn i $\mathbf{R}^m$ av idel ettor, och x står som tidigare för n-tipeln $(x_1, x_2, \ldots, x_n)$.

Låt X' beteckna polyedern av tillåtna punkter i problemet (LP_M). Eftersom $(x, x_{n+1}) = (0, 1)$ satisfierar alla bivillkor med strikt olikhet, är $(0, 1)$ en inre punkt i X'.

För inputlängden $\ell(X')$ hos polyedern X' och inputlängden $L(M)$ hos

problemet (LP_M) erhåller vi följande uppskattningar:

$$\begin{aligned}(18.17)\qquad \ell(X') &= \ell(A) + \sum_{i=1}^{m} \bigl\lceil \log_2\bigl(|b_i - 1| + 1\bigr)\bigr\rceil + 1 + 1 + \ell(b) + 2 \\ &\le \ell(X) + 4 + \sum_{i=1}^{m} \bigl(1 + \bigl\lceil \log_2\bigl(1 + |b_i|\bigr)\bigr\rceil\bigr) \\ &= \ell(X) + 4 + m + \ell(b) \le 2\ell(X) + 4 \le 2L - 4\end{aligned}$$

och

$$\begin{aligned}(18.18)\qquad L(M) &= \ell(X') + \ell(c) + \lceil \log_2(M+1)\rceil + m + n + 3 \\ &\le 2\ell(X) + 2\ell(c) + \lceil \log_2 M\rceil + m + n + 8 \\ &= 2L + \lceil \log_2 M\rceil - (m+n) + 8 \le 2L + \lceil \log_2 M\rceil + 4.\end{aligned}$$

Motivet för att studera problemet (LP_M) framgår av nästa lemma.

Lemma 18.4.4. *Antag att problemet* (LP) *har ett ändligt värde. Då gäller:*

(i) För varje heltal $M > 0$ har problemet (LP_M) *ett ändligt värde.*

(ii) Om $(\hat{x}, 0)$ är en optimal lösning till problemet (LP_M)*, så är $\hat{x}$ en optimal lösning till det ursprungliga problemet* (LP)*.*

(iii) Antag att $M \ge 2^{4L}$ och att problemet (LP_M) *antar sitt optimala värde i extremalpunkten $(\hat{x}, \hat{x}_{n+1})$ till X'. Då är $\hat{x}_{n+1} = 0$, och $\hat{x}$ är följaktligen en optimal lösning till problemet* (LP)*.*

Bevis. (i) Antagandet om ändligt värde innebär att polyedern X inte är tom och att målfunktionen $\langle c, x\rangle$ är nedåt begränsad på X, och enligt sats 12.1.1 betyder detta att vektorn c ligger i dualkonen till recessionskonen $\operatorname{recc} X$. Eftersom

$$\begin{aligned}\operatorname{recc} X' &= \{(x, x_{n+1}) \mid Ax + (b - \mathbf{1})x_{n+1} \le 0,\ \ x_{n+1} = 0\} \\ &= \operatorname{recc} X \times \{0\},\end{aligned}$$

är dualkonen till $\operatorname{recc} X'$ lika med $(\operatorname{recc} X)^+ \times \mathbf{R}$. Vektorn (c, M) ligger följaktligen i dualkonen $(\operatorname{recc} X')^+$, vilket betyder att målfunktionen i problemet (LP_M) är nedåt begränsad på den icke-tomma mängden X'. Det artificiella problemet har således ett ändligt värde.

Polyedern X' är vidare linjefri eftersom

$$\begin{aligned}\operatorname{lin} X' &= \{(x, x_{n+1}) \mid Ax + (b - \mathbf{1})x_{n+1} = 0,\ \ x_{n+1} = 0\} \\ &= \operatorname{lin} X \times \{0\} = \{(0, 0)\}.\end{aligned}$$

(ii) En punkt $(x, 0)$ är tillåten för problemet (LP_M) om och endast om x ligger i X, dvs. är tillåten för problemet (LP). Om $(\hat{x}, 0)$ är en optimal lösning till det artificiella problemet, så är därför speciellt

$$\langle c, \hat{x}\rangle = \langle c, \hat{x}\rangle + M \cdot 0 \leq \langle c, x\rangle + M \cdot 0 = \langle c, x\rangle$$

för alla $x \in X$, vilket visar att $\hat{x}$ är en optimal lösning till problemet (LP).

(iii) Antag att extremalpunkten $(\hat{x}, \hat{x}_{n+1})$ i polyedern X' är en optimal punkt i problemet (LP_M). Lemma 18.4.3 tillämpat på polyedern X' och uppskattningen (18.17) ger då att

$$\|\hat{x}\|_\infty \leq 2^{\ell(X')} \leq 2^{2\ell(X)+4} \leq 2^{2L-4}, \tag{18.19}$$

så det följer med hjälp av lemma 18.4.1 att

$$\begin{aligned} |\langle c, \hat{x}\rangle| &\leq \sum_{j=1}^{n} |c_j||\hat{x}_j| \leq \|c\|_1 \|\hat{x}\|_\infty \leq 2^{\ell(c)} \cdot 2^{2\ell(X)+4} \leq 2^{2\ell(X)+2\ell(c)+4} \\ &\leq 2^{2L-2m-2n+4} \leq 2^{2L-4}. \end{aligned}$$

Antag att $\hat{x}_{n+1} \neq 0$; då är $\hat{x}_{n+1} \geq 2^{-\ell(X')} \geq 2^{-2L}$ på grund av lemma 18.4.3. För det optimala värdet $\hat{v}_M$ till det artificiella problemet (LP_M) gäller därför att

$$\hat{v}_M = \langle c, \hat{x}\rangle + M\hat{x}_{n+1} \geq M\hat{x}_{n+1} - |\langle c, \hat{x}\rangle| \geq M \cdot 2^{-2L} - 2^{2L-4}.$$

Låt nu x vara en godtycklig extremalpunkt till polyedern X. Eftersom $(x, 0)$ är en tillåten punkt för problemet (LP_M) och $\|x\|_\infty \leq 2^{\ell(X)}$ på grund av lemma 18.4.3, satisfierar det optimala värdet $\hat{v}_M$ också olikheten

$$\hat{v}_M \leq \langle c, x\rangle + M \cdot 0 \leq |\langle c, x\rangle| \leq \|c\|_1 \cdot \|x\|_\infty \leq 2^{\ell(c)+\ell(X)} = 2^{L-m-n} \leq 2^{L-4}.$$

Genom att kombinera de båda olikheterna för $\hat{v}_M$ erhåller vi olikheten

$$2^{L-4} \geq M \cdot 2^{-2L} - 2^{2L-4},$$

som medför att

$$M \leq 2^{3L-4} + 2^{4L-4} < 2^{4L}.$$

För $M \geq 2^{4L}$ är således $\hat{x}_{n+1} = 0$. □

V. Vi är nu redo för huvudsteget i beviset för sats 18.4.2, som utgörs av följande lemma.

Lemma 18.4.5. *Antag att problemet* (LP) *har ett ändligt värde. Den vägföljande algoritmen, tillämpad på LP-problemet* (LP_M) *med* $\|x\|_\infty \leq 2^{2L}$ *som extra bivillkor,* $M = 2^{4L}$, $\epsilon = 2^{-4L}$ *och* $(0, 1)$ *som startpunkt för fas 1, kompletterad med en efterföljande reningsoperation, genererar då en optimal lösning till problemet* (LP) *efter högst* $O(m^{3/2}n^2L)$ *aritmetiska operationer.*

Bevis. Det följer av föregående lemma och uppskattningen (18.19) att LP-problemet (LP_M) för $M = 2^{4L}$ har en optimal lösning $(\hat{x}, 0)$ som uppfyller tilläggsvillkoret $\|\hat{x}\|_\infty \leq 2^{2L}$. Det LP-problem som erhålles genom att till problemet (LP_M) addera de $2n$ stycken bivillkoren

$$x_j \leq 2^{2L} \quad \text{och} \quad -x_j \leq 2^{2L}, \qquad j = 1, 2, \ldots, n,$$

har därför samma optimala värde som problemet (LP_M).

Det utvidgade problemet har $m + 2n + 2 = O(m)$ linjära bivillkor, och punkten $\overline{z} = (\overline{x}, \overline{x}_{n+1}) = (0, 1)$ är en inre punkt i den kompakta polyedern av tillåtna punkter, som vi fortsättningsvis betecknar Z. För $\epsilon = 2^{-4L}$ och med $\overline{z}$ som startpunkt stoppar därför den den vägföljande metoden enligt sats 18.3.1 efter

$$O((m + 2n + 2)^{3/2} n^2) \ln((m + 2n + 2)\Phi/\epsilon + 1) = O(m^{3/2} n^2) \ln(m 2^{4L} \Phi + 1)$$

operationer i en punkt i polyedern X', och målfunktionens värde i stoppunkten approximerar det optimala värdet $\hat{v}_M$ med ett fel som är mindre än eller lika med 2^{-4L}.

En efterföljande rening enligt metoden i sats 18.3.2 leder till en extremalpunkt $(\hat{x}, \hat{x}_{n+1})$ i polyedern X' med ett målfunktionsvärde som också är mindre än eller lika med $\hat{v}_M + 2^{-4L}$, och eftersom $2^{-4L} = 4^{-2L} < 4^{-\ell(X')}$, följer det av lemma 18.4.3 att $(\hat{x}, \hat{x}_{n+1})$ är en optimal lösning till (LP_M). På grund av lemma 18.4.4 betyder detta i sin tur att $\hat{x}$ är en optimal lösning till det ursprungliga problemet (LP).

För reningsoperationen behövs det $O(mn^2)$ aritmetiska operationer, så den totala aritmetiska kostnaden är

$$O(mn^2) + O(m^{3/2} n^2) \ln(m 2^{4L} \Phi + 1) = O(m^{3/2} n^2) \ln(m 2^{4L} \Phi + 1)$$

operationer, och det återstår nu endast att visa att $\ln(m 2^{4L} \Phi + 1) = O(L)$.

Eftersom $m \leq L$, följer detta om vi visar att $\ln \Phi = O(L)$. Per definition är

$$\Phi = \mathrm{Var}_Z(c, M) \cdot \frac{1}{1 - \pi_{\hat{z}_F}(\overline{z})},$$

där $\hat{z}_F$ är Z:s analytiska centrum med avseende på den relevanta logaritmiska barriären F. För beloppet av målfunktionens värde i en punkt $(x, x_{n+1}) \in Z$ erhålls uppskattningen

$$|\langle c, x \rangle + M x_{n+1}| \leq \|c\|_1 \|x\|_\infty + 2M \leq 2^{\ell(c)+2L} + 2 \cdot 2^{4L} \leq 2^{4L+2},$$

och funktionens maximala variation är högst dubbelt så stor, varför

$$\mathrm{Var}_Z(c, M) \leq 2^{4L+3}.$$

Den andra komponenten i Φ uppskattar vi med hjälp av sats sats 18.1.7. Låt $\overline{B}_\infty(a, a_{n+1}; r)$ beteckna den slutna bollen i $\mathbf{R}^{n+1} = \mathbf{R}^n \times \mathbf{R}$ med radie

r och centrum i punkten (a, a_{n+1}) när avståndet ges av maxnormen i $\mathbf{R}^{n+1}$, dvs.

$$\overline{B}_\infty(a, a_{n+1}; r) = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x - a\|_\infty \le r,\ \ |x_{n+1} - a_{n+1}| \le r\}.$$

Per definition inkluderas polyedern Z i $\overline{B}_\infty(0, 0; 2^{2L})$. Å andra sidan ligger den lilla bollen $\overline{B}_\infty(\overline{z}; 2^{-L})$ helt i Z, ty om $\|x\|_\infty \le 2^{-L}$ och $|x_{n+1} - 1| \le 2^{-L}$, så är

$$\begin{aligned}
&\sum_{j=1}^{n} a_{ij}x_j + (b_i - 1)x_{n+1} - b_i = \sum_{j=1}^{n} a_{ij}x_j + b_i(x_{n+1} - 1) - x_{n+1} \\
&\quad \le \sum_{j=1}^{n} |a_{ij}||x_j| + |b_i||x_{n+1} - 1| - x_{n+1} \le 2^{-L}\Big(\sum_{j=1}^{n} |a_{ij}| + |b_i|\Big) - (1 - 2^{-L}) \\
&\quad \le 2^{-L+\ell(X)} + 2^{-L} - 1 \le 2^{-4} + 2^{-L} - 1 < 0,
\end{aligned}$$

vilket visar att den i:te olikheten i systemet $Ax + (b - 1)x_{n+1} \le b$ gäller med strikt olikhet för $i = 1, 2, \ldots, m$, och övriga olikheter som definierar polyedern Z är uppenbarligen strikt uppfyllda.

Det följer därför av sats 18.1.7 att

$$\pi_{\hat{z}_F}(\overline{z}) \le \frac{2 \cdot 2^{2L}}{2 \cdot 2^{2L} + 2^{-L}},$$

och att följaktligen

$$\frac{1}{1 - \pi_{\hat{z}_F}(\overline{z})} \le 2 \cdot 2^{3L} + 1 < 2^{3L+2}.$$

Detta medför att $\Phi \le 2^{4L+3} \cdot 2^{3L+2} = 2^{7L+5}$ och att således $\ln \Phi = O(L)$. Därmed är beviset för lemmat klart. □

VI. Det återstår nu att visa att man med $O(m^{7/2}L)$ operationer kan avgöra om det ursprungliga problemet (LP):s värde är $+\infty$, $-\infty$ eller ändligt.

För att avgöra om värdet är $+\infty$ eller ej, dvs. om polyedern X är tom eller ej, betraktar vi det artificiella LP-problemet

$$\begin{aligned}
&\min \ \ x_{n+1} \\
&\text{då} \quad \begin{cases} Ax - \mathbf{1}x_{n+1} \le b \\ \qquad -x_{n+1} \le 0 \end{cases}
\end{aligned}$$

Detta problem har garanterat tillåtna punkter eftersom $(0, t)$ uppfyller samtliga bivillkor för tillräckligt stora positiva tal t. Vidare är uppenbarligen problemets optimala värde större än eller lika med noll, och värdet är lika med noll om och endast om $X \neq \emptyset$.

Vi kan således avgöra om polyedern X är tom eller ej genom att bestämma en optimal lösning till det artificiella problemet. Det artificiella problemets inputlängd är lika med $\ell(X) + 2m + n + 4$, och eftersom detta tal är $\leq 2L$, följer det av lemma 18.4.5 att vi kan avgöra huruvida polyedern X är tom eller ej med $O(m^{3/2}n^2L)$ aritmetiska operationer.

Observera att vi inte behöver lösa det artificiella problemet exakt. Om värdet är större än noll, så är det nämligen på grund av lemma 18.4.3 större än eller lika med 2^{-2L}. Det är därför tillräckligt att bestämma en punkt som approximerar värdet med ett fel som understiger 2^{-2L} för att veta om värdet är noll eller ej.

VII. Om polyedern X inte är tom, så behöver vi som nästa punkt avgöra om målfunktionen är nedåt begränsad, och så är fallet om och endast om det duala problemet till (LP) har tillåtna punkter. Det duala maximeringsproblemet är ekvivalent med minimeringsproblemet

$$\begin{array}{ll} \min & \langle -b, y\rangle \\ \text{då} & \begin{cases} A^Ty \leq c \\ -A^Ty \leq -c \\ -y \leq 0, \end{cases} \end{array}$$

som har m variabler, $2n + m$ $(= O(m))$ bivillkor och inputlängd

$$2\ell(A) + m + 2\ell(c) + \ell(b) + m + (2n + m) \leq 2L + m \leq 3L.$$

Det följer därför av steg VI att vi kan avgöra om det duala problemet har några tillåtna punkter med $O(m^{7/2}L)$ operationer.

Därmed är beviset för sats 18.4.2 komplett. □

Övningar

18.1 Visa att om funktionerna f_i är ν_i-självkonkordanta barriärer till delmängderna X_i av $\mathbf{R}^{n_i}$, så är funktionen $f(x_1, \ldots, x_m) = f_1(x_1) + \cdots + f_m(x_m)$ en $(\nu_1 + \cdots + \nu_m)$-självkonkordant barriär till produktmängden $X_1 \times \cdots \times X_m$.

18.2 Visa att den till funktionen f hörande duala lokala normen $\|v\|_x^*$ är ändlig om och endast om v är en vektor i $\mathcal{N}(f''(x))^\perp$ och att restriktionen av $\|\cdot\|_x^*$ till $\mathcal{N}(f''(x))^\perp$ är en äkta norm.

18.3 Låt X vara en sluten äkta konvex kon med icke-tomt inre, låt $\nu \geq 1$ vara ett reellt tal, och antag att funktionen $f\colon \operatorname{int} X \to \mathbf{R}$ är sluten och självkonkordant och att $f(tx) = f(x) - \nu \ln t$ för alla $x \in \operatorname{int} X$ och alla $t > 0$. Visa att

a) $f'(tx) = t^{-1}f'(x)$ b) $f'(x) = -f''(x)x$ c) $\lambda(f,x) = \sqrt{\nu}$.

Funktionen f är således en ν-självkonkordant barriär till X.

18.4 Visa att icke-negativa ortanten $X = \mathbf{R}^n_+$, $\nu = n$ och den logaritmiska barriären $f(x) = -\sum_{i=1}^n \ln x_i$ uppfyller förutsättningarna i föregående övning.

18.5 Sätt $X = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid x_{n+1} \geq \|x\|_2\}$.

a) Visa att funktionen $f(x) = -\ln(x_{n+1}^2 - (x_1^2 + \cdots + x_n^2))$ är självkonkordant på mängden $\operatorname{int} X$.

b) Visa att X, $\nu = 2$ och f uppfyller förutsättningarna i övning 18.3. Funktionen f är således en 2-självkonkordant barriär till X.

18.6 Antag att funktionen $f\colon \mathbf{R}_{++} \to \mathbf{R}$ är konvex och tre gånger kontinuerligt deriverbar samt att

$$|f'''(x)| \leq 3\frac{f''(x)}{x}$$

för alla $x > 0$. Funktionen

$$F(x,y) = -\ln(y - f(x)) - \ln x$$

med $X = \{(x,y) \in \mathbf{R}^2 \mid x > 0,\, y > f(x)\}$ som definitionsmängd är självkonkordant enligt övning 16.3. Visa att funktionen är en 2-självkonkordant barriär till slutna höljet $\operatorname{cl} X$.

18.7 Visa att funktionen

$$F(x,y) = -\ln(y - x\ln x) - \ln x$$

är en 2-självkonkordant barriär till epigrafen

$$\{(x,y) \in \mathbf{R}^2 \mid y \geq x\ln x,\, x \geq 0\}.$$

18.8 Visa att funktionen

$$G(x,y) = -\ln(\ln y - x) - \ln y$$

är en 2-självkonkordant barriär till epigrafen $\{(x,y) \in \mathbf{R}^2 \mid y \geq \mathrm{e}^x\}$.

Bibliografiska och historiska notiser

Standardverk i konvex analys är Rockafellar [1] från 1970 och Hiriart-Urutty–Lemarechal [1] från 1993. I Rockafellars bok kan så gott som samtliga resultat från kapitel 1–10 i den här boken återfinnas i en eller annan form. Rockafellars bok innehåller också en historisk översikt med referenser till originalarbeten inom området.

En mer lättillgänglig bok i samma ämne är Webster [1]. Bland läroböcker i konvexitet med tonvikten lagd på polyedrar kan nämnas Stoer–Witzgall [1] och den mer kombinatoriskt inriktade Grünbaum [1].

En modern lärobok i konvex optimering är Boyd & Vandenberghe [1], som förutom teori och algoritmer också innehåller intressanta tillämpningar från en mängd olika områden.

Del I

Den allmänna konvexitetsteorin grundlades kring sekelskiftet 1900 av Hermann Minkowski [1, 2] som en biprodukt till hans talteoretiska studier. Bland annat inför Minkowski begreppen separation och extremalpunkt, och han visar varje kompakt konvex mängd är lika med konvexa höljet av sina extremalpunkter och att varje polyeder är ändligt genererad, dvs. den ena riktningen av sats 5.3.1 − omvändningen noterades senare av Weyl [1].

Begreppet dualkon introducerades av Steinitz [1], som även visade grundläggande resultat för recessionskonen.

Teorin för linjära olikheter är förvånansvärt ung − ett specialfall av sats 3.3.7 (övning 3.11a) visades av Gordan [1], den algebraiska versionen av Farkas lemma, dvs. korollarium 3.3.3, finns i Farkas [1], och ett närbesläktat resultat (övning 3.11b) ges av Stiemke [1]. För den första systematiska framställningen av teorin står Weyl [1] och Motzkin [1]. Betydelsefulla bidrag har också lämnats av Tucker [1]. Beviset i kapitel 3 för Farkas lemma är geometriskt till sin natur; ett alternativt algebraiskt induktionsbevis för lemmat

finns i Kuhn [1].

Extremalpunkter och fasader behandlas utförligt i Klee [1,2].

Jensen [1] studerar konvexa funktioner av en reell variabel och visar att konvexa funktioner med $\mathbf{R}$ som definitionsmängd är kontinuerliga och har enkelsidiga derivator överallt. Jensens olikhet visades dock tidigare av Hölder [1] för funktioner med positiv andraderivata.

Konjugatfunktionen införs av Fenchel [1], och en modern framställning av teorin för konvexa koner, mängder och funktioner ges i Fenchel [2], som bl.a. innehåller originalresultat om tillslutning av konvexa funktioner och subdifferentialen.

Del II

Det tidigast kända exemplet på linjär programmering återfinns i Fouriers arbeten från 1820-talet (Fourier [1]) och handlar om problemet att bestämma den i maximumnorm bästa anpassningen till ett överbestämt linjärt ekvationssystem. Fourier reducerade detta problem till att minimera en linjär form över en polyeder och antydde också en med simplexalgoritmen ekvivalent metod för att beräkna minimum.

Det skulle emellertid dröja ända fram 1940-talet innan man i större skala började formulera praktiska problem som linjär programmering. Transportproblemet formulerades av Hitchcock [1], som också angav en konstruktiv lösningsmetod, och dietproblemet studerades av Stigler [1], som dock inte lyckades ange någon fullständig lösning. Den ryske matematikern och ekonomen Kantorovich [1] hade något tidigare formulerat och löst LP-problem inom produktionsplanering, men hans arbeten uppmärksammades inte utanför Sovjetunionen och kom därför inte att påverka den fortsatta utvecklingen.

I samband med andra världskriget hade det uppstått behov av matematiska metoder för att lösa militära planeringsproblem, och 1947 arbetade en grupp matematiker under ledning av George Dantzig och Marshall Wood vid U.S. Department of the Air Force med sådana problem. Gruppens arbete resulterade i en insikt om betydelsen av linjär programmering, och den första versionen av simplexalgoritmen beskrevs av Dantzig [1] och Wood–Dantzig [1].

Att simplexalgoritmen kom nästan samtidigt med de första datorerna gjorde det plötsligt möjligt att behandla stora problem numeriskt och bidrog till genombrottet för linjär programmering. Ett viktigt led i populariseringen blev också den konferens om linjär programmering, som Tjalling Koopmans arrangerade 1949 i Chicago. Under konferensen presenterade ekonomer, matematiker och statistiker uppsatser om linjär programmering, som senare

publicerades i Koopmans [1], och denna bok blev startskottet för en snabbt växande litteratur om linjär programmering.

Teorin för konvexa program har sitt ursprung i en uppsats av Kuhn–Tucker [1], som behandlar nödvändiga och tillräckliga villkor för optimalitet i icke-linjära problem. Kuhn–Tucker noterade sambandet mellan Lagrangemultiplikatorer och sadelpunkter, och de fokuserade på konvexitetens roll i sammanhanget. Ett besläktat resultat med Lagrangemultiplikatorvillkor hade annars visats tidigare av John [1] för allmänna differentierbara olikhetsbivillkor, och KKT-villkoren förekommer redan i en opublicerad masteruppsats av Karush [1]. Sats 11.2.1 visades av Uzawa [1].

Dualitetssatsen i linjär programmering var känd som ett resultat inom spelteori av John von Neumann, men det först publicerade beviset för denna sats finns i Gale–Kuhn–Tucker [1].

Bland tidiga läroböcker inom linjär programmering kan nämnas Dantzig [4], som förutom det matematiska materialet även innehåller en grundlig historisk översikt, många tillämpningar och en omfattande bibliografi, och Gale [1], som ger en koncis men matematiskt stringent framställning av linjär programmering med tonvikten lagd på ekonomiska tillämpningar. Bland lite nyare böcker kan nämnas Chvatal [1] och Luenberger [1].

Del III

Dantzig [2] behandlade i sin grundläggande artikel 1951 det icke-degenererade fallet av simplexalgoritmen, och möjligheten av cykling vid degeneration vållade till en början en del bekymmer. Det första exemplet på cykling konstruerades av Hoffman [1], men redan före denna upptäckt hade Charnes [1] föreslagit en metod för att undvika cykling. Andra sådana metoder gavs sedan av Dantzig–Orden–Wolfe [1] och Wolfe [2]. Blands [1] enkla pivoteringsregel är av relativt sent datum.

Det är enkelt att modifiera simplexalgoritmen så att den är direkt applicerbar på problem med begränsade variabler, vilket först noterades av Charnes–Lemke [1] och Dantzig [3].

Den duala simplexalgoritmen utvecklades av Beale [1] och Lemke [1]. De idag effektivaste varianterna av simplexalgoritmen är primala-duala algoritmer.

Konvex-kvadratiska program kan lösas med en variant av simplexalgoritmen, som formulerats av Wolfe [1].

Khachiyans [1] komplexitetsresultat bygger på ellipsoidalgoritmen, som först föreslogs av Shor [1] som metod inom allmän konvex optimering. Se Bland–Goldfarb–Todd [1] för en översikt över ellipsoidmetoden.

Efter publiceringen av Karmarkars [1] algoritm utvecklades många varianter. Algoritmer för LP-problem med $O(n^3L)$ som komplexitetsgräns har beskrivits av Gonzaga [1] och Ye [1].

Del IV

Newtons metod är en klassisk iterativ algoritm för att hitta kritiska punkter till differentierbara funktioner. Att Newtons metod är lokalt kvadratiskt konvergent för funktioner med Lipschitzkontinuerlig, positivt definit andraderivata i en omgivning av den kritiska punkten visades av Kantorovich [2].

Under 1950-talet började man använda barriärmetoder för att lösa ickelinjära optimeringsproblem. Centrala vägen för logaritmiska barriärer studerades av Fiacco och McCormick, och deras bok om sekventiell minimeringsteknik – Fiacco–McCormick [1], först publicerad 1968 – är ett standardverk inom området. Metoderna fungerade oftast bra i praktiken, men det saknades teoretiska komplexitetsresultat. De förlorade i popularitet på 1970-talet för att sedan få en renässans i efterdyningarna till Karmarkars upptäckt.

I Karmarkars [1] polynomiella algoritm för linjär programmering avbildas i varje iteration polyedern av tillåtna punkter och den aktuella approximativa lösningen x_k först med hjälp av en projektiv skalningstransformation på en ny polyeder och en ny punkt x'_k som ligger nära centrum av den nya polyedern. Därefter utförs ett steg i det transformerade rummet som resulterar i en punkt x_{k+1} med lägre målfunktionsvärde. Framstegen mäts med hjälp av en logaritmisk potentialfunktion.

Man fann snart att Karmarkars potentialreducerande algoritm var besläktad med tidigare studerade vägföljande metoder, och Renegar [1] och Gonzaga [1] lyckades visa att den vägföljande metoden med logaritmisk barriär är polynomiell för LP-problem.

En allmän presentation av linjär programmering och utvecklingen på algoritmområdet fram till slutet av 1980-talet (ellipsoidmetoden, Karmarkars algoritm, m. m.) ges av Goldfarb–Todd [1]. En översikt över potentialreducerande algoritmer ges av Todd [1], medan Gonzaga [2] beskriver utvecklingen av vägföljande algoritmer fram till år 1992.

Ett genombrott inom konvex optimering skedde i slutet av 1980-talet, när Yurii Nesterov upptäckte att Gonzagas och Renegars bevis bara utnyttjade två egenskaper hos den logaritmiska barriärfunktionen, nämligen att den uppfyller de två differentialolikheter som med Nesterovs terminologi innebär att barriären är självkonkordant med ändlig parameter ν. Eftersom det går att konstruera explicita, beräkningsbara självkonkordanta barriärer för ett antal viktiga typer av konvexa mängder, kunde de teoretiska komplexitetsresultaten för linjär programmering nu utvidgas till att också gälla

för en stor klass av konvexa optimeringsproblem, och tillsammans med Nemirovskii utvecklade Nesterov algoritmer för konvex optimering som bygger på självkonkordanta barriärer. Se Nesterov–Nesterovski [1].

Referenser

Adler, I. & Megiddo, N.

[1] A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension. *J. ACM* 32 (1985), 871–895.

Beale, E.M.L

[1] An alternative method of linear programming, *Proc. Cambridge Philos. Soc.* 50 (1954), 513–523.

Bland, R.G.

[1] New Finite Pivoting Rules for the Simplex Method, *Math. Oper. Res.* 2 (1977), 103–107.

Bland, R.G., Goldfarb, D. & Todd, M.J.

[1] The ellipsoid method: A survey, *Oper. Res.* 29 (1981), 1039–1091.

Borgwardt, K.H.

[1] *The Simplex Method − a probabilistic analysis,* Springer-Verlag, 1987.

Boyd, S. & Vandenberghe, L.

[1] *Convex Optimization*, Cambridge Univ. Press, Cambridge, UK, 2004.

Charnes, A.

[1] Optimality and Degeneracy in Linear Programming, *Econometrica* 20 (1952), 160–170.

Charnes, A. & Lemke, C.E.

[1] *The bounded variable problem.* ONR memorandum 10, Carnegie Institute of Technology, 1954.

Chvátal, V.

[1] *Linear Programming.* W.H. Freeman, 1983.

Dantzig, G.B.

[1] Programming of Interdependent Activities. II. Mathematical Model, *Econometrica* 17 (1949), 200–211.

[2] Maximization of Linear Functions of Variables Subject to Linear Inequalities. Sid 339–347 i T.C. Koopmans (ed.), *Activity Analysis of Production and Allocation*, John Wiley, 1951.

[3] Upper Bounds, Secondary Constraints and Block Triangularity in Linear Programming, *Econometrica* 23 (1955), 174–183.

[4] *Linear Programming and Extensions.* Princeton University Press, 1963.

Dantzig, G.B., Orden, A. & Wolfe, P.
[1] The genereralized simplex method for minimizing a linear form under linear inequality constraints, *Pacific J. Math.* 5 (1955), 183–195.

Farkas, J.
[1] Theorie der einfachen Ungleichungen, *J. reine angew. Math.* 124 (1902), 1–27.

Fenchel, W.
[1] On conjugate convex functions. *Canad. J. Math.* 1 (1949), 73-77.
[2] *Convex Cones, Sets and Functions.* Lecture Notes, Princeton University, 1951.

Fiacco, A.V. & McCormick, G.P.
[1] *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Society for Industrial and Applied Mathematics, 1990. (Först publicerad 1968 av Research Analysis Corporation.)

Fourier, J.-B.
[1] Solution d'une question particulière du calcul des inégalités. Sid 317–328 i *Oeuvres de Fourier II*, 1890.

Gale, D.
[1] *The Theory of Linear Economic Models.* McGraw–Hill, 1960.

Gale, D., Kuhn, H.W. & Tucker, A.W.
[1] Linear programming and the theory of games. Sid 317–329 i Koopmans, T.C. (ed.), *Activity Analysis of Production and Allocation*, John Wiley & Sons, 1951.

Goldfarb, D.G. & Todd, M.J.
[1] Linear programming. Kapitel 2 i Nemhauser, G.L. et al. (eds.), *Handbooks in Operations Research and Management Science, vol. 1: Optimization*, North-Holland, 1989.

Gonzaga, C.C.
[1] An algorithm for solving linear programming problems in $O(n^3L)$ operations. Sid 1–28 i Megiddo, N. (ed.), *Progress in Mathematical Programming: Interior-Point and Related Methods,* Springer-Verlag, 1988.
[2] Path-Following Methods for Linear Programming, *SIAM Rev.* 34 (1992), 167–224.

Gordan, P.
[1] Über die Auflösung linearer Gleichungen mit reellen Coefficienten, *Math. Ann.* 6 (1873), 23–28.

Grünbaum, B.
[1] *Convex Polytopes.* Interscience publishers, New York, 1967.

Hiriart-Urruty, J.-B. & Lemaréchal, C.
[1] *Convex Analysis and Minimization Algorithms.* Springer, 1993.

Hitchcock, F.L.
[1] The distribution of a product from several socurces to numerous localities, *J. Math. Phys.* 20 (1941), 224–230.

Hoffman, A. J.
[1] *Cycling in the Simplex Algorithm.* Report No. 2974, National Bureau of Standards, Gaithersburg, MD, USA, 1953.

Hölder, O.
[1] Über einen Mittelwertsatz, *Nachr. Ges. Wiss. Göttingen*, 38–47, 1989.

Jensen, J.L.W.V.
[1] Sur les fonctions convexes et les inégalités entre les valeur moyennes, *Acta Math.* 30 (1906), 175–193.

John, F.
[1] Extremum problems with inequalities as subsidiary conditions, 1948. Sid 543–560 i Moser J. (ed.), *Fritz John, Collected Papers*, Birkhäuser Verlag, 1985.

Kantorovich, L.V.
[1] Mathematical methods of organizing and planning production, Leningrad State Univ. (på ryska), 1939. Eng. översättning i *Management Sci.* 6 (1960), 366-422.
[2] *Functional Analysis and Applied Mathematics.* National Bureau of Standards, 1952. (Först publicerad på ryska 1948.)

Karmarkar, N.
[1] A new polynomial-time algorithm for linear programming, *Combinatorica* 4 (1984), 373–395.

Karush, W.
[1] *Minima of Functions of Several Variables with Inequalities as Side Constraints.* M. Sc. Disseratation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

Khachiyan, L.G.
[1] A polynomial algorithm in linear programming, *Dokl. Akad. Nauk SSSR* 244 (1979), 1093–1096. Eng. översättning i *Soviet Math. Dokl.* 20 (1979), 191-194.

Klee, V.
[1] Extremal structure of convex sets, *Arch. Math.* 8 (1957), 234–240.
[2] Extremal structure of convex sets, II, *Math. Z.* 69 (1958), 90–104.

Klee, V. & Minty, G.J.
[1] How Good is the Simplex Algorithm? Sid 159–175 i Shisha, O. (ed.), *Inequalities, III,* Academic Press, 1972.

Koopmans, T.C., ed.
[1] *Activity Analysis of Production and Allocation.* John Wiley & Sons, 1951.

Kuhn, H.W.
[1] Solvability and Consistency for Linear Equations and Inequalities, *Amer. Math. Monthly* 63 (1956), 217–232.

Kuhn, H.W. & Tucker, A.W.
[1] Nonlinear programming. Sid 481–492 i *Proc. of the second Berkeley Symposium on Mathematical Statistics and Probability.* Univ. of California Press, 1951.

Lemke, C.E.
[1] The dual method of solving the linear programming problem, *Naval Res. Logist. Quart.* 1 (1954), 36–47.

Luenberger, D.G.
[1] *Linear and Nonlinear Programming.* Addison–Wesley, 1984

Minkowski, H.
[1] *Geometrie der Zahlen.* Teubner, Leipzig, 1910.
[2] *Gesammelte Abhandlungen von Hermann Minkowski, Vol. 1, 2.* Teubner, Leibzig, 1911

Motzkin, T.
[1] *Beiträge zur Theorie der linearen Ungleichungen.* Azviel, Jerusalem, 1936.

Nesterov, Y. & Nemirovskii, A.
[1] *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

Renegar, J.
[1] A polynomial-time algorithm based on Newton's method for linear programming, *Math. Programm.* 40 (1988), 59–94.

Rockafellar, R.T.
[1] *Convex Analysis.* Princeton Univ. Press., 1970

Shor, N.Z.
[1] Utilization of the operation of space dilation in the minimization of convex functions, *Cybernet. System Anal.* 6 (1970), 7–15.

Smale, S.
[1] On the average number of steps in the simplex method of linear programming, *Math. Program.* 27 (1983), 241–262.

Steinitz, E.
[1] Bedingt konvergente Reihen und konvexe Systeme, I, II, III, *J. Reine Angew. Math.* 143 (1913), 128–175; 144 (1914), 1–40; 146 (1916), 1–52.

Stiemke, E.
[1] Über positive Lösungen homogener linearer Gleichungen, *Math. Ann.* 76 (1915), 340–342.

Stigler, G.J.
[1] The Cost of Subsistence, *J. Farm Econ.* 27 (1945), 303–314.

Stoer, J. & Witzgall, C.
[1] *Convexity and Optimization in Finite Dimensions I.* Springer-Verlag, 1970.

Todd, M.
[1] Potential-reduction methods in mathematical programming, *Math. Program.* 76 (1997), 3–45.

Tucker, A.W.
[1] Dual Systems of Homogeneous Linear Relations. Sid 3–18 i Kuhn, H.W. & Tucker, A.W. (eds.), *Linear Inequalities and Related Systems*, Princeton Univ. Press, 1956.

Uzawa, H.
[1] The Kuhn–Tucker theorem in concave programming. I Arrow, K.J., Hurwicz, L. & H. Uzawa, H. (eds.), *Studies in Linear and Non-linear Programming,* Stanford Univ. Press, 1958.

Webster, R.
[1] *Convexity.* Oxford University Press, 1954.

Weyl, H.
[1] Elementare Theorie der konvexen Polyeder, *Comment. Math. Helv.* 7 (1935), 290–306.

Wolfe, P.
[1] The Simplex Method for Quadratic Programming, *Econometrica* 27 (1959), 382–398.
[2] A Technique for Resolving Degeneracy in Linear Programming, *J. of the Soc. for Industrial and Applied Mathematics* 11 (1963), 205–211.

Wood, M.K. & Dantzig, G.B.
[1] Programming of Interdependent Activities. I. General discussion, *Econometrica* 17 (1949), 193–199.

Wright, S.
[1] *Primal-dual interior-point methods.* SIAM Publications, 1997.

Ye, Y.

[1] An $O(n^3L)$ potential reduction algorithm for linear programming. *Math. Program.* 50 (1991), 239–258.

[2] *Interior point algorithms.* John Wiley and Sons, 1997.

Svar och lösningar till övningarna

Kapitel 2

2.2 a) $\{x \in \mathbf{R}^2 \mid 0 \le x_1 + x_2 \le 1,\ x_1, x_2 \ge 0\}$
b) $\{x \in \mathbf{R}^2 \mid \|x\| \le 1\}$
c) $\mathbf{R}^2_{++} \cup \{(0,0)\}$

2.3 T. ex. $\{(0,1)\} \cup (\mathbf{R} \times \{0\})$ i $\mathbf{R}^2$.

2.4 $\{x \in \mathbf{R}^3_{++} \mid x_3^2 \le x_1 x_2\}$

2.5 Utnyttja triangelolikheten

$$\left(\sum_1^n (x_j + y_j)^2\right)^{1/2} \le \left(\sum_1^n x_j^2\right)^{1/2} + \left(\sum_1^n y_j^2\right)^{1/2}$$

för att visa att mängden är sluten under addition av vektorer. Alternativt kan man utnyttja perspektivavbildningen; se exempel 2.3.4.

2.6 Följer av att $-\mathbf{e}_k$ är en konisk kombination av vektorerna $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_n$.

2.7 Låt X vara halvrummet $\{x \in \mathbf{R}^n \mid \langle c, x\rangle \ge 0\}$. För varje $x \in X$ ligger vektorn $y = x - \langle c, x\rangle \|c\|^{-2} c$ i det $(n-1)$-dimensionella delrummet $Y = \{x \in \mathbf{R}^n \mid \langle c, x\rangle = 0\}$, som enligt föregående övning genereras som kon av n vektorer. Vektorn x är en konisk kombination av dessa n vektorer och vektorn c.

2.8 Snittet mellan konen X och enhetscirkeln är en sluten cirkelbåge med ändpunkter i punkterna x och y, säg. Cirkelbågens längd är antingen mindre än π, lika med π eller lika med 2π. I det förstnämnda fallet är konen X äkta och genereras av de två vektorerna x och y. I de båda övriga fallen är X ett halvplan resp. hela $\mathbf{R}^2$ och genereras av tre vektorer.

2.9 Utnyttja övning 2.8.

2.10 a) $\operatorname{recc} X = \{x \in \mathbf{R}^2 \mid x_1 \ge x_2 \ge 0\}$, $\operatorname{lin} X = \{(0,0)\}$
b) $\operatorname{recc} X = \operatorname{lin} X = \{(0,0)\}$

2.10 c) $\operatorname{recc} X = \{x \in \mathbf{R}^3 \mid 2x_1 + x_2 + x_3 \le 0,\ x_1 + 2x_2 + x_3 \le 0\}$,
$\operatorname{lin} X = \{(t, t, -3t) \mid t \in \mathbf{R}\}$

d) $\operatorname{recc} X = \{x \in \mathbf{R}^3 \mid x_1 \ge |x_2|\}$,
$\operatorname{lin} X = \{x \in \mathbf{R}^3 \mid x_1 = x_2 = 0\}$.

2.12 b) (i) $c(X) = \{x \in \mathbf{R}^2 \mid 0 \le \frac{1}{3}x_1 \le x_2 \le \frac{1}{2}x_1\} = \operatorname{cl}(c(X))$

(ii) $c(X) = \{x \in \mathbf{R}^2 \mid 0 < x_2 \le \frac{1}{2}x_1\} \cup \{(0,0)\}$,
$\operatorname{cl}(c(X) = \{x \in \mathbf{R}^2 \mid 0 \le x_2 \le \frac{1}{2}x_1\}$,

(iii) $c(X) = \{x \in \mathbf{R}^3 \mid x_1 x_3 \ge x_2^2,\ x_3 > 0\} \cup \{(0,0,0)\}$,
$\operatorname{cl}(c(X)) = c(X) \cup \{(x_1, 0, 0) \mid x_1 \ge 0\}$.

c) $c(X) = \{(x, x_{n+1}) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\| \le x_{n+1}\}$.

2.14 Låt z_0 vara en randpunkt till $X + Y$. Då finns det en följd $(z_n)_1^\infty$ av punkter $z_n \in X + Y$ sådana att $z_n \to z_0$ då $n \to \infty$. Sätt $z_n = x_n + y_n$ med $x_n \in X$ och $y_n \in Y$. Eftersom mängden Y är kompakt, har följden $(y_n)_1^\infty$ enligt Bolzano–Weierstrass sats en delföljd $(y_{n_k})_{k=1}^\infty$ som konvergerar mot en punkt $y_0 \in Y$.

Betrakta nu motsvarande delföljd $(x_{n_k})_{k=1}^\infty$; eftersom $x_{n_k} = z_{n_k} - y_{n_k}$, är också denna delföljd konvergent med gränsvärde $x_0 = z_0 - y_0$, och eftersom mängden X är sluten, ligger gränspunkten x_0 i X. Detta visar att $z_0 = x_0 + y_0$ ligger i $X + Y$. Mängden $X + Y$ innehåller således alla sina randpunkter och är därför en sluten mängd.

Kapitel 3

3.1 T. ex. $\{x \in \mathbf{R}^2 \mid x_2 \le 0\}$ och $\{x \in \mathbf{R}^2 \mid x_2 \ge \mathrm{e}^{x_1}\}$.

3.2 Följer av sats 3.1.3 för slutna mängder och av sats 3.1.5 för öppna mängder.

3.4 a) $\mathbf{R}_+ \times \mathbf{R}$ b) $\{0\} \times \mathbf{R}$ c) $\{0\} \times \mathbf{R}_+$ d) $\mathbf{R}_+ \times \mathbf{R}_+$
e) $\{x \in \mathbf{R}^2 \mid x_2 \ge x_1 \ge 0\}$

3.6 a) $X = X^{++} = \{x \in \mathbf{R}^2 \mid x_1 + x_2 \ge 0,\ x_2 \ge 0\}$,
$X^+ = \{x \in \mathbf{R}^2 \mid x_2 \ge x_1 \ge 0\}$

b) $X = X^{++} = \mathbf{R}^2$, $X^+ = \{(0,0)\}$

c) $X = \mathbf{R}^2_{++} \cup \{(0,0)\}$, $X^+ = X^{++} = \mathbf{R}^2_+$

3.7 (i) $\Rightarrow$ (iii): Eftersom $-a_j \notin \operatorname{con} A$, finns det för varje j en vektor c_j så att $-\langle c_j, a_j\rangle < 0$ och $\langle c_j, x\rangle \ge 0$ för alla $x \in \operatorname{con} A$, vilket medför att $\langle c_j, a_j\rangle > 0$ och $\langle c_j, a_k\rangle \ge 0$ för alla k. Det följer att $c = c_1 + c_2 + \cdots + c_m$ duger.

(iii) $\Rightarrow$ (ii): Antag att $\langle c, a_j\rangle > 0$ för alla j. Då medför $\sum_1^m \lambda_j a_j = 0$ att $0 = \langle c, 0\rangle = \sum_1^m \lambda_j \langle c, a_j\rangle$, och om $\lambda_j \ge 0$ för alla j är därför $\lambda_j \langle c, a_j\rangle = 0$ för alla j med slutsatsen att $\lambda_j = 0$.

(ii) $\Rightarrow$ (i): Om det finns en vektor x så att $x = \sum_1^m \lambda_j a_j$ och $-x = \sum_1^m \mu_j a_j$ med icke-negativa skalärer λ_j, μ_j, så får vi genom addition $\sum_1^m (\lambda_j + \mu_j) a_j = 0$ med slutsatserna $\lambda_j + \mu_j = 0$, $\lambda_j = \mu_j = 0$ och $x = 0$.

3.8 Saknar lösning.

3.10 Lösbart för $\alpha \leq -2$, $-1 < \alpha < 1$ och $\alpha > 1$.

3.11 a) Systemen (S) och (S*) är ekvivalenta med systemen

$$\begin{cases} Ax \geq 0 \\ -Ax \geq 0 \\ Ex \geq 0 \\ \mathbf{1}^T x > 0 \end{cases} \qquad \text{resp.} \qquad \begin{cases} A^T(y' - y'') + Ez + \mathbf{1}t = 0 \\ y', y'', z \geq 0,\ t > 0, \end{cases}$$

där y motsvaras av $y'' - y'$. Sats 3.3.7 kan nu tillämpas.

b) Systemen (S) och (S*) är ekvivalenta med systemen

$$\begin{cases} Ax \geq 0 \\ -Ax \geq 0 \\ Ex > 0 \end{cases} \qquad \text{resp.} \qquad \begin{cases} A^T(y' - y'') + Ez = 0 \\ y', y'', z \geq 0,\ z \neq 0, \end{cases}$$

där y motsvaras av $y'' - y'$. Påståendet följer därför av sats 3.3.7.

3.12 Enligt sats 3.3.7 är systemet lösbart om och endast om det duala systemet

$$\begin{cases} A^T(y' - y'') + z + u = 0 \\ A(w + u) = 0 \\ y', y'', z, w, u \geq 0,\ u \neq 0 \end{cases}$$

saknar lösning. Av det duala systemets båda ekvationer följer:

$$\begin{aligned} 0 &= -(w+u)^T A^T = -(w+u)^T A^T (y' - y'') = (w+u)^T (z+u) \\ &= w^T z + w^T u + u^T z + u^T u, \end{aligned}$$

och eftersom samtliga fyra termer i den sista summan är icke-negativa, är speciellt $u^T u = 0$, vilket medför att $u = 0$. Det duala systemet är således olösbart.

Kapitel 4

4.1 a) $\operatorname{ext} X = \{(1,0), (0,1)\}$ b) $\operatorname{ext} X = \{(0,0), (1,0), (0,1), (\frac{1}{2}, 1)\}$
c) $\operatorname{ext} X = \{(0,0,1), (0,0,-1)\} \cup \{(x_1, x_2, 0) \mid (x_1 - 1)^2 + x_2^2 = 1\} \setminus \{(0,0,0)\}$.

4.2 Antag att $x \in \operatorname{cvx} A \setminus A$; då är $x = \lambda a + (1-\lambda)y$ där $a \in A$, $y \in \operatorname{cvx} A$ och $0 < \lambda < 1$. Det följer att $x \notin \operatorname{ext}(\operatorname{cvx} A)$.

4.3 Enligt föregående övning är $\operatorname{ext} X \subseteq A$. Antag $a \in A \setminus \operatorname{ext} X$. Då är $a = \lambda x_1 + (1-\lambda)x_2$, där $x_1, x_2 \in X$, $x_1 \neq x_2$ och $0 < \lambda < 1$. Sätt $x_i = \mu_i a + (1-\mu_i)y_i$, där $0 \leq \mu_i < 1$ och $y_i \in \operatorname{cvx}(A \setminus \{a\})$. Det följer nu av likheten $a = (1-\lambda\mu_1-(1-\lambda)\mu_2)^{-1}(\lambda(1-\mu_1)y_1+(1-\lambda)(1-\mu_2)y_2)$ att a ligger i $\operatorname{cvx}(A\setminus\{a\})$, och därför är $\operatorname{cvx}(A\setminus\{a\}) = \operatorname{cvx} A = X$, vilket strider mot minimalitetsantagandet för A. Denna motsägelse visar att $\operatorname{ext} X = A$.

4.4 Mängden $X \setminus \{x_0\}$ är konvex om och endast om $]a,b[\subseteq X \setminus \{x_0\}$ för alla $a, b \in X \setminus \{x_0\}$, dvs. om och endast om $x_0 \notin]a,b[$ för alla $a, b \in X \setminus \{x_0\}$, dvs. om och endast om $x_0 \in \operatorname{ext} X$.

4.5 T. ex. mängden i övning 4.1 c).

4.6 a) Följer direkt av sats 4.1.3.

b) Den slutna konvexa mängden $\left\{x \in \mathbf{R}^2 \mid x_2 \leq \sqrt{1-x_1^2},\ |x_1| \leq 1\right\}$ har två icke-utsatta extremalpunkter, $(-1,0)$ och $(1,0)$.

4.7 b) En nolldimensionell generell fasad är en extremalpunkt, och en nolldimensionell utsatt fasad är en utsatt punkt. Övning 4.6 b) ger därför ett exempel på en generell fasad som inte är en utsatt fasad.

c) Antag att $a, b \in X$ och att den öppna sträckan $]a,b[$ skär F'. Eftersom $F' \subseteq F$, skär samma sträcka också F, så det följer att $a, b \in F$. Men eftersom F' är en generell fasad till F följer härav att $a, b \in F'$, vilket visar att F' är en generell fasad till X.
I mängden X i övning 4.6 b) är $F = \{1\}\times]-\infty, 0]$ en utsatt fasad, och $F' = \{(1,0)\}$ är en utsatt fasad till F men inte till X.

d) Fixera en punkt $x_0 \in F \cap \operatorname{rint} C$. Till varje $x \in C$ kan vi då hitta en punkt $y \in C$ sådan att x_0 ligger på den öppna sträckan $]x,y[$, och det följer nu av definitionen av generell fasad att $x \in F$.

e) Använd resultatet i d) på mängden $C = X \cap \operatorname{cl} F$. Eftersom $\operatorname{rint} C$ innehåller $\operatorname{rint} F$, är säkert $F \cap \operatorname{rint} C \neq \emptyset$, så det följer att $C \subseteq F$. Den omvända inklusionen är förstås trivial.

f) Använd resultatet i d) med $F = F_1$ och $C = F_2$, vilket ger inklusionen $F_2 \subseteq F_1$, och den omvända inklusionen fås analogt.

g) Om F är en generell fasad och $F \cap \operatorname{rint} X \neq \emptyset$, så följer av d) att $X \subseteq F$. För fasader $F \neq X$ är därför $F \cap \operatorname{rint} X = \emptyset$, vilket innebär att $F \subseteq \operatorname{rbdry} X$.

Kapitel 5

5.1 a) $(-\frac{2}{3}, \frac{4}{3})$ och $(4, -1)$ b) $(-\frac{2}{3}, \frac{4}{3})$, $(4, -1)$ och $(-3, -1)$
c) $(0,0,0)$, $(2,0,0)$, $(0,2,0)$, $(0,0,4)$ och $(\frac{4}{3}, \frac{4}{3}, 0)$
d) $(0,4,0,0)$, $(0, \frac{5}{2}, 0, 0)$, $(\frac{3}{2}, \frac{5}{2}, 0, 0)$, $(0,1,1,0)$ och $(0, \frac{5}{2}, 0, \frac{3}{2})$

5.2 Extremalstrålarna genereras av vektorerna $(-2,4,3)$, $(1,1,0)$, $(4,-1,1)$ och $(1,0,0)$.

5.3 $C = \begin{bmatrix} 1 & -2 & 1 \\ -1 & 2 & 3 \\ -3 & 2 & 5 \end{bmatrix}$

5.4 a) $A = \{(1,0), (0,1)\}$, $B = \{(-\frac{2}{3}, \frac{4}{3}), (4,-1)\}$
b) $A = \emptyset$, $B = \{(-\frac{2}{3}, \frac{4}{3}), (4,-1), (-3,-1)\}$
c) $A = \{(1,1,-3), (-1,-1,3), (4,-7,-1), (-7,4,-1)\}$,
$B = \{(0,0,0), (2,0,0), (0,2,0), (0,0,4), (\frac{4}{3}, \frac{4}{3}, 0)\}$
d) $A = \emptyset$,
$B = \{(0,4,0,0), (0, \frac{5}{2}, 0, 0), (\frac{3}{2}, \frac{5}{2}, 0, 0), (0,1,1,0), (0, \frac{5}{2}, 0, \frac{3}{2})\}$.

5.5 Inklusionen $X = \operatorname{cvx} A + \operatorname{con} B \subseteq \operatorname{con} A + \operatorname{con} B = \operatorname{con}(A \cup B)$ medför att $\operatorname{con} X \subseteq \operatorname{con}(A \cup B)$. Uppenbarligen är $A \subseteq \operatorname{cvx} A \subseteq X$. Eftersom $\operatorname{cvx} A$ är kompakt, är $\operatorname{recc} X = \operatorname{con} B$, så antagandet $0 \in X$ medför att $B \subseteq \operatorname{con} B \subseteq X$. Så $A \cup B \subseteq X$ med $\operatorname{con}(A \cup B) \subseteq \operatorname{con} X$ som slutsats.

Kapitel 6

6.1 T. ex. $f_1(x) = x - |x|$ och $f_2(x) = -x - |x|$.

6.3 $a \geq 5$ resp. $a > 5$.

6.4 Använd resultatet i övning 2.1.

6.5 Följer av att $f(x) = \max(x_{i_1} + x_{i_2} + \cdots + x_{i_k})$, där maximum tas över alla delmängder $\{i_1, i_2, \ldots, i_k\}$ av $\{1, 2, \ldots, n\}$ med k stycken element.

6.6 För $x_1 + x_2 + \cdots + x_n = 0$ är olikheten trivial, och för $x_1 + \cdots + x_n > 0$ fås den genom addition av de n olikheterna

$$f(x_i) \leq \frac{x_i}{x_1 + \cdots + x_n} f(x_1 + \cdots + x_n) + \big(1 - \frac{x_i}{x_1 + \cdots + x_n}\big) f(0).$$

6.7 Välj

$$c = \frac{f(x_2) - f(x_1)}{\|x_2 - x_1\|^2} (x_1 - x_2);$$

då är $f(x_1) + \langle c, x_1 \rangle = f(x_2) + \langle c, x_2 \rangle$, och kvasikonvexitetsantagandet medför att $f(\lambda x_1 + (1-\lambda)x_2) + \langle c, \lambda x_1 + (1-\lambda)x_2 \rangle \leq f(x_1) + \langle c, x_1 \rangle$, vilket efter förenkling leder till olikheten

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

6.8 Definiera $f\colon \mathbf{R}^n \times \mathbf{R} \to \overline{\mathbf{R}}$ genom att sätta

$$f(x,t) = \begin{cases} t & \text{om } (x,t) \in C, \\ +\infty & \text{om } (x,t) \notin C. \end{cases}$$

Då är $\inf\{t \in \mathbf{R} \mid (x,t) \in C\} = \inf\{f(x,t) \mid t \in \mathbf{R}\}$, och sats 6.2.6 följer nu ur korollarium 6.2.7.

6.9 Utnyttja att om $x, y \in X$ så finns det följder $(x_k)_1^\infty$, $(y_k)_1^\infty$ av punkter $x_k, y_k \in \operatorname{int} X$ så att $x_k \to x$ och $y_k \to y$ då $k \to \infty$. Punkterna $\lambda x_k + (1-\lambda)y_k$ tillhör också $\operatorname{int} X$ och gränsövergång i olikheten

$$f(\lambda x_k + (1-\lambda)y_k) \le \lambda f(x_k) + (1-\lambda)f(y_k)$$

visar att f är konvex på X.

6.10 Sätt $m = \inf\{f(x) \mid x \in \operatorname{rint}(\operatorname{dom} f)\}$ och fixera en relativt inre punkt x_0 i $\operatorname{dom} f$. Om x är en godtycklig punkt i $\operatorname{dom} f$ och $0 < \lambda < 1$ så är $\lambda x + (1-\lambda)x_0$ en relativt inre punkt i $\operatorname{dom} f$, varför det följer att

$$m \le f(\lambda x + (1-\lambda)x_0) \le \lambda f(x) + (1-\lambda)f(x_0).$$

Genom att låta $\lambda \to 1$ erhåller man olikheten $f(x) \ge m$.

6.11 Minimum 8 fås för $x = (\frac{1}{8}, 2)$.

6.12 a) $\|x\|_p$ b) $\max(x_1, 0)$.

Kapitel 7

7.2 Ja.

7.5 Låt J vara ett delintervall av I. Om $f'_+(x) \ge 0$ för alla $x \in J$, så är $f(y) - f(x) \ge f'_+(x)(y-x) \ge 0$ för alla $y > x$ i intervallet J, dvs. funktionen f är växande på J. Om istället $f'_+(x) \le 0$ för alla $x \in J$, så är $f(y) - f(x) \ge f'_+(x)(y-x) \ge 0$ för alla $y < x$, dvs. funktionen är avtagande på J.
Eftersom högerderivatan $f'_+(x)$ är växande på I, är antingen $f'_+(x) \ge 0$ för alla $x \in I$, och då är f växande på I, eller också är $f'_+(x) \le 0$ för alla $x \in I$, och då är f avtagande på I, eller också finns det en punkt $c \in I$ så att $f'_+(x) \le 0$ till vänster om c och $f'_+(x) > 0$ till höger om c, och då är f avtagande till vänster om c och växande till höger om c.

7.6 a) Att gränsvärdena existerar följer av resultatet i föregående övning.
b) Betrakta den utvidgade funktionens epigraf.

7.7 Följer direkt av övning 7.6 b).

7.8 Antag att $f \in \mathcal{F}$. Låt $x_0 \in \mathbf{R}^n$ vara en godtycklig punkt, och betrakta funktionen $g(x) = f(x) - \langle f'(x_0), x - x_0\rangle$. Funktionen g ligger i $\mathcal{F}$

och $g'(x_0) = 0$. Det följer att $g(x) \geq g(x_0)$ för alla x, vilket betyder att $f(x) \geq f(x_0) + \langle f'(x_0), x - x_0\rangle$ för alla x. Funktionen f är därför konvex på grund av sats 7.2.1.

7.9 Enligt sats 6.7.1 är $\phi(t) = f(x + tv) = f(x) + t\langle f'(x), v\rangle$ för $v \in V_f$. Genom att derivera två gånger får vi $D^2 f(x)[v, v] = \phi''(0) = 0$, varav följer att $f''(x)v = 0$.

7.13 Sats 7.3.1 (i) med x bytt mot $\hat{x}$ och $v = x - \hat{x}$ ger i kombination med Cauchy-Schwarz olikhet: $\mu\|x - \hat{x}\|^2 \leq \langle f'(x), x - \hat{x}\rangle \leq \|f'(x)\|\|x - \hat{x}\|$.

Kapitel 8

8.1 Antag att f är μ-starkt konvex, där $\mu > 0$, och låt c vara en subgradient i 0 till den konvexa funktionen $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$. Då är $f(x) \geq f(0) + \langle c, x\rangle + \frac{1}{2}\mu\|x\|^2$ för alla x, och högerledet går mot ∞ då $\|x\| \to \infty$. Påståendet följer också enkelt av sats 8.1.6.

8.2 Sträckan $[-\mathbf{e}_1, \mathbf{e}_2]$, där $\mathbf{e}_1 = (1, 0)$ och $\mathbf{e}_2 = (0, 1)$.

8.3 a) $\overline{B}_2(0; 1) = \{x \mid \|x\|_2 \leq 1\}$ b) $\overline{B}_1(0; 1) = \{x \mid \|x\|_1 \leq 1\}$
c) $\overline{B}_\infty(0; 1) = \{x \mid \|x\|_\infty \leq 1\}$.

8.4 a) $\operatorname{dom} f^* = \{a\}$, $f^*(a) = b$
b) $\operatorname{dom} f^* = \{x \mid x < 0\}$, $f^*(x) = -1 - \ln(-x)$
c) $\operatorname{dom} f^* = \mathbf{R}_+$, $f^*(x) = x \ln x - x$, $f^*(0) = 0$
d) $\operatorname{dom} f^* = \mathbf{R}$, $f^*(x) = \mathrm{e}^{x-1}$
e) $\operatorname{dom} f^* = \mathbf{R}_-$, $f^*(x) = -2\sqrt{-x}$.

Kapitel 9

9.1
$$\min\ 5000x_1 + 4000x_2 + 3000x_3 + 4000x_4$$
$$\text{då}\ \begin{cases} -x_1 + 2x_2 + 2x_3 + x_4 \geq 16 \\ 4x_1 + x_2 + 2x_4 \geq 40 \\ 3x_1 + x_2 + 2x_3 + x_4 \geq 24,\ x \geq 0 \end{cases}$$

9.2
$$\max\ v$$
$$\text{då}\ \begin{cases} 2x_1 + x_2 - 4x_3 \geq v \\ x_1 + 2x_2 - 2x_3 \geq v \\ -2x_1 - x_2 + 2x_3 \geq v \\ x_1 + x_2 + x_3 = 1,\ x \geq 0 \end{cases}$$

9.3 Radspelaren skall välja rad 2 och kolonnspelaren kolonn 1.

9.4 Utbetalningsmatrisen är

	Sp E	Ru E	Ru 2
Sp E	-1	1	-1
Ru E	1	-1	-2
Sp 2	-1	2	2

och kolonnspelarens problem kan skrivas

$$\begin{aligned} &\min\ u \\ &\text{då}\ \begin{cases} -y_1 + y_2 + y_3 \le u \\ y_1 - y_2 - 2y_3 \le u \\ -y_1 + 2y_2 + 2y_3 \le u \\ y_1 + y_2 + y_3 = 1,\ y \ge 0 \end{cases} \end{aligned}$$

9.5 a) $(\frac{4}{5}, \frac{13}{15})$

9.6 a) $\max r$ då $\begin{cases} -x_1 + x_2 + r\sqrt{2} \le 0 \\ x_1 - 2x_2 + r\sqrt{5} \le 0 \\ x_1 + x_2 + r\sqrt{2} \le 1 \end{cases}$ b) $\max r$ då $\begin{cases} -x_1 + x_2 + 2r \le 0 \\ x_1 - 2x_2 + 3r \le 0 \\ x_1 + x_2 + 2r \le 1 \end{cases}$

Kapitel 10

10.1 $\phi(\lambda) = 2\lambda - \frac{1}{2}\lambda^2$

10.2 För de båda problemens duala funktioner ϕ_a resp. ϕ_b gäller:

$$\phi_a(\lambda) = 0 \text{ för alla } \lambda \ge 0 \text{ och } \phi_b(\lambda) = \begin{cases} 0 & \text{om } \lambda = 0, \\ \lambda - \lambda \ln \lambda & \text{om } 0 < \lambda < 1, \\ 1 & \text{om } \lambda \ge 1. \end{cases}$$

10.5 För alla $i \in I(\hat{x})$ är $g_i(x_0) \ge g_i(\hat{x}) + \langle g_i'(\hat{x}), x_0 - \hat{x}\rangle = \langle g_i'(\hat{x}), x_0 - \hat{x}\rangle$. För $i \in I_{\text{övr}}(\hat{x})$ är därför $\langle g_i'(\hat{x}), \hat{x} - x_0\rangle \ge -g_i(x_0) > 0$, och för $i \in I_{\text{aff}}(\hat{x})$ är $\langle g_i'(\hat{x}), \hat{x} - x_0\rangle \ge -g_i(x_0) \ge 0$.

10.6 a) $v_{\min} = -1$ för $x = (-1, 0)$ b) $v_{\max} = 2 + \frac{\pi}{4}$ för $x = (1, 1)$
c) $v_{\min} = -\frac{1}{3}$ för $x = \pm(\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}})$ d) $v_{\max} = \frac{1}{54}$ för $x = (\frac{1}{6},\ 2,\ \frac{1}{3})$

Kapitel 11

11.1 $\hat{\lambda} = 2b$

11.3 b) Låt $L\colon \Omega \times \Lambda \to \mathbf{R}$ och $L_1\colon (\mathbf{R} \times \Omega) \times (\mathbf{R}_+ \times \Lambda) \to \mathbf{R}$ vara Lagrangefunktionerna till problemen (P) resp. (P′), och låt ϕ och ϕ_1 vara

problemens duala funktioner. Följande samband gäller för Lagrangefunktionerna:

$$L_1(t, x, \lambda_0, \lambda) = (1 - \lambda_0)(t - f(x)) + L(x, \lambda).$$

För ett givet $\lambda \in \Lambda$ är Lagrangefunktionen L_1 nedåt begränsad om och endast om $\lambda_0 = 1$ och $\lambda \in \operatorname{dom} \phi$. Följaktligen är $\operatorname{dom} \phi_1 = \{1\} \times \operatorname{dom} \phi$. Vidare är $\phi_1(1, \lambda) = \phi(\lambda)$ för alla $\lambda \in \operatorname{dom} \phi$.

11.4 Låt I vara indexmängden för de icke-affina olikhetsbivillkoren, låt k vara antalet element i I och sätt $\overline{x} = k^{-1} \sum_{i \in I} \overline{x}_i$. Punkten $\overline{x}$ uppfyller Slaters villkor.

11.5 Låt $b^{(1)}$ och $b^{(2)}$ vara två punkter i U och välj givet $\epsilon > 0$ tillåtna punkter $x^{(i)}$ i respektive problem $(\mathrm{P}_{b^{(i)}})$ som uppfyller $f(x^{(i)}) < v_{\min}(b^{(i)}) + \epsilon$. För $0 < \lambda < 1$ och $b = \lambda b^{(1)} + (1 - \lambda) b^{(2)}$ är $x = \lambda x^{(1)} + (1 - \lambda) x^{(2)}$ en tillåten punkt i problemet (P_b). Det följer att

$$\begin{aligned} v_{\min}(b) &\le f(x) \le \lambda f(x^{(1)}) + (1 - \lambda) f(x^{(2)}) \\ &< \lambda v_{\min}(b^{(1)}) + (1 - \lambda) v_{\min}(b^{(2)}) + \epsilon, \end{aligned}$$

och eftersom $\epsilon > 0$ är godtyckligt, visar detta att funktionen $v_{\min}$ är konvex på U.

11.6 a) $v_{\min} = 2$ för $x = (0, 0)$ b) $v_{\min} = 2$ för $x = (0, 0)$
c) $v_{\min} = \ln 2 - 1$ för $x = (-\ln 2, \frac{1}{2})$ d) $v_{\min} = -5$ för $x = (-1, -2)$
e) $v_{\min} = 1$ för $x = (1, 0)$ f) $v_{\min} = 2\,\mathrm{e}^{1/2} + \frac{1}{4}$ för $x = (\frac{1}{2}, \frac{1}{2})$

11.7 $v_{\min} = 2 - \ln 2$ för $x = (1, 1)$

11.9 $\min \; 50x_1^2 + 80x_1x_2 + 40x_2^2 + 10x_3^2$
då $\begin{cases} 0.2x_1 + 0.12x_2 + 0.04x_3 \ge 0.12 \\ x_1 + \quad x_2 + \quad x_3 = 1, \quad x \ge 0 \end{cases}$
Optimum fås för $x_1 = x_3 = 0.5$ milj kr.

Kapitel 12

12.1 Alla icke-tomma mängder $X(b) = \{x \mid Ax \ge b\}$ av tillåtna punkter har samma recessionskon eftersom $\operatorname{recc} X(b) = \{x \mid Ax \ge 0\}$ om $X(b) \ne \emptyset$. Det följer därför av sats 12.1.1 att det optimala värdet $v(b)$ är ändligt för alla högerled b för vilka $X(b) \ne \emptyset$. På grund av dualitetssatsen är vidare

$$v(b) = \min\{\langle -b, y \rangle \mid A^T y \le c,\, y \ge 0\},$$

så det följer också av samma sats att värdefunktionen v är konvex.

12.2 T. ex. $\min\ x_1 - x_2$ då $\begin{cases} -x_1 \geq 1 \\ x_2 \geq 1,\ x \geq 0 \end{cases}$ och $\max\ y_1 + y_2$ då $\begin{cases} -y_1 \leq 1 \\ y_2 \leq -1,\ y \geq 0 \end{cases}$

12.5 $v_{\max} = \begin{cases} \dfrac{t-3}{t+1} & \text{för } x = \left(-\dfrac{2}{t+1}, \dfrac{t-1}{t+1}\right) \text{ om } t < -2, \\ 5 & \text{för } x = (2,3) \text{ om } t \geq -2. \end{cases}$

Kapitel 13

13.1 a) $\min\ 2x_1 - 2x_2 + x_3$ då $\begin{cases} x_1 + x_2 - x_3 - s_1 = 3 \\ x_1 + x_2 - x_3 + s_2 = 2 \\ x_1, x_2, x_3, s_1, s_2 \geq 0 \end{cases}$

b) $\min\ x_1 + 2x_2' - 2x_2''$ då $\begin{cases} x_1 + x_2' - x_2'' - s_1 = 1 \\ x_2' - x_2'' - s_2 = -2 \\ x_1, x_2', x_2'', s_1, s_2 \geq 0 \end{cases}$

13.2 a) $(5,5,0)$ och $(7\frac{1}{2}, 0, 2\frac{1}{2})$ b) $(3,0,0,0)$ och $(0,0,0,3)$

13.3 $\max\ y_1 + 7y_2$ då $\begin{cases} y_1 + y_2 \leq 1 \\ 2y_2 \leq 1 \\ -y_1 + 7y_2 \leq 4 \end{cases}$

13.4 a) $v_{\min} = -1$ för $x = (0,0,4,1)$ b) $v_{\max} = 56$ för $x = (24,0,0,1,11)$
c) $v_{\max} = 30\frac{6}{7}$ för $x = (1\frac{5}{7}, \frac{3}{7}, 0)$ d) $v_{\max} = 23$ för $x = (2,0,3,0,5)$
e) $v_{\min} = -\infty$ f) $v_{\min} = -1\frac{13}{15}$ för $x = (0, \frac{2}{3}, 0, \frac{2}{5})$

13.5 $v_{\min} = -2$ antas i punkterna på sträckan mellan $(0,3,1,1,0)$ och $(0,2,2,0,1)$.

13.6 $v_{\max} = 15$ för $x = (2\frac{1}{2}, 2\frac{1}{2}, 2\frac{1}{2}, 0)$

13.8 $v_{\min} = 9$ för $x = (\frac{2}{3}, 1\frac{2}{3}, 1\frac{2}{3})$

13.9 $v_{\min} = -40\frac{3}{5}$ för $x = (-3\frac{3}{5}, 11\frac{4}{5})$

13.10 a) $v_{\min} = 4\frac{1}{4}$ för $x = (\frac{3}{4}, \frac{1}{2}, \frac{3}{4})$ b) $v_{\min} = \frac{4}{5}$ för $x = (0, \frac{2}{5}, 0)$
c) $v_{\min} = 5\frac{7}{12}$ för $x = (1\frac{1}{4}, \frac{11}{12}, 0)$

13.12 $v_{\max} = \begin{cases} 7 & \text{för } x = (3\frac{1}{2}, 0) \text{ om } t \leq 1, \\ 4 + 3t & \text{för } x = (2,3) \text{ om } 1 < t < 2, \\ 5t & \text{för } x = (0,5) \text{ om } t \geq 2. \end{cases}$

13.13 500 par av modell A och 700 par av modell B.

13.14 4 liter mjölk och 1 limpa. Mjölkpriset kan stiga till 10 kr/l.

13.17 Använd först algoritmen $\mathcal{A}$ på det system som består av de linjära olikheterna $Ax \geq b$, $x \geq 0$, $A^T y \leq c$, $y \geq 0$, $\langle c, x\rangle \leq \langle b, y\rangle$. Om algoritmen resulterar i lösningen $(\overline{x}, \overline{y})$, så är på grund av komplementaritetssatsen $\overline{x}$ den optimala lösningen till minimeringsproblemet. Om algoritmen istället visar att systemet saknar lösning, så använder vi algoritmen på systemet $Ax \geq b$, $x \geq 0$ för att avgöra huruvida minimeringsproblemet har tillåtna punkter eller ej. I det förstnämnda fallet är målfunktionen nedåt obegränsad på grund av dualitetssatsen eftersom det duala problemet måste sakna tillåtna punkter.

Kapitel 14

14.1 $x_1 = (\frac{4}{9}, -\frac{1}{9})$, $x_2 = (\frac{2}{27}, \frac{2}{27})$, $x_3 = (\frac{8}{243}, -\frac{2}{243})$.

14.3 Följer av att $hf'(x_k) = f(x_k) - f(x_{k+1}) \to f(\hat{x}) - f(\hat{x}) = 0$ och $hf'(x_k) \to hf'(\hat{x})$.

Kapitel 15

15.1 $\Delta x_{\text{nt}} = -x \ln x$, $\lambda(f, x) = \sqrt{x} \ln x$, $\|v\|_x = |v|/\sqrt{x}$.

15.2 a) $\Delta x_{\text{nt}} = (\frac{1}{3}, \frac{1}{3})$, $\lambda(f, x) = \sqrt{\frac{1}{3}}$, $\|v\|_x = \frac{1}{2}\sqrt{5v_1^2 + 2v_1v_2 + 5v_2^2}$

b) $\Delta x_{\text{nt}} = (\frac{1}{3}, -\frac{2}{3})$, $\lambda(f, x) = \sqrt{\frac{1}{3}}$, $\|v\|_x = \frac{1}{2}\sqrt{8v_1^2 + 8v_1v_2 + 5v_2^2}$.

15.3 $\Delta x_{\text{nt}} = (v_1, v_2)$, där $v_1 + v_2 = -1 - \mathrm{e}^{-(x_1+x_2)}$,
$\lambda(f, x) = \mathrm{e}^{(x_1+x_2)/2} + \mathrm{e}^{-(x_1+x_2)/2}$, $\|v\|_x = \mathrm{e}^{(x_1+x_2)/2}|v_1 + v_2|$.

15.4 Om rang $A < m$, så är rang $M < m+n$, och om $\mathcal{N}(A) \cap \mathcal{N}(P)$ innehåller en nollskild vektor x, så är $M \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Matrisen M saknar således invers i dessa fall.

Antag omvänt att rang $A = m$, dvs. att $\mathcal{N}(A^T) = \{0\}$, och att snittet $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$. Vi visar att koefficientmatrisen M är inverterbar genom att visa att det homogena systemet

$$\begin{cases} Px + A^T y = 0 \\ Ax \qquad\quad = 0 \end{cases}$$

bara har den triviala lösningen $x = 0$, $y = 0$.

Genom att multiplicera den första ekvationen från vänster med x^T fås

$$0 = x^T Px + x^T A^T y = x^T Px + (Ax)^T y = x^T Px,$$

och eftersom P är positivt semidefinit, följer det att $Px = 0$. Den första ekvationen ger nu $A^T y = 0$. Så $x \in \mathcal{N}(A) \cap \mathcal{N}(P)$ och $y \in \mathcal{N}(A^T)$, vilket enligt förutsättningarna betyder att $x = 0$ och $y = 0$.

15.5 a) Förutsättningarna innebär att $\langle v, f''(x)v\rangle \geq \mu\|v\|^2$ om $Av = 0$. För funktionen $\tilde{f}$ innebär detta, beroende på att $AC = 0$, att

$$\begin{aligned}\langle w, \tilde{f}''(z)w\rangle &= \langle w, C^T f''(x)Cw\rangle = \langle Cw, f''(x)Cw\rangle \geq \mu\|Cw\|^2 \\ &= \mu\langle w, C^T Cw\rangle \geq \mu\sigma\|w\|^2\end{aligned}$$

för alla $w \in \mathbf{R}^p$, vilket visar att funktionen $\tilde{f}$ är $\mu\sigma$-starkt konvex.

b) Påståendet följer av a) om vi visar att restriktionen av f till X är $K^{-2}M^{-1}$-starkt konvex. Så antag att $x \in X$ och att $Av = 0$. Då är

$$\begin{bmatrix} f''(x) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} v \\ 0 \end{bmatrix} = \begin{bmatrix} f''(x)v \\ 0 \end{bmatrix}$$

så det följer av begränsningen på den inversa matrisens norm att

$$\|v\| \leq K\|f''(x)v\|.$$

Den positivt semidefinita andraderivatan $f''(x)$ har en positivt semidefinit kvadratrot $f''(x)^{1/2}$ och $\|f''(x)^{1/2}\| = \|f''(x)\|^{1/2} \leq M^{1/2}$. Det följer att

$$\begin{aligned}\|f''(x)v\|^2 &= \|f''(x)^{1/2}f''(x)^{1/2}v\|^2 \leq \|f''(x)^{1/2}\|^2\|f''(x)^{1/2}v\|^2 \\ &\leq M\|f''(x)^{1/2}v\|^2 = M\langle v, f''(x)v\rangle,\end{aligned}$$

vilket insatt i olikheten ovan ger att

$$\langle v, f''(x)v\rangle \geq K^{-2}M^{-1}\|v\|^2.$$

Kapitel 16

16.2 Om P_i är projektionen av $\mathbf{R}^{n_1} \times \cdots \times \mathbf{R}^{n_m}$ på den i:te faktorn $\mathbf{R}^{n_i}$, så är $f(x) = \sum_{i=1}^m f_i(P_i x)$. Det följer därför av satserna 16.1.5 och 16.1.6 att f är självkonkordant.

16.3 a) $g''(x) = \dfrac{f'(x)^2}{f(x)^2} - \dfrac{f''(x)}{f(x)} + \dfrac{1}{x^2} \geq 0$, så funktionen g är konvex.

$g'''(x) = -\dfrac{f'''(x)}{f(x)} + 3\dfrac{f'(x)f''(x)}{f(x)^2} - 2\dfrac{f'(x)^3}{f(x)^3} - \dfrac{2}{x^3}$, så

$|g'''(x)| \leq 3\dfrac{f''(x)}{x|f(x)|} + 3\dfrac{|f'(x)|f''(x)}{f(x)^2} + 2\dfrac{|f'(x)|^3}{|f(x)|^3} + 2\dfrac{1}{x^3}$.

Genom att välja $a = \sqrt{f''(x)/|f(x)|}$, $b = |f'(x)|/|f(x)|$ och $c = 1/x$ i olikheten

$$(*) \qquad 3a^2b + 3a^2c + 2b^3 + 2c^3 \leq 2(a^2 + b^2 + c^2)^{3/2}$$

erhålls olikheten $|g'''(x)| \le 2g''(x)^{3/2}$, som visar att funktionen g är självkonkordant.
För att visa olikheten $(*)$ kan man på grund av homogeniteten anta att $a^2+b^2+c^2=1$. Då är $a^2 = 1-b^2-c^2$, och olikheten reduceras efter lite räknande till olikheten $(b+c)(3-(b+c)^2) \le 2$, som gäller eftersom $x(3-x^2) \le 2$ för $x \ge 0$.

16.3 b) Låt $\phi(t) = F(x_0+\alpha t, y_0+\beta t)$ vara restriktionen av F till en godtycklig linje genom punkten (x_0, y_0) i $\operatorname{dom} F$. För varje val av konstanterna A och B uppfyller funktionen $f(x) - Ax - B$ förutsättningarna i övningen, så funktionen $h(x) = -\ln(Ax+B-f(x)) - \ln x$ är på grund av resultatet i a) också självkonkordant. Om $\alpha \ne 0$, så är $\phi(t) = h(\alpha t + x_0)$ för $A = \beta/\alpha$ och $B = y_0 - \beta x_0/\alpha$, och om $\alpha = 0$ är istället $\phi(t) = -\ln(\beta t + a) + b$ för $a = y_0 - f(x_0)$ och $b = -\ln x_0$. I båda fallen är ϕ självkonkordant. Funktionen F är följaktligen självkonkordant.

16.6 a) Sätt $\lambda = \lambda(f,x)$ och använd olikheterna (16.7) och (16.6) i sats 16.3.2 med $y = x^+$ och $v = x^+ - x = (1+\lambda)^{-1}\Delta x_{\mathrm{nt}}$. Som resultat erhålls olikheten

$$\begin{aligned}
\langle f'(x^+), w\rangle &\le \langle f'(x), w\rangle + \frac{1}{1+\lambda}\langle f''(x)\Delta x_{\mathrm{nt}}, w\rangle + \frac{\lambda^2\|w\|_x}{(1+\lambda)^2(1-\lambda/(1+\lambda))} \\
&= \langle f'(x), w\rangle - \frac{1}{1+\lambda}\langle f'(x), w\rangle + \frac{\lambda^2}{1+\lambda}\|w\|_x \\
&= \frac{\lambda}{1+\lambda}\langle f'(x), w\rangle + \frac{\lambda^2}{1+\lambda}\|w\|_x \\
&\le \frac{\lambda}{1+\lambda}\lambda\|w\|_x + \frac{\lambda^2}{1+\lambda}\|w\|_x = \frac{2\lambda^2}{1+\lambda}\|w\|_x \\
&\le \frac{2\lambda^2\|w\|_{x^+}}{(1+\lambda)(1-\lambda/(1+\lambda))} = 2\lambda^2\|w\|_{x^+}
\end{aligned}$$

med $\lambda(f, x^+) \le 2\lambda^2$ som slutsats.

Kapitel 18

18.1 Följer av satserna 18.1.3 och 18.1.2.

18.2 För att visa implikationen $\|v\|_x^* < \infty \Rightarrow v \in \mathcal{N}(f''(x))^\perp$ skriver vi v på formen $v = v_1 + v_2$ med $v_1 \in \mathcal{N}(f''(x))$ och $v_2 \in \mathcal{N}(f''(c))^\perp$, och noterar att $\|v_1\|_x = 0$. Det följer att $\|v\|_1^2 = \langle v_1, v_1\rangle = \langle v, v_1\rangle \le \|v\|_x^*\|v_1\|_x = 0$, vilket innebär att $v_1 = 0$ och att följaktligen v tillhör $\mathcal{N}(f''(x))^\perp$.
För varje $v \in \mathcal{N}(f''(x))^\perp$ finns det vektor u sådan att $v = f''(x)u$, och

vi skall visa att $\|v\|_x^* = \|u\|_x$. Detta medför att $\|v\|_x^* < \infty$ och att $\|\cdot\|_x^*$ är en norm på delrummet $\mathcal{N}(f''(x))^\perp$ av $\mathbf{R}^n$.
För godtycklig vektor $w \in \mathbf{R}^n$ är på grund av Cauchy–Schwarz olikhet

$$\begin{aligned}\langle v, w\rangle &= \langle f''(x)u, w\rangle = \langle f''(x)^{1/2}u, f''(x)^{1/2}w\rangle \\ &\le \|f''(x)^{1/2}u\|\|f''(x)^{1/2}w\| = \|u\|_x\|v\|_x,\end{aligned}$$

och detta medför att $\|v\|_x^* \le \|u\|_x$. Om $v \neq 0$ så ligger vektorn u inte i $\mathcal{N}(f''(x))$, vilket betyder att $\|u\|_x \neq 0$. För $w = u/\|u\|_x$ får vi nu likheten

$$\langle v, w\rangle = \|u\|_x^{-1}\langle f''(x)^{1/2}u, f''(x)^{1/2}u\rangle = \|u\|_x^{-1}\|f''(x)^{1/2}u\|^2 = \|u\|_x,$$

som visar att är $\|v\|_x^* = \|u\|_x$. Och om $v = 0$, så är u en vektor i $\mathcal{N}(f''(x))$ och vi har $\|v\|_x^* = \|u\|_x$ även i detta fall.

18.3 a) Derivera likheten $f(tx) = f(x) - \nu \ln t$ med avseende på x.

b) Derivera den resulterande likheten i a) med avseende på t och sätt $t = 1$.

c) Eftersom X inte innehåller någon linje, är den självkonkordanta funktionen f icke-degenererad, och det följer av resultatet i b) att funktionens unika Newtonriktning i punkten x är lika med x. Genom att derivera likheten $f(tx) = f(x) - \nu \ln t$ med avseende på t och sedan sätta $t = 1$ erhålles $\langle f'(x), x\rangle = -\nu$. Följaktligen är

$$\nu = -\langle f'(x), x\rangle = -\langle f'(x), \Delta x_{\mathrm{nt}}\rangle = \lambda(f, x)^2.$$

18.5 Sätt $g(x, x_{n+1}) = (x_1^2 + \cdots + x_n^2) - x_{n+1}^2 = \|x\|^2 - x_{n+1}^2$, så att

$$f(x) = -\ln(-g(x, x_{n+1})),$$

och sätt $w = (v, v_{n+1})$. Då är

$$\begin{aligned}Dg &= Dg(x, x_{n+1})[w] = 2(\langle x, v\rangle - x_{n+1}v_{n+1}),\\ D^2g &= D^2g(x, x_{n+1})[w, w] = 2(\|v\|^2 - v_{n+1}^2),\\ D^3g &= D^3g(x, x_{n+1})[w, w, w] = 0,\\ Df &= Df(x, x_{n+1})[w] = -\frac{1}{g}Dg\\ D^2f &= D^2f(x, x_{n+1})[w, w] = \frac{1}{g^2}\big((Dg)^2 - gD^2g\big),\\ D^3f &= D^3f(x, x_{n+1})[w, w, w] = \frac{1}{g^3}\big(-2(Dg)^3 + 3gDgD^2g\big).\end{aligned}$$

Betrakta differensen

$$\Delta = (Dg)^2 - gD^2g = 4(\langle x, v\rangle - x_{n+1}v_{n+1})^2 + 2(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2).$$

Eftersom $x_{n+1} > \|x\|$, är $\Delta \geq 0$ om $|v_{n+1}| \leq \|v\|$. Så antag att $|v_{n+1}| > \|v\|$. Då är

$$\begin{aligned} |x_{n+1}v_{n+1} - \langle x, v\rangle| &\geq x_{n+1}|v_{n+1}| - |\langle x, v\rangle| \\ &\geq x_{n+1}|v_{n+1}| - \|x\|\|v\| \geq 0 \end{aligned}$$

och följaktligen

$$\begin{aligned} \Delta &\geq 4(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 2(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \\ &= 2(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 2(x_{n+1}\|v\| - \|x\||v_{n+1}|)^2 \geq 0. \end{aligned}$$

Detta visar att $D^2f = \Delta/g^2 \geq 0$, så funktionen f är konvex.
För att visa att funktionen är självkonkordant, skall vi visa att

$$4(D^2f)^3 - (D^3f)^2 \geq 0.$$

Efter förenkling fås

$$4(D^2f)^3 - (D^3f)^2 = g^{-4}(D^2g)^2(3(Dg)^2 - 4gD^2g),$$

och problemet har därmed reducerats till att visa att differensen

$$\begin{aligned} \Delta' &= 3(Dg)^2 - 4gD^2g \\ &= 12(\langle x, v\rangle - x_{n+1}v_{n+1})^2 + 8(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \end{aligned}$$

är icke-negativ. För $|v_{n+1}| \leq \|v\|$ är detta uppenbart, och för $|v_{n+1}| > \|v\|$ fås på motsvarande sätt som ovan

$$\begin{aligned} \Delta' &\geq 12(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 8(x_{n+1}^2 - \|x\|^2)(\|v\|^2 - v_{n+1}^2) \\ &= 4(x_{n+1}|v_{n+1}| - \|x\|\|v\|)^2 + 8(x_{n+1}\|v\| - \|x\||v_{n+1}|)^2 \geq 0. \end{aligned}$$

18.6 Låt $w = (u, v)$ vara en godtycklig vektor i $\mathbf{R}^2$. Med beteckningarna $a = 1/(y - f(x))$, $b = -1/x$, $A = f'(x)$ och $B = f''(x)$, där $a > 0$ och $B \geq 0$, blir

$$\begin{aligned} DF(x, y)[w] &= (aA + b)u - av \\ D^2F(x, y)[w, w] &= (aB + a^2A^2 + b^2)u^2 - 2a^2Auv + a^2v^2. \end{aligned}$$

Efter förenkling fås olikheten

$$\begin{aligned} &2D^2F(x, y)[w, w] - \big(DF(x, y)[w]\big)^2 \\ &\quad = a^2A^2u^2 + b^2u^2 + a^2v^2 + 2abuv - 2a^2Auv - 2abAu^2 + 2aBu^2 \\ &\quad = (aAu - bu - av)^2 + 2aBu^2 \geq 0, \end{aligned}$$

som visar att funktionen F är 2-självkonkordant.

18.7 Använd resultatet i övning 18.5 med $f(x) = x \ln x$.

18.8 Resultatet i övning 18.5 ger för $f(x) = -\ln x$ att

$$F(x, y) = -\ln(\ln x + y) - \ln x$$

är en 2-självkonkordant barriär till slutna höljet av området $-y < \ln x$. Eftersom $G(x, y) = F(y, -x)$, följer det sedan av sats 18.1.3 att G är en 2-självkonkordant barriär till området $y \geq \mathrm{e}^x$.

Sakregister

affin
  avbildning, 25
  dimension, 23
  hölje, 22
  kombination, 21
  mängd, 21
  styckvis — funktion, 100
aktivt bivillkor, 169
analytiskt centrum, 354
andraderivata, 18
Armijos regel, 293
artificiell variabel, 271
avstånd, 10

barriär, 354
basindexmängd, 246
  tillåten —, 253
baslösning, 248
  degenererad —, 249
  tillåten —, 253
basvariabel, 248
begränsad mängd, 12
bidualkon, 59
bild, 5
  invers —, 5
bivillkor
  aktivt —, 169
  implicit —, 169
Blands regel, 267
boll, 10
brantaste lutningsmetoden, 292, 296

centrala vägen, 355
cykling, 257

degenererad baslösning, 249
derivata, 16
descentmetod, 291
dietproblemet, 176
differential, 16
differentierbar, 16
dimension, 23
domän, 5
dual
  funktion, 192
  lokal norm, 368
duala
  priser, 177
  problem, 194, 223, 238
  simplexalgoritmen, 280
dualitet, 194, 223, 238
  stark —, 194
  svag —, 194, 225, 238
dualitetssatsen, 206, 226, 275
dualkon, 58
dämpade Newtonmetoden, 309

effektiv domän, 5
ellipsoidmetoden, 283
epigraf, 91
euklidisk norm, 10
explicit bivillkor, 169
extremalpunkt, 67
extremalstråle, 68

Farkas lemma, 62
fas 1, 270, 359
fasad, 69, 77
Fenchels olikhet, 151

Fencheltransform, 150
form
    kvadratisk —, 9
    linjär —, 8
fri variabel, 248

generator, 40
gradient, 16

halvrum, 28
    koniskt —, 37
heltalsprogrammering, 171
hessian, 18
hyperplan, 24
    separerande —, 51
    stödhyperplan, 54
högerderivata, 125
Hölders olikhet, 108
hölje
    affint —, 22
    koniskt —, 40
    konvext —, 32

icke-degenererad
    självkonkordant funktion, 329
implicit bivillkor, 169
indikatorfunktion, 151
inputlängd, 386, 387
inre, 11
    punkt, 11
invers bild, 5

Jensens olikhet, 97
Johns sats, 200

Karush–Kuhn–Tuckers
    sats, 159, 207
    villkor, 198
kompakt mängd, 12
komplementaritet, 197
komplementaritetssatsen, 229
kon, 37
    bidualkon, 59
    dualkon, 58
    polyedrisk —, 39
    recessionskon, 42
    äkta —, 38
    ändligt genererad —, 41
konditionstal, 136
konisk
    kombination, 39
    polyeder, 39
koniskt
    halvrum, 37
    hölje, 40
konjugatfunktion, 150
konkav funktion, 92
    strikt —, 96
kontinuerlig funktion, 13
konvergens
    kvadratisk —, 295, 296
    linjär —, 295, 296
konvex
    funktion, 92
    hölje, 32
    kombination, 26
    kvadratisk programmering, 171
    mängd, 26
    optimering, 170, 205
konvex funktion, 92
    starkt —, 133
    strikt —, 96
kvadratisk
    form, 9
    konvergens, 295, 296
kvalificerande villkor, 199
kvasikonkav, 93
    strikt —, 96
kvasikonvex, 93
    strikt —, 96
känslighetsanalys, 220

ℓ^1-norm, 10
ℓ^p-norm, 96
Lagrangefunktionen, 191
Lagrangemultiplikator, 191
ligga mellan, 68
linjefri mängd, 46
linjesökning, 292
linjär
    avbildning, 8
    form, 8
    konvergens, 295, 296
    operator, 8
    programmering, 170
Lipschitzkontinuerlig, 13
lokal seminorm, 305

maxnorm, 10
medelvärde, 106
medelvärdessatsen, 17
Minkowskifunktionalen, 121
Minkowskis olikhet, 109
minsta kvadratlösningen, 187
målfunktion, 166

ν-självkonkordant barriär, 361
Newtondekrement, 304, 319
Newtonriktning, 303, 319
Newtons metod, 292, 309, 320, 346
norm, 10, 96
    euklidisk, 10
    ℓ^1-norm, 10

operatornorm, 14
optimal
    lösning, 166
    punkt, 166
optimalitetskriteriet, 193, 226, 239
optimalt värde, 166
optimering
    icke-linjär —, 171
    konvex —, 170, 205
    linjär —, 170
ortant, 7

perspektiv, 103
perspektivavbildningen, 30
pivotelement, 242
polyeder, 29
polyedrisk kon, 39
polynomiell algoritm, 283
positivt
    definit, 9
    homogen, 95
    semidefinit, 9

rand, 11
randpunkt, 11
recedera, 42
recessionskon, 42
recessionsvektor, 42
recessivt delrum, 46
    till funktion, 114
reducerad kostnad, 254
relativ
    rand, 34
    randpunkt, 34
relativt
    inre, 34
    inre punkt, 34
rena Newtonmetoden, 309
rening, 383
riktningsderivata, 155

sadelpunkt, 196
seminorm, 95
separerande hyperplan, 51
simplexalgoritmen, 256
    duala —, 280
    fas 1, 270
simplextabell, 242
självkonkordant funktion, 326
    icke-degenererad —, 329
    med parameter ν, 361
slackvariabel, 173

Slaters villkor, 159, 205
sluten
    boll, 11
    konvex funktion, 116
    mängd, 12
slutna höljet, 12
snitt, 4
standardform, 237, 370
standardskalärprodukt, 6
stark dualitet, 194
starkt konvex, 133
steglängd, 292
strikt
    konvex, 96
    kvasikonkav, 96
    kvasikonvex, 96
stråle, 37
sträcka, 7
styckvis affin, 100
stödfunktion, 118
stödhyperplan, 54
stödlinje, 129
subadditiv, 95
subdifferential, 141
subgradient, 141
subnivåmängd, 91
surplusvariabel, 173
svag dualitet, 194, 225, 238
symmetrisk avbildning, 8
sökriktning, 292

tillslutning av funktion, 148
tillåten
    lösning, 166
    punkt, 166
translat, 7
transponerad avbildning, 8
transportproblemet, 179
tvåpersoners nollsummespel, 181

union, 4
utsatt punkt, 77

vägföljande metoden, 358
vänsterderivata, 125
värde, 166

yttre punkt, 11

äkta
    fasad, 69
    kon, 38
ändligt genererad kon, 41

öppen
    boll, 10
    mängd, 12
    sträcka, 7