Konvexe Optimierung

Prof. Dr. Sven Rahmann

LS 11, Fakultät für Informatik, TU Dortmund

2009–2010
Entwurf vom 17. Mai 2010

Vorbemerkungen

Dieses Dokument enthält Notizen zu meiner Vorlesung *Konvexe Optimierung*, die ich im Wintersemester 2009/10 an der TU Dortmund gehalten habe. Sie richtet sich stark nach dem Buch *Convex Optimization* von Stephen Boyd und Lieven Vandenberghe, das bei Cambridge University Press erschienen und sehr lesenswert ist.

Vielen wird der Begriff der *linearen Optimierung* vertraut sein. Dabei geht es darum, eine lineare (oder affine) Funktion unter linearen (affinen) Nebenbedingungen zu optimieren; die Nebenbedingungen sind Gleichungen und/oder Ungleichungen. Die bekannteste Methode hierfür ist das *Simplex-Verfahren*, doch es gibt andere, in der Theorie effizientere Verfahren.

Das Buch *Convex Optimization* von Stephen Boyd und Lieven Vandenberghe, auf das diese Vorlesung aufbaut, stellt die These auf, dass im Grunde eine weitaus größere Klasse an Optimierungsproblemen, die *konvexen Optimierungsprobleme*, fast genauso einfach und effizient zu lösen ist wie lineare Probleme.

Bei einem konvexen Problem ist die Zielfunktion konvex, und die Nebenbedingungen haben die Form $f(x) \leq 0$ mit konvexer Funktion f. Eine Funktion ist konvex, wenn sie stets unterhalb der Strecken verläuft, die Punkte auf ihrem Graphen miteinander verbinden. Die große Bedeutung der Konvexität in der Optimierung wird klar, wenn man sich überlegt, dass ein lokales Minimum einer konvexen Funktion gleichzeitig auch globales Minimum ist.

Zur Zeit befindet sich dieses Skript noch im Aufbau; mit Fehlern und Unvollständigkeiten ist daher leider noch zu rechnen. Für Hinweise dazu bin ich jederzeit dankbar.

– Prof. Dr. Sven Rahmann, TU Dortmund

Inhaltsverzeichnis

KAPITEL 1

Einleitung

In der konvexen Optimierung geht es um die Minimierung konvexer Funktionen unter konvexen Nebenbedingungen. Die lineare Optimierung ist ein Spezialfall davon. Ein anderer wichtiger Spezialfall ist die Optimierung quadratischer Funktionen unter quadratischen und/oder linearen Nebenbedingungen. Viele Anwendungsprobleme lassen sich konvex formulieren. Wir geben hier einen Vorgeschmack und wiederholen dabei und danach grundlegende Begriffe, vor allem aus der linearen Algebra.

1.1 Ein geometrisches Problem

Wir betrachten examplarisch folgendes geometrisches Problem. Gegeben seien n Punkte im $\mathbb{R}^2$, nämlich (x_i, y_i) mit $i = 1, \ldots, n$. Wir suchen den kleinsten Kreis, also Mittelpunkt (x, y) und Radius $r \geq 0$, der alle Punkte enthält. Offensichtlich lässt sich dieses Problem mit $q = r^2$ wie folgt formulieren.

$$\begin{array}{lll} \text{Minimiere} & q & \\ \text{so dass} & (x_i - x)^2 + (y_i - y)^2 \leq q & \text{für } i = 1, \ldots, n. \end{array}$$

Mit den Variablen $z = (x, y, q)$ lässt sich dies schreiben als

$$\begin{array}{lll} \text{Minimiere} & f(z) & \\ \text{so dass} & f_i(z) \leq 0 & \text{für } i = 1, \ldots, n, \end{array}$$

mit der linearen (und damit konvexen) Funktion $f(z) = q$ und den konvexen (nichtlinearen) Funktionen $f_i(z) = (x_i - x)^2 + (y_i - y)^2 - q$ für $i = 1, \ldots, n$.

In diesem Skript geht es darum, wie man aus gegebenen Problemstellungen (aus verschiedenen Anwendungen) ein solches abstraktes konvexes Optimierungsproblem formuliert, und wie man solche Probleme algorithmisch löst.

Man beachte, dass wir schon einiges an Arbeit geleistet haben, um das geometrische Problem in der obigen Form zu formulieren: Wir haben Variablen (x, y) für den Mittelpunkt eingeführt, sowie eine Variable q für das Quadrat des gesuchten Radius r. Eigentlich sollten die Nebenbedingungen ja $d_{(x,y)}(x_i, y_i) \leq r$ lauten mit $d_{x,y}(u, v) = \sqrt{(x-u)^2 + (y-v)^2}$, aber die obige Formulierung scheint auf den ersten Blick natürlicher. Die (im Grunde) äquivalente Formulierung mit den Variablen (x, y, r) mit $(x - x_i)^2 + (y - y_i)^2 - r^2 \leq 0$ ist kein konvexes Problem (warum nicht?).

Hier wird schon deutlich, dass man sich bei der Formulierung eines (angewandten) Problems als konvexes Optimierungsproblem Mühe geben muss. Ist diese Arbeit geschafft, greifen eine Reihe von Formalismen und Algorithmen, die in diesem Skript vorgestellt werden.

1.2 Ein Beispiel aus dem Maschinellen Lernen: SVMs

Gegeben seien n Datenpunkte in einem d-dimensionalen Raum, die aus zwei verschiedenen Klassen stammen. Die Klassen bezeichnen wir oBdA mit $+1$ und -1. Beispielsweise kennen wir zu $n = 100$ Patienten die Genexpressionswerte von $d = 20000$ Genen und wissen, ob diese an einer bestimmten Krankheit leiden (Klasse $+1$) oder nicht (Klasse -1). Gegeben sind also n Vektoren $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ für $i = 1, \ldots, n$.

Gesucht ist nun eine Hyperebene h in $\mathbb{R}^d$, die die beiden Klassen trennt, d.h., eine Punktmenge $h = \{ y \in \mathbb{R}^d \mid \langle w|y\rangle = b \}$ mit $w \in \mathbb{R}^d$ und $b \in \mathbb{R}$, so dass im Halbraum $H^- := \{ y \in \mathbb{R}^d \mid \langle w|y\rangle \leq b \}$ die Punkte der Klasse -1 liegen und im Halbraum $H^+ := \{ y \in \mathbb{R}^d \mid \langle w|y\rangle \geq b \}$ die Punkte der Klasse $+1$. Sofern solche Hyperebenen für die gegebenen Punkte überhaupt existieren (die Punkte also linear trennbar sind), suchen wir idealerweise eine Hyperebene, die den Abstand zu den nächstgelegenen Punkten aus jeder Klasse maximiert (eine sog. *maximum margin hyperplane*).

Ist in der Darstellung von h der Vektor $w \in \mathbb{R}^d$ ein normierter Vektor ($\|w\|_2 = 1$), dann ist der Abstand eines Punktes x zu h genau $\langle w|x\rangle - b$. Die Forderung, der kleinste Abstand eines Punkte x_i von h mindestens c beträgt, ist äquivalent dazu, dass alle Punkte x_i mindestens den Abstand c von h haben sollen, also $|\langle w|x_i\rangle - b| \geq c$. Berücksichtigt man die Klassenlabel y_i lässt sich dies als

$$y_i \cdot (\langle w|x_i\rangle - b) - c \geq 0 \quad \text{für } i = 1, \ldots, n$$

schreiben. Um die Anzahl der Variablen zu reduzieren (und die Beschränkung von w auf normierte Vektoren aufzuheben), dividieren wir durch c und setzen $w' := w/c$ und $b' := b/c$. Dies führt auf die Bedingungen

$$y_i \cdot (\langle w'|x_i\rangle - b') - 1 \geq 0 \quad \text{für } i = 1, \ldots, n$$

Der geforderte Mindestabstand eines Punktes x_i von h ist damit entsprechend $1/\|w'\|_2$. Also gilt es, diesen zu maximieren, oder äquivalent $\|w'\|_2^2$ zu minimieren.

Wir nennen jetzt w' wieder w und b' wieder b und haben also das (quadratische konvexe) Optimierungsproblem

Minimiere	$\|w\|_2^2$	
so dass	$y_i \cdot (\langle w \| x_i \rangle - b) - 1 \geq 0$	für $i = 1, \ldots, n$.

Die Variablen sind $w = (w_1, \ldots, w_d)$ und $b \in \mathbb{R}$. Die Zielfunktion ist quadratisch in den w_j, die Nebenbedingungen sind linear in allen Variablen.

Dieses Problem lässt sich im Prinzip so wie es jetzt aufgestellt ist lösen. Wenn die Klassen überhaupt linear trennbar sind, finden wir eine optimale Lösung (w, b). Bekommen wir neue zu klassifizierende Punkte x, betrachten wir einfach das Vorzeichen von $\langle w|x\rangle - b$, um uns für eine Klasse zu entscheiden.

Wir werden in Kapitel 7.1 sehen, dass es sich lohnt, die *duale Formulierung* (Kapitel 6 dieses Problems anzuschauen; diese wird es uns leicht machen, auch nichtlineare Klassifikationen zuzulassen. Außerdem werden wir das Problem dahin verallgemeinern, dass wir eine geringe Anzahl an Fehlklassifikationen zulassen, wenn dadurch der Abstand (margin) stark vergrößert werden kann.

1.3 Aufbau des Skripts

Wir beginnen mit einer ausführlichen Definition der Konvexität. Kapitel 2 stellt konvexe Mengen und ihre Eigenschaften vor. Auf konvexen Mengen werden in Kapitel 3 konvexe Funktionen definiert und ihre Eigenschaften beleuchtet. Als Spezialfälle betrachten wir auch stets durch lineare (Un-)Gleichungen beschränkte Mengen und lineare bzw. affine Funktionen darauf. Mit diesen Begriffen untersuchen wir in Kapitel 5 Methoden zur Optimierung ohne Nebenbedingungen, die die Grundlage für spätere Probleme mit Nebenbedingungen bilden.

1.4 Wiederholung von Elementen der Linearen Algebra

Reelle n-Tupel. Es sei $\mathbb{R}$ die Menge der reellen Zahlen, $\mathbb{R}_+$ die Menge der nichtnegativen reellen Zahlen, und $\mathbb{R}_{++}$ die Menge der echt positiven reellen Zahlen.

Den Vektorraum der d- bzw. n-Tupel reeller Zahlen bezeichnen wir mit $\mathbb{R}^d$ bzw. $\mathbb{R}^n$. Auf $\mathbb{R}^n$ ist für $p > 1$ die p-Norm $\|\cdot\|_p : \mathbb{R}^n \to \mathbb{R}$ definiert mit $x = (x_1, \ldots, x_n) \mapsto (\sum_{i=1}^n |x_i|^p)^{1/p}$. Für $p \to \infty$ erhält man die max-Norm $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$.

Für $p = 2$ ergibt sich die übliche Euklidische Norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ und das zugehörige Skalarprodukt, das wir in der bra-ket Notation schreiben: Für $x, y \in \mathbb{R}^n$ ist $\langle x|y\rangle := \sum_{i=1}^n x_i y_i$. Fassen wir x und y als Spaltenvektoren bzw. als $n \times 1$-Matrizen auf, dann ist mit dem üblichen Matrizenprodukt $x^T y = \langle x|y\rangle$, wobei x^T den Vektor x transponiert darstellt (also als $1 \times n$-Zeilenvektor).

Je nach Gewohnheit ist die Notation mit transponierten Vektoren oder die bra-ket-Notation leichter zu lesen. Dem mit der bra-ket-Notation nicht vertrauten Leser wird geraten, die Sätze und Gleichungen in eine vertraute Notation zu übertragen. Beispielsweise ist noch

$x^T Ay = \langle x|A|y\rangle$, wenn x und y Spaltenvektoren sind und A eine passende Matrix ist. Es ist $x^T x = \langle x|x\rangle = \|x\|_2^2 \in \mathbb{R}$; andererseits ist $xx^T = |x\rangle\langle x|$ eine $n \times n$-Matrix mit Rang 1. Ein Vorteil der bra-ket-Notation liegt darin, dass man sich nicht merken muss, ob x als Spalten- oder Zeilenvektor gegeben ist. Stets ist $\langle x|$ ein Zeilenvektor und $|x\rangle$ ein Spaltenvektor, wenn $x \in \mathbb{R}^n$.

Ein endlich-dimensionaler Vektorraum mit Skalarprodukt wird auch *Euklidischer Raum* genannt. In ihm gilt die *Cauchy-Schwarz'sche Ungleichung* $\langle x|y\rangle^2 \leq \langle x|x\rangle\langle y|y\rangle$ oder nach Wurzelziehen $\langle x|y\rangle \leq \|x\|\|y\|$, wobei die Norm die aus dem Skalarprodukt gewonnenene Norm ist.

Symmetrische Matrizen. Es sei $\mathbb{S}^n$ die Menge der symmetrischen $n \times n$-Matrizen, ein Vektorraum der Dimension $\binom{n+1}{2}$. Weiter sei $\mathbb{S}_+$ die Menge der positiv semidefiniten Matrizen in $\mathbb{S}^n$, also die Menge derjenigen $X \in \mathbb{S}^n$, für die $\langle z|X|z\rangle \geq 0$ für alle Vektoren $z \in \mathbb{R}^n$ ist. Analog sei $\mathbb{S}_{++}$ die Menge der positiv definiten Matrizen in $\mathbb{S}^n$, also die Menge derjenigen $X \in \mathbb{S}^n$, für die $\langle z|X|z\rangle > 0$ für alle Vektoren $z \neq 0$ ist.

Auf $\mathbb{S}^n$ ist ebenfalls ein Skalarprodukt definiert, für das wir ebenso die bra-ket-Notation benutzen(!), nämlich $\langle X|Y\rangle := \operatorname{tr}(X^T Y)$. Damit wird $\mathbb{S}^n$ ebenfalls zu einem Euklidischen Raum. Bei dieser Definition ist $\operatorname{tr}(A) = \sum_{i=1}^n A_{ii}$ die Spur (engl. *trace*) von A, also die Summe der Diagonalelemente, die gleich der Summe der Eigenwerte von A ist. Daher ist $\langle X|Y\rangle = \sum_{i=1}^n \sum_{j=1}^n X_{ij}Y_{ij}$; das entspricht dem Euklidischen Skalarprodukt, wenn man eine Matrix als Vektor im $\mathbb{R}^{n^2}$ auffasst. Die zugehörige Norm ist die *Frobenius-Norm* mit $\|X\|_F^2 = \langle X|X\rangle = \operatorname{tr}(X^T X) = \sum_{i,j=1}^n X_{ij}^2$.

Neben der Spur einer Matrix X ist auch ihre *Determinante* $\det X$ definiert; sie lässt sich beispielsweise als Produkt ihrer Eigenwerte schreiben. Bei einer Diagonal- oder (nicht symmetrischen) Dreiecksmatrix ist die Determinante gleich dem Produkt der Diagonalelemente.

Orthogonalität. Ein Euklidischer Raum induziert den Begriff der Orthogonalität. Zwei Vektoren x, y heißen orthogonal, wenn $\langle x|y\rangle = 0$. Ist U ein Untervektorraum von V, so heißt die Menge aller Vektoren, die orthogonal zu *jedem* Element von U ist, *orthogonales Komplement* von U. Es wird als $U^\perp := \{\, x \mid \langle x|y\rangle = 0 \text{ für alle } y \in U \,\}$ geschrieben. Ist $0 \neq a \in V$ ein einzelner Vektor und U der von a aufgespannte eindimensionale Unterraum von V, dann ist $a^\perp := U^\perp = \{\, x \mid \langle x|a\rangle = 0 \,\}$.

KAPITEL 2

Konvexe Mengen

2.1 Lineare Räume, affine Mengen, Kegel und konvexe Mengen

2.1 Definition. Seien $x_1, \ldots, x_k$ Punkte in einem reellen Vektorraum V.

1. Eine *Linearkombination* der $x_1, \ldots, x_k$ ist eine Summe der Form $\sum_{i=1}^{k} \theta_i x_i$ mit $\theta_i \in \mathbb{R}$ für alle $i = 1, \ldots, k$.
2. Eine *affine Kombination* der $x_1, \ldots, x_k$ ist eine Summe der Form $\sum_{i=1}^{k} \theta_i x_i$ mit $\theta_i \in \mathbb{R}$ für alle $i = 1, \ldots, k$ und $\sum_{i=1}^{k} \theta_i = 1$.
3. Eine *konische Kombination* der $x_1, \ldots, x_k$ ist eine Summe der Form $\sum_{i=1}^{k} \theta_i x_i$ mit $\theta_i \in \mathbb{R}$ und $\theta_i \geq 0$ für alle $i = 1, \ldots, k$.
4. Eine *Konvexkombination* der $x_1, \ldots, x_k$ ist eine Summe der Form $\sum_{i=1}^{k} \theta_i x_i$ mit $\theta_i \in \mathbb{R}$ und $\theta_i \geq 0$ für alle $i = 1, \ldots, k$ und $\sum_{i=1}^{k} \theta_i = 1$.

2.2 Beispiel (Ein Punkt)**.** Wir betrachten den Spezialfall eines einzelnen Punktes $x \neq 0$.

1. Die Menge der Linearkombinationen aus x ist die Gerade durch den Nullpunkt und x.
2. Die Menge der affinen Kombinationen ist $\{x\}$.
3. Die Menge der konischen Kombinationen ist $r(x) := \{\theta \cdot x : \theta \geq 0\}$; eine solche Menge nennt man den (vom Nullpunkt ausgehenden) *Strahl (engl.* ray*)* durch x.
4. Die Menge der Konvexkombinationen ist wieder $\{x\}$.

♡

2.3 Beispiel (Zwei Punkte)**.** Wir betrachten den Spezialfall von $k = 2$ Punkten x_1, x_2. Wie sehen die Mengen aus, die man aus zwei Punkten kombinieren kann?

1. Die Menge aller Linearkombinationen von x_1 und x_2 ist der kleinste Untervektorraum, der x_1 und x_2 enthält. Sind x_1 und x_2 linear unabhängig, ist dies die Ebene durch den Nullpunkt, x_1 und x_2.
2. Die Menge aller affinen Kombinationen von x_1 und x_2 ist die Gerade durch x_1 und x_2.
3. Die Menge aller konischen Kombinationen von x_1 und x_2 ist die Menge von Strahlen durch x_1 und x_2 und jeden Punkt auf der Strecke zwischen x_1 und x_2.
4. Die Menge aller Konvexkombinationen von x_1 und x_2 ist die Verbindungsstrecke zwischen x_1 und x_2; wir können sie auch als $\{\theta x + (1-\theta)x_2 : 0 \leq \theta \leq 1\}$ schreiben.

♡

2.4 Definition. Sei V ein reeller Vektorraum (z.B. $V = \mathbb{R}^n$); sei $U \subset V$.

1. U heißt *linearer Raum* oder *Unterraum* von V, wenn mit je zwei Punkten auch jede ihrer Linearkombinationen in U liegt.
2. U heißt *affiner Raum* oder *affine Menge*, wenn mit je zwei Punkten auch jede ihrer affinen Kombinationen in U liegt.
3. U heißt *konvexer Kegel*, wenn mit je zwei Punkten auch jede ihrer konischen Kombinationen in U liegt. (U heißt (nicht notwendig konvexer) *Kegel*, wenn mit jedem Punkt auch der hindurchgehende Strahl in U liegt.)
4. U heißt *konvexe Menge*, wenn mit je zwei Punkten auch jede ihrer Konvexkombinationen in U liegt.

Alternativ kann man in Definition 2.4 statt Kombinationen von je zwei Punkten solche von endlich vielen Punkte betrachten; die Äquivalenz zeigt man durch vollständige Induktion.

Anschaulich bedeutet Konvexität, dass mit je zwei Punkten auch ihre Verbindungsstrecke in der Menge enthalten ist.

In der Wahrscheinlichkeitsrechnung ist noch eine andere Schreibweise von Konvexkombinationen gängig. Wir betrachten die Koeffizienten $\theta = (\theta_1, \ldots, \theta_k)$ der Konvexkombination als Wahrscheinlichkeitsverteilung auf k Punkten und x als Zufallsvektor, der gemäß θ verteilt ist; dann ist gerade der Erwartungswert $\mathbb{E}_\theta[x] = \sum_{i=1}^k \theta_i x_i$ die Konvexkombination. Dies lässt sich mit Hilfe der Maßtheorie auf abzählbar und überabzählbar viele Punkte ausdehnen. Am allgemeinsten ist folgende Definition von Konvexität.

2.5 Definition (Wahrscheinlichkeitstheoretische Defintion von Konvexität)**.** Eine Teilmenge U eines Vektorraums heißt konvex, wenn für jede Verteilung $\mathbb{P}$ mit $\mathbb{P}(U) = 1$ gilt, dass $\mathbb{E}\,\mathbb{P} \in U$.

2.6 Lemma (Abgeschlossenheit unter Schnitten)**.** *Seien X, Y lineare Räume (affine Mengen; konvexe Kegel; konvexe Mengen). Dann ist ihr Schnitt $X \cap Y$ ebenfalls ein linearer Raum (affine Menge; konvexer Kegel; konvexe Menge).*

Sind X_i, $i \in I$ (eine beliebige Indexmenge) lineare Räume (affine Mengen; konvexe Kegel; konvexe Mengen), dann auch der Schnitt $\bigcap_{i\in I} X_i$.

Beweis. Die Aussagen folgen elementar aus der Definition der Konvexität. □

Ist eine Eigenschaft abgeschlossen unter Schnitten (z.B. linear, affin, konvexer Kegel, konvex) und ist eine endliche Menge X gegeben, dann kann man nach dem Schnitt U aller Mengen fragen, die X enthalten und die genannte Eigenschaft haben. Ein solches U nennt man dann *Hülle*. Man spricht in dem Zusammenhang dann von der *kleinsten* Menge mit der genannten Eigenschaft, die X enthält.

2.7 Definition (Hüllen)**.** Gegeben sei eine beliebige Teilmenge X eines Vektorraums V.

1. Die *lineare Hülle* von X, auch $\operatorname{span}(X)$, ist der kleinste Untervektorraum von V, der X enthält, also der Schnitt aller Untervektorräume, die X enthalten.
2. Die *affine Hülle* von X, auch $\operatorname{aff}(X)$, ist die kleinste affine Menge in V, die X enthält, also der Schnitt aller affinen Mengen, die X enthalten.
3. Die *konische Hülle* von X, auch $\operatorname{cone}(X)$, ist der kleinste konvexe Kegel in V, der X enthält, also der Schnitt aller konvexen Kegel, die X enthalten.
4. Die *konvexe Hülle* von X, auch $\operatorname{conv}(X)$, ist die kleinste konvexe Menge in V, die X enthält, also der Schnitt aller konvexen Mengen, die X enthalten.

Ist X endlich, kann man häufig eine konkretere Charakterisierung geben: In den genannten Fällen kann man nachweisen, dass die Hülle gleich der Menge aus den entsprechenden Kombinationen von Punkten aus X ist. Die Äquivalenz für unendliche Mengen nachzuweisen, kann subtiler sein. Wir zeigen dies nur am Beispiel der konvexen Hülle und Konvexkombinationen endlich vieler Punkte.

2.8 Satz. *Sei $X \subset V$ eine endliche Teilmenge eines reellen Vektorraums V. Dann ist* $\operatorname{conv}(X)$ *gleich der Menge K aller Konvexkombinationen von Punkten aus X. Entsprechendes gilt für* $\operatorname{span}(X)$, $\operatorname{aff}(X)$ *und* $\operatorname{cone}(X)$.

Beweis. Zuerst zeigen wir $\operatorname{conv}(X) \subset K$; dazu genügt es zu zeigen, dass K konvex ist und X enthält. Dass $X \subset K$ gilt, ist klar (triviale Konvexkombinationen). Die Konvexität von K weist man elementar anhand der Definition nach.

Jetzt zeigen wir $K \subset \operatorname{conv}(X)$; dazu nehmen wir ein beliebiges konvexes Y, das X enthält, und zeigen $K \subset Y$ durch vollständige Induktion über die Mächtigkeit von $X = \{x_0, \ldots, x_m\}$. Ist $m = 0$, dann ist $K = \{x_0\} = X = \operatorname{conv}(X)$ und nichts weiter zu zeigen. Sei also $m > 0$ und $y := \sum_{i=0}^{m} \theta_i x_i$ eine beliebige Konvexkombination aus X, also $y \in K$. Sei oBdA $\theta_0 \neq 1$ (ansonsten ist $y = x_0 \in X \subset Y$ klar). Dann ist nach Induktionsvoraussetzung die kleinere Konvexkombination $y' = \sum_{i=1}^{m} \frac{\theta_i}{1-\theta_0} x_i \in Y$. Da Y konvex ist, ist auch die Konvexkombination $\theta_0 x_0 + (1-\theta_0) y' = y$ in Y. Damit ist $K \subset Y$ gezeigt. Da $\operatorname{conv}(X)$ der Schnitt aller konvexen Y ist, die X enthalten und Y sobeben beliebig mit dieser Eigenschaft war, haben wir nun auch $K \subset \operatorname{conv}(X)$ gezeigt. □

2.9 Definition (affine Dimension, affine Unabhängigkeit)**.** Die *affine Dimension* einer Menge $X \subset \mathbb{R}^n$ ist die Dimension ihrer affinen Hülle. Wir nennen $k+1$ Punkte $x_0, \ldots, x_k$ *affin unabhängig*, wenn ihre affine Hülle die Dimension k hat. Dies ist genau dann der Fall, wenn $x_1 - x_0, \ldots, x_k - x_0$ linear unabhängig sind.

2.2 Einfache Beispiele für konvexe Mengen

Wir betrachten einige Beispiele für konvexe Mengen. Beweise werden, solange sie elementar sind, nicht aufgeführt: Man rechnet einfach die Konvexitätseigenschaft laut Definition nach.

Aufgabe 2.1. Beweise folgende Aussagen: Unterräume, affine Mengen und konvexe Kegel sind konvex. Unterräume sind affine Mengen und konvexe Kegel.

2.10 Definition (Positiv (semi)definite Matrizen)**.** Eine symmetrische Matrix $S \in \mathbb{R}^{n\times n}$ heißt *positiv semidefinit*, wenn $\langle x|S|x\rangle \geq 0$ für alle $x \in \mathbb{R}^n$. Sie heißt *positiv definit*, wenn $\langle x|S|x\rangle > 0$ für alle $x \neq 0$. Wir schreiben $S \trianglerighteq 0$, wenn S positiv semidefinit ist und $S \triangleright 0$, wenn S positiv definit ist. Weiter sei $\mathbb{S}^n$ die Menge aller symmetrischen $n \times n$-Matrizen, $\mathbb{S}^n_+$ die Menge aller positiv semidefiniten Matrizen in $\mathbb{S}^n$ und $\mathbb{S}^n_{++}$ die Menge aller positiv definiten Matrizen in $\mathbb{S}^n$.

2.11 Beispiel (Kegel der positiv semidefiniten Matrizen)**.** Die Menge aller positiv semidefiniten Matrizen bildet einen konvexen Kegel und ist damit konvex: Seien S_1, S_2 positiv semidefinit. Dann ist mit $\theta_1 \geq 0$ und $\theta_2 \geq 0$ auch $\langle x|\theta_1 S_1 + \theta_2 S_2|x\rangle = \theta_1 \langle x|S_1|x\rangle + \theta_2 \langle x|S_2|x\rangle \geq 0$ für alle $x \in \mathbb{R}^n$, also die konische Kombination $\theta_1 S_1 + \theta_2 S_2$ positiv semidefinit. ♡

Aufgabe 2.2. Welche Bedingungen muss eine symmetrische 2×2-Matrix erfüllen, um positiv semidefinit zu sein?

Aufgabe 2.3. Bilden die positiv definiten Matrizen einen konvexen Kegel? eine konvexe Menge?

2.12 Definition (Hyperebenen und Halbräume)**.** Eine *Hyperebene* im $\mathbb{R}^n$ ist eine Menge der Form $H := \{\, x \mid \langle a|x\rangle = b \,\}$ mit $0 \neq a \in \mathbb{R}^n$ und $b \in \mathbb{R}$. Alternativ kann $H = \{\, x \mid \langle a|x - x_0\rangle = 0 \,\}$ mit einem x_0 mit $\langle a|x_0\rangle = b$ geschrieben werden. Eine weitere Schreibweise ist daher $x_0 + a^\perp$.

Eine Hyperebene H teilt $\mathbb{R}^n$ in zwei (abgeschlossene) *Halbräume* $\{\, x \mid \langle a|x\rangle \leq b \,\}$ und $\{\, x \mid \langle a|x\rangle \geq b \,\}$, die sich in H schneiden.

Aufgabe 2.4. Beweise: Hyperebenen sind affine Mengen und daher konvex. Ein Halbraum ist konvex, aber nicht affin.

2.13 Beispiel (Euklidische Kugeln und Ellipsoide)**.** Eine (Euklidische) *Kugel* im $\mathbb{R}^n$ hat die Form

$$B(x_c, r) := \{\, x \mid \|x - x_c\|_2 \leq r \,\} = \{\, x \mid \langle x - x_c|x - x_c\rangle \leq r^2 \,\}$$

mit $x_c \in \mathbb{R}^n$ und $r \geq 0$. Dabei heißt x_c Mittelpunkt oder Zentrum der Kugel und r Radius. Die Kugel beinhaltet alle Punkt mit Abstand höchstens r vom Zentrum. Eine andere Schreibweise ist

$$B(x_c, r) = \{\, x_c + ru \mid \|u\|_2 \leq 1 \,\}.$$

Verzerrt man die Koordinaten mit einer positiv definiten symmetrischen Matrix A, so erhält man ein *Ellipsoid*

$$B(x_c, A) := \{\, x_c + Au \mid \|u\|_2 \leq 1 \,\}.$$

Eine andere Darstellung davon ist

$$B(x_c, A) = \{\, x \mid \langle x - x_c| A^{-2} |x - x_c\rangle \leq 1 \,\}.$$

Die Länge der Halbachsen des Ellipsoids sind $1/\mu_i$, wobei μ_i die Eigenwerte von A sind. Ist A positiv semidefinit aber singulär, so handelt es sich um einen degenerierten Ellipsoiden (mindestens eine Halbachse hat Länge Null). ♡

Aufgabe 2.5. Beweise: Euklidische Kugeln und Ellipsoide sind konvex.

2.14 Beispiel (Normkugeln und Normkegel)**.** Sei $\|\cdot\|$ eine beliebige Norm im $\mathbb{R}^n$, nicht notwendig die Euklidische. Wir definieren *Normkugel* mit Mittelpunkt x_c und Radius r als

$$B(x_c, r) := \{\, x \mid \|x - x_c\| \leq r \,\}$$

und den *Normkegel*

$$C := \{\, (x,t) \mid \|x\| \leq t \,\} \subset \mathbb{R}^{n+1}.$$

♡

Aufgabe 2.6. Beweise: Normkugeln und Normkegel sind konvex. Zeichne im $\mathbb{R}^2$ die Normkugeln zur 1, 2, und ∞-Norm und auch die entsprechenden Normkegel. Zeichne weiter den Ellipsoiden, der durch $A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$ beschrieben wird.

2.15 Definition (Simplex)**.** Ein *Simplex* ist eine Menge, das als konvexe Hülle von $k+1$ affin unabhängigen Punkten, den Eckpunkten, beschrieben werden kann, und ist daher konvex. Seien also $v_0, \ldots, v_k$ affin unabhängig (d.h., $v_1 - v_0, \ldots, v_k - v_0$ linear unabhängig); dann ist $\operatorname{conv}\{\, v_0, \ldots, v_k \,\}$ ein Simplex mit affiner Dimension k.

2.16 Beispiel (Simplexe)**.** Beispiele sind das n-dimensionale *Einheitssimplex* im $\mathbb{R}^n$ mit 0 und den n Einheitsvektoren als Eckpunkten, und das $(n-1)$-dimensionale *Wahrscheinlichkeitssimplex* im $\mathbb{R}^n$ mit den Einheitsvektoren als Eckpunkten. Jeder Punkt im Wahrscheinlichkeitssimplex entspricht einer Wahrscheinlichkeitsverteilung auf n Punkten; die Ecken sind die Dirac-Verteilungen. ♡

2.17 Definition (Polyeder, Polytop)**.** Ein *Polyeder* $\mathcal{P}$ im $\mathbb{R}^n$ ist die Schnittmenge von endlich vielen Halbräumen und Hyperebenen und daher konvex. Handelt es sich um m Halbräume $\{\, x \mid \langle a_j | x \rangle \leq b_j \,\}$ und p Hyperebenen $\{\, x \mid \langle c_j | x \rangle = d_j \,\}$, so lässt sich $\mathcal{P}$ schreiben als

$$\mathcal{P} = \{\, x \mid Ax \leq b,\ Cx = d \,\}. \tag{2.1}$$

Dabei ist A die $m \times n$-Matrix mit Zeilen a_j und C die $p \times n$-Matrix mit Zeilen c_j. Ist ein Polyeder beschränkt, so spricht man auch von einem *Polytop*.

2.3 Konvexitätserhaltende Operationen

Bereits in Lemma 2.6 hatten wir gesehen, dass beliebige *Schnitte* von konvexen Mengen konvex sind. Hiermit ergibt sich ein alternativer Beweis für die Konvexität (und auch Kegeleigenschaft) in Beispiel 2.11. Ebenso gilt (mit ebenso einfachem Beweis) das folgende Lemma.

2.18 Lemma (Cartesische Produkte)**.** *Das Cartesische Produkt zweier konvexer Mengen ist konvex.*

2.19 Lemma (Bilder und Urbilder unter Funktionen, die Strecken auf Strecken abbilden)**.** *Sei $f : \mathbb{R}^n \to \mathbb{R}^m$ eine Funktion, die Strecken auf Strecken abbildet. Seien $C \subset \mathbb{R}^n$ und $D \subset \mathbb{R}^m$ konvex. Dann sind auch das Bild $f(C) \subset \mathbb{R}^m$ und das Urbild $f^{-1}(D) \subset \mathbb{R}^n$ konvex.*

Beweis. Die Aussage für das Bild ist unmittelbar klar. Wir beweisen die Aussage für das Urbild. Seien also $x, y \in f^{-1}(D)$ und sei $z = \theta x + (1-\theta)y$; zu zeigen ist also $z \in f^{-1}(D)$ oder $f(z) \in D$. Nun sind nach Voraussetzung $f(x) \in D$ und $f(y) \in D$, und z liegt auf der Strecke zwischen x und y, die nach Voraussetzung auf die Strecke zwischen $f(x)$ und $f(y)$ abgebildet wird. Also liegt $f(z)$ auf dieser Strecke, und da D konvex ist, liegt $f(z)$ in D, was zu zeigen war. □

2.20 Lemma (Bilder und Urbilder affiner Funktionen)**.** *Sei $f : \mathbb{R}^n \to \mathbb{R}^m$ eine affine Funktion, d.h. $f(x) = Ax+b$ mit einer $m \times n$-Matrix A und einem m-Vektor b. Seien $C \subset \mathbb{R}^n$ und $D \subset \mathbb{R}^m$ konvex. Dann sind auch das Bild $f(C) \subset \mathbb{R}^m$ und das Urbild $f^{-1}(D) \subset \mathbb{R}^n$ konvex.*

Beweis. Affine Funktionen bilden Strecken auf Strecken ab; wir wenden das vorige Lemma an. □

2.21 Beispiel (Spezielle affine Abbildungen)**.** Einfache Beispiele für affine Abbildungen sind Skalierungen, Verschiebungen, Projektionen auf ein Teil der Koordinaten, sowie Permutationen der Koordinaten. ♡

2.22 Beispiel (Summe zweier Mengen)**.** Als Anwendung der Lemmas 2.18 und 2.20 sehen wir: Sind $C_1, C_2 \subset \mathbb{R}^n$ konvex, so auch ihre Summe $C_1 + C_2 = \{\, y + z \mid y \in C_1, z \in C_2 \,\}$. Denn: $C = C_1 \times C_2 \subset \mathbb{R}^{2n}$ ist konvex, darauf wenden wir die affine (sogar lineare) Funktion $f : \mathbb{R}^{2n} \to \mathbb{R}^n : (y, z) \mapsto y + z$ an. ♡

2.23 Beispiel (Polyeder)**.** Die Menge $\mathbb{R}^m_+ \times \{0\}$ ist offensichtlich konvex. Ihr Urbild unter der Abbildung $f : x \mapsto (b - Ax, d - Cx)$ mit geeignet dimensionierten Matrizen A, C und Vektoren b, d ist das Polyeder aus Definition 2.17. ♡

Aufgabe 2.7. Zeige, dass ein Ellipsoid konvex ist, ausgehend von der Konvexität der Einheitskugel, mit Hilfe einer geeigneten affinen Abbildung.

Wir betrachten nun linear-fraktionale Funktionen; diese erhalten, wie wir sehen werden, ebenfalls die Konvexität. Ein einfacher Spezialfall ist die Perspekiv-Funktion.

2.24 Definition (Perspektiv-Funktion)**.** Die *Perspektiv-Funktion* oder *Perspektive* ist die Funktion $P : \mathbb{R}^{n+1} \to \mathbb{R}^n$, definiert auf $\mathbb{R}^n \times \mathbb{R}_{++}$ als $P : (z, t) \mapsto z/t$.

Die Perspektiv-Funktion kann man sich anhand einer Lochkamera vorstellen. Die letzte (positive) Komponente wird auf 1 normalisiert und dann abgeschnitten.

2.25 Lemma (Bilder und Urbilder unter der Perspektiv-Funktion)**.** *Bilder und Urbilder konvexer Mengen unter der Perspektiv-Funktion sind konvex.*

Beweis. Wir weisen nach, dass P Strecken auf Strecken abbildet, genauer: sind $x = (x', x'')$ und $y = (y', y'')$ die Endpunkte der Strecke $[x, y]$, dann wird $z = \theta x + (1 - \theta)y$ auf $\mu P(x) + (1-\mu)P(y)$ abgebildet, mit $\mu = \theta \cdot x''/(\theta \cdot x'' + (1-\theta) \cdot y'') \in [0, 1]$, und dieser Zusammenhang $\theta \mapsto \mu$ ist monoton auf $[0, 1]$. Daher ist $P([x, y]) = [P(x), P(y)]$. Damit liegt mit je zwei Punkten auch die Strecke zwischen ihnen im Bild von P. Die Aussage des Lemmas folgt nun aus Lemma 2.19. □

2.26 Definition (Linear-fraktionale Funktion)**.** Eine linear-fraktionale Funktion oder projektive Funktion ist eine Funktion der Form $f : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto f(x) := \frac{Ax+b}{\langle c|x\rangle+d}$, definiert wo immer $\langle c|x\rangle + d > 0$.

2.27 Lemma (Bilder und Urbilder linear-fraktionaler Funktionen)**.** *Bilder und Urbilder konvexer Mengen unter linear-fraktionalen Funktionen sind konvex.*

Beweis. Eine linear-Fraktionale Funktion ist die Komposition $f = P \circ g$ mit der affinen Funktion $g : x \mapsto \begin{pmatrix} A \\ c \end{pmatrix} x + \begin{pmatrix} b \\ d \end{pmatrix}$ mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^{1\times n}$ und $d \in \mathbb{R}$. □

2.4 Eigenschaften konvexer Mengen

2.28 Definition (Extremalpunkt)**.** Ein Punkt x einer konvexen Menge C heißt *Extremalpunkt* oder *Ecke*, wenn er sich nicht als Konvexkombination anderer Punkte darstellen lässt. Ist also $x = \sum_i \theta_i x_i$ eine Darstellung mit paarweise verschiedenen x_i, dann ist für genau ein i gerade $x = x_i$ und $\theta_i = 1$, entsprechend $\theta_j = 0$ für $j \neq i$.

2.29 Satz (Satz von Carathéodory)**.** *Sei S eine Menge von affiner Dimension k. Sei $C =$* conv *S die konvexe Hülle. Dann lässt sich jeder Punkt $x \in C$ als Konvexkombination von höchstens $k + 1$ Punkten aus S schreiben. (Die benötigten Punkte sind im allgemeinen von x abhängig.)*

Beweis. [Skizze] Wir nutzen die affine Abhängigkeit von $k + 2$ Punkten in einer entsprechenden Konvexkombination und konstruieren daraus eine kleinere Konvexkombination. □

2.5 Polyeder, Polytope und Simplexe

Jedes Simplex ist ein Polytop und lässt sich in die Form (2.1) bringen. **TODO: Beweis?** Aber nicht jedes Polytop ist ein Simplex.

2.30 Satz (Satz von Minkowski)**.** *Ist $\mathcal{P} \subset \mathbb{R}^n$ ein Polytop, so besteht $\mathcal{P}$ genau aus den Konvexkombinationen seiner endlich vielen Extremalpunkte.*

2.31 Beispiel (Würfel)**.** Der Würfel $\{ x \in \mathbb{R}^n \mid |x_i| \leq 1 \text{ für } i = 1, \ldots, n \}$ ist ein Polytop mit den 2^n Extremalpunkten mit den Komponenten ± 1. ♡

2.32 Satz (Satz von Carathéodory)**.** *Ist $\mathcal{P} \subset \mathbb{R}^n$ ein Polytop mit affiner Dimension k, dann lässt sich jeder Punkt in $\mathcal{P}$ als Konvexkombination von $k + 1$ seiner Extremalpunkte schreiben. (Für jeden Punkt muss man ggf. verschiedene Extremalpunkte heranziehen.)*

2.33 Satz. *Zu jedem Polyeder gibt es endlich viele Punkte x_i $(i = 1, \dots, m)$ und Richtungen v_i $(i = 1, \dots, p)$, so dass*

$$\mathcal{P} = \operatorname{conv}\{ x_i \mid i = 1, \dots, m \} + \operatorname{cone}\{ v_i \mid i = 1, \dots, p \}.$$

TODO: Verschiedene Darstellungen von Polyedern, Polytopen. Idee des Simplex-Verfahrens (später)

KAPITEL 3

Konvexe Funktionen

3.1 Definitionen

Für eine Funktion, die auf einer Menge $C \subset \mathbb{R}^n$ definiert ist und Werte in $\mathbb{R}$ hat, schreiben wir $f : \mathbb{R}^n \to \mathbb{R}$, dom $f = C$. Die Menge vor dem Pfeil ($\to$) gibt also nicht den Definitionsbereich, sondern den (eventuell größeren) Vektorraum an, der den Definitionsbereich enthält.

3.1 Definition (konvexe Funktion)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *konvex*, wenn

- ihr Definitionsbereich dom f eine konvexe Menge $C \subset \mathbb{R}^n$ ist, und
- für je zwei Punkte $x, y \in C$ und alle $\lambda \in [0, 1]$ gilt

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y),$$

  wenn also der Funktionswert an der Konvexkombination zweier Punkte unterhalb der Konvexkombination der Funktionswerte an den beiden Punkten liegt.

3.2 Definition (strikt konvexe Funktion)**.** Eine konvexe Funktion f heißt *strikt konvex*, wenn sogar für je zwei Punkte $x \neq y \in C$ und alle $\lambda \in (0, 1)$ gilt

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y).$$

Strikte Konvexität ist eine etwas stärkere Aussage als Konvexität alleine. Wenn sich die Differenz in der obigen Ungleichung noch sinnvoll beschränken lässt (was zum Beispiel auf kompakten Mengen immer der Fall ist), erhält man den Begriff der starken Konvexität.

3.3 Definition (stark konvexe Funktion)**.** Eine konvexe Funktion f heißt *stark konvex* oder auch *gleichmäßig konvex* mit Konstante $c > 0$, für je zwei Punkte $x \neq y \in C$ und alle $\lambda \in (0,1)$ gilt

$$\lambda f(x) + (1-\lambda)f(y) - f(\lambda x + (1-\lambda)y) \geq \frac{c}{2}\lambda(1-\lambda)\|x-y\|_2^2.$$

Für $c = 0$ erhält man offensichtlich wieder einfache Konvexität.

3.4 Lemma (Beziehung zwischen konvexen und stark konvexen Funktionen)**.** *Es ist f genau dann stark konvex mit Konstante $c > 0$, wenn $f - \frac{c}{2}\|x\|_2^2$ konvex ist.*

Beweis. Sei $f - \frac{c}{2}\|x\|_2^2$ konvex. Dann ist

$$\lambda\left[f(x) - \frac{c}{2}\|x\|_2^2\right] + (1-\lambda)\left[f(y) - \frac{c}{2}\|y\|_2^2\right] - f(\lambda x + (1-\lambda)y) + \frac{c}{2}\|\lambda x + (1-\lambda)y\|_2^2 \geq 0$$

oder äquivalent

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) - f(\lambda x + (1-\lambda)y) &\geq \frac{c}{2}\left[\lambda\|x\|_2^2 + (1-\lambda)\|y\|_2^2 - \|\lambda x + (1-\lambda)y\|_2^2\right] \\ &= \frac{c}{2}\lambda(1-\lambda)\|x-y\|_2^2\end{aligned}$$

durch bilineares Ausrechnen der Normterme. Umgekehrt funktioniert die Rechnung genauso. □

3.5 Definition (Konkavität)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *konkav* bzw. *strikt konkav* bzw. *stark konkav mit Konstante $c > 0$*, wenn $-f$ konvex bzw. strikt konvex bzw. stark konvex mit Konstante c ist.

3.6 Lemma. *Eine Funktion ist genau dann konvex (konkav), wenn sie auf jeder Strecke innerhalb ihres Definitionsbereichs konvex (konkav) ist.*

Beweis. Die Definition der Konvexität beruht nur auf zwei beliebigen Punkten des Definitionsbereichs und der Strecke dazwischen. □

3.7 Beispiel (affine und lineare Funktionen)**.** Bekanntermaßen heißt eine Funktion *affine Funktion*, wenn in Definition 3.1 sogar Gleichheit für alle $\lambda \in \mathbb{R}$ statt der Ungleichung für $0 \leq \lambda \leq 1$ gilt. Affine Funktionen, insbesondere lineare Funktionen, sind also gleichzeitig konvex und konkav. Affine Funktionen sind die einzigen Funktionen, die gleichzeitig konvex und konkav sind. ♡

3.8 Beispiel (quadratische Funktionen)**.** Die Funktion $f : \mathbb{R} \to \mathbb{R}$; $x \mapsto ax^2$ ist konvex genau dann wenn $a \geq 0$. Im Fall $a = 0$ ist die Funktion konstant und daher konvex. Ist $a > 0$, dann ist $f - ax^2$ die konvexe Nullfunktion; wegen Lemma 3.4 ist f daher stark konvex mit Konstante $2a$. ♡

3.9 Beispiel (Maximum)**.** Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$; $(x_1, \ldots, x_n) \mapsto \max_{i=1,\ldots,n} x_i$ ist konvex: Es ist $\max_i(\lambda x_i + (1-\lambda)y_i) = \lambda x_j + (1-\lambda)y_j$ für ein geeignetes j; dies ist $\leq \lambda \max_i x_i + (1-\lambda)\max_i y_i$. ♡

Ist eine konvexe Funktion nicht auf ganz $\mathbb{R}^n$ (oder dem entsprechenden gesamten Vektorraum) definiert, kann es zweckmäßig sein, sie auf dem ganzen Raum zu definieren, aber dort auf $+\infty$ zu setzen, wo sie ursprünglich nicht definiert ist. Die Konvexitätseigenschaft wird davon nicht beeinflusst. Ebenso kann man den Definitionsbereich einer konkaven Funktion auf den ganzen Vektorraum ausweiten, indem man sie an den noch nicht definierten Stellen auf $-\infty$ setzt.

3.10 Definition (Graph; Epigraph; Hypograph)**.** Der *Graph* einer Funktion $f : \mathbb{R} \to \mathbb{R}$ ist die Punktmenge $\operatorname{graph} f := \{\, (x, f(x)) \mid x \in \operatorname{dom} f \,\} \subset \mathbb{R}^{n+1}$. Der *Epigraph* von f ist die Punktmenge $\operatorname{epi} f := \{\, (x,t) \mid x \in \operatorname{dom} f, f(x) \leq t \,\} \subset \mathbb{R}^{n+1}$. Der *Hypograph* von f ist die Punktmenge $\operatorname{hypo} f := \{\, (x,t) \mid x \in \operatorname{dom} f, t \leq f(x) \,\} \subset \mathbb{R}^{n+1}$.

Aufgabe 3.1. Eine Funktion f ist genau dann konvex, wenn $\operatorname{dom} f$ und $\operatorname{epi} f$ konvexe Mengen sind.

3.11 Definition (Niveaumenge)**.** Die *α-Niveaumenge* einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist die Menge $\{\, x \in \operatorname{dom} f \mid f(x) \leq \alpha \,\}$.

Aufgabe 3.2. Die Niveaumengen einer konvexen Funktion sind konvex. Die Umkehrung gilt nicht: Selbst wenn alle Niveaumengen einer Funktion konvexe Mengen sind, muss die Funktion nicht konvex sein. Dies führt später auf den Begriff der Quasikonvexität.

3.12 Lemma (Jensen'sche Ungleichung)**.** *Aus der Konvexitätsbedingung für zwei Punkte folgt allgemeiner die Jensen'sche Ungleichung für beliebige Konvexkombinationen: Sei f konvex; dann ist*

$$f(\sum_{i=1}^{k} \theta_i x_i) \leq \sum_{i=1}^{k} \theta_i f(x_i) \text{ oder}$$
$$f(\mathbb{E}\, x) \leq \mathbb{E}\, f(x).$$

3.2 Eigenschaften konvexer Funktionen

Konvexität ist eine sehr starke Eigenschaft, aus der weitere Eigenschaften folgen. Beispielsweise ist eine konvexe Funktion auf einem offenen Definitionsbereich stets stetig, überall richtungsdifferenzierbar und fast überall differenzierbar. Ist sie überall differenzierbar, so ist ihre Ableitung automatisch stetig. In diesem Abschnitt beweisen wir einige diese Eigenschaften.

3.13 Definition (Richtungsableitung)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist im Punkt x *richtungsdifferenzierbar* in Richtung p, wenn der Grenzwert $\lim_{t \searrow 0} [f(x+tp) - f(x)]/t$ existiert und endlich ist. Der Grenzwert heißt dann *Richtungsableitung* von f im Punkt x in Richtung p.

3.14 Satz (Richtungsableitungen konvexer Funktionen)**.** *Sei f auf einer offenen konvexen Menge $C \subset \mathbb{R}^n$ definiert und konvex. Dann existiert in jedem $x \in C$ die Richtungsableitung $f'(x;p)$ für alle Richtungen $p \in \mathbb{R}^n$.*

Beweis. Wir betrachten ein Stück des von x ausgehenden Strahls in Richtung p und nehmen oBdA an, dass $x+p$ und $x-p$ noch in C enthalten sind (ansonsten verkürzen wir p zunächst entsprechend und nutzen die Homogenität aus; s.u.). Wir definieren $\phi(t) := [f(x+tp) - f(x)]/t$ für $0 < t \leq 1$ und zeigen die Existenz des Grenzwerts $\lim_{t \searrow 0} \phi(t)$. Dazu zeigen wir: (1) ϕ ist beschränkt; (2) ϕ ist monoton wachsend.

Sei $0 < t \leq 1$ beliebig. Wir schreiben x als Konvexkombination von $x + tp$ und $x - p$:

$$x = \frac{1}{1+t}(x+tp) + \frac{t}{1+t}(x-p).$$

Jetzt nutzen wir die Konvexität von f und folgern

$$f(x) \leq \frac{1}{1+t} f(x+tp) + \frac{t}{1+t} f(x-p),$$

was äquivalent zu $f(x) - f(x-p) \leq \phi(t)$ ist. Daher ist ϕ nach unten beschränkt.

Um die Monotonie in t zu zeigen, betrachten wir $0 < s \leq t \leq 1$ und schreiben $x + sp$ als Konvexkombination von $x + tp$ und x: Es ist $x + sp = \frac{s}{t}(x+tp) + \frac{t-s}{t}x$. Daher ist wiederum nach Ausnutzen der Konvexität von f

$$f(x+sp) - f(x) \leq \frac{s}{t}[f(x+tp) - f(x)],$$

was äquivalent zu $\phi(s) \leq \phi(t)$ ist; damit ist ϕ monoton. □

3.15 Definition (Gateaux-Variation in x)**.** Die Funktion $G_x : \mathbb{R}^n \to \mathbb{R}^n$, $p \mapsto f'(x;p)$, die jeder Richtung p die Richtungsableitung von f im Punkt x zuordnet, heißt *Gateaux-Variation* von f im Punkt x.

3.16 Lemma (Positive Homogenität und Subadditivität der Gateaux-Variation)**.** *Die Gateaux-Variation einer konvexen Funktion f auf einer offenen Menge C im Punkt x ist positiv homogen und subadditiv, d.h.*

$$\begin{aligned} f'(x;tp) &= t f'(x;p) && \text{für alle } t \geq 0,\ p \in \mathbb{R}^n \\ f'(x;p+q) &\leq f'(x;p) + f'(x;q) && \text{für alle } p, q \in \mathbb{R}^n \end{aligned}$$

Beweis. Elementares Nachrechnen; Aufgabe. □

Bekanntlich ist eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ *stetig*, wenn für alle Folgen $(x_n) \to x$ gilt, dass stets auch $f(x_n) \to f(x)$ gilt.

3.17 Satz. *Ist f auf einer offenen Menge C konvex, dann ist f dort stetig.*

Beweis. Es genügt jetzt zu zeigen, dass aus der Existenz aller Richtungsableitungen in einem Punkt bereits die Stetigkeit in diesem Punkt folgt; dies stellen wir als Aufgabe. Dann folgt der Satz aus Satz 3.14. □

Aufgabe 3.3. Existieren alle Richtungsableitungen in x, dann ist f in x stetig.

3.3 Konvexitätskriterien für differenzierbare Funktionen

Wir setzen nun voraus, dass f auf einer *offenen* konvexen Menge C definiert und differenzierbar ist, so dass inbesondere zu jedem Punkt in C eine ganze Umgebung existiert, auf der f definiert und differenzierbar ist. Die Ableitung ist aufgrund der Konvexität dann sogar stetig (was wir nicht beweisen).

Wir schreiben ∇f für die Funktion $\mathbb{R}^n \to \mathbb{R}^n$, die x_0 den Zeilenvektor der Ableitungen oder *Gradienten* $(\partial f/\partial x_1, \ldots, \partial f/\partial x_n)(x_0)$ zuordnet.

Für die Richtungsableitung in Richtung p gilt dann, dass sie als Skalarprodukt zwischen Gradient und p berechnet werden kann: Es ist

$$f'(x;p) = \langle \nabla f(x) | p \rangle \,.$$

Wir leiten nun Kriterien her, anhand derer man mit Hilfe des Gradienten einer differenzierbaren Funktion entscheiden kann, ob sie konvex ist.

3.18 Satz (Konvexitätskriterien erster Ordnung)**.** *Sei $C \subset \mathbb{R}^n$ konvex und offen und f auf C stetig differenzierbar. Dann gilt:*

1. *f ist auf C konvex, genau dann wenn für alle x_0 und x aus C gilt*

$$f(x) \geq f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle \,. \tag{3.1}$$

2. *f ist auf C strikt konvex, genau dann wenn für alle $x \neq x_0$ aus C gilt*

$$f(x) > f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle \,.$$

3. *f ist auf C stark konvex mit Konstante $c > 0$, genau dann wenn für alle x_0 und x aus C gilt*

$$f(x) \geq f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + c/2 \cdot \|x - x_0\|_2^2.$$

Beweis. Wir beweisen zunächst die Richtung $\Rightarrow$ der ersten Aussage. Sei also f auf C konvex und seien x, x_0 beliebig aus C. Aus der Konvexität von f folgt

$$\begin{aligned} f(\lambda x + (1-\lambda)x_0) &\leq \lambda f(x) + (1-\lambda) f(x_0) \\ \text{oder } f(\lambda x + (1-\lambda)x_0) - f(x_0) &\leq \lambda [f(x) - f(x_0)]. \end{aligned}$$

Wir dividieren durch λ und lassen $\lambda \searrow 0$ gehen. Damit geht die linke Seite gegen die Richtungsableitung $\langle \nabla f(x_0) | x - x_0 \rangle$ und die rechte Seite ist konstant $f(x) - f(x_0)$. Damit ist die Behauptung bewiesen.

Umgekehrt ($\Leftarrow$) seien x_1 und x_2 aus C; wir setzen $x_0 := \lambda x_1 + (1-\lambda)x_2$ für ein beliebiges $0 < \lambda < 1$. Nun ist nach Voraussetzung

$$\begin{aligned} f(x_1) &\geq f(x_0) + \langle \nabla f(x_0) | x_1 - x_0 \rangle \,, && | \cdot \lambda \\ f(x_2) &\geq f(x_0) + \langle \nabla f(x_0) | x_2 - x_0 \rangle \,. && | \cdot (1-\lambda) \\ \rightsquigarrow \quad \lambda f(x_1) + (1-\lambda) f(x_2) &\geq f(x_0) + \langle \nabla f(x_0) | \lambda x_1 + (1-\lambda)x_2 - x_0 \rangle \\ &= f(x_0) = f(\lambda x_1 + (1-\lambda)x_2). \end{aligned}$$

Damit ist die Konvexität von f auf C bewiesen.

Zum Beweis der Aussage über strikte Konvexität ($\Rightarrow$) sei f strikt konvex, seien $x \neq x_0$ in C, und sei $0 < \lambda < 1$. Dann gilt

$$\begin{aligned}\langle \nabla f(x_0) | \lambda(x - x_0) \rangle &\leq f(x_0 + \lambda(x - x_0)) - f(x_0) \\ &< \lambda[f(x) - f(x_0)].\end{aligned}$$

Zum Beweis der Rückrichtung ($\Leftarrow$) gehen wir vor wie bei (nichtstrikter) Konvexität, nutzen aber die strikten Ungleichungen.

Zum Beweis der Aussage über starke Konvexität wenden wir Teil 1 auf die Funktion $h := f - \frac{c}{2}\|\cdot\|_2^2$ an, mit $\nabla h(x) = \nabla f(x) - cx$. □

Aus (3.1) leiten wir die Darstellung von Tangentialhyperebenen her: Ist $(x, t) \in \operatorname{epi} f$, so ist $t \geq f(x) \geq f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle$, oder

$$\left\langle \begin{pmatrix} \nabla f(x_0) \\ -1 \end{pmatrix} \middle| \begin{pmatrix} x - x_0 \\ t - f(x_0) \end{pmatrix} \right\rangle \leq 0.$$

3.19 Definition (Tangentialhyperebene; unterstützende Hyperebene)**.** Dem Graphen (bzw. Epigraphen) einer differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist in jedem Punkt $(x_0, f(x_0)) \in \mathbb{R}^{n+1}$ des Definitionsbereichs die *Tangentialhyperebene* bzw. *unterstützende Hyperebene* $T_f(x_0) \subset \mathbb{R}^{n+1}$ zugeordnet; das ist die Hyperebene durch $(x_0, f(x_0))$, die auf $(\nabla f(x_0), -1)$ senkrecht steht, also $T_f(x_0) := \{\, (x, t) \mid \langle (\nabla f(x_0), -1) | (x - x_0, t - f(x_0)) \rangle = 0 \,\}$.

Wir merken uns: Die differenzierbare Funktion f ist konvex, wenn sie oberhalb ihrer Tangentialhyperebenen liegt. Sie ist strikt konvex, wenn jede Tangentialhyperebene sie nur in einem Punkt berührt. Sie ist stark konvex, wenn es zusätzlich konvexe quadratische Funktionen gibt, die f minorisieren.

In der Charakterisierung in Lemma 3.18 geht es um f, den Gradienten ∇f, und insbesondere um die Tangentialhyperebenen an f. Die folgende Charakterisierung der Konvexität benutzt nur den Gradienten ∇f, nicht mehr f selbst. Dazu jedoch erst eine Definition.

3.20 Definition (Monotonie)**.** Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$ heißt auf C

1. *monoton*, wenn für alle $x, x' \in C$ gilt $\langle F(x) - F(x') | x - x' \rangle \geq 0$;
2. *strikt monoton*, wenn für alle $x \neq x' \in C$ gilt $\langle F(x) - F(x') | x - x' \rangle > 0$;
3. *stark monoton mit Konstante $c > 0$*, wenn für alle $x, x' \in C$ gilt $\langle F(x) - F(x') | x - x' \rangle \geq c\|x - x'\|^2$.

3.21 Satz (Konvexität über Monotonie des Gradienten)**.** *Sei $C \subset \mathbb{R}^n$ konvex und offen und f auf C stetig differenzierbar. Dann gilt:*

1. *f ist auf C konvex, genau dann wenn ∇f auf C monoton ist.*
2. *f ist auf C strikt konvex, genau dann wenn ∇f auf C strikt monoton ist.*
3. *f ist auf C stark konvex mit Konstante $c > 0$, genau dann wenn ∇f auf C stark monoton mit Konstante $c > 0$ ist.*

Beweis. Wir beweisen die Aussage ($\Rightarrow$) für Konvexität und starke Konvexität gleichzeitig, indem wir $c \geq 0$ erlauben: Sei f auf C stark konvex mit Konstante $c \geq 0$. Dann bestehen nach Satz 3.18 für $x, x_0 \in C$ die Ungleichungen

$$f(x) \geq f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + \frac{c}{2}\|x - x_0\|_2^2,$$
$$f(x_0) \geq f(x) + \langle \nabla f(x) | x_0 - x \rangle + \frac{c}{2}\|x_0 - x\|_2^2.$$

Addition zeigt die (starke) Monotonie von ∇f. Ebenso verfahren wir für die strikte Konvexität. In der Tat ist Monotonie genau so definiert, dass die Ableitung konvexer Funktionen monoton ist. Interessanter ist daher die Umkehrung, die wir jetzt beweisen.

Zum Beweis der Rückrichtungen ($\Leftarrow$) betrachten wir zu gegebenen x_0, x_1 aus C die univariate Funktion ϕ, $t \mapsto f(x_t)$ mit $x_t := x_0 + t(x_1 - x_0)$, dabei komme t aus einem offenen Intervall, das $[0, 1]$ enthält. Es ist $\phi'(t) = \langle \nabla f(x_t) | x_1 - x_0 \rangle$. Daher ist nach Voraussetzung der (starken) Monotonie von ∇f für $0 < t \leq 1$ stets

$$\begin{aligned}\phi'(t) - \phi'(0) &= \langle \nabla f(x_t) - \nabla f(x_0) | x_1 - x_0 \rangle \\ &= \frac{1}{t} \langle \nabla f(x_t) - \nabla f(x_0) | x_t - x_0 \rangle \\ &\geq \frac{1}{t} c\|x_t - x_0\|^2 = tc\|x_1 - x_0\|^2.\end{aligned}$$

Mittels Integraldarstellung berechnen wir

$$\begin{aligned}f(x_1) - f(x_0) - \langle \nabla f(x_0) | x_1 - x_0 \rangle &= \phi(1) - \phi(0) - \phi'(0) \\ &= \int_0^1 [\phi'(t) - \phi'(0)] \, \mathrm{d}t \\ &\geq c\|x_1 - x_0\|^2 \int_0^1 t \, \mathrm{d}t \\ &= \frac{c}{2}\|x_1 - x_0\|^2;\end{aligned}$$

damit ist f (stark) konvex. Für strikte Konvexität geht man genauso vor. □

3.22 Beispiel (Exponentialfunktion)**.** Die Exponentialfunktion $f(x) = \exp(x)$ erfüllt $f'(x) = f''(x) = \exp(x) > 0$; damit ist $(f'(x_1) - f'(x_0))(x_1 - x_0) = \frac{f'(x_1) - f'(x_0)}{x_1 - x_0} \cdot (x_1 - x_0)^2 \to f''(x_0) \cdot (x_1 - x_0)^2 > 0$ für $x_1 \to x_0$. Das heißt, die Exponentialfunktion ist strikt monoton, aber da $\exp(x)$ für $x \to -\infty$ gegen Null geht, gibt es kein $c > 0$, für das die Exponentialfunktion stark konvex ist. Auf jedem kompakten Intervall $[a, b]$ ist $\exp(x)$ aber stark konvex mit Konstante $c = \exp(a)$. ♡

Wir kommen nun zu den meist leichter anwendbaren Kriterien zweiter Ordnung, wobei wir voraussetzen dass f nun mindestens zwei mal stetig differenzierbar ist.

Wie üblich sei $\nabla^2 f(x) := \left(\frac{\partial f}{\partial x_i \partial x_j}\right)(x)_{i,j=1,\ldots,n}$ die *Hesse-Matrix* der zweiten Ableitungen zu $f : \mathbb{R}^n \to \mathbb{R}$. Zur Abkürzung schreiben wir H_0 für die Hesse-Matrix in einem beliebigen Punkt x_0. Die Hesse-Matrix ist symmetrisch; ihren kleinsten Eigenwert im Punkt x_0 bezeichnen wir mit $\lambda_{\min}(H_0)$.

3.23 Satz (Konvexitätskriterien zweiter Ordnung)**.** *Sei $\Omega \subset \mathbb{R}^n$ offen; sei $C \subset \Omega$ konvex und f auf Ω zweimal stetig differenzierbar. Zu $x_0 \in C$ sei $H_0 := \nabla^2 f(x_0)$ die Hesse-Matrix im Punkt x_0. Dann gilt:*

1. *Ist H_0 positiv semidefinit für alle $x_0 \in C$, dann ist f ist auf C konvex.*
2. *Ist H_0 sogar positiv definit für alle $x_0 \in C$, dann ist f strikt konvex.*
3. *Ist $\lambda_{\min}(H_0) \geq c > 0$ für alle $x_0 \in C$, dann ist f auf C stark konvex mit Konstante $c > 0$.*
4. *Ist die konvexe Menge C offen, so gelten in 1. und 3. (aber nicht in 2.) auch die Umkehrungen.*

Beweis. Die Aussagen 1 und 3 werden zusammen bewiesen (mit $c = 0$ für einfache Konvexität). Sei also $c = \lambda_{\min}(H_0) \geq 0$, und seien x, y zwei beliebige Punkte aus C. Wir definieren $\phi(t) := f(x + t(y - x))$ für $t \in [0, 1]$, die Einschränkung von f auf die Strecke zwischen x und y, die natürlich in C liegt. Aufgrund des Taylor-Satzes ist $\phi(1) = \phi(0) + \phi'(0) + \frac{1}{2}\phi''(t)$ mit einem $t \in (0, 1)$. Angewendet auf f heißt das

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x) | y - x \rangle &= \frac{1}{2} \langle y - x | \nabla^2 f(x + t(y - x)) | y - x \rangle \\ &\geq \frac{c}{2} \|y - x\|_2^2, \end{aligned}$$

da alle Eigenwerte der Hessematrix in C nicht kleiner als c sind. Zu Aussage 2 argumentiert man genauso und stellt fest, dass für $y \neq x$ und der strikten Positivität der Eigenwerte die Abschätzung in der letzten Zeile echt größer Null lautet.

Zu den Umkehrungen zu 1. und 3. benutzen wir Satz 3.21: Ist f (stark) konvex mit Konstante $c > 0$ und C offen, dann ist ∇f (stark) monoton, also $\langle \nabla f(x + tp) - \nabla f(x) | tp \rangle \geq c\|tp\|_2^2$ für beliebige x und p und alle hinreichend kleinen $|t|$. Ist H die Hesse-Matrix in x, dann ist

$$\langle p | H | p \rangle = \lim_{t \to 0} \frac{\langle \nabla f(x + tp) - \nabla f(x) | p \rangle}{t} = \lim_{t \to 0} \frac{\langle \nabla f(x + tp) - \nabla f(x) | tp \rangle}{t^2} \geq c\|p\|_2^2$$

für alle p und alle x, so dass der kleinste Eigenwert von H in ganz C mindestens c beträgt. □

Aufgabe 3.4. Betrachte auf $\mathbb{R}$ die Funktion $f : x \mapsto x^4$; es ist $f''(0) = 0$; trotzdem ist f strikt konvex. Dies zeigt, dass die Umkehrung von 2. im obigen Satz nicht gilt. Betrachte weiter $f : x \mapsto 1/x^2$ auf $\mathbb{R} \setminus \{0\}$; hier ist stets $f''(x) > 0$; trotzdem ist f nicht konvex. Wieso ist dies kein Widerspruch zum obigen Satz?

Aufgabe 3.5 (Rosenbrock-Funktion)**.** Die Rosenbrock-Funktion f ist auf ganz $\mathbb{R}^2$ definiert durch $f(x, y) := 100(y - x^2)^2 + (1 - x)^2$. Ist sie auf ganz $\mathbb{R}^2$ konvex? Ist sie in einer Umgebung von $(1, 1)$ konvex? Versuche möglichst genau zu charakterisieren, wo f konvex ist.

3.4 Beispiele für konvexe Funktionen

3.24 Beispiel (Beispiele auf $\mathbb{R}$)**.** Die Konvexität der folgenden Beispiele weist man leicht mit Hilfe der zweiten Ableitung nach.

- Affine (konstante, lineare) Funktionen sind konvex und konkav, da ihre zweite Ableitung auf $\mathbb{R}$ verschwindet.
- Quadratische Funktionen $x \mapsto ax^2 + bx + c$ sind konvex auf $\mathbb{R}$ genau dann, wenn $a \geq 0$, und stark konvex mit Konstante a, wenn $a > 0$.
- Potenzen $x \mapsto x^a$ auf $\mathbb{R}_{++}$ sind konvex, wenn $a \leq 0$ oder $a \geq 1$.
- $x \mapsto -\log x$ ist konvex auf $\mathbb{R}_{++}$.
- Die negative Entropie $x \mapsto x \log x$ mit der stetigen Fortsetzung $0 \log 0 := 0$ auf $\mathbb{R}_+$ ist konvex.
- Potenzen von Absolutwerten auf $\mathbb{R}$, $x \mapsto |x|^p$, sind für $p \geq 1$ konvex. Zum exakten Nachweis mache man Fallunterscheidungen zwischen positiven und negativen Werten von x.

♡

Aufgabe 3.6. Als Anwendung der Konvexität von $-\log x$ und der Jensen'schen Ungleichung lassen sich weitere nützliche Ungleichungen beweisen.

1. Für $a, b \geq 0$ und $0 \leq \theta \leq 1$ ist $a^\theta b^{1-\theta} \leq \theta a + (1-\theta)b$. Mit $\theta = 1/2$ bedeutet dies $\sqrt{ab} \leq \frac{a+b}{2}$.
2. Die Hölder'sche Ungleichung für $x, y \in \mathbb{R}^n$ besagt $\langle x|y\rangle \leq \|x\|_p \cdot \|y\|_q$ mit $p > 1$ und $1/p + 1/q = 1$. Ausgeschrieben lautet sie

$$\sum_{i=1}^n x_i y_i \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \cdot \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}.$$

Wir kommen zu Beispielen für konvexe Funktionen im Mehrdimensionalen. Das Beispiel der max-Funktion wurde bereits diskutiert.

3.25 Beispiel (Normen)**.** Jede Norm ist eine konvexe Funktion ihrer Argumente. Dies folgt aus der Dreiecksungleichung. ♡

3.26 Beispiel (Quadrat durch linear)**.** Wir betrachten $f : \mathbb{R}^2 \to \mathbb{R}$ mit $\operatorname{dom} f = \mathbb{R} \times \mathbb{R}_{++}$, definiert durch $(x, y) \mapsto x^2/y$. Die Hesse-Matrix lässt sich schreiben als $\frac{2}{y^3}\begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix}$. Diese ist positiv semidefinit, da sie sich als äußeres Produkt schreiben lässt.

Als Verallgemeinerung lässt sich beweisen: $f : \mathbb{R}^n \times \mathbb{S}^n \to \mathbb{R}$ mit $\operatorname{dom} f = \mathbb{R}^n \times \mathbb{S}^n_{++}$, definiert durch $f(x, Y) := \langle x|Y^{-1}|x\rangle$ ist konvex. ♡

3.27 Beispiel (Log-sum-exp)**.** Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) := \log \sum_{i=1}^n \exp(x_i)$ ist konvex. Sie kann als eine glatte (analytische) Approximation der nicht differenzierbaren max-Funktion aufgefasst werden: Es ist $\max_i x_i \leq f(x) \leq \max_i x_i + \log n$.

Wir berechnen die Hesse-Matrix H: Sei $z := \exp(x)$ komponentenweise. Es ist (Übung!)

$$H = \frac{1}{\langle 1|z\rangle^2} \cdot [\langle 1|z\rangle \operatorname{diag}(z) - |z\rangle \langle z|].$$

Zu zeigen ist, dass $\langle v|H|v\rangle \geq 0$ für alle Vektoren v, oder (ohne den Vorfaktor)

$$\langle 1|z\rangle \langle v|v\rangle \operatorname{diag}(z) - \langle v|z\rangle \langle z|v\rangle = \left(\sum_i z_i\right)\left(\sum_i v_i^2 z_i\right) - \left(\sum_i v_i z_i\right)^2 \geq 0.$$

Dies folgt aber sofort aus der Cauchy-Schwarz'schen Ungleichung $\langle a|b\rangle^2 \leq \langle a|b\rangle \langle b|b\rangle$ mit $a = v\sqrt{z}$ und $b = \sqrt{z}$, jeweils komponentenweise. ♡

3.28 Beispiel (Geometrisches Mittel ist konkav)**.** Es sei $A(x) := \frac{1}{n}\sum_i x_i$ das *arithmetische Mittel* und $G(x) := \sqrt[n]{\prod_i x_i}$ das *geometrische Mittel* für $x \in \mathbb{R}^n_+$. Aus der Jensen'schen Ungleichung und der Konvexität von $-\log x$ folgt bereits, dass stets $G(x) \leq A(x)$ gilt. Die Gleichheit gilt genau dann, wenn alle Komponenten von x gleich sind.

Wir werden gleich zeigen, dass die Funktion $G(x)$ konkav in x ist. Als Anwendung beweisen wir: Sei $\alpha \in (0,1)$ und sei $S_\alpha := \{\, x \mid G(x) \geq \alpha A(x) \,\}$. Diese Menge ist konvex, denn es ist $S_\alpha = \{\, x \mid \alpha A(x) - G(x) \leq 0 \,\}$ die 0-Niveaumenge der konvexen Funktion $\alpha A(x) - G(x)$. Da mit x auch λx für $\lambda \geq 0$ in S_α enthalten ist, ist sie sogar ein konvexer Kegel.

Zum Beweis, dass $G(x)$ konkav ist, berechnen wir wieder die Hesse-Matrix (Übung) und weisen mit der Cauchy-Schwarz'schen Ungleichung für geeignete Vektoren nach, dass sie positiv semidefinit ist. ♡

3.29 Beispiel (logdet ist konkav)**.** Wir betrachten $f : \mathbb{S}^n \to \mathbb{R}$ auf $\mathbb{S}^n_{++}$, definiert durch $f(X) := \log\det X$. Die Konkavität von f weisen wir nach, indem wir zeigen, dass f auf jeder Strecke innerhalb des Definitionsbereichs konkav ist; dazu schreiben wir $X_t = Z + tV$ mit $Z, V \in \mathbb{S}^n$, so dass $X_0 = Z$ für $t = 0$ in $\mathbb{S}_{++}$ liegt, und definieren $g(t) := f(X_t) = f(Z + tV)$. Wir weisen die Konkavität von g nach. Es ist

$$\begin{aligned} g(t) &= \log\det(Z + tV) \\ &= \log\det(Z^{1/2}(I + tZ^{-1/2}VZ^{-1/2})Z^{1/2}) \\ &= \sum_{i=1}^n \log(1 + t\lambda_i) + \log\det Z, \end{aligned}$$

wobei die λ_i die Eigenwerte von $Z^{-1/2}VZ^{1/2}$ sind. Damit ist $g'(t) = \sum_i \frac{\lambda_i}{1+t\lambda_i}$ und $g''(t) = -\sum_i \frac{\lambda_i^2}{(1+t\lambda_i)^2} \leq 0$, also ist f konkav. ♡

3.5 Konvexitätserhaltende Operationen

Wir untersuchen, unter welchen Operationen die Konvexität von Funktionen erhalten bleibt. Fundamental ist der folgende Satz.

3.30 Satz. *Die Menge der konvexen Funktionen auf einer konvexen Menge C bildet einen konvexen Kegel, d.h., mit f, g sind auch αf für $\alpha \geq 0$ und $f + g$ konvex.*

Allgemeiner sei $f(x, y)$ für jedes y aus einer Indexmenge $\mathcal{Y}$ konvex in x, und es sei $w(y) \geq 0$. Dann ist auch $\int_{\mathcal{Y}} w(y) f(x, y)\,\mathrm{d}y$ konvex.

Beweis. Übungsaufgabe (z.B. auch über Epigraph). □

3.31 Satz (Komposition mit affinen Abbildungen)**.** *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex auf $\operatorname{dom} f$. Sei $A \in \mathbb{R}^{n\times m}$ und $b \in \mathbb{R}^n$. Sei $g : \mathbb{R}^m \to \mathbb{R}$ definiert durch $g(x) := f(Ax + b)$, wo immer $Ax + b \in \operatorname{dom} f$ ist. Ist f konvex, dann auch g. Ist f konkav, dann auch g.*

Beweis. Nachrechnen: Sei f konvex. Es ist

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= f(A(\lambda x + (1-\lambda)y) + b) \\ &= f(\lambda(Ax+b) + (1-\lambda)(Ay+b)) \\ &\le \lambda f(Ax+b) + (1-\lambda) f(Ay+b) \\ &= \lambda g(x) + (1-\lambda) g(y). \end{aligned}$$

Für Konkavität genauso. □

3.32 Satz (Punktweise Maxima und Suprema)**.** *Sind f_1, f_2 konvex auf C, dann auch $g := \max f_1, f_2$. Allgemeiner: Ist f_i konvex für alle $i \in I$ (beliebige Indexmenge), dann auch $g := \sup_{i\in I} f_i$.*

Beweis. Es ist $\operatorname{epi} g = \cap_{i\in I} \operatorname{epi} f_i$ als Schnitt konvexer Mengen konvex. □

3.33 Beispiel (Distanz zum weitest entfernten Punkt einer Menge)**.** Sei $S \subset \mathbb{R}^n$ eine beliebige Menge. Definiere $f(x)$ für $x \in \mathbb{R}^n$ als Abstand von x zum am weitesten entfernten Punkt in S (sofern ein solcher existiert), also $f(x) := \sup_{y\in S} \|x - y\|$. Dann ist f konvex, denn $\|x - y\|$ ist konvex in x für jedes y. ♡

Aufgabe 3.7. Sei $X \in \mathbb{S}^n$ eine symmetrische Matrix und $\lambda_{\max}(X)$ der größte Eigenwert von X. Zeige, dass $\lambda_{\max} : \mathbb{S}^n \to \mathbb{R}$ konvex ist.

Aufgabe 3.8. Zu $x \in \mathbb{R}^n$ sei $x_{[i]}$ die i-t größte Komponente, so dass $x_{[1]} \ge x_{[2]} \ge \dots$. Sei $r \in \{0, 1, \dots, n\}$ und sei $f_r : \mathbb{R}^n \to \mathbb{R}$, die Funktion, die x die Summe seiner r größten Einträge zuordnet, also $f(x) = \sum_{i=1}^{r} x_{[i]}$. Dann ist f konvex.

Kompositionen. Wir betrachten jetzt Bedingungen, unter denen die Komposition zweier Funktionen konvex ist. Es seien $g : \mathbb{R}^n \to \mathbb{R}^k$ und $h : \mathbb{R}^k \to \mathbb{R}$ Funktionen und $f : \mathbb{R}^n \to \mathbb{R}$ die Komposition $f = h \circ g$; $f(x) := h(g(x))$ mit $\operatorname{dom} f = \{ x \in \operatorname{dom} g \mid g(x) \in \operatorname{dom} h \}$.

Es seien $\tilde{g}, \tilde{h}$ die Funktionen g, h mit erweitertem Wertebereich $\pm\infty$. (Erinnerung: Eine konvexe Funktion setzen wir für nicht definierte Werte auf $+\infty$; eine konkave Funktion setzen wir für nicht definierte Werte auf $-\infty$.) Zunächst sei $k = 1$; danach betrachten wir allgemein $k \ge 1$.

3.34 Satz (Komposition $f(x) = h(g(x))$ mit reellwertiger Funktion $g : \mathbb{R}^n \to \mathbb{R}$)**.** *Sei eine der folgenden Bedingungen erfüllt:*

1. *g konvex, h konvex, und $\tilde{h}$ wachsend;*
2. *g konkav, h konvex, und $\tilde{h}$ fallend.*

Dann ist $f = h \circ g$ konvex.

Sei eine der folgenden Bedingungen erfüllt:

1. *g konkav, h konkav, und $\tilde{h}$ wachsend;*

2. *g konvex, h konkav, und $\tilde{h}$ fallend.*

Dann ist $f = h \circ g$ konkav.

Beweis. Wir zeigen nur den ersten Fall (f konvex); die anderen Fälle sind ähnlich.

Seien $x, y \in \operatorname{dom} f$, also insbesondere $x, y \in \operatorname{dom} g$ und $g(x), g(y) \in \operatorname{dom} h$. Da g konvex ist, ist

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y).$$

Da h konvex ist, ist die rechte Seite in $\operatorname{dom} h$. Da $\tilde{h}$ wachsend ist, muss auch die linke Seite in $\operatorname{dom} h$ liegen; damit liegt $tx + (1-t)y$ zumindest in $\operatorname{dom} f$ und $\operatorname{dom} f$ ist konvex.

Da h wachsend ist, können wir h auf obige Ungleichung anwenden und dann noch die Konvexität von h nutzen:

$$h(g(tx + (1-t)y)) \leq h(tg(x) + (1-t)g(y)) \leq th(g(x)) + (1-t)h(g(y));$$

das ist genau die gesuchte Konvexitätsaussage zu f. □

3.35 Beispiel (Komposition mit skalaren Funktionen)**.**

- Ist g konvex, dann auch $\exp g$, denn exp ist konvex und wachsend.
- Ist g konkav und positiv, dann ist $\log g$ konkav, denn log ist konkav und wachsend.
- Ist g konkav und positiv, dann ist $1/g$ konvex, denn $1/\cdot$ ist konvex und fallend.
- Ist g konvex und positiv und $p \geq 1$, dann ist g^p konvex.
- Ist g konvex, dann ist $-\log(-g)$ auf $\{\, x \mid g(x) < 0 \,\}$ konvex.

♡

Wir betrachten jetzt den allgemeineren Fall, dass $g : \mathbb{R}^n \to \mathbb{R}^k$ mit $k \geq 1$ abbildet und $h : \mathbb{R}^k \to \mathbb{R}$. Wir schreiben $g = (g_1, \ldots, g_k)$ für die k Komponenten von g.

3.36 Satz (Komposition $f(x) = h(g(x))$ mit vektorwertiger Funktion $g : \mathbb{R}^n \to \mathbb{R}^k$)**.** *Sei eine der folgenden Bedingungen erfüllt:*

1. *Alle g_i konvex, h konvex, und $\tilde{h}$ wachsend in jedem Argument;*
2. *Alle g_i konkav, h konvex, und $\tilde{h}$ fallend in jedem Argument.*

Dann ist $f = h \circ g$ konvex.

Sei eine der folgenden Bedingungen erfüllt:

1. *Alle g_i konkav, h konkav, und $\tilde{h}$ wachsend in jedem Argument;*
2. *Alle g_i konvex, h konkav, und $\tilde{h}$ fallend in jedem Argument.*

Dann ist $f = h \circ g$ konkav.

Beweis. Betrachte die einzelnen Komponenten von g wie in Satz 3.34. □

3.37 Beispiel (log-sum-exp)**.** Sind die Funktionen g_i konvex, dann auch $\log \sum_{i=1}^k \exp(g_i)$, denn $h(z) = \log \sum_{i=1}^k \exp(z_i)$ ist konvex und wachsend in jedem Argument. ♡

Minimierung. Das punktweise Maximum bzw. Supremum einer Familie von konvexen Funktionen haben wir bereits (trivial) als konvex nachgewiesen. Interessanter ist der Fall der Minimierung über einzelne Koordinaten.

3.38 Satz. *Sei f konvex in $(x, y) \in C \subset \mathbb{R}^n \times \mathbb{R}^m$, und sei $Q \subset \mathbb{R}^m$ konvex und nicht leer. Dann ist die Funktion*

$$g(x) := \inf_{y \in Q} f(x, y)$$

auf $\operatorname{dom} g = \{\, x \mid \exists\, y \in Q : (x, y) \in \operatorname{dom} f \,\}$ *konvex, sofern überhaupt ein x mit $g(x) > -\infty$ existiert.*

Beweis. Seien $x_1, x_2 \in \operatorname{dom} g$ und sei $\varepsilon > 0$. Es existieren $y_1, y_2 \in Q$ so dass $f(x_i, y_i) \leq g(x_i) + \varepsilon$ für $i = 1, 2$. Sei $t \in [0, 1]$; wir betrachten g an der Konvexkombination:

$$\begin{aligned} g(tx_1 + (1-t)x_2) &= \inf_{y \in Q} f(tx_1 + (1-t)x_2, y) \\ &\leq f(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \\ &\leq tf(x_1, y_1) + (1-t)f(x_2, y_2) \\ &\leq tg(x_1) + (1-t)g(x_2) + \varepsilon. \end{aligned}$$

Da die Ungleichung für beliebig kleine $\varepsilon > 0$ gilt, gilt sie auch für $\varepsilon = 0$; damit ist g als konvex nachgewiesen. Sofern es überhaupt ein x mit $g(x) > -\infty$ gibt, ist $g(x) > -\infty$ für alle x und damit sinnvoll definiert. □

Aufgabe 3.9. Gib einen alternativen Beweis für Satz 3.38 mit Hilfe der Epigraphen von f und g.

3.39 Beispiel (Distanz eines Punktes zu einer konvexen Menge)**.** Sei $C \subset \mathbb{R}^n$ konvex und $x \in \mathbb{R}^n$. Der Abstand von x zur Menge C ist definiert als

$$d(x, C) := \inf_{y \in C} \|x - y\|$$

(für eine beliebige Norm). Da $\|x - y\|$ konvex in (x, y) ist, ist $d(x, C)$ konvex in x.

Achtung: Im Gegensatz zur maximalen Distanz (Beispiel 3.33) wird hier ausgenutzt, dass die Menge C konvex ist (sonst stimmt es nicht!) und dass $\|x - y\|$ nicht nur in x konvex ist (für festes y), sondern sogar in (x, y) konvex ist. Dies sollte man noch einmal zeigen. ♡

3.40 Beispiel. Sei h konvex. Definiere $g(x) = \inf \{\, h(y) \mid Ay = x \,\}$. Dann ist g konvex. Wir definieren dazu $f(x, y) := h(y)$, wenn $Ay = x$, und $f(x, y) := 0$ andernfalls, und zeigen dass f in (x, y) konvex ist. Damit ist $g(x) := \inf_y f(x, y)$ und daher konvex. ♡

Perspektive. Jeder Funktion $f : \mathbb{R}^n \to \mathbb{R}$ lässt eine andere Funktion $\mathbb{R}^{n+1} \to \mathbb{R}$, die Perspektive von f, zuordnen.

3.41 Definition (Perspektive)**.** Sei $f : \mathbb{R}^n \to \mathbb{R}$. Die *Perspektive* von f ist die Funktion $g : \mathbb{R}^{n+1} \to \mathbb{R}$, $g(x, t) := tf(x/t)$ mit $\operatorname{dom} g = \{\, (x, t) \mid t > 0 \text{ und } x/t \in \operatorname{dom} f \,\}$.

3.42 Lemma (Perspektive erhält Konvexität)**.** *Ist f konvex (konkav), dann auch ihre Perspektive.*

Beweis. Der Beweis wird über die Epigraphen von f und ihrer Perspektive g geführt. Sei $x \in \mathbb{R}^n$, $t > 0$ und $s \in \mathbb{R}$. Es ist

$$(x,t,s) \in \operatorname{epi} g \iff tf(x/t) \le s \iff f(x/t) \le s/t \iff (x/t, s/t) \in \operatorname{epi} f.$$

Ist P die Abbildung $(u,v,w) \mapsto (u,w)/v$, dann ist $\operatorname{epi} g = P^{-1}(\operatorname{epi} f)$ das Urbild von $\operatorname{epi} f$. Nun ist P aber (im wesentlichen, bis auf Koordinatenpermutation) die Perspektivfunktion aus Definition 2.24, die wir bereits als konvexitäterhaltend für Urbilder erkannt haben. Da nach Voraussetzug $\operatorname{epi} f$ konvex ist, ist auch $\operatorname{epi} g$ konvex und damit g. □

3.43 Beispiel (Anwendungen der Perspektive)**.**

- Die Funktion $g(x,t) := \langle x|x\rangle / t$ ist konvex in (x,t) für $t > 0$, denn sie ist die Perspektive von $f(x) = \langle x|x\rangle = \|x\|_2^2$.
- Die *relative Entropie* zwischen $t > 0$ und $x > 0$, $g(x,t) = t\log(t/x)$ ist konvex auf $\mathbb{R}^2_{++}$, denn sie ist die Perspektive von $f(x) = -\log x$.
- Sei $f : \mathbb{R}^m \to \mathbb{R}$ konvex, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, und $d \in \mathbb{R}$. Wir definieren $g : \mathbb{R}^n \to \mathbb{R}$ durch

  $$g(x) := (\langle c|x\rangle + d) \cdot f((Ax+b)/(\langle c|x\rangle + d))$$

  mit $\operatorname{dom} g = \{\, x \mid \langle c|x\rangle + d > 0,\ (Ax+b)/(\langle c|x\rangle + d) \in \operatorname{dom} f \,\}$. Dann ist g konvex.

♡

3.44 Beispiel (Relative Entropie und Kullback-Leibler-Divergenz)**.** Die *relative Entropie* zwischen zwei Vektoren $u, v \in \mathbb{R}^n_{++}$,

$$H(u\|v) := \sum_{i=1}^{n} u_i \log(u_i/v_i),$$

ist konvex in (u,v), da sie die Summe der relativen Entropien der Komponenten ist.

Die *Kullback-Leibler-Divergenz* zwischen $u, v \in \mathbb{R}^n_{++}$,

$$D(u,v) := \sum_{i=1}^{n} [u_i \log(u_i/v_i) - u_i + v_i]$$

ist konvex in (u,v), denn sie ist die relative Entropie plus lineare Funktionen. ♡

Aufgabe 3.10. Zeige: Es ist stets $D(u,v) \ge 0$ und $D(u,v) = 0$ genau dann, wenn $u = v$.

KAPITEL 4

Konvexe Optimierungsprobleme

In diesem Kapitel geben wir eine Einführung in die Terminologie bei Optimierungsproblemen, zunächst ganz allgemein (ohne Konvexität vorauszusetzen) und studieren dann insbesondere konvexe Optimierungsprobleme, bei denen sowohl die Zielfunktion als auch die Nebenbedingungen konvex sein müssen. Wir untersuchen dann noch kurz die spezielleren Problemstellungen der quadratischen und der linearen Optimierung.

Statt "Problem" oder "Optimierungsproblem" spricht man auch häufig von "Programm" (linear, quadratic program).

4.1 Optimierungsprobleme

4.1 Definition (allgemeines Optimierungsproblem in Standard-Form)**.** Ein *allgemeines Optimierungsproblem in Standard-Form* hat die Form

$$\begin{aligned} &\text{Minimiere } f_0(x) \\ &\quad \text{so dass } f_i(x) \leq 0 && \text{für } i = 1, \ldots, m, \\ &\qquad\quad h_i(x) = 0 && \text{für } i = 1, \ldots, p. \end{aligned}$$

Dabei sind die Elemente von $x \in \mathbb{R}^n$ die *Variablen* (oder *Optimierungsvariablen*), $f_0 : \mathbb{R}^n \to \mathbb{R}$ ist die *Zielfunktion* (oder *Kostenfunktion*). die $f_i : \mathbb{R}^n \to \mathbb{R}$ bilden die m *Ungleichungs-Restriktions-Funktionen* zu den m *Ungleichungs-Restriktionen* $f_i(x) \leq 0$, und die $h_i : \mathbb{R}^n \to \mathbb{R}$ bilden die p *Gleichungs-Restriktions-Funktionen* zu den p *Gleichungs-Restriktionen* $h_i(x) \leq 0$. Ein Problem mit $m = p = 0$ heißt *unrestringiert*.

Der *Definitionsbereich* (die *Definitionsmenge*) des Problems ist $\mathcal{D} = \operatorname{dom} f_0 \cap \bigcap_{i=1}^m \operatorname{dom} f_i \cap \bigcap_{i=1}^p \operatorname{dom} h_i$, also der Schnitt der Definitionsbereiche aller auftretenden Funktionen.

Der *zulässige Bereich* (die *zulässige Menge*) ist die Menge der Punkte, für die das Problem definiert ist und die alle Restriktionen erfüllen, also

$$\mathcal{F} = \{\, x \in \mathcal{D} \mid f_i(x) \leq 0 \text{ für } i = 0, \ldots, m \text{ und } h_i(x) = 0 \text{ für } i = 0, \ldots, p \,\}.$$

Das Problem heißt *zulässig* oder *erfüllbar*, wenn $\mathcal{F}$ nicht leer ist.

Die Restriktion $f_i(x) \leq 0$ heißt *aktiv* in $x \in \mathcal{F}$, wenn $f_i(x) = 0$ und *inaktiv*, wenn $f_i(x) < 0$.

Einige Bemerkungen dazu: Nahezu jedes praktisch vorkommende Optimierungsproblem (mit einer einzelnen Zielfunktion) lässt sich in die hier definierte Standard-Form bringen.

- Maximierungs-Probleme werden durch negieren der Zielfunktion zu Minimierungsproblemen.
- Größer-gleich-Ungleichungen werden durch Negieren zu Kleiner-gleich-Ungleichungen.
- Steht auf der rechten Seite von Gleichungen keine Null, so zieht man die rechte Seite ab.
- Man könnte auf die Gleichungen $h_i(x) = 0$ verzichten und sie durch zwei Ungleichungen $h_i(x) \leq 0$ und $-h_i(x) \leq 0$ ersetzen; es ist jedoch aus algorithmischer Sicht sinnvoll, Gleichungen gesondert zu behandeln.
- Man kann im allgemeinen auf Ungleichungen nicht verzichten, aber man kann sie beschränken auf Ungleichungen der Form $x_i \geq 0$. Dazu führt man für jede Ungleichung $f_i(x) \leq 0$ eine neue Variable s_i ein und ersetzt die Ungleichung durch das Gleichungs-, Ungleichungspaar $f_i(x) + s_i = 0$ und $s_i \geq 0$. Es ergibt sich ein Problem in den Variablen (x, s). Die s-Variablen nennt man auch *Schlupfvariablen*. Diese Formulierung ist insbesondere in der linearen Optimierung von Interesse.

Die Standard-Form erlaubt keine strikten Ungleichungen (außer implizit über den Definitionsbereich $\mathcal{D}$). Diese sind aber in der Praxis eher selten; wie soll man z.B. in der Praxis bei einer Ungleichung wie $x > 0$ ein beliebig kleines $x > 0$ von $x = 0$ unterscheiden? Hier ist oft eine Ungleichung der Form $x \geq c$ (oder $c - x \leq 0$ in Standard-Form) für ein geeignetes $c > 0$ angemessener.

4.2 Definition (Optimalität und Lösbarkeit)**.** Der *optimale Wert* (auch: global optimaler Wert) des Problems aus Defintion 4.1 ist $p^* := \inf \{\, f_0(x) \mid x \in \mathcal{F} \,\}$, der auch die Werte $\pm\infty$ annehmen kann.

Ist $p^* = -\infty$, dann findet sich zu jeder Lösung eine bessere und wir sagen, das Problem ist (nach unten) *unbeschränkt*. Es ist $p^* = +\infty$ genau dann, wenn $\mathcal{F}$ leer und das Problem nicht erfüllbar ist.

Ein $x^* \in \mathcal{F}$ mit $f_0(x^*) = p^*$ heißt *optimaler Punkt* oder *global optimaler Punkt*. Die Menge aller optimalen Punkte ist die *optimale Menge* $X^* = \{\, x \in \mathcal{F} \mid f_0(x) = p^* \,\}$. Ist X^* nicht leer, dann sagen wir, dass der optimale Wert *erreicht* wird und dass das Problem *lösbar* ist.

Ein $x \in \mathcal{F}$ mit $f_0(x) \leq p^* + \varepsilon$ heißt *ε-suboptimaler Punkt*. Die Menge aller ε-suboptimalen Punkte ist die *ε-suboptimale Menge*.

Dazu wieder einige Bemerkungen:

- Ist $p^* \in \{\, \pm\infty \,\}$, so kann es per Definiton nicht erreicht werden.

- Auch wenn p^* endlich ist, muss es nicht erreicht werden, beispielsweise wenn $\mathcal{F}$ unbeschränkt (oder offen) ist. Man denke an $f_0(x) = 1/x$ mit $x \geq 1$. Das Infimum $p^* = 0$ wird nicht angenommen.
- In allen oben genannten Fällen ist das Problem *nicht lösbar* in dem Sinne, dass sich keine Variablen x angeben lassen, für die $f_0(x) = p^*$. Die Gründe für die Nichtlösbarkeit (unbeschränkt oder nicht erfüllbar oder Optimum endlich aber nicht angenommen) sind jedoch grundverschieden.

Aufgabe 4.1. Untersuche die folgenden unrestringierten Probleme auf $\mathcal{D} = \mathbb{R}_{++}$ auf Lösbarkeit.

1. $f_0(x) = 1/x^2$
2. $f_0(x) = \log x$
3. $f_0(x) = x \log x$

Bei allgemeinen Funktionen ist die Suche nach optimalen Punkten in der Regel sehr schwierig, da sich das Verhalten der Funktion (fast) nicht vorhersagen lässt: Woher will man wissen, wie sich die Funktion in weit entfernten Punkten verhält, wenn man sie aktuell in einem Punkt x betrachtet? Bei konvexen Funktionen ist das glücklicherweise anders; hier werden wir sehen, dass die Konvexität uns von vielen dieser Probleme befreit. Dennoch definieren wir zunächst eine Abschwächung des Optimalitätsbegriffs, um auch bei allgemeinen Funktionen noch "ein wenig" optimieren zu können: In Definition 4.2 hatten wir global optimale Punkte definiert.

4.3 Definition (Lokales Optimum)**.** Ein Punkt $x \in \mathbb{R}^n$ heißt *lokales Optimum* für das Optimierungsproblem aus Definition 4.1, wenn es ein $R > 0$ gibt, so dass $f_0(x) \leq f_0(z)$ für alle $z \in \mathcal{F}$ mit $\|z - x\|_2 \leq R$ gilt, wenn also x das Optimierungsproblem eingeschränkt auf eine hinreichend kleine Umgebung von x löst.

Ein lokales Optimum x heißt *striktes lokales Optimum*, wenn ein $R > 0$ existiert, so dass $f_0(x) < f_0(z)$ für alle $z \in \mathcal{F}$ mit $\|z - x\|_2 \leq R$ gilt.

Ein lokales Optimum x heißt *isoliertes lokales Optimum*, wenn eine $R > 0$ existiert so dass in der R-Umgebung von x kein anderes lokales Optimum liegt.

Aufgabe 4.2. Jedes isolierte lokale Optimum ist strikt. Aber nicht jedes strikte lokale Optimum ist isoliert, wie die Funktion $f_0(x) = x^4 \cos(1/x) + 2x^4$ mit $f(0) := 0$ zeigt (diese ist zweimal stetig differenzierbar in 0): Zunächst ist $x = 0$ ein striktes lokales Minimum, aber um $x = 0$ herum liegen beliebig viele weitere strikte lokale Minima.

Die Tatsache, dass ein lokales Optimum ein Häufungspunkt anderer lokaler Optima sein kann, zeigt, dass allgemeine Optimierungsprobleme sehr schwierig sein können.

Wir betrachten daher jetzt speziellere Optimierungsprobleme (insbesondere konvexe), bei denen jeweils bestimmte Einschränkungen an die Ziel- und Restriktionsfunktionen gelten müssen.

4.4 Definition (Konvexes Optimierungsproblem)**.** Das allgemeine Optimierungsproblem in Standard-Form aus Definition 4.1

$$\begin{aligned} &\text{Minimiere } f_0(x) \\ &\quad \text{so dass } f_i(x) \leq 0 && \text{für } i = 1, \ldots, m, \\ &\qquad\quad h_i(x) = 0 && \text{für } i = 1, \ldots, p. \end{aligned}$$

heißt *konvexes Optimierungsproblem*, wenn f_0 und f_i für $i = 1, \ldots, m$ konvex und h_i für $i = 1, \ldots, p$ affin sind.

Einige Erläuterungen:

- Per Definition muss die Definitionsmenge $\mathcal{D}$ als Schnitt der Definitionsmengen konvexer Funktionen konvex sein.
- Auch die zulässige Menge $\mathcal{F}$ ist als Schnitt von Niveaumengen konvexer Funktionen konvex.
- Ein Maximierungsproblem mit konkavem f_0 ist ebenfalls ein konvexes Optimierungsproblem.
- Die Gleichungs-Restriktionsfunktionen müssen auf affin eingeschränkt werden, da sie äquiavalent zu zwei Ungleichungen sind: $h_i(x) \leq 0$ und $-h_i(x) \leq 0$. Sollen sowohl h_i als auch $-h_i$ konvex sein, dann ist h_i gleichzeitig konvex und konkav, also affin.
- Die $h_i(x) = 0$ lassen sich als $\langle a_i | x \rangle = b_i$ schreiben; oBdA ist dabei $a_i \neq 0$ (ansonsten liegt die Gleichung $0 = b_i$ vor, die für $b_i = 0$ redundant und für $b_i \neq 0$ unerfüllbar ist).

Die große Bedeutung konvexer Probleme ergibt sich aus folgendem Lemma.

4.5 Lemma (Lokale Optima konvexer Probleme sind global optimal.)**.** *Sei x lokales Optimum eines konvexen Optimierungsproblems (Definition 4.4. Dann ist x auch (global) optimaler Punkt.*

Beweis. Angenommen, es gäbe ein zulässiges y mit $f_0(y) < f_0(x)$. Da x lokal optimal ist, gibt es $R > 0$, so dass x in seiner R-Umgebung optimal ist, also ist $\|y - x\|_2 > R$. Wir betrachten einen Zwischenpunkt $z = \theta x + (1 - \theta)y$, so dass $\|z - x\|_2 = R/2$, wählen also $\theta := 1 - \frac{R}{2\|y-x\|_2}$. Da $\mathcal{F}$ konvex ist, ist z zulässig. Da f_0 konvex ist, rechnet man nun $f_0(z) < f_0(x)$ nach, aber z liegt in der R-Umgebung von x, wo x optimal ist. Widerspruch. □

Weitere Spezialfälle ergeben sich, wenn man die Konvexitätsforderung noch weiter einschränkt. Wir betrachten einige häufig auftretende Fälle, ohne Anspruch auf Vollständigkeit zu erheben.

4.6 Definition (quadratisch restringiertes quadratisches Problem, QCQP)**.** Ein *quadratisch restringiertes quadratisches Optimierungsproblem* (quadratically constrained quadratic program, QCQP) ist ein konvexes Optimierungsproblem nach Definition 4.4, bei dem f_0 und f_i konvex quadratisch sind, also die Form $f_i(x) = \langle x | P_i | x \rangle + \langle q_i | x \rangle + r_i$ hat mit $P_i \in \mathbb{S}^n_+$ (positiv semidefinit), $q_i \in \mathbb{R}^n$, $r_i \in \mathbb{R}$ für $i = 0, \ldots, m$.

4.7 Definition (quadratisches Problem, QP)**.** Ein *quadratisches Problem* (QP) ist ein konvexes Optimierungsproblem nach Definition 4.4, bei dem f_0 konvex quadratisch ist und die f_i affin sind ($i = 1, \ldots, m$).

4.8 Definition (lineares Problem, LP)**.** Ein *lineares Problem* (LP) ist ein konvexes Optimierungsproblem nach Definition 4.4, bei dem f_0 und die f_i affin sind ($i = 1, \ldots, m$).

Wir fassen zusammen:

- Bei einem linearen Problem wird eine lineare Funktion über einem Polyeder minimiert.
- Bei einem konvexen Problem wird eine konvexe Funktion über einer konvexen Menge minimiert.

4.2 Elementare Beispiele für Optimierungsprobleme

4.9 Beispiel (Diätproblem)**.** Gesunde Ernährung besteht darin, m verschiedene Nährstoffe in gewissen Mindestmengen $b_1, \ldots, b_m$ zu sich zu nehmen. Dazu können wir die Mengen $x_1, \ldots, x_n$ von n verschiedenen Speisen wählen. Eine Einheit der Speise j enthält die Menge a_{ij} des Nährstoffs i und kostet c_j. Welcher Ernährungsplan (unter allen gesunden) minimiert unsere Kosten? Offenbar lässt sich das Problem wie folgt formulieren:

$$\begin{aligned} &\text{Minimiere } \langle c|x\rangle \\ &\quad \text{so dass } Ax \geq b, \\ &\qquad x \geq 0. \end{aligned}$$

Dies lässt sich leicht in Standardform schreiben und ist offensichtlich ein LP. ♡

4.10 Beispiel (Chebyshev-Zentrum eines Polyeders)**.** Gegeben sei ein Polyeder durch m affine Ungleichungen: $\mathcal{P} = \{\, x \mid Ax \leq b \,\} = \{\, x \mid \langle a_i|x\rangle \leq b_i,\ i = 1, \ldots, m \,\}$. Wir suchen den Punkt in $\mathcal{P}$, der am weitesten vom Rand entfernt ist; dieser heißt *Chebyshev-Zentrum* von $\mathcal{P}$.

Dazu bestimmen wir Mittelpunkt $c \in \mathbb{R}^n$ und Radius $r \geq 0$ der größten Kugel $\mathcal{B} = \{\, c + u \mid \|u\|_2 \leq r \,\}$, die ganz in $\mathcal{P}$ liegt. Es müssen also c und r so gewählt werden, dass aus $\|u\|_2 \leq r$ folgt $\langle a_i|c + u\rangle \leq b_i$ für alle $i = 1, \ldots, m$.

Wie groß kann $\langle a_i|u\rangle$ für $\|u\|_2 \leq r$ maximal werden? Nach der Cauchy-Schwarz'schen Ungleichung wird das Skalarprodukt maximal, wenn u und a_i in dieselbe Richtung zeigen (und u maximal lang ist), wenn also $u^* = ra_i/\|a_i\|_2$.

Damit liegt $\mathcal{B}$ genau dann ganz in $\mathcal{P}$, wenn $\langle a_i|c + u^*\rangle = \langle a_i|c\rangle + r\|a_i\|_2 \leq b_i$ für alle i. Unter diesen Bedingungen soll r maximiert werden.

Offensichtlich führt das auf ein LP in den Variablen (c, r), das sich auch in Standard-Form bringen lässt:

$$\begin{aligned} &\text{Maximiere } r \\ &\quad \text{so dass } \langle a_i|c\rangle + r\|a_i\|_2 \leq b_i \qquad\qquad i = 1, \ldots, m. \end{aligned}$$

♡

4.11 Beispiel (Kleinste Quadrate und lineare Regression)**.** Ist das Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$ nicht lösbar, kann man zumindest der "besten" Lösung fragen in dem Sinne, dass $\|Ax - b\|_2$ minimal werden soll. Das ist genau dann der Fall, wenn auch das Quadrat der Zielfunktion minimal wird: Minimiere $\langle Ax - b|Ax - b\rangle = \langle Ax|Ax\rangle - 2\langle b|Ax\rangle + \langle b|b\rangle = \langle x|A^TA|x\rangle - 2\langle b|A|x\rangle + \langle b|b\rangle$. Wenn gewünscht, können noch affine Nebenbedingungen hinzukommen, beispielsweise untere und obere Schranken für die Variablen: $l_i \leq x_i \leq u_i$ für $i = 1, \ldots, n$. Dies ist offensichtlich ein QP.

Wir bemerken, dass das unrestringierte Problem eine wohlbekannte analytische Lösung hat: $x = A^-b$, wobei A^- die Pseudoinverse von A ist; dies führt stets auf die Lösung x mit kleinstem Abstand vom Nullpunkt. ♡

4.12 Beispiel (Abstand zwischen Polyedern)**.** Gegeben sind zwei Polyeder $\mathcal{P}$ und $\mathcal{Q}$ in Ungleichungsform: $\mathcal{P} = \{ x \mid Ax \leq b \}$ und $\mathcal{Q} = \{ x \mid Cx \leq d \}$. Die Distanz zwischen $\mathcal{P}$ und $\mathcal{Q}$ ist die Euklidische Distanz zwischen ihren nächsten Punkten (null, wenn sie sich schneiden). Um diese zu finden, lösen wir das folgende Minimierungsproblem: Minimiere $\|x - y\|_2$ so dass $x \in \mathcal{P}$ und $y \in \mathcal{Q}$. Nach Quadrieren der Zielfunktion ist das ein QP in (x, y):

$$\begin{aligned} \text{Minimiere } & \langle x - y | x - y \rangle \\ \text{so dass } & Ax \leq b, \\ & Cy \leq d. \end{aligned}$$

♡

4.13 Beispiel (Schranken für Erwartungswert und Varianz von Verteilungen)**.** Sei $p \in \mathbb{R}^n$ eine unbekannte Wahrscheinlichkeitsverteilung auf einer Menge $\{ u_1, \ldots, u_n \}$ mit $p_j = \mathbb{P}(\{ u_j \})$. Insbesondere sei $p \geq 0$ und $\langle 1 | p \rangle = 1$. Es sei X eine Zufallsvariable mit Verteilung p.

Angenommen, es sind zumindest gewisse Schranken für verschiedene Erwartungswerte von Funktionen von X bekannt, etwa $\alpha_i \leq \mathbb{E}[f_i(X)] \leq \beta_i$ für $i = 1, \ldots, m$. Da $\mathbb{E}[f_i(X)] = \sum_{j=1}^n p_j f_i(u_j) = \langle a_i | p \rangle$ mit $a_{ij} = f_i(u_j)$, lassen sich diese als affine Ungleichungen $\alpha_i \leq \langle a_i | p \rangle \leq \beta_i$ schreiben.

Wir fragen nun nach dem minimal und maximal möglichen Erwartungswert einer weiteren Funktion f_0, also nach $\mathbb{E}[f_0(X)]$. Insbesondere können wir nach $\mathbb{E}[X]$ selbst fragen. Da dies wieder eine lineare Funktion von p in der Form $\langle a_0 | p \rangle$ ist, erhalten wir ein LP. Indem wir minimieren, erhalten wir eine untere Schranke; indem wir maximieren, erhalten wir eine obere Schranke.

Ferner können wir nach einer *oberen* Schranke der Varianz (einer Funktion) von X fragen, denn die Varianz lässt sich schreiben als $\mathrm{Var}[f(X)] = \mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2$, also mit $f_j := f(u_j)$ als $\mathrm{Var}[X] = \langle f^2 | p \rangle - (\langle f | p \rangle)^2$; dies ist eine konkave Funktion von p. Maximieren führt auf ein QP. ♡

Wir erinnern auch noch einmal an die Beispiele im Einführungskapitel.

4.3 Äquivalenz von Optimierungsproblemen und Beispiele

Wir geben keine formale Definition der Äquivalenz von Optimierungsproblemen. Informell sind zwei Probleme äquivalent, wenn man aus der Lösung des einen effizient die Lösung des jeweils anderen erhalten kann. Typische Äquivalenzumformungen sind dabei:

- bijektive Variablen-Transformationen
- bijektive Zielfunktions- und Restriktionsfunktions-Transformationen
- Epigraph-Formulierung
- Einführen von Schlupfvariablen
- Eliminieren und Hinzufügen von Gleichungsrestriktionen

- Analytische Optimierung über einige der Variablen
- Umwandlung zwischen impliziten und expliziten Restriktionen

Wir diskutieren diese Transformationen nicht im einzelnen abstrakt, sondern an konkreten Beispielen.

Interessant dabei ist vor allem, dass sich dabei manchmal zunächst nicht konvexe Probleme in konvexe Probleme verwandeln lassen. Wir betrachten zunächst ein einfaches künstliches Beispiel und dann das (auch praktisch wichtige) Beispiel der geometrischen Probleme.

4.14 Beispiel (Äquivalente Probleme)**.** Das Problem

$$\begin{aligned} \text{Minimiere } & f_0(x) = x_1^2 + x_2^2, \\ \text{so dass } & f_1(x) = x_1/(1+x_2^2) \leq 0, \\ & h_1(x) = (x_1+x_2)^2 = 0 \end{aligned}$$

ist nicht konvex, da f_1 nicht konvex und h_1 nicht affin ist. Es ist aber offensichtlich äquivalent zu folgendem quadratischen Problem (QP):

$$\begin{aligned} \text{Minimiere } & f_0(x) = x_1^2 + x_2^2, \\ \text{so dass } & f_1(x) = x_1 \leq 0 \\ & h_1(x) = x_1 + x_2 = 0. \end{aligned}$$

Die zulässige Menge ist unverändert. ♡

4.15 Definition (geometrisches Optimierungsproblem)**.** Ein *Monom* ist eine Funktion $h : \mathbb{R}^n \to \mathbb{R}$ mit $\operatorname{dom} h = \mathbb{R}^n_{++}$ der Form $h(x) = cx_1^{a_1}x_2^{a_2}\cdot\cdots\cdot x_n^{a_n}$ mit $c > 0$ und $a_i \in \mathbb{R}$ (beachte: es ist $x_i > 0$, aber $a_i \in \mathbb{R}$ beliebig). Ein *Posynom* ist eine endliche Summe von Monomen.

Ein Optimierungsproblem der Form

$$\begin{aligned} \text{Minimiere } & f_0(x) & & \\ \text{so dass } & f_i(x) \leq 1 & & \text{für } i = 1,\dots,m, \\ & h_i(x) = 1 & & \text{für } i = 1,\dots,p, \end{aligned}$$

mit Posynomen $f_0, f_1, \dots, f_m$ und Monomen $h_1, \dots h_p$ heißt *geometrisches Optimierungsproblem.* (Die Bedingung $x > 0$ ist implizit durch die Definitionsmenge gegeben.)

Man überzeugt sich leicht, dass ein geometrisches Problem nicht konvex ist (z.B. ist die konkave Funktion $\sqrt{x_1}$ als Zielfunktion möglich).

Es lässt sich jedoch ein äquivalentes konvexes Problem angeben; dies nennen wir ein *geometrisches Problem in konvexer Form.* Die Idee dabei ist, zuerst eine bijektive Variablentransformation $y_j = \log x_j$ durchzuführen und die Ziel- und Restriktionsfunktionen ebenfalls zu logarithmieren.

4.16 Lemma (Geometrisches Optimierungsproblem in konvexer Form)**.** *Setzt man in einem geometrischen Optimierungsproblem nach Definition 4.15 komponentenweise $y := \log x$, so dass $x = \exp y$ und definiert $\tilde{f}_i(y) := \log f_i(\exp(y))$ und $\tilde{h}_i(y) := \log h_i(\exp(y))$ für alle i, dann ergibt sich ein konvexes Problem.*

Beweis. Wir betrachten ein Monom: Sei $h = cx_1^{a_1} \cdot \dots \cdot x_n^{a_n} = c\exp(y_1)^{a_1} \cdot \dots \cdot \exp(y_n)^{a_n} = \exp(\log c + a_1 y_1 + \dots + a_n y_n) = \exp(\langle a|y\rangle + \tilde{c})$ mit $\tilde{c} = \log c$. Damit ist $\tilde{h} = \log h = \langle a|y\rangle + \tilde{c}$ affin in y; die entsprechende Restriktion lautet $\tilde{h}(y) = 0$.

Wir betrachten ein Posynom mit $K \geq 2$ Termen: Sei $f = \sum_{k=1}^{K} c_k x_1^{a_{k1}} \cdot \dots \cdot x_n^{a_{n1}} = \sum_{k=1}^{K} \exp(\langle a_k|y\rangle + \tilde{c_k})$ mit $\tilde{c_k} = \log c_k$. Damit ist $\tilde{f} = \log f = \log \sum_{k=1}^{K} \exp(\langle a_k|y\rangle + \tilde{c_k})$ als Komposition einer affinen Funktion mit log-sum-exp als konvex in y erkannt. Die entsprechende Restriktion lautet $\tilde{f}(y) \leq 0$. □

Ein weiterer nützlicher Umformulierungstrick ist die Epigraph-Form, die auch zeigt, dass bei konvexen Problemen oBdA die Zielfunktion linear gewählt werden kann.

4.17 Lemma (Epigraph-Form eines Problems)**.** *Sei ein allgemeines Optimierungsproblem in $x \in \mathbb{R}^n$ nach Definition 4.1 gegeben. Dazu äquivalent ist das folgende Problem in $(x,t) \in \mathbb{R}^{n+1}$ ("*Epigraph-Form*"):*

$$\begin{aligned} &\textit{Minimiere } t \\ &\quad \textit{so dass } f_0(x) - t \leq 0, \\ &\qquad\quad f_i(x) \leq 0 \qquad\qquad \textit{für } i = 1, \dots, m, \\ &\qquad\quad h_i(x) = 0 \qquad\qquad \textit{für } i = 1, \dots, p. \end{aligned}$$

Das Problem in Epigraph-Form ist genau dann konvex, wenn das ursprüngliche Problem konvex ist.

Beweis. Zunächst ist klar: x ist genau dann zulässig für das Originalproblem, wenn (x,t) mit $t \geq f_0(x)$ zulässig ist.

Wir nehmen oBdA an, dass der optimale Wert angenommen wird. Sei t^* der optimale Wert des Epigraph-Problems, p^* der optimale Wert des ursprünglichen Problems, und x^* ein optimaler Punkt des ursprünglichen Problems. Wir zeigen $t^* \leq p^*$: Dies gilt, da (x^*, p^*) im Epigraph-Problem zulässig ist. Wir zeigen $p^* \leq t^*$: In jedem zulässigen Punkt (x,t) ist $t \geq f_0(x)$, also insbesondere $t \geq \inf\{ f_0(x) \mid x \in \mathcal{F} \} = p^*$ für alle zulässigen (x,t).

Die Konvexitätsaussage ist klar. □

Als Beispiel der Nützlichkeit führen wir an, dass sich damit nicht differenzierbare Zielfunktionen häufig elegant umgehen lassen.

4.18 Beispiel (Maximums-Norm-Zielfunktion)**.** Ist das Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ nicht lösbar, kann man zumindest der "besten" Lösung fragen. Häufig wird das in dem Sinne interpretiert, dass $\|Ax - b\|$ minimal werden soll; insbesondere mit der 2-Norm führt dies auf das kleinste-Quadrate-Problem.

Eine andere Interpretation ist, dass die größte Abweichung in allen Komponenten minimiert werden soll, also $\|Ax - b\|_\infty = \max_i |\langle a_i|x\rangle - b_i|$ minimiert wird. Dies ist ein unrestringiertes Optimierungsproblem; die Zielfunktion ist aber nicht glatt (differenzierbar) in x.

Umformulieren in die Epigraph-Form führt zunächst auf

$$\begin{aligned} &\text{Minimiere } t \\ &\quad \text{so dass } \max_{i=1,\dots,m} |\langle a_i|x\rangle - b_i| \leq t. \end{aligned}$$

Weiteres Nachdenken zeigt, dass das Betragsmaximum genau dann unter t bleibt, wenn *alle* Komponenten es tun, also macht man aus der einen Ungleichung m Paare, indem man die Betragsstriche auflöst.

$$\begin{aligned} &\text{Minimiere } t \\ &\quad \text{so dass } -t \le \langle a_i | x \rangle - b_i \le t \qquad\qquad i = 1, \dots, m \end{aligned}$$

Bringt man dies in Standard-Form, hat man $2m$ affine Ungleichungen. Jetzt sind Ziel- und Restriktionsfunktionen differenzierbar (sogar affin), aber das Problem ist nicht mehr unrestringiert. ♡

Da Ausgleichsprobleme der Form "Minimiere eine konvexe Funktion von $Ax - b$" häufig vorkommen, betrachten wir hierzu noch ein weiteres Beispiel.

4.19 Definition (Huber-Funktion, Huber-Ausgleichsproblem)**.** Die *Huber-Funktion* (engl. Huber penalty function) mit Parameter $M \ge 0$ ist definiert als $H : \mathbb{R} \to \mathbb{R}$ mit

$$H(u) := \begin{cases} u^2 & |u| \le M, \\ M(2|u| - M) & |u| \ge M. \end{cases}$$

Die Huber-Funktion ist zwischen $-M$ und M quadratisch und jenseits davon affin.

Das Huber-Ausgleichsproblem zu Daten $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und Parameter $M \ge 0$ ist definiert als das unrestringierte Optimierungsproblem

$$\text{Minimiere } \sum_{i=1}^{m} H(\langle a_i | x \rangle - b_i).$$

Da H konvex ist, ist das Huber-Ausgleichsproblem ein konvexes unrestringiertes Optimierungsproblem. Es ergibt sich wieder das Problem, dass H und damit die Zielfunktion nicht differenzierbar ist.

Aufgabe 4.3. Zeige, dass das Huber-Ausgleichsproblem zu folgendem QP in den $2n$ Variablen $(u_i), (v_i)$ mit $5m$ affinen Ungleichungen äquivalent ist:

$$\begin{aligned} &\text{Minimiere } \sum_{i=1}^{n} (u_i^2 + 2Mv_i) \\ &\quad \text{so dass } -u - v \le Ax - b \le u + v && (2m \text{ Ungleichungen}), \\ &\qquad\qquad 0 \le u_i \le M && (2m \text{ Ungleichungen}), \\ &\qquad\qquad 0 \le v_i && (m \text{ Ungleichungen}). \end{aligned}$$

Aufgabe 4.4. Zeige, dass das Huber-Ausgleichsproblem auch zu folgendem Problem äquivalent ist und dass dieses Problem konvex ist. Die Variablen sind $(x, w) \in \mathbb{R}^{n+m}$ mit Definitionsbereich $\mathcal{D} = R^n \times \{\, w \in \mathbb{R}^m \mid w_i > -1 \ \forall i \,\}$.

$$\begin{aligned} &\text{Minimiere } \sum_{i=1}^{m} (\langle a_i | x \rangle - b_i)^2 / (w_i + 1) + M^2 \langle 1 | w \rangle \\ &\quad \text{so dass } w_i \ge 0 && (2m \text{ Ungleichungen}). \end{aligned}$$

4.4 Optimalitätsbedingungen für konvexe Probleme

Wir setzen jetzt voraus, dass der Definitionsbereich $\mathcal{D}$ des Problems offen, die Zielfunktion f_0 auf $\mathcal{D}$ differenzierbar ist und $\mathcal{F} \subset \mathcal{D}$ gilt. Wir geben jetzt ein notwendiges und hinreichendes Optimalitätskriterium für einen Punkt x^* an.

4.20 Satz (Optimalitätskriterium)**.** *Ein Punkt x^* ist optimal für ein konvexes Problem nach Definition 4.4 mit $\mathcal{D}$ offen, f_0 differenzierbar genau dann, wenn*

$$\langle \nabla f_0(x^*) | x - x^* \rangle \geq 0 \text{ für alle } x \in \mathcal{F}.$$

Beweis. "⇐": Da f_0 konvex und differenzierbar ist, gilt das Konvexitätskriterium nach Satz 3.18, und für alle $x \in \mathcal{F}$ ist

$$f_0(x) \geq f_0(x^*) + \langle \nabla f_0(x^*) | x - x^* \rangle \geq f_0(x^*).$$

Damit ist x^* in $\mathcal{F}$ optimal.

"⇒": Es sei x^* optimal. Angenommen, es gäbe ein $x \in \mathcal{F}$ mit $\langle \nabla f_0(x^*) | x - x^* \rangle < 0$; dann wäre $f_0(x^* + \varepsilon(x - x^*)) < f_0(x^*)$ für hinreichend kleines $\varepsilon > 0$ nach der Taylor-Entwicklung, also war x^* nicht optimal; Widerspruch. □

Wir leiten jetzt ein paar Spezialfälle her. Systematisch behandeln wir Optimalitätsbedingungen in Kapitel 6 über Dualität.

Unrestringierte konvexe Probleme. Wenn es keine Nebenbedingungen gibt und $\mathcal{D}$ offen ist, kann man von x^* aus ein kleines Stück in jede Richtung p gehen, insbesondere zu $x = x^* + \varepsilon p$ mit hinreichend kleinem $\varepsilon > 0$. Speziell für die Richtung $p := -\nabla f_0(x^*)$ erhält man die Bedingung $\langle \nabla f_0(x^*) | - \varepsilon \nabla f_0(x^*) \rangle = -\varepsilon \|\nabla f_0(x^*)\|_2^2 \stackrel{!}{=} 0$. Diese Bedingung ist genau dann erfüllt, wenn $\nabla f_0(x^*) = 0$. Insgesamt erhält man so: Notwendig und hinreichend für die Lösung eines unrestringierten differenzierbaren konvexen Problems ist $\nabla f_0(x^*) = 0$.

Konvexe Probleme mit Gleichungs-Restriktionen. Lautet das Problem "Minimiere $f_0(x)$ so dass $Ax = b$", dann besagt die Optimalitätsbedingung, dass $\langle \nabla f_0(x^*) | x - x^* \rangle \geq 0$ für alle x mit $Ax = b$ gelten muss.

Jedes solche x lässt sich als $x^* + v$ mit $v \in \mathrm{Kern}(A)$ schreiben. Daher lautet die Optimalitätsbedingung $\langle \nabla f_0(x^*) | v \rangle \geq 0$ für alle $v \in \mathrm{Kern}(A)$. Da $\mathrm{Kern}(A)$ ein Unterraum ist, der mit v auch $-v$ enthält, muss sogar $\langle \nabla f_0(x^*) | v \rangle = 0$ für alle $v \in \mathrm{Kern}(A)$ gelten, also muss $\nabla f_0(x^*)$ auf $\mathrm{Kern}(A)$ senkrecht stehen.

Da das orthogonale Komplement von $\mathrm{Kern}(A)$ gleich dem Bild von A^T ist, kann man die Optimalitätsbedingung auch als $\nabla f_0(x^*) \in \mathrm{Bild}(A^T)$ formulieren oder: Es gibt ein $\nu \in \mathbb{R}^n$ mit $\nabla f_0(x^*) + A^T \nu = 0$. Zusammen mit der Zulässigkeitsbedingung $Ax^* = b$ liefert diese Optimalitätsbedingung notwendige und hinreichende Bedingungen für das Optimum.

Nichtnegative Variablen. Das Problem "Minimiere $f_0(x)$ so dass $x \geq 0$", dann besagt die Optimalitätsbedingung: $\langle \nabla f_0(x^*) | x - x^* \rangle \geq 0$ für alle $x \geq 0$.

Ist die Komponente $x_j^* > 0$, kann man entlang der j-ten Achse sowohl in die positive als auch negative Richtung gehen; daher muss $\nabla f_0(x^*)_j = 0$ sein.

Ist aber $x_j^* = 0$, so muss $\nabla f_0(x^*)_j \geq 0$ sein; negative Werte würden bedeuten, dass man die Zielfunktion durch kleine Schritte in positive Richtung auf der j-ten Achse verbessern kann.

Insgesamt ergeben sich die Bedingungen

$$x^* \geq 0, \quad \nabla f_0(x^*) \geq 0, \quad x_j^* \nabla f_0(x^*)_j = 0 \text{ für alle } j.$$

Insbesondere verschwindet also die j-te Komponente des Gradienten, oder x_j^* verschwindet.

4.21 Beispiel (Kleinste Quadrate)**.** Wir betrachten das Problem "Minimiere $\|Ax - b\|_2$ so dass $x \geq 0$" unter der Voraussetzung, dass $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ vollen Rang n hat.

Der Gradient der quadrierten Zielfunktion $f_0(x)^2 = \langle x | A^T A | x \rangle - 2 \langle b | Ax \rangle + \langle b | b \rangle$ ist $\nabla f_0(x) = 2A^T Ax - 2A^T b$. Unter den Voraussetzungen ist $A^T A$ eine invertierbare $n \times n$-Matrix, und die Bedingungen lauten: $x \geq 0$, $A^T Ax \geq A^T b$ und für alle $j = 1, \ldots, n$ gilt: $x_j = 0$ oder $(A^T Ax)_j = (A^T b)_j$. ♡

KAPITEL 5

Lösung unrestringierter konvexe Optimierungsprobleme

Konvexe Funktionen sind deshalb besonders "optimierungsfreundlich", weil sie, wie wir in den vorigen Kapiteln gesehen haben,

- von selbst relativ glatt sind (stetig im Inneren ihres Definitionsbereiches; alle Richtungsableitungen existieren dort),
- garantieren, dass alle lokalen Optima auch globale Optima sind. (Das folgt schon daraus, dass die Niveaumengen einer konvexen Funktion konvex sind.)

Wenn jetzt der glückliche (aber leider anwendungsferne) Fall eintritt, dass keine Restriktionen (Nebenbedingungen) vorliegen, wird die Optimierung besonders einfach: Man muss lediglich sicherstellen, den Definitionsbereich nicht zu verlassen und immer "abwärts" zu laufen. Dann landet man (wenn man nicht stehenbleibt) irgendwann in einem lokalen Minimum (sofern eins existiert), und das ist dann ein globales Minimum.

Auch wenn es hierzu nicht viele praktische Anwendungen gibt, bilden die Elemente dieses Kapitels dennoch die Grundlage für die aufwändigeren Algorithmen unter Nebenbedingungen. Daher ziehen wir es vor, die zum Einsatz kommenden Methoden zunächst ohne diese zusätzlichen Schwierigkeiten zu betrachten.

5.1 Beispiele

5.1 Beispiel (Unrestringierte approximative Lösung eines linearen Gleichungssystems)**.** Zu gegebenen Daten A, b, finde x, das $\|Ax - b\|$ (bezüglich irgend einer gegebenen Norm) minimiert. Dies ist stets ein konvexes Optimierungsproblem, führt aber (außer in wenigen

Ausnahmen wie der Euklidischen Norm) auf eine nicht überall differenzierbare Zielfunktion. Will man aber die Differenzierbarkeit der Zielfunktion erreichen, so ist oft eine Umformulierung notwendig, die allerdings dann i.d.R. zu Nebenbedingungen führt (ein Beispiel dafür ist das Huber-Ausgleichsproblem). ♡

5.2 Beispiel (Analytisches Zentrum)**.** Gegeben ist ein Polytop der Form $\{ x \mid Ax \leq b \}$, von dem wir annehmen, dass sein Inneres nicht leer ist, dass also Punkte mit $Ax < b$ existieren. Gegeben sei nun die Zielfunktion $f_0(x) := -\sum_{i=1}^m \log(b_i - \langle a_i | x \rangle)$ mit $\operatorname{dom} f_0 = \{ x \mid Ax < b \}$. Da f_0 auf $\operatorname{dom} f_0$ differenzierbar ist, lautet die entsprechende Optimalitätsbedingung

$$\nabla f_0(x) = \sum_{i=1}^m \frac{1}{b_i - \langle a_i | x \rangle} a_i = 0, \quad Ax < b.$$

Gibt es keine Lösung, dann gibt es keine optimalen Punkte. Das ist genau dann der Fall, wenn entweder $Ax < b$ unerfüllbar oder das Problem unbeschränkt ist. Gibt es genau eine Lösung, dann gibt es genau einen optimalen Punkt. Das ist genau dann der Fall, wenn das Polyeder $Ax \leq b$ nicht leer und beschränkt ist. Gibt es mehrere Lösungen, dann bilden sie eine affine Menge (hier ohne Beweis).

Wenn es sich bei $\{ x \mid Ax \leq b \}$ um ein kompaktes Polyeder handelt mit nichtleerem Inneren handelt, kann man nachrechnen, dass der eindeutige optimale Punkt x^* derjenige ist, für den das Produkt der Abstände zu den begrenzenden Hyperebenen maximal wird. Er heißt "analytisches Zentrum" des Polyeders. ♡

5.3 Beispiel (Re-skalierung von Variablen bei Endomorphismen)**.** Gegeben sei ein linearer Endomorphismus $M : \mathbb{R}^n \to \mathbb{R}^n$, dargestellt durch eine Matrix M, der x nach $y = Mx$ abbildet. Wenn wir die Koordinatenachsen von $\mathbb{R}^n$ unabhängig reskalieren, also $\tilde{x}_i = d_i x_i$ und $\tilde{y}_i = d_i y_i$ setzen und $D = \operatorname{diag}(d)$ schreiben, dann ist $\tilde{y} = Dy = DMx = DMD^{-1}\tilde{x}$.

Die Skalierung soll so gewählt werden, dass die resultierende Matrix DMD^{-1} möglichst kleine Einträge im Sinne der Frobenius-Norm $\|DMD^{-1}\|_F$ hat (Wurzel aus der Summe der quadrierten Einträge). Man rechnet schnell aus, dass $\|DMD^{-1}\|_F^2 = \sum_{i,j=1}^n M_{ij}^2 d_i^2 / d_j^2$. Dies ist ein Posynom in d, die Minimierung führt also auf ein unrestringiertes geometrisches Problem, das in ein äquivalentes konvexes Problem transformiert werden kann (Übung!). Erfreulicherweise ist dessen Zielfunktion auch differenzierbar. ♡

5.2 Ein Modellverfahren für unrestringierte Probleme

Da im Normalfall nicht davon ausgegangen werden kann, dass die unregstringierte Optimalitätsbedingung $\nabla f(x) = 0$ durch eine einfache Formel gelöst werden kann, sind iterative Verfahren, die gegen eine Lösung konvergieren, notwendig. Die (globale) Konvergenz dieser Verfahren kann in der Regel nachgewiesen werden, wenn man die Existenz einer Lösung und starke Konvexität voraussetzt.

Wir betrachten zunächst einen allgemeinen Rahmen für Abstiegsverfahren (die beim konvexen Funktionen immer zum Ziel führen) und dann einige wichtige Spezialfälle (Gradientenverfahren, Newton-Verfahren, BFGS-Verfahren), die wir unterschiedlich genau analysieren. In der Praxis wendet man im Normalfall das BFGS-Verfahren an.

5.4 Definition (Abstiegsverfahren)**.** Ein *Abstiegsverfahren* ist ein Iterationsverfahren, bei dem ausgehend von einem Startwert $x_0 \in \mathcal{F}$ entweder eine endliche Folge $(x_0, x_1, \ldots, x_n)$ gebildet wird, so dass $f(x_0) > f(x_1) > \cdots > f(x_n) = p^*$, oder eine Folge $(x_0, x_1, x_2 \ldots)$ mit $f(x_0) > f(x_1) > f(x_2) > \cdots \to p^*$.

Wir erinnern jetzt noch einmal an den grundlegenden Satz von Taylor in n Dimensionen (Entwicklung einer Funktion f um einen Punkt x_0): Es ist

$$f(x) = f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + o(\|x - x_0\|);$$
$$f(x) = f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + \frac{1}{2} \langle x - x_0 | \nabla^2 f(x_0) | x - x_0 \rangle + o(\|x - x_0\|^2),$$

stetige zweite Ableitungen von f vorausgesetzt. Hierbei wird das Restglied abgeschätzt. Eine alternative Formulierung, bei der die letzte Ableitung nicht in x_0, sondern in einem Zwischenpunkt $x_0 + t(x - x_0)$ mit einem geeigneten $0 \leq t \leq 1$ ausgewertet wird, ergibt die folgende Darstellung:

$$f(x) = f(x_0) + \langle \nabla f(x_0 + t(x - x_0)) | x - x_0 \rangle \,; \tag{5.1}$$
$$f(x) = f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + \frac{1}{2} \langle x - x_0 | \nabla^2 f(x_0 + t(x - x_0)) | x - x_0 \rangle \,. \tag{5.2}$$

Modellalgorithmus. Wir betrachten jetzt Verfahren, die zunächst eine *Abstiegsrichtung* p bestimmen und dann eine Schrittweite $t > 0$, so dass $x = x_0 + tp$ die nächste Iteration bildet und $f(x) < f(x_0)$ ist (es sei denn x_0 ist schon ein optimaler Punkt).

Sei x_0 der aktuelle Punkt; wir berechnen den nächsten Punkt x_+.

1. Teste auf Konvergenz.
2. Bestimme eine Abstiegsrichtung p in x_0.
3. Bestimme eine Schrittweite $t > 0$ für die Richtung p.
4. Setze $x_+ := x_0 + tp$.
5. Gehe zu 1. mit $x_0 := x_+$.

Im Folgenden überlegen wir uns Details zu den einzelnen Schritten.

Bestimmung einer Abstiegsrichtung.

5.5 Definition (Abstiegsrichtung)**.** Eine *Abstiegsrichtung* p für $f : \mathbb{R}^n \to \mathbb{R}$ im Punkt $x_0 \in \mathbb{R}^n$ ist eine Richtung, für die $f'(x_0; p) < 0$ gilt. (Hier ist $f'(x_0; p) = \langle \nabla f(x_0) | p \rangle$ die Richtungsableitung.)

Für hinreichend kleine $t > 0$ ist also $f(x_0 + tp) < f(x_0)$ aufgrund des Taylor-Satzes: $f(x_0 + tp) = f(x_0) + t \langle \nabla f(x_0) | p \rangle + o(t)$.

Abstiegsrichtungen bekommt man aus dem folgenden Lemma.

5.6 Lemma (Abstiegsrichtungen)**.** *Sei W positiv definit. Dann ist $p := -W \cdot \nabla f(x_0)$ eine Abstiegsrichtung für f in x_0, es sei denn x_0 ist optimal.*

Beweis. Sei x_0 nicht optimal. Dann ist $\nabla f(x_0) \neq 0$. Es folgt $f'(x_0; p) = \langle \nabla f(x_0) | p \rangle = -\langle \nabla f(x_0) | W | \nabla f(x_0) \rangle < 0$, da W positiv definit ist. □

Insbesondere ist mit der Einheitsmatrix $W = \mathrm{Id}$ der negative Gradient selbst eine Abstiegsrichtung; dies ist sogar die Richtung des steilsten Abstiegs (in der Euklidischen Norm).

Aufgabe 5.1. Sei p eine Richtung mit $\|p\| = \|\nabla f(x_0)\|$ (dies ist nur eine Normierungsbedingung). Dann ist $f'(x_0; p) \geq f'(x_0; -\nabla f(x_0))$; der Abstieg ist also weniger "steil".

Die Wahl $p := -\nabla f(x_0)$ führt auf das Gradientenverfahren, das wir im nächsten Abschnitt genauer untersuchen. Obwohl naheliegend, ist diese nicht immer die beste, wie sich aus den Konvergenzanalysen ergeben wird.

Eine andere sinnvolle Wahl ist zum Beispiel die *Newton-Richtung* $p := -[\nabla^2 f(x_0)]^{-1} \nabla f(x_0)$, die sich wie folgt motivieren lässt: Wir betrachten die quadratische Taylor-Approximation m an f, entwickelt in x_0, und so verschoben, dass x_0 dem Nullpunkt für m entspricht:

$$m(p) = f(x_0) + \langle \nabla f(x_0) | p \rangle + \langle p | \nabla^2 f(x_0) | p \rangle \,.$$

Dies ist eine quadratische konvexe Funktion in p, deren optimales p (strikte Konvexität der Eindeutigkeit wegen vorausgesetzt) sich als die Newton-Richtung ergibt (Übung!). Da die (Inverse der) Hessematrix in jedem Punkt x_0 positiv definit ist, handelt es sich um eine Abstiegsrichtung.

Bestimmung einer Schrittweite. Als nächstes stellt sich die Frage nach einer sinnvollen *Schrittweite*. Beim Newton-Verfahren liegt es beispielsweise nahe, die Schrittweite $t = 1$ zu wählen, da man auf diese Art und Weise ja die Modellfunktion m minimiert. Allerdings muss m nicht überall eine gute Approximation an f sein. Beim Gradientenverfahren und im Allgemeinen ist die Frage ohnehin schwieriger zu beantworten. Am besten ist es, wenn man in der gewählten Richtung zum Minimum (in dieser Richtung) vordringen kann.

5.7 Definition (Exakte Schrittweite)**.** Die Schrittweite $t^* > 0$ heißt *exakte Schrittweite* zu Richtung p im Punkt x_0 für Funktion f, wenn $\phi(t^*) \leq \phi(t) := f(x_0 + tp)$ für alle $t \geq 0$.

Eine exakte Schrittweite existiert immer, denn ϕ ist konvex in t, da f konvex ist. (Allerdings kann, sofern das Problem unbeschränkt ist, $t^* = \infty$ werden. Diesen Fall schließen wir in diesem Abschnitt aus, da sich ansonsten ohnehin keine Konvergenzaussagen machen lassen.) Hier muss nur ein eindimensionales Minimierungsproblem (in t) gelöst werden. Dies funktioniert numerisch stets mit einer Intervallsuche:

1. In einem ersten Schritt wird ein $C > 0$ bestimmt, so dass das Minimum sicher in $[0, C]$ liegt. Dies kann z.B. geschenen, indem man mit $C := 1$ beginnt und solange verdoppelt, bis $\phi(C) > \phi(0)$ ist (oder aufeinander folgende ϕ-Werte steigen).
2. Im folgenden Schritt wird mittels binärer Suche (oder durch ein anderes Verfahren, z.B. Intervallsuche mit goldenem Schnitt) in $[0, C]$ das optimale (minimierende) t^* bestimmt.

Es werden allerdings vergleichsweise viele Funktionsauswertungen benötigt, um das optimale t^* zu bestimmen. Da sich danach ohnehin die Richtung ändert und man (normalerweise) noch nicht das globale Minimum von f erreicht hat, macht dieser Aufwand meist nicht viel Sinn.

Statt dessen verlangt man nur, dass ϕ (im Vergleich zur Schrittweite t) hinreichend stark fällt. Da ϕ konvex ist, gilt in jedem Fall $\phi(t) \geq \phi(0)+t\phi'(0)$ (eine konvexe Funktion liegt stets oberhalb ihrer Tangenten). Wählt man aber ein beliebiges $0 < \alpha < 1$, so ist die Bedingung $\phi(t) \leq \phi(0)+\alpha t\phi'(0)$ zumindest in einem kleinen Intervall $[0, t_0]$ erfüllbar (dies folgt wiederum sofort aus dem Satz von Taylor). Dies garantiert, dass "pro Schritteinheit" der Funktionswert um mindestens $\alpha\phi'(0)$ verbessert wird: Es ist dann $[\phi(t) - \phi(0)]/t \leq \alpha\phi'(0) < 0$.

Im Folgenden wählen wir sogar ein $0 < \alpha < 1/2$ und suchen nach einer nicht zu kleinen Schrittweite, die die Bedingung erfüllt. Dies erreichen wir zum Beispiel, indem wir zunächst $t := 1$ prüfen. Verletzt ein t diese Bedingung, versuchen wir es mit $t \leftarrow \beta t$ mit einem $0 < \beta < 1$ erneut, bis die Bedingung schließlich erfüllt ist. Diese Strategie heißt *Armijo-Schrittweite* oder *Backtracking-Schrittweite* mit Parametern $0 < \alpha < 1/2$ und $0 < \beta < 1$.

5.8 Definition (Armijo-Schrittweite, Backtracking-Schrittweite)**.** Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex, sei $x_0 \in \mathbb{R}^n$, sei $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_0, und sei $\phi(t) := f(x_0 + tp)$. Seien $0 < \alpha < 1/2$ und $0 < \beta < 1$ zwei Parameter. Die Armijo-Schrittweite oder Backtracking-Schrittweite zu (α, β) ist durch folgenden Algorithmus gegeben:

- Sei $t := 1$.
- Solange $f(x_0 + tp) > f(x_0) + \alpha t \langle \nabla f(x_0) | p \rangle$: Setze $t := \beta t$.
- Gib t zurück.

Es gibt weitere Bedingungen, die man an eine Schrittweite stellen kann (Wolfe-Bedingungen; Powell-Schrittweite); diese sind vor allem für nicht konvexe Funktionen wichtig, um Konvergenzaussagen zu machen. Für unsere Zwecke genügt zunächst die Armijo-Schrittweite.

Voraussetzungen. Für die weitere Analyse machen wir folgende Voraussetzungen.

5.9 Voraussetzung (Analyse von Abstiegsverfahren)**.**

- Sei die initiale Niveaumenge $S := \{ x \mid f(x) \leq f(x_0) \}$ kompakt. (Dies garantiert, dass das Problem beschränkt ist.)
- Sei f auf S stark konvex; insbesondere gebe es $0 < m \leq M$, so dass für alle $x \in S$ die Eigenwerte der Hesse-Matrix $\nabla^2 f(x)$ in $[m, M]$ liegen. (Die Existenz eines $0 < M < \infty$ folgt schon aus der Kompaktheit von S; die Existenz eines $m > 0$ garantiert die Eindeutigkeit des optimalen Punktes.)
- Es gebe eine Konstante $L \geq 0$, so dass $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2^2 \leq L\|x - y\|_2^2$ für alle $x, y \in S$. (Dies ist z.B. immer dann gegeben, wenn die dritten Ableitungen existieren und auf S durch L beschränkt sind. Bei quadratischen Funktionen kann $L = 0$ gewählt werden. Der Wert von L ist ein Maß dafür, wie schwer sich f quadratisch approximieren lässt und spielt bei der Analyse des Newton-Verfahrens eine Rolle.)

5.3 Das Gradientenverfahren

Das Gradientenverfahren erhalten wir, wenn wir als Abstiegsrichtung im Punkt x_0 den negativen Gradienten $p := -\nabla f(x_0)$ wählen. Das Ziel dieses Abschnittes ist, die globale Konvergenz des Gradientenverfahrens für die exakte Schrittweite und die Armijo-Schrittweite zu beweisen. Wir nehmen an, dass die Voraussetzungen 5.9 gelten. Zunächst beweisen wir ein nützliches Lemma, das auch einen Hinweis auf eine Abbruchbedingung gibt.

5.10 Lemma (Kleiner Gradient heißt nahe Optimum)**.** *Es gelten die Voraussetzungen 5.9. Dann ist*

$$f(x) - p^* \leq \frac{\|\nabla f(x)\|_2^2}{2m}.$$

Beweis. Aus den Voraussetzungen und aus (5.2) folgt, dass für alle x, y gilt

$$f(y) \geq f(x) + \langle \nabla f(x) | y - x \rangle + \frac{m}{2}\|y - x\|_2^2.$$

Die rechte Seite ist eine konvexe Funktion in y; die Ungleichung bleibt richtig, wenn wir sie durch ihr Minimum in y ersetzen. Nachrechnen zeigt, dass dies für $\tilde{y} = x - \frac{1}{m}\nabla f(x)$ angenommen wird. Einsetzen von $\tilde{y}$ auf der rechten Seite führt auf

$$f(y) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \text{für alle } x, y.$$

Insbesondere kann man jetzt für y den optimalen Punkt wählen; dies zeigt

$$p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2 \quad \text{für alle } x;$$

das ist die Behauptung. □

Aufgrund des Lemmas ist es sinnvoll, zum Abbruch des Verfahrens vor jedem Schritt $\|\nabla f(x)\|_2^2 \leq 2m\varepsilon$ zu testen. Ist diese Bedingung erfüllt, ist auch $f(x) - p^* \leq \|\nabla f(x)\|_2^2/(2m) \leq \varepsilon$. (Ein praktisches Hindernis dabei ist lediglich, dass die Konstante m der starken Konvexität im Normalfall nicht bekannt sein wird. Trotzdem ist die Aussage qualitativ interessant.)

5.11 Satz (Gradientenverfahren mit exakter Schrittweite)**.** *Sei $(x_0, x_1, \ldots, x_k, \ldots)$ die durch das Gradientenverfahren mit exakter Schrittweite erzeugte Folge von Punkten mit Startwert x_0, so dass die Voraussetzungen 5.9 gelten. Dann gibt es eine Zahl $c := 1 - m/M < 1$, so dass*

$$(f(x_k) - p^*) \leq c^k \cdot (f(x_0) - p^*),$$

d.h. das Verfahren konvergiert linear mit Faktor $c < 1$.

Beweis. Wir betrachten den Übergang von x zu x_+ in einem Iterationsschritt; es sei $\phi(t) := f(x + tp)$ mit $p = -\nabla f(x)$. Zunächst folgt aus den Voraussetzungen und aus (5.2), dass

$$\phi(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2. \tag{5.3}$$

Die obere Schranke wird für $t_M = 1/M$ minimiert, wie man nachrechnet. Für die exakte Schrittweite t^* ist $\phi(t^*) \leq \phi(t)$ für alle $t \geq 0$, also auch

$$\phi(t^*) \leq \phi(t_M) = f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2.$$

Abziehen von p^* liefert eine additive Aussage:

$$(f(x_+) - p^*) \leq (f(x) - p^*) - \frac{1}{2M}\|\nabla f(x)\|_2^2. \tag{5.4}$$

Aus Lemma 5.10 erhält man die Aussage $\|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^*)$; kombiniert man dies mit (5.4), so ergibt sich die Aussage

$$(f(x_+) - p^*) \leq (f(x) - p^*) \cdot \big(1 - \frac{2m}{2M}\big);$$

die Aussage des Satzes folgt durch k-maliges Iterieren. □

Aus Satz 5.11 lässt sich die Anzahl der benötigten Iterationen bestimmen, wenn $\varepsilon_0 := f(x_0) - p^*$ der Anfangsabstand zum Optimum ist und ein Endabstand von höchstens ε gewünscht ist: Es muss $\varepsilon_0 \cdot c^k \leq \varepsilon$ oder

$$k \geq \frac{\log(\varepsilon_0/\varepsilon)}{\log(1/c)}$$

gelten. Dies ist einfach zu interpretieren:

- $\varepsilon_0/\varepsilon$ ist ein Maß für die gewünschte Verbesserung; dies geht nur logarithmisch in den Aufwand ein.
- Es ist $\log(1/c) = -\log(1 - m/M) \approx m/M$, wenn $m/M \ll 1$. Also wächst die Anzahl der benötigten Schritte in diesem Fall mit M/m, d.h. dem Verhältnis von grösstem Eigenwert der Hessematrizen auf der initialen Niveaumenge S und kleinstem Eigenwert der Hessematrizen auf S.

Für die Armijo-Schrittweite mit Parametern α, β ergibt sich eine ganz ähnlich Aussage. In der Praxis wählt man hier häufig $\alpha \in [0.01, 0.3]$ und $\beta \in [0.1, 0.9]$.

5.12 Satz (Gradientenverfahren mit Armijo-Schrittweite)**.** *Sei $(x_0, x_1, \dots, x_k, \dots)$ die durch das Gradientenverfahren mit Armijo-Schrittweite mit $0 < \alpha < 1/2$ und $0 < \beta < 1$ erzeugte Folge von Punkten mit Startwert x_0, so dass die Voraussetzungen 5.9 gelten. Dann gibt es eine Zahl $c := 1 - \min\{2m\alpha, 2m\alpha\beta/M\} < 1$, so dass*

$$f(x_k) - p^* \leq c^k \cdot (f(x_0) - p^*),$$

d.h. das Verfahren konvergiert linear mit Faktor $c < 1$.

Beweis. Wir erinnern daran, dass die Abbruchbedingung für die Armijo-Schrittweite $f(x + tp) \leq f(x) + \alpha t \langle \nabla f(x) | p \rangle$ lautet, hier also $f(x + tp) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$. Wir werden nachweisen, dass für die Schrittweite $t = \min\{1, \beta/M\}$ gilt, also

$$f(x_+) \leq f(x) - \alpha \min\{1, \beta/M\}\|\nabla f(x)\|_2^2 \text{ oder}$$
$$(f(x_+) - p^*) \leq (f(x) - p^*) - \alpha \min\{1, \beta/M\}\|\nabla f(x)\|_2^2$$

analog zu (5.4). Mit Lemma 5.10 folgt dann schon die Behauptung.

Zum Beweis der Schrittweiten-Behauptung weisen wir nach, dass die Armijo-Bedingung auf jeden Fall für $0 \leq t \leq 1/M$ erfüllt ist. Dort gilt nämlich $-t + Mt^2/2 \leq -t/2$. Mit (5.3) folgt daraus

$$\begin{aligned}\phi(t) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \alpha t\|\nabla f(x)\|_2^2,\end{aligned}$$

da $\alpha < 1/2$. Daher terminiert die Armijo-Strategie in jedem Fall (wenn nicht schon initial mit $t = 1$) mit einem $t \geq \beta/M$ (im Schritt davor könnte die Abbruchbedingung gerade noch verletzt gewesen sein). □

Die Aussagen dieses Abschnitts lassen vermuten, dass das Gradientenverfahren für Probleme, deren Variablen sehr unterschiedlich skaliert sind, relativ langsam konvergiert ($c \approx 1$). Dies ist in der Tat der Fall, wie man an einem einfachen Beispiel (z.B. Beispiel 5.29) sehen kann, auch wenn die Abschätzung in diesem Abschnitt insgesamt pessimistisch ist. Das Problem entsteht dadurch, dass die Richtung des negativen Gradienten bei stark verzerrten Höhnelinien sehr weit von der Richtung des Minimums abweichen kann. Das Verfahren ist aber gut, wenn $m \approx M$ gilt, wenn also die Höhenlinien in etwa kreisförmig sind.

5.4 Das Newton-Verfahren

Definition und Eigenschaften. Beim Newton-Verfahren wird als Abstiegsrichtung $p := -[\nabla^2 f(x)]^{-1}\nabla f(x)$ gewählt; eine Motivation dafür wurde oben schon angegeben (Minimierung eines quadratischen Modells von f in x für Schrittweite 1). Eine andere Motivation ergibt sich wie folgt.

Angenommen, wir führen eine Koordinatentransformation auf $\mathbb{R}^n$ durch, und zwar mit einer positiv definiten $n \times n$-Matrix W, indem wir $y := W^{1/2}x$ setzen. (Die Quadratwurzel einer positiv definiten Matrix erhält man zum Beispiel, indem man sie als $W = Q^{-1}DQ$ diagonalisiert mit D diagonal und die Quadratwurzeln aus den Einträgen von D zieht; dann ist $W^{1/2} = Q^{-1}D^{1/2}Q$.) Wir führen nun den Gradientenabstieg nach der Koordinatentransformation durch. Dazu definieren wir $g(y) := f(W^{-1/2}y) = f(x)$. Damit ist $\nabla g(y) = W^{-1/2}\nabla f(W^{-1/2}y) = W^{-1/2}\nabla f(x)$ und $\nabla^2 g(y) = W^{-1/2}\nabla^2 f(W^{-1/2}y)W^{-1/2} = W^{-1/2}\nabla^2 f(x)W^{-1/2}$.

Die Wahl $W := \nabla^2 f(x)$ führt also dazu, dass $\nabla^2 g(y) = \mathrm{Id}$ wird und das Verhältnis von größtem zu kleinstem Eigenwert der Hessematrix Eins ist. Die Konvergenzanalyse des Gradientenverfahrens hat gezeigt, dass es in diesem Fall schnell konvergiert (allerdings muss man der worst case über die ganze initiale Niveaumenge betrachten).

Drückt man die Richtung $\nabla g(y) = W^{-1/2}\nabla f(x)$ wieder im x-Koordinatensystem aus, muss man noch mit $W^{-1/2}$ multiplizieren und bekommt als Abstiegsrichtung $p = -W^{-1}\nabla f(x) = -\nabla^2 f(x)^{-1}\nabla f(x)$. Zusammenfassend kann man daher festhalten, dass man die Newtonrichtung erhält, wenn man das Gradientenverfahren in einem transformierten Koordinatensystem anwendet, dass lokal die Hessematrix zur Einheitsmatrix macht.

Der Vorteil am Newton-Verfahren ist, dass die Folge der berechneten Punkte invariant unter affinen Transformationen ist, wie wir im folgenden Lemma (im wesentlichen) zeigen.

5.13 Lemma. *Sei $T \in \mathbb{R}^{n\times n}$ eine nichtsinguläre Matrix. Sei f konvex und p_x die Newton-Richtung für f in x. Definiere $g(y) := f(Ty) = f(x)$ mit $x = Ty$, und sei q_y die Newton-Richtung für g in y. Dann gilt $x + p_x = T(y + q_y)$.*

Beweis. Es ist $\nabla g(y) = T^T \nabla f(x)$ und $\nabla^2 g(y) = T^T \nabla^2 f(x) T$. Daher ergibt sich die Newtonrichtung q_y als

$$\begin{aligned} q_y &= -(T^T \nabla^2 f(x) T)^{-1} (T^T \nabla f(x)) \\ &= -T^{-1} \nabla^2 f(x)^{-1} \nabla f(x) \\ &= T^{-1} p_x, \end{aligned}$$

also $p_x = T q_y$. Mit $x = Ty$ und der Linearität von T folgt die Behauptung. □

Eine wichtige Rolle bei der Analyse des Newton-Verfahrens spielt das sogenannte Newton-Dekrement.

5.14 Definition (Newton-Dekrement)**.** Zu einem Punkt x einer hinreichend glatten konvexen Funktion f sei

$$\lambda(x) := \sqrt{\langle \nabla f(x) | \nabla^2 f(x)^{-1} | \nabla f(x) \rangle}$$

das sogenannte *Newton-Dekrement*.

Für die Richtungsableitung in Newtonrichtung $p = -\nabla^2 f(x) \nabla f(x)$ gilt: $f'(x; p) = \langle \nabla f(x) | p \rangle = -\lambda(x)^2$; also ist $\lambda(x)$ ein Maß dafür, wie stark f in Newton-Richtung fällt. Es ist genau dann $\lambda(x) = 0$, wenn x optimal ist.

Konvergenzaussage zum Newton-Verfahren. Über das Newton-Verfahren lassen sich für konvexe Funktionen unter den Voraussetzungen 5.9 eine Reihe von Aussagen machen. Leider ergeben sich jedoch keine von m und M unabhängigen Aussagen (was man jetzt vielleicht erwarten könnte). In der Praxis verhält sich das Newton-Verfahren meist besser, als im folgenden Satz angegeben.

5.15 Satz. *Sei $(x_0, x_1, \ldots, x_k, \ldots)$ die durch das Newton-Verfahren mit Armijo-Schrittweite mit $0 < \alpha < 1/2$ und $0 < \beta < 1$ erzeugte Folge von Punkten mit Startwert x_0, so dass die Voraussetzungen 5.9 gelten. Dann gibt es Konstanten $0 < \eta := \min\{1, 3(1-2\alpha)\} m^2/L \leq m^2/L$ und $\gamma := \alpha\beta\eta^2 m/M^2 > 0$, so dass folgendes gilt:*

1. *Ist $\|\nabla f(x_k)\|_2 \geq \eta$, dann ist $f(x_{k+1}) - f(x_k) \leq -\gamma$.*
2. *Ist $\|\nabla f(x_k)\|_2 < \eta$, dann wird in dieser und in allen folgenden Iterationen die Schrittweite $t = 1$ gewählt, und es ist*

$$\frac{L}{2m^2} \|\nabla f(x_{\ell+1})\|_2 \leq \left(\frac{L}{2m^2} \|\nabla f(x_\ell)\|_2 \right)^2$$

für alle $\ell \geq k$ und

$$f(x_\ell) - p^* \leq \frac{2m^3}{L^2} \left(\frac{1}{2}\right)^{2^{\ell-k+1}},$$

d.h. das Verfahren konvergiert ab einer gewissen Iteration k quadratisch.

3. *Eine praktische obere Schranke für die Anzahl der Iterationen ist* $6 + (f(x_0) - p^*)/\gamma = 6 + \frac{M^2L^2/m^5}{\alpha\beta\min\{1,9(1-2\alpha)^2\}}(f(x_0) - p^*)$.

Beweis. Siehe (**?**, Abschnitt 9.5); wir gehen auf die Details nicht ein. □

Der Satz ist insofern unbefriedigend, weil die Konstanten m, M, L, die für die Anwendung meist nicht bekannt sind, auftreten, und wir eigentlich erwarten, dass das Verhalten des Newton-Verfahren nicht (sehr) von M, m abhängen sollte.

Selbst-konkordante Funktionen und ihre Eigenschaften. Eine elegantere Analyse ist möglich, wenn man *selbst-konkordante* Funktionen betrachtet, was wir jetzt tun werden. Damit soll nicht behauptet werden, dass das Newton-Verfahren auf selbst-konkordanten Funktionen besser ist als auf anderen konvexen Funktionen; nur die Analyse ist "schöner". Die Idee dabei ist, die Größe der dritten Ableitung durch die der zweiten Ableitung zu beschränken.

5.16 Definition (Selbst-Konkordanz)**.** Eine konvexe Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *schwach selbst-konkordant*, wenn sie dreimal stetig differenzierbar ist und es ein $k \geq 0$ gibt, so dass für alle $x \in \operatorname{dom} f$ gilt $|f'''(x)| \leq kf''(x)^{3/2}$. Eine schwach selbst-konkordante Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *selbst-konkordant*, wenn für alle $x \in \operatorname{dom} f$ gilt $|f'''(x)| \leq 2f''(x)^{3/2}$ (wenn also die obige Konstante $k = 2$ gewählt werden kann). Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt [schwach] selbst-konkordant, wenn die auf jeder Geraden in ihrem Definitionsbereich [schwach] selbst-konkordant ist.

5.17 Beispiel (Einfache selbst-konkordante Funktionen)**.** Die folgenden Funktionen $f : \mathbb{R} \to \mathbb{R}$ sind selbst-konkordant.

1. konstante, affine und quadratische Funktionen (dritte Ableitung verschwindet).
2. negativer Logarithmus $f(x) = -\log x$ für $x > 0$.
3. $f(x) = x \log x - \log x$

♡

Wir zeigen zunächst, dass jede schwach selbst-konkordante Funktion durch Reskalierung selbst-konkordant gemacht werden kann, dass Selbst-Konkordanz affin-invariant und unter Summen und gewissen Skalierungen abgeschlossen ist.

5.18 Lemma. *Ist f schwach selbst-konkordant mit Konstante $k > 2$, dann ist $g := k^2/4 \cdot f$ selbst-konkordant. (Daher müssen schwach selbst-konkordante Funktionen nicht weiter betrachtet werden.)*

Aufgabe 5.2. Beweise Lemma 5.18.

5.19 Lemma. *Sei $f : \mathbb{R} \to \mathbb{R}$ selbst-konkordant und $g(x) := f(ax + b)$ mit $a, b \in \mathbb{R}$. Dann ist g selbst-konkordant. Sei $f : \mathbb{R}^n \to \mathbb{R}$ selbst-konkordant und $g(x) := f(Ax + b)$ mit einer Matrix A und einem Vektor b. Dann ist g selbst-konkordant.*

5.20 Lemma. *Seien f, g selbst-konkordant und sei $c \geq 1$. Dann sind $f + g$ und cf selbst-konkordant.*

Aufgabe 5.3. Beweise Lemmas 5.19 und 5.20.

5.21 Beispiel (Zusammengesetze selbst-konkordante Funktionen)**.**

1. Betrachte die zu einem Polyeder $\{ x \mid Ax \leq b \}$ gehörenden log-Barriere-Funktion $f(x) := -\sum_{i=1}^{m} \log(b_i - \langle a_i | x \rangle)$. Diese ist selbst-konkordant.
2. Betrachte $f(x) = -\log\det X$ auf $\mathbb{S}^n_{++}$. Diese ist selbst-konkordant. Dies folgt durch Reduzierung auf den Eindimensionalen Fall durch $X = X_0 + tV$ wie in Beispiel 3.29: Seien $X_0 \in \mathbb{S}^n_{++}, V$ beliebig und sei dazu $\phi(t) := f(X_0 + tV)$. Es ist $\phi(t) = -\log\det X_0 - \sum_{i=1}^{n} \log(1 + t\lambda_i)$, wobei λ_i die Eigenwerte von $X_0^{-1/2} V X_0^{-1/2}$ sind. Jeder Summand ist selbst-konkordant in t, also auch ϕ, also f.
3. Betrachte $f(x) = -\log(\langle x|P|x\rangle + \langle q|x\rangle + r)$, wobei P negativ semidefinit sei. Wir zeigen f ist auf $\operatorname{dom} f = \{ x \mid \langle x|P|x\rangle + \langle q|x\rangle + r > 0 \}$ selbst-konkordant. Durch Reduktion auf den eindimensionalen Fall bekommen wir oBdA $f(x) = -\log(px^2 + px + r) - -\log(-p(x-a)(b-x)$ mit Nullstellen a, b und $\operatorname{dom} f = (a, b)$. Damit ist $f(x) = -\log(-p) - \log(x-a) - \log(b-x)$.

♡

Wir halten jetzt einige weitere Eigenschaften von selbst-konkordanten Funktionen fest.

5.22 Lemma (Alternative Charakterisierung selbst-konkordanter Funktionen)**.** *Sei $\phi : \mathbb{R} \to \mathbb{R}$ strikt konvex. Dann ist ϕ genau dann selbst-konkordant, wenn für alle $x \in \operatorname{dom}\phi$ gilt, dass*

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \big(\phi''(t)^{-1/2} \big) \right| \leq 1. \tag{5.5}$$

Beweis. Elementar: Es ist $\frac{\mathrm{d}}{\mathrm{d}t}\big(\phi''(t)^{-1/2}\big) = -1/2\phi''(t)^{-3/2}\phi'''(t)$. Dank der vorausgesetzten strikten Konvexität kann man $f'' > 0$ nutzen. □

5.23 Lemma (Schranken für selbst-konkordante Funktionen)**.** *Sei $\phi : \mathbb{R} \to \mathbb{R}$ strikt konvex und selbst-konkordant. Dann ist*

$$\frac{\phi''(0)}{(1 + t\phi''(0)^{1/2})^2} \leq \phi''(t) \leq \frac{\phi''(0)}{(1 - t\phi''(0)^{1/2})^2}, \tag{5.6}$$

$$\phi''(0)^{1/2} - \frac{\phi''(0)^{1/2}}{1 + t\phi''(0)^{1/2}} \leq \phi'(t) - \phi'(0) \leq -\phi''(0)^{1/2} - \frac{\phi''(0)^{1/2}}{1 - t\phi''(0)^{1/2}}, \tag{5.7}$$

$$t\phi'(0) + t\phi''(0)^{1/2} - \log(1 + t\phi''(0)^{1/2}) \leq \phi(t) - \phi(0) \leq t\phi'(0) - t\phi''(0)^{1/2} - \log(1 - t\phi''(0)^{1/2}). \tag{5.8}$$

Die unteren Schranken gelten für alle $t \geq 0$ in $\operatorname{dom}\phi$; die oberen nur für $0 \leq t < \phi''(0)^{-1/2}$.

Beweis. Integrieren von (5.5) liefert $-t \leq \phi''(t)^{-1/2} - \phi''(0)^{-1/2} \leq t$. Umstellen liefert die Behauptung (5.6). Erneutes Integrieren der unteren Schranke liefert dann (5.7), und abermaliges Integrieren (5.8). □

5.24 Lemma. *Sei f selbst-konkordant und $x \in \operatorname{dom} f$ ein Punkt mit $\lambda(x) \leq 0.68$. Dann ist $f(x) - p^* \leq \lambda(x)^2$. Insbesondere garantiert die Abbruchbedingung $\lambda(x)^2 \varepsilon$ mit einem $\varepsilon \leq 0.68$, dass auch das Minimum bis auf ε genau bestimmt wurde.*

Beweis. Sei p irgendeine Abstiegsrichtung für f in x. Sei $\phi(t) := f(x + tp)$. Es ist $\lambda(x) \geq -\phi''(0)^{-1/2}\phi'(0)$ mit Gleichheit genau dann, wenn p die Newtonrichtung ist. Das folgt daher, dass die Newtonrichtung die steilste Abstiegsrichtung in einem entsprechend transformierten Koordinatensystem ist (siehe oben).

Wir minimieren nun die untere Schranke aus (5.8) für $\phi(t)$. Das Minimum wird angenommen für $t_{\min} = \frac{-\phi'(0)}{\phi''(0)+\phi''(0)^{1/2}\phi'(0)}$, so dass

$$\inf_{t\geq 0} \phi(t) \geq \phi(t_{\min}) = \phi(0) - \phi'(0)\phi''(0)^{-1/2} + \log(1 + \phi'(0)\phi''(0)^{-1/2})$$

folgt. Jetzt zeigt obige Ungleichung für $\lambda(x)$, dass

$$\inf_{t\geq 0} \phi(t) \geq \phi(0) + \lambda(x) + \log(1 - \lambda(x)).$$

Da dies für jedes ϕ zu einer Abstiegsrichtung gilt, gilt es auch für die Richtung, in der der optimale Punkt liegt. Wir erhalten

$$p^* \geq f(x) + \lambda(x) + \log(1 - \lambda(x)).$$

Für $\lambda \leq 0.68$ gilt $-(\lambda + \log(1 - \lambda) \leq \lambda^2$; damit folgt die Behauptung. □

Konvergenz des Newton-Verfahrens für selbst-konkordante Funktionen.

5.25 Satz. *Sei f eine selbst-konkordante Funktion. Sei $(x_0, x_1, \ldots, x_k, \ldots)$ die durch das Newton-Verfahren mit Armijo-Schrittweite mit $0 < \alpha < 1/2$ und $0 < \beta < 1$ erzeugte Folge von Punkten mit Startwert x_0. Dann gibt es Konstanten $0 < \eta := (1 - 2\alpha)/4 < 1/4$ und $\gamma := \alpha\beta\eta^2/(1 + \eta) > 0$, so dass folgendes gilt:*

1. *Ist $\lambda(x_k) > \eta$, dann ist $f(x_{k+1}) - f(x_k) \leq -\gamma$.*
2. *Ist $\lambda(x_k) \leq \eta$, dann wird in dieser und in allen folgenden Iterationen die Schrittweite $t = 1$ gewählt, und es ist*
$$2\lambda(x_{\ell+1}) \leq (2\lambda(x_\ell))^2$$
für alle $\ell \geq k$ und
$$f(x_\ell) - p^* \leq \left(\frac{1}{2}\right)^{2^{\ell-k+1}},$$
d.h. das Verfahren konvergiert ab einer gewissen Iteration k quadratisch.
3. *Eine obere Schranke für die Anzahl der notwendigen Iterationen bis zur Genauigkeit 10^{-10} ist $6 + \frac{20-8\alpha}{\alpha\beta(1-2\alpha)^2}(f(x_0) - p^*)$, etwa $6 + 375(f(x_0) - p^*)$ für $\alpha = 0.1$ und $\beta = 0.8$.*

Beweis. Aussage 1. Wir definieren wieder $\phi(t) := f(x + tp)$ mit der Newtonrichtung $p = -\nabla^2 f(x)^{-1}\nabla f(x)$ für $t \geq 0$; diese Funktion ist konvex und selbst-konkordant.

Wir zeigen, dass unter den gemachten Voraussetzungen die Schrittweite $t_0 := 1/(1 + \lambda(x)$ die Armijo-Bedingungen erfüllt. Es ist $\phi'(0) = -\lambda^2$ (siehe Definition 5.14) und $\phi''(0) = \lambda^2$, wie man nachrechnet. Aus der oberen Schranke (5.8) folgt daher

$$\phi(t) \leq \phi(0) - t\lambda(x)^2 - t\lambda(x) - \log(1 - t\lambda(x)) \quad (0 \leq t < 1/\lambda(x)).$$

Da $t_0 < 1/\lambda(x)$ ist, lässt sich die Ungleichung für t_0 anwenden:

$$\phi(t_0) \leq \phi(0) - \lambda(x) + \log(1 + \lambda(x)).$$

Für $z \geq 0$ ist $\log(1+z) - z \leq \frac{-z^2}{2(1+z)}$. Mit $z = \lambda(x)$ und $\alpha < 1/2$ ergibt sich

$$\phi(t_0) \leq \phi(0) - \alpha\frac{\lambda(x)^2}{1+\lambda(x)} = \phi(0) - \alpha\lambda(x)^2 t_0;$$

damit ist für t_0 die Armijo-Abbruchbedingung erfüllt.

Die vom Armijo-Verfahren gewählte Schrittweite ist also ein $t^* \geq \beta/(1+\lambda(x))$. Damit ist

$$\phi(t^*) - \phi(0) \leq -\alpha\beta\frac{\lambda(x)^2}{1+\lambda(x)}.$$

Ist $\lambda(x) > \eta$, dann ist insbesondere

$$\phi(t^*) - \phi(0) \leq -\alpha\beta\frac{\eta^2}{1+\eta} =: -\gamma.$$

Dies gilt zunächst für alle Wahlen von $\eta > 0$.

Aussage 2. Wir zeigen: Ist $\lambda(x) < (1-2\alpha)/2$, dann wird die Schrittweite 1 akzeptiert. Denn es ist nach der Abschätzung aus Teil 1

$$\phi(1) \leq \phi(0) - \lambda(x)^2 - \lambda(x) - \log(1 - \lambda(x).$$

Da $-z - \log(1-x) \leq z^2/2 + z^3$ für $0 \leq z \leq 0.81$ und $\lambda(x) < 1/2$, folgt

$$\phi(1) \;\leq\; \phi(0) - \lambda^2(1/2 - \lambda(x)) \;\leq\; \phi(0) - \alpha\lambda(x)^2.$$

Damit wird $t^* = 1$ als Armijo-Schrittweite gewählt.

Als nächstes zeigen wir: Ist $\lambda(x) \leq 1/4$, dann ist $\lambda(x^+) \leq 2\lambda(x)^2$. Dabei ist $x^+ := x - \nabla^2 f(x)^{-1}\nabla f(x)$ der nächste Iterationswert.

TODO $\lambda(x^+) \leq \frac{\lambda(x)^2}{(1-\lambda(x))^2}$

Wählen wir daher $\eta := (1-2\alpha)/4$ und ist $\lambda(x) \leq \eta$, so sind beide obigen Aussagen gültig und es ist $2\lambda(x^+) \leq (2\lambda(x))^2 < \eta$, und die Aussage ist rekursiv ab einer gewissen Iteration k anwendbar. Es folgt für alle $\ell \geq k$

$$2\lambda(x_\ell) \leq (2\lambda(x_k))^{2^{\ell-k}} \leq (2\eta)^{2^{\ell-k}} \leq (1/2)^{2^{\ell-k}}.$$

Aus Lemma 5.24 folgt für alle $\ell \geq k$ dann

$$f(x_l) - p^* \leq \lambda(x_\ell)^2 \leq \frac{1}{4}\left(\frac{1}{2}\right)^{2^{\ell-k+1}} \leq \left(\frac{1}{2}\right)^{2^{\ell-k+1}},$$

wie behauptet.

Aussage 3. Da sich der Funktionswert anfangs ($\lambda(x) > \eta$) in jeder Iteration um γ verbessert, ist man nach spätestens $k := (f(x_0) - p^*)/\gamma$ Iterationen entweder im Optimum, oder es ist $\lambda(x) \leq \eta$. In dieser Phase gilt laut Aussage 2, dass zu gegebenem $\varepsilon > 0$ in jedem Fall nach $\ell - k \geq \log_2\log_2(1/\varepsilon)$ Iterationen gilt, dass $f(x_\ell) - p^* \leq \varepsilon$. Für $\varepsilon = 10^{-10}$ lässt sich dies durch die Konstante 6 abschätzen. Insgesamt werden also höchstens $6 + (f(x_0) - p^*)/\gamma$ Iterationen benötigt. □

5.5 Das BFGS-Verfahren als Quasi-Newton-Verfahren

Das Newton-Verfahren verfügt zwar über lokal quadratische Konverzenz (d.h., ist man nahe genug an der Lösung x^*, dann wird in jeder Iteration der schon kleine Fehler im Zielfunktionswert $f(x_k) - p^*$ sogar quadriert); allerdings ist es sehr aufwändig: In jedem Schritt muss die $n \times n$ Hesse-Matrix $\nabla^2 f(x_k)$ zunächst beschafft werden, um dann das Gleichungssystem $\nabla^2 f(x_k) \cdot p = -\nabla f(x_k)$ zu lösen. Im Normalfall bedeutet dies einen Aufwand von $O(n^3)$ Rechenoperationen pro Iteration (der ggf. auch höher sein kann; dies hängt davon ab, wie $\nabla^2 f(x)$ berechnet werden kann). Das Gradientenverfahren ist dagegen einfach, es benötigt nur den n-Vektor $\nabla f(x_k)$ und muss kein Gleichungssystem lösen. Daher reicht hier normalerweise in jedem Iterationsschritt ein Aufwand von $O(n^2)$ Operationen aus. Im Gegenzug erhält man im nur lineare Konvergenz, die schleichend langsam verläuft, wenn die Hessematrix der Zielfunktion auf der initialen Niveaumenge über stark unterschiedlich große Eigenwerte verfügt.

Es stellt sich die Frage, ob man nicht die Vorzüge beider Verfahren kombinieren kann. Eine Idee ist, statt der exakten Hesse-Matrix in jedem Schritt eine Approximation zu verwenden, die aufgrund der Funktions-Informationen, die man in jedem Schritt berechnet, aktualisiert wird. Solange man sicherstellen kann, dass die Approximation stets positiv definit ist, lässt sich zumindest der nächste Iterationswert eindeutig als Minimum der so gebildeten Modellfunktion berechnen.

Es stellen sich folgende offensichtliche Fragen an solch ein Verfahren: Wie wählt man die initiale Approximation? Wie wird in jedem Schritt die Aktualisierung effizient durchgeführt? Wie vermeidet man $O(n^3)$ Aufwand zum Lösen eines Gleichungssystems zur Bestimmung der Abstiegsrichtung? Wie ist das globale und lokale Konvergenzverhalten?

Seit den 1950er Jahren sind sehr viele Verfahren erfunden worden, die auf die oben stehenden Fragen verschiedene Antworten geben. Wir stellen hier eingangs das DFP-Verfahren (nach Davidon, Fletcher, Powell) und insbesondere das verwandte BFGS-Verfahren (nach Broyden, Fletcher, Goldfarb, Shanno) vor. Für das letztere lässt sich globale superlineare Konvergenz beweisen, und es verhält sich robust in der Praxis bei in der Regel $O(n^2)$ Aufwand pro Iteration. Das BFGS-Verfahren ist daher oft die Methode der Wahl in der Praxis.

Wir verwenden in diesem Abschnitt folgende Bezeichnungen: $x_0, x_1, \ldots$ ist die Folge der Lösungs-Kandidaten. Sei $f_k := f(x_k)$ und $\nabla f_k := \nabla f(x_k)$, sowie B_k eine positiv definite Matrix, die wir als Approximation an die Hesse-Matrix $\nabla^2 f(x_k)$ auffassen.

Herleitung des DFP- und BFGS-Verfahrens. Wir bilden für $k = 0, 1, 2, \ldots$ die Modellfunktionen

$$m_k(p) := f_k + \langle \nabla f_k | p \rangle + \frac{1}{2} \langle p | B_k | p \rangle \approx f(x_k + p)$$
$$\text{mit } \nabla m_k(p) = \nabla f_k + B_k \cdot p.$$

Offensichtlich ist

$$m_k(0) = f_k, \quad \nabla m_k(0) = \nabla f_k,$$

d.h. Funktionswert und Ableitung des Modells in Null stimmen mit Funktionswert und Ableitung des "Originals" in x_k überein.

Wie wir leicht nachrechnen können, wird das eindeutige globale Minimum (B_k wurde ja als positiv definit vorausgesetzt) für $p = -B_k^{-1} \cdot \nabla f_k$ angenommen; dort ist $\nabla m_k(p) = 0$. Da B_k positiv definit ist, ist p insbesondere eine Abstiegsrichtung für f im Punkt x_k.

Der Folgepunkt x_{k+1} ergibt sich dann als

$$x_{k+1} = x_k + t_k \cdot B_k^{-1} \cdot \nabla f_k$$

mit der Armijo-Schrittweite t_k (zu gegebenen Parametern $0 < \alpha < 1/2$ und $0 < \beta < 1$. Wir bilden die nächste Modellfunktion m_{k+1} entsprechend.

Welche Information können wir über $\nabla^2 f_{k+1}$ in dem gerade gemachten Schritt gewinnen? Wir können sinnvollerweise fordern, dass am alten Punkt der Gradient der Modellfunktion wieder mit ∇f_k übereinstimmt. Dazu müssen wir den Schritt $t_k p_k$, den wir gerade gemacht haben, in m_{k+1} zurück gehen und erhalten die Bedingung

$$\nabla m_{k+1}(-t_k p_k) = \nabla f_k, \quad \text{also } \nabla f_{k+1} - t_k B_{k+1} p_k = \nabla f_k.$$

Setzen wir $y_k := \nabla f_{k+1} - \nabla f_k$ als Differenz der Gradienten und $s_k := x_{k+1} - x_k = t_k p_k$ als Differenz der x-Werte, lautet die Bedingung

$$y_k = B_{k+1} s_k \quad (\text{“Sekantengleichung”}). \tag{5.9}$$

Im Eindimensionalen wäre dies äquivalent zu $f'(x_{k+1}) - f'(x_k) = b \cdot (x_{k+1} - x_k)$, womit man durch den Differenzenquotienten klar erkennen kann, dass mit b eine Approximation an die zweite Ableitung irgendwo zwischen x_k und x_{k+1} berechnet wird.

Es stellt sich die Frage: Gibt es überhaupt Matrizen B_{k+1}, die die Sekantengleichung (5.9) erfüllen? Wenn ja, gibt es darunter eine besonders ausgezeichnete?

Da ja B_{k+1} positiv definit sein soll, erhält man durch Multiplizieren mit s_k von links, dass

$$\langle s_k | y_k \rangle = \langle s_k | B_{k+1} | s_k \rangle > 0$$

notwendig ist, sofern nicht $s_k = 0$ und damit $x_{k+1} = x_k$ gilt und damit ein optimaler Punkt gefunden wäre.

In der Tat ist die Bedingung $\langle s_k | y_k \rangle = (x_{k+1} - x_k) \cdot (\nabla f_{k+1} - \nabla f_k) \overset{!}{>} 0$ für alle *strikt* konvexen Funktionen f wegen der Konvexitätscharakterisierung über die Monotonie des Gradienten (Satz 3.21) erfüllt. (Für nicht strikt konvexe Funktionen kann man die Bedingung durch weitere Bedingungen an die Schrittweiten-Strategie sicherstellen.)

Leider bestimmt die Sekantengleichung (5.9) die Matrix B_{k+1} aber nicht eindeutig. Um ein konkretes Verfahren zu bekommen, muss man daher weitere Bedingungen vorgeben, z.B. verlangen, dass B_{k+1} sich möglichst wenig von der vorigen Iteration B_k unterscheidet (bezüglich irgend einer Norm). Wir finden also B_{k+1}, indem wir in B das Optimierungsproblem

$$\min_{B \text{ symmetrisch}} \|B_k - B\|, \quad \text{so dass } y_k = B s_k$$

lösen. Die Wahl der Norm bestimmt, welches Verfahren wir erhalten. Sinnvoll ist, dass die gewählte Norm invariant unter affinen Transformationen der Variablen ist.

Durch eine geeignete Norm-Wahl erhält man das DFP-Verfahren mit

$$B_{k+1} = \big(\mathrm{Id} - \rho_k \left|y_k\right\rangle \left\langle s_k\right|\big) \cdot B_k \cdot \big(\mathrm{Id} - \rho_k \left|s_k\right\rangle \left\langle y_k\right|\big) \; + \; \rho_k \left|y_k\right\rangle \left\langle s_k\right| \qquad \text{(DFP)}$$

mit $\rho_k := 1/\left\langle y_k | s_k \right\rangle > 0$ (hier ohne Details und Beweis).

Das Problem ist, dass wir nicht B_{k+1}, sondern die Inverse $H_{k+1} := B_{k+1}^{-1}$ benötigen, um $B_{k+1} \cdot p = \nabla f_k$ zu lösen. Eine elegante Lösung, dieses Problem nicht zu lösen, aber zu umgehen, ist, das Optimierungsproblem direkt für H_k formulieren statt für B_k. Dies führt zum BFGS-Verfahren.

Wir schreiben also die Sekantengleichung invers auf, nämlich als

$$s_k = H_{k+1} y_k, \tag{5.10}$$

und lösen das Optimierungsproblem

$$\min_{H \text{ symmetrisch}} \|H_k - H\|, \qquad \text{so dass } s_k = H y_k$$

mit der Lösung (bezüglich der selben geeignet gewählten Norm wie oben)

$$H_{k+1} = \big(\mathrm{Id} - \rho_k \left|s_k\right\rangle \left\langle y_k\right|\big) \cdot H_k \cdot \big(\mathrm{Id} - \rho_k \left|y_k\right\rangle \left\langle s_k\right|\big) \; + \; \rho_k \left|s_k\right\rangle \left\langle y_k\right| \qquad \text{(BFGS)}$$

mit $\rho_k := 1/\left\langle y_k | s_k \right\rangle > 0$. Diese Aktualisierungsformel ergibt sich formal auch, wenn man in der DFP-Formel s_k mit y_k und H_k mit B_k vertauscht. Sie lässt sich mit $O(n^2)$ Operationen durchführen.

Woher bekommt man jeweils die inverse Matrix, also H_{k+1} im DFP-Verfahren (die benötigt wird) bzw. B_{k+1} im BFGS-Verfahren (die nicht notwendig ist)? Die Sherman-Morrison-Woodbury-Formel erlaubt, aus einer Update-Formel für eine Matrix A eine Update-Formel für A^{-1} zu bekommen.

5.26 Lemma (Sherman-Morrison-Woodbury)**.** *Sei A eine invertierbare $n \times n$-Matrix. Sei $A' := A + UV^T$ mit $n \times p$-Matrizen U, V (Rang-p-Update). Dann ist A'^{-1} genau dann invertierbar, wenn die $p \times p$-Matrix $\mathrm{Id} + V^T A^{-1} U$ invertierbar ist; in diesem Fall ist*

$$A'^{-1} = A^{-1} - A^{-1}U(\mathrm{Id} + V^T A^{-1} U)^{-1} V^T A^{-1}.$$

Beweis. Nachrechnen. Siehe auch Lehrbücher über Matrix-Algebra. □

Die Bedeutung dieser Formel liegt darin, dass zur Berechnung von A'^{-1} nur eine $p \times p$-Matrix invertiert werden muss. Ist $p = O(1)$, lässt sich das gesamte Update mit $O(n^2)$ Operationen durchführen, wenn A^{-1} bereits vorliegt. Es ergeben sich folgende Formeln:

$$H_{k+1} = H_k - \frac{H_k \left|y_k\right\rangle \left\langle y_k\right| H_k}{\left\langle y_k | H_k | y_k \right\rangle} + \rho_k \left|s_k\right\rangle \left\langle s_k\right|, \qquad \text{(DFP)}$$

$$B_{k+1} = B_k - \frac{B_k \left|s_k\right\rangle \left\langle s_k\right| B_k}{\left\langle s_k | B_k | s_k \right\rangle} + \rho_k \left|y_k\right\rangle \left\langle y_k\right|. \qquad \text{(BFGS)}$$

Wir beweisen folgende Aussage.

Eingabe: Berechnungsvorschrift für ∇f, Startpunkt x_0, Initialisierung H_0 für inverse Hessematrix in x_0, Toleranz $\varepsilon > 0$, Schrittweitenparameter $0 < \alpha < 1/2$, $0 < \beta < 1$.
Ausgabe: Approximativer optimaler Punkt x mit $\|\nabla f(x)\| \leq \varepsilon$.

- $k = 0$
- Solange $\|\nabla f_k\| > \varepsilon$:
  - Berechne Richtung $p_k = -H_k \cdot \nabla f_k$.
  - Berechne Schrittweite t_k mit Armijo-Strategie zu α, β.
  - Setze $x_{k+1} = x_k + t_k p_k$.
  - Berechne H_{k+1} laut (BFGS) aus $s_k = x_{k+1} - x_k$ und $y_k = \nabla f_{k+1} - \nabla f_k$.
  - $k = k + 1$
- Gib x_k zurück.

Tabelle 5.1: BFGS-Verfahren mit Armijo-Schrittweite (Backtracking-Schrittweite)

5.27 Satz. *Ist im BFGS-Verfahren die initiale Matrix H_0 positiv definit, dann auch alle H_k für $k \geq 1$.*

Beweis. Wir rechnen nach, dass $\langle z|H_{k+1}|z\rangle > 0$ für alle $z \neq 0$: Es ist

$$\langle z|H_{k+1}|z\rangle = \langle w|H_k|w\rangle + \rho_k \langle z|s_k\rangle^2$$

mit $w = z - \rho_k y_k \langle s_k|z\rangle$ (nachrechnen!). Da $\rho_k > 0$ und H_k positiv definit ist, ist der erste Term immer nichtnegativ. Jetzt unterscheiden wir zwei Fälle. Entweder es ist $\langle s_k|z\rangle > 0$, dann ist der zweite Term positiv und damit die ganze Summe. Ist aber $\langle s_k|z\rangle = 0$, dann ist $w = z$ und damit $w \neq 0$ und damit der erste Term positiv und damit die Summe. □

Zusammenfassung des BFGS-Algorithmus. Zur Initialisierung von H_0 wird eine möglichst einfache positiv definite Matrix gewählt, etwa $H_0 := \beta\mathrm{Id}$ mit geeignetem Skalierungsfaktor β, der einen Eigenwert von $\nabla^2 f(x_0)$ approximieren sollte. Algorithmus 5.1 zeigt das Verfahren insgesamt. Es ist zu beachten, dass alle Aktualisierungs-Operationen (ohne Funktions- und Gradientenauswertungen) in $O(n^2)$ Zeit implementiert werden können; insbesondere wird keine komplette Matrix-Multiplikation benötigt. Die Abbruchbedingung ist wegen Lemma 5.10 sinnvoll: Ist der Gradient klein, ist man nahe am Ziel.

Konvergenzverhalten. Wir führen keine genaue Analyse des BFGS-Verfahrens durch. Es lässt sich aber Folgendes zeigen.

5.28 Satz. *Wenn die Voraussetzungen von Satz 5.15 (der analoge Satz für das Newton-Verfahren) erfüllt sind, dann konvergiert das BFGS-Verfahren global (d.h. für jeden Startwert x_0). Es konvergiert lokal (d.h. hinreichend nahe am Optimum) superlinear.*

Im Gegensatz zum Newtonverfahren kann man lokal quadratische Konvergenz nicht beweisen, aber man kann mehr (superlineare statt lineare Konvergenz) als beim Gradientenverfahren beweisen. Vom DFP-Verfahren konnte bisher keine globale Konvergenz bewiesen werden. Hinzu kommt, dass sich das BFGS-Verfahren in der Praxis als robust bewährt hat (d.h. es gleicht auftretende numerische Fehler in den nächsten Schritten häufig wieder aus, so dass sich die Fehler nicht aufschaukeln).

5.6 Die Verfahren im Vergleich

Wir fassen die wichtigsten Ergebnisse noch einmal zusammen. Wir setzen voraus, dass die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ stark konvex ist und dass auf der initialen Niveaumenge $\{ x \mid f(x) \leq f(x_0) \}$ die Eigenwerte der Hessematrizen in $[m, M]$ liegen mit $0 < m \leq M$. Wir setzen weiter voraus, dass stets die exakte Schrittweite oder die Armijo-Schrittweite zu Parametern $0 < \alpha < 1/2$ und $0 < \beta < 1$ gewählt wird.

- Das Gradientenverfahren wählt als Abstiegsrichtung $-\nabla f(x)$. Es konvergiert global linear mit einem Faktor, der höchstens $1 - m/M$ ist. Diese Abschätzung ist realistisch; insbesondere ist das Gradientenverfahren für schlecht konditionierte Probleme ($m/M \ll 1$) langsam.
- Das Newtonverfahren wählt als Abstiegsrichtung $-\nabla^2 f(x)^{-1} \nabla f(x)$. Es konvergiert (mit exakter oder mit Armijo-Schrittweite) global. Es konvergiert lokal quadratisch schnell. Das Newton-Verfahren ist invariant gegenüber affinen Koordinatentransformationen. Eine elegante Analyse ist insbesondere für selbst-konkordante Funktionen möglich. Obwohl in der Regel wenige Iterationen benötigt werden, ist das Newton-Verfahren in der Praxis weniger geeignet, da die Berechnung und Invertierung von $\nabla^2 f(x)$ aufwändig ist.
- Das BFGS-Verfahren wählt als Abstiegsrichtung $-H \nabla f(x)$ mit einer positiv definiten Matrix H, die eine Approximation an $\nabla^2 f(x)^{-1}$ darstellt und im Lauf der Iteration gewonnen wird. Man kann zeigen, dass das BFGS-Verfahren global konvergiert **?** und lokal superlinear schnell konvergiert **?**. Da keine Hesse-Matrizen berechnet werden müssen, ist jeder Schritt des BFGS-Verfahren effizient durchführbar. In der Praxis zeichnet sich das BFGS-Verfahren auch durch numerische Robustheit aus; es ist daher oft die Methode der Wahl.

5.29 Beispiel (Zweidimensionale quadratische Funktion)**.** Betrachte die Funktion $f(x) = x_1^2 + \gamma x_2^2$ auf $\mathbb{R}^2$ mit $\gamma > 1$. Offensichtlich liegt ihr Minimum in $(0, 0)$ mit dem Wert $p^* = f(0, 0) = 0$. Starte jedes der in diesem Kapitel diskutierten Verfahren in $(\gamma, 1)$. Beobachte das Konvergenzverhalten der Verfahren für $\gamma \approx 1$ und $\gamma \gg 1$ (entweder durch analytisches Ausrechnen oder numerisch), und gib eine Interpretation der Beobachtungen an.

Hinweis: Für das exakte Gradientenverfahren ergibt sich die k-te Iteration

$$x^{(k)} = \left(\gamma \Big(\frac{\gamma - 1}{\gamma + 1} \Big)^k, \Big(- \frac{\gamma - 1}{\gamma + 1} \Big)^k \right), \quad f(x^{(k)}) = \Big(\frac{\gamma - 1}{\gamma + 1} \Big)^{2k} \cdot f(x^{(0)}).$$

♡

KAPITEL 6

Dualität

Wir betrachten wieder das allgemeine (nicht notwendig konvexe!) Optimierungsproblem nach Definition 4.1 mit Variablen $x \in \mathbb{R}^n$:

$$\begin{aligned} &\text{Minimiere } f_0(x) && && (6.1)\\ &\quad \text{so dass } f_i(x) \leq 0 && \text{für } i = 1, \ldots, m,\\ &\qquad\quad h_i(x) = 0 && \text{für } i = 1, \ldots, p. \end{aligned}$$

Wir setzen voraus, dass die Definitionsmenge $\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom} f_i \cap \bigcap_{i=i}^{p} \operatorname{dom} h_i$ nicht leer ist; die zulässige Menge sei $\mathcal{F}$, der optimale Wert sei p^* (dabei ist $-\infty$ möglich).

Die Grundidee in diesem Kapitel ist, die Nebenbedingungen mit Gewichten (λ, ν) in die Zielfunktion einzubauen, um so die *Lagrange-Funktion* zu erhalten; die Gewichte heißen dann *Langrange'sche Multiplikatoren*. Minimiert man die Lagrange-Funktion für feste Gewichte über alle x, erhält man die *duale Funktion* zum Optimierungsproblem; diese ist eine Funktion der Gewichte (λ, ν), die auch *duale Variablen* genannt werden.

Wir werden sehen, dass (für geeignete Gewichte) die duale Funktion stets eine untere Schranke $d \leq p^*$ liefert. Es macht daher Sinn, nach der *besten* unteren Schranke zu fragen, also die duale Funktion (unter geeigneten Einschränkungen an die Gewichte) zu maximieren. Dies führt auf das *duale Problem* zum gegebenen Optimierungsproblem mit optimalem Wert d^*. Während stets $d^* \leq p^*$ gilt, ist eine wichtige Frage, unter welchen Bedingungen Gleichheit gilt, wann also eine Lösung des dualen Problems als (im wesentlichen) gleichbedeutend mit einer Lösung des ursprünglichen (oder auch primalen) Problems angesehen werden kann. Bei konvexen Problemen ist dies unter gewissen Bedingungen der Fall. Damit lassen sich geeignete Optimalitätsbedingungen in den primalen und dualen Variablen (x, λ, ν) formulieren, die *KKT-Bedingungen* (nach Karush-Kuhn-Tucker).

Verschiedene Beispiele werden die Nützlichkeit dieses Ansatzes verdeutlichen.

6.1 Lagrange-Funktion, duale Funktion und duales Problem

6.1 Definition (Lagrange-Funktion)**.** Zum Problem (6.1) definieren wir die *Lagrange-Funktion* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ mit $\operatorname{dom} L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$ durch

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x).$$

Wir nennen die Variable λ_i den Lagrange'schen Multiplikator zur i-ten Ungleichung $f_i(x) \leq 0$ und ν_i den Lagrange'schen Multiplikator zur i-ten Gleichung $h_i(x) = 0$. Zusammen heißen (λ, ν) auch *duale Variablen*. Dabei wird x auch *primale Variable* genannt.

Wir sehen: Für jedes x ist L affin in (λ, ν). Ist das Problem konvex, dann ist L auch konvex in x für alle $\lambda \geq 0$ und alle $\nu \in \mathbb{R}^p$.

6.2 Definition (Duale Funktion)**.** Zum Problem (6.1) definieren wir die *duale Funktion* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ mit $\operatorname{dom} g = \mathbb{R}^m \times \mathbb{R}^p$ durch

$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu),$$

also durch Minimierung der Lagrange-Funktion über x für festes (λ, ν); dabei ist der Wert $-\infty$ möglich, aber es ist $\operatorname{dom} g = \{\, (\lambda, \nu) \mid g(\lambda, \nu) \text{ endlich} \,\}$.

Die duale Funktion hat zwei wichtige Eigenschaften: Sie ist konkav, und sie liefert untere Schranken für den primalen optimalen Wert p^*.

6.3 Lemma (Konkavität der dualen Funktion)**.** *Die duale Funktion ist stets konkav (auch wenn das Originalproblem nicht konvex ist).*

Beweis. Die Lagrange-Funktion $L(x, \lambda, \nu)$ ist nach Definition affin in (λ, ν) für jedes x. Die duale Funktion ist daher ein punktweises Infimum (über x) affiner Funktionen und damit konkav. □

6.4 Lemma (Untere Schranken durch duale Funktion)**.** *Sei p^* der optimale Wert des Problems 6.1 und $g(\lambda, \nu)$ die duale Funktion dazu. Dann gilt für* alle nichtnegativen $\lambda \in \mathbb{R}^m$ *und* alle $\nu \in \mathbb{R}^p$

$$g(\lambda, \nu) \leq p^*.$$

Beweis. Ist die zulässige Menge $\mathcal{F}$ leer, dann ist $p^* = +\infty$ und die Ungleichung gilt trivialerweise. Sei $x \in \mathcal{F}$ ein zulässiger Punkt, so dass also $f_i(x) \leq 0$ für alle i und $h_i(x) = 0$ für alle i. Da $\lambda \geq 0$, ist $\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \leq 0$ und damit $L(x, \lambda, \nu) \leq f_0(x)$. Insbesondere gilt dies auch, wenn man auf der linken Seite das Infimum anwendet: $g(\lambda, \nu) \leq f_0(x)$. Dies gilt für alle zulässigen x, insbesondere auch für den optimalen Punkt x^* mit $f_0(x^*) = p^*$. □

6.5 Definition (dual zulässig)**.** Ein Punkt $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^p$ heißt *dual zulässig*, wenn $\lambda \geq 0$ und $g(\lambda, \nu) \neq -\infty$ (wenn also $(\lambda, \nu) \in \operatorname{dom} g$). Wir bezeichnen die dual zulässige Menge mit $\mathcal{G}$.

Die Frage nach der *besten* unteren Schranke, die man so erhalten kann, führt auf das duale Problem.

6.6 Definition (duales Problem)**.** Zum primalen Problem (6.1) sei $g(\lambda, \nu)$ die duale Funktion. Das Problem

$$\begin{aligned} &\text{Maximiere } g(\lambda, \nu) && && (6.2)\\ &\quad \text{so dass } \lambda_i \geq 0 && \text{für } i = 1, \ldots, m, \\ &\qquad\quad g(\lambda, \nu) > -\infty \end{aligned}$$

heißt *duales Problem* zu (6.1).

Die Nebenbedingung $g(\lambda, \nu) > -\infty$ verdient eine Bemerkung: Typischerweise nimmt die duale Funktion $g(\lambda, \nu)$ für gewisse Punkte (λ, ν) den Wert $-\infty$ an; diese Punkte kommen (wegen der Formulierung als Maximierungsproblem) ohnehin nicht als optimale Punkte in Frage, so dass man sich die Bedingung im Grunde sparen kann. Allerdings ist es häufig nützlich und wichtig, die Punkte, die tatsächlich zulässig sind, zu charakterisieren, und die Bedingung explizit hinzuschreiben.

6.7 Lemma (Konvexität des dualen Problems)**.** *Das zu einem beliebigen Optimierungsproblem 6.1 gehörige duale Problem ist ein konvexes Optimierungsproblem.*

Beweis. Die duale Funktion ist konkav und soll maximiert werden. Die explizite Nebenbedingung $\lambda \geq 0$ ist affin. Die Menge der Werte ungleich $-\infty$ einer konkaven Funktion ist konvex. □

Die Lösung des dualen Problems liefert die mit dem gemachten Ansatz bestmögliche untere Schranke für das primale Problem.

6.8 Satz (Schwacher Dualitätssatz)**.** *Sei* (6.1) *ein primales Problem mit optimalem Wert* p^**. Sei* (6.2) *das zugehörige duale Problem mit optimalem Wert* d^**. Dann ist* $d^* \leq p^*$*. Diese Beziehung nennt man auch* schwache Dualität*.*

Beweis. Nach Lemma 6.4 ist $g(\lambda, \nu) \leq p^*$ für alle (λ, ν), also auch für das Supremum über alle $\lambda \geq 0$. □

Von Interesse wird die Frage sein, wann sogar $d^* = p^*$ (starke Dualität) gilt. Bevor wir uns dieser Frage widmen, betrachten wir aber einige Beispiele für duale Probleme.

6.2 Einfache Beispiele

6.9 Beispiel (Lösung minimaler Norm eines linearen Gleichungssystems)**.** Das lineare Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{p \times n}$ sei unterbestimmt und lösbar. Wir setzen voraus, dass die Matrix A vollen Rang $p < n$ hat. Gesucht ist die Lösung minimaler euklidischer Norm. Dies lässt sich als quadratisches Problem wie folgt formulieren:

$$\begin{aligned} &\text{Minimiere } \langle x | x \rangle \\ &\quad \text{so dass } Ax = b. \end{aligned}$$

In der Lagrange-Funktion taucht λ nicht auf, da nur p Gleichungen $Ax = b$ vorliegen. Es ist

$$L(x, \nu) = \langle x|x\rangle + \langle \nu|Ax - b\rangle \,,$$

diese ist eine konvexe quadratische Funktion in x und lässt sich durch Ableiten explizit in x minimieren. Es ist $\nabla_x L(x, \nu) = 2x + A^T\nu$, so dass $x = -(1/2)A^T\nu$ der Minimierer ist. Daher ist

$$g(\nu) = -(1/4)\,\langle \nu|AA^T|\nu\rangle - \langle b|\nu\rangle \,,$$

eine konkave quadratische Funktion auf $\mathbb{R}^p$, denn $-AA^T$ ist negativ semidefinit. *Jede* Wahl von $\nu \in \mathbb{R}^p$ liefert mit $g(\nu)$ eine untere Schranke für die Lösung des primalen Problems. Das duale Problem ist das unrestringierte konvexe Problem

$$\text{Maximiere } -(1/4)\,\langle \nu|AA^T|\nu\rangle - \langle b|\nu\rangle \,,$$

das sich explizit lösen lässt mit Wert $d^* = -2(AA^T)^{-1}b$. Es stellt sich die Frage, ob $p^* = d^*$. Wir werden sehen, dass das hier zutrifft. ♡

6.10 Beispiel (Lineares Programm in Standard-Form)**.** Wir betrachten das Standard-LP

$$\begin{aligned} &\text{Minimiere } \langle c|x\rangle \\ &\text{so dass } Ax = b, \\ &\qquad x \geq 0. \end{aligned}$$

Hier ist die Anzahl der Ungleichungen m gleich der Anzahl der Variablen n und $f_i(x) = -x_i$ für $i = 1, \ldots, n$, und wieder $A \in \mathbb{R}^{p\times n}$. Die Lagrange-Funktion lautet

$$L(x, \lambda, \nu) = \langle c|x\rangle - \langle \lambda|x\rangle + \langle \nu|Ax - b\rangle = -\langle b|\nu\rangle + \langle c + A^T\nu - \lambda|x\rangle$$

Die duale Funktion ist

$$g(\lambda, \nu) = -\langle b|\nu\rangle + \inf_x \langle c + A^T\nu - \lambda|x\rangle \,.$$

Die zu minimierende Funktion ist linear in x. Da eine lineare Funktion nur beschränkt ist, wenn ihr Koeffizient Null ist, folgt

$$g(\lambda, \nu) = \begin{cases} -\langle b|\nu\rangle & \text{wenn } A^T\nu - \lambda + c = 0, \\ -\infty & \text{sonst.} \end{cases}$$

Das duale Problem ist daher formal

$$\begin{aligned} &\text{Maximiere } g(\lambda, \nu) = \begin{cases} -\langle b|\nu\rangle & \text{wenn } A^T\nu - \lambda + c = 0, \\ -\infty & \text{sonst,} \end{cases} \\ &\text{so dass } \lambda \geq 0. \end{aligned}$$

Indem man die Bedingung explizit als $\lambda = A^T\nu + c$ schreibt, erhält man das äquivalente Problem

$$\begin{aligned} &\text{Maximiere } -\langle b|\nu\rangle \,, \\ &\text{so dass } \lambda = A^T\nu + c, \\ &\qquad \lambda \geq 0, \end{aligned}$$

Darin kann man λ komplett eliminieren und das Problem wiederum äquivalent schreiben als

$$\begin{aligned} &\text{Maximiere } -\langle b|\nu\rangle \\ &\text{so dass } A^T\nu + c \geq 0. \end{aligned}$$

Dies ist wiederum ein lineares Programm. Während das primale Problem nur Gleichungs- und Positivitäts-Bedingungen hatte, hat das duale Problem nur Ungleichungs-Bedingungen.

Bei der letzten Version handelt es sich formal nicht um das eigentliche duale Problem (da λ eliminiert wurde). Da es jedoch in offensichtlicher Weise äquivalent zum dualen Problem ist, wird es auch als "das duale Problem" zum Standard-LP bezeichnet. Im allgemeinen kann es schwierig sein, festzustellen, ob zwei Optimierungsprobleme äquivalent sind. ♡

6.11 Beispiel (LP in Ungleichungs-Form)**.** Wir betrachten das LP in Ungleichungs-Form

$$\begin{aligned} &\text{Minimiere } \langle c|x\rangle \\ &\text{so dass } Ax \leq b. \end{aligned}$$

Die Lagrange-Funktion lautet

$$L(x,\lambda,\nu) \;=\; \langle c|x\rangle + \langle \lambda|Ax-b\rangle \;=\; -\langle b|\lambda\rangle + \langle A^T\lambda + c|x\rangle$$

Wie eben ergibt sich als duale Funktion

$$g(\lambda) = \begin{cases} -\langle b|\lambda\rangle & \text{wenn } A^T\lambda + c = 0, \\ -\infty & \text{sonst.} \end{cases}$$

Das duale Problem (in expliziter Form) lautet damit

$$\begin{aligned} &\text{Maximiere } -\langle b|\lambda\rangle\,, \\ &\text{so dass } A^T\lambda + c = 0, \\ &\lambda \geq 0. \end{aligned}$$

Dies ist ein LP in Standard-Form (nur Gleichungen und Positivität). ♡

Aufgabe 6.1. Formuliere das eben hergeleitete duale Problem zum LP in Ungleichungs-Form als Minimierungs-Problem. Bilde dazu wieder das duale Problem. Zeige, dass es zum ursprünglichen LP in Ungleichungs-Form äquivalent ist. Warum folgt hieraus nicht die starke Dualität in allen Fällen?

6.12 Beispiel (Bipartitionierungsproblem)**.** Wir betrachten das folgende *diskrete* Optimierungsproblem: Es sollen n Gegenstände in zwei Klassen $+1$ und -1 eingeteilt werden. Sind Gegenstand i und j in derselben Klasse, entstehen Kosten $2w_{ij}$, sind sie in verschiedenen Klassen entstehen Kosten $-2w_{ij}$. Selbtredend wollen wir die Kosten minimieren. Wir modellieren die Klassenzugehörigkeit durch einen Vektor $x = (x_1,\ldots,x_n) \in \{\pm 1\}^n$. Die Gesamtkosten sind dann durch $\langle x|W|x\rangle$ mit einer vorgegebenen symmetrischen Matrix $W \in \mathbb{S}^n$ beschrieben. Die Bedingung $x_i \in \{\pm 1\}$ kann durch $x_i^2 = 1$ beschrieben werden. Dies ist kein konvexes Problem (warum nicht?)! Es lässt sich zeigen, dass das Problem NP-schwer ist.

Die Lagrange-Funktion des Problems ist

$$L(x,\nu) = \langle x|W|x\rangle + \sum_{i=1}^{n} \nu_i(x_i^2 - 1) = \langle x|W + \operatorname{diag}(\nu)|x\rangle - \langle 1|\nu\rangle\,.$$

Wir minimieren dies über x. Das Infimum einer quadratischen Form ist entweder Null (wenn nämlich die Matrix positiv semidefinit ist) oder $-\infty$ (sonst). Daher ist

$$g(\nu) = \begin{cases} \langle 1|\nu\rangle & \text{wenn } W + \operatorname{diag}(\nu) \text{ positiv semidefinit,} \\ -\infty & \text{sonst.} \end{cases}$$

Setzen wir beispielsweise alle Komponenten von ν auf $-\lambda_{\min}(W)$ (das Negative des kleinstens Eigenwerts von W), dann ist ν dual zulässig, da $W - \lambda_{\min}(W)\mathrm{Id}$ immer positiv semidefinit (der kleinste Eigenwert ist dann genau Null). Dabei ist dann $\langle 1|\nu\rangle = n\lambda_{\min}(W)$ eine untere Schranke für das kombinatorische Optimierungsproblem. Auf das genaue duale Problem gehen wir nicht ein. ♡

6.13 Beispiel (Entropie-Maximierung)**.** Sei $x = (x_1, \ldots, x_n)$ eine Wahrscheinlichkeitsverteilung auf n Elementen. Die Entropie von x ist definiert als $\sum_{i=1}^n x_i \log(1/x_i) = -\sum_i x_i \log x_i$ und misst die in x inhärente Unsicherheit. Die Basis des Logarithmus ist nicht entscheidend; üblicherweise wird 2 als Basis gewählt und die Entropie in Bits gemessen (oder die Eulersche Zahl e, dann spricht man von Nats). Die Entropie wird maximal für die Gleichverteilung und minimal (null) für Dirac-Verteilungen, bei denen ein Element Wahrscheinlichkeit 1 hat. Häufig ist es interessant, eine Verteilung maximaler Entropie unter (linearen) Nebenbedingungen zu finden. Dies lässt sich dann als das konvexe Problem

$$\begin{aligned} \text{Minimiere } f_0(x) &= \sum_{i=1}^n x_i \log x_i, \\ \text{so dass } Ax &\le b, \\ \langle 1|x\rangle &= 1 \end{aligned}$$

formulieren. Die Ungleichung $x \ge 0$ ist schon durch den Definitionsbereich $\operatorname{dom} f_0 = \mathbb{R}^n_{++}$ abgedeckt; ggf. lässt sich noch aus Stetigkeitsgründen $0 \log 0 := 0$ definieren.

Die Lagrange-Funktion ist konvex:

$$L(x, \lambda, \nu) = \sum_{i=1}^n x_i \log x_i + \langle \lambda | Ax - b\rangle + \nu(\sum_i x_i - 1).$$

Minimieren (Aufgabe!) führt auf die duale Funktion

$$g(\lambda, \nu) = -\langle b|\lambda\rangle - \nu - \sum_{i=1}^n \exp(-\langle a_i|\lambda\rangle - \nu - 1),$$

wobei a_i hier die i-te Spalte von A ist. Das duale Problem lautet daher

$$\begin{aligned} \text{Maximiere } & -\langle b|\lambda\rangle - \nu - \sum_{i=1}^n \exp(-\langle a_i|\lambda\rangle - \nu - 1), \\ \text{so dass } & \lambda \ge 0. \end{aligned}$$

Es lässt sich vereinfachen, indem wir analytisch über ν maximieren. Ableiten und Nullsetzen führt auf $\nu = \log \sum_i \exp(-\langle a_i|\lambda\rangle) - 1$; Einsetzen auf das vereinfachte duale Problem

$$\begin{aligned} \text{Maximiere } & -\langle b|\lambda\rangle - \log\big(\sum_{i=1}^n \exp(-\langle a_i|\lambda\rangle)\big), \\ \text{so dass } & \lambda \ge 0. \end{aligned}$$

Dies ist ein geometrisches Programm (in konvexer Form) mit Positivitätsbedingungen. ♡

6.3 Konjugierte Funktion und weitere Beispiele

Wenn man beim primalen Problem von einer Formulierung zu einer äquivalenten übergeht, kann sich das duale Problem dabei stark ändern. Es ist häufig eine Kunst herauszufinden, welche Formulierung am einfachsten oder nützlichsten ist. Wir werden dies an einigen Beispielen illustrieren.

Hierbei wird deutlich werden, dass im dualen Problem häufig die zu einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ konjugierte Funktion $f^* : \mathbb{R}^n \to \mathbb{R}$ auftritt. Diese ist wie folgt definiert.

6.14 Definition (konjugierte Funktion)**.** Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion (nicht notwendig konvex). Die *konjugierte Funktion* zu f ist definiert als $f^* : \mathbb{R}^n \to \mathbb{R}$,

$$f^*(y) := \sup_{x \in \operatorname{dom} f} (\langle y|x\rangle - f(x)).$$

Ihr Definitionsbereich $\operatorname{dom} f^*$ ist die Menge, auf der das Supremum endlich ist.

Für $n = 1$ ist $f^*(y)$ der maximale (vorzeichenbehaftete) Abstand zwischen der linearen Funktion $x \mapsto y \cdot x$ und $f(x)$.

Eine ökonomische Interpretation der konjugierten Funktion lautet wie folgt: Ein Unternehmen verbraucht n Ressourcen in Mengen $r = (r_1, \ldots, r_n)$. Der Einkauf einer Einheit der Ressource i kostet p_i; wir setzen $p = (p_1, \ldots, p_n)$. Die Preise seien fest (das Unternehmen kann sie nicht beeinflussen). Setzt das Unternehmen die Ressourcenmengen r ein, kann es einen Verkaufspreis $S(r)$ erzielen. (Wir nehmen an, dass diese Funktion dem Unternehmen bekannt ist; in der Praxis ist diese schwierig vorherzusagen!) Der Profit ist dann $S(r) - \langle p|r\rangle$. Natürlich will das Unternehmen den Profit maximieren und fragt daher nach dem erreichbaren Maximum (bei gegebenen Einkaufspreisen)

$$M(p) = \sup_r (S(r) - \langle p|r\rangle).$$

Dies ist fast die konjugierte Funktion (bis auf Vorzeichenprobleme). Der genaue Zusammenhang ist

$$M(p) = (-S)^*(-p).$$

6.15 Lemma (Konvexität der konjugierten Funktion)**.** *Die konjugierte Funktion f^* ist stets konvex.*

Beweis. Nach Definition ist f^* das Supremum über eine Familie von affinen Funktionen von y. □

6.16 Beispiel (Berechnung eindimensionaler konjugierter Funktionen)**.** Wir berechnen einige konjugierte Funktionen im Eindimensionalen; das Ausfüllen der Details ist eine gute Übungsaufgabe.

1. Sei $f(x) = ax + b$. Wir betrachten für festes y die Funktion $x \mapsto yx - ax - b$. Diese ist nur dann beschränkt, wenn $y = a$, und dann konstant $-b$. Also ist $\operatorname{dom} f^* = \{a\}$ und $f^*(a) = b$. Für $y \neq a$ ist f^* nicht definiert (bzw. ∞).

2. Sei $f(x) = -\log x$ mit $\operatorname{dom} f = \mathbb{R}_{++}$. Die Funktion $x \mapsto yx + \log x$ ist für $y \geq 0$ unbeschränkt. Für $y < 0$ nimmt sie ihr Maximum für $x = -1/y > 0$ an. Daher ist $f^*(y) = -\log(-y) - 1$ für $y < 0$ mit $\operatorname{dom} f^* = -\mathbb{R}_{++}$.
3. Sei $f(x) = \exp(x)$ auf $\mathbb{R}$. Die Funktion $x \mapsto yx - \exp(x)$ ist für $y < 0$ unbeschränkt. Für $y > 0$ nimmt sie ihr Maximum bei $x = \log y$ an; damit ist $f^*(y) = y \log y - y$. Für $y = 0$ ist das Maximum 0, so dass die Formel mit der Definition $0 \log 0 = 0$ gültig bleibt. Insgesamt ist $f^*(y) = y \log y - y$ auf $\mathbb{R}_+$.
4. Sei $f(x) = x \log x$ die negative Entropiefunktion mit $f(0) = 0$ und $\operatorname{dom} f = \mathbb{R}_+$. Die Funktion $x \mapsto yx - x \log x$ ist auf $\mathbb{R}_+$ für alle y beschränkt, daher ist $\operatorname{dom} f^* = \mathbb{R}$. Das Maximum wird für $x = \exp(y-1)$ erreicht; damit ist $f^*(y) = y \exp(y-1) - \exp(y-1) \cdot (y-1) = \exp(y-1)$.
5. Sei $f(x) = 1/x$ auf $\mathbb{R}_{++}$. Die Funktion $x \mapsto yx - 1/x$ ist für $y > 0$ unbeschränkt. Für $y = 0$ ist das Supremum 0. Für $y < 0$ wird das Maximum bei $x = (-y)^{-1/2}$ erreicht. Einsetzen liefert $f^*(y) = -2\sqrt{y}$ mit $\operatorname{dom} f^* = -\mathbb{R}_+$.

♡

6.17 Beispiel (Berechnung mehrdimensionaler konjugierter Funktionen)**.** Es folgen Beispiele für $f : \mathbb{R}^n \to \mathbb{R}$.

1. Sei $f(x) = \frac{1}{2} \langle x|Q|x\rangle$ mit $Q \in \mathbb{S}^n_{++}$ eine strikt konvexe quadratische Funktion. Die Funktion $x \mapsto \langle y|x\rangle - \frac{1}{2} \langle x|Q|x\rangle$ ist für alle y nach oben beschränkt. Das Maximum wird für $x = Q^{-1}y$ angenommen. Damit ist $f^*(y) = \frac{1}{2} \langle y|Q^{-1}|y\rangle$.
2. Sei $f(x) = \log(\sum_{i=1}^n \exp(x_i))$ die log-sum-exp-Funktion. Wir fragen, für welche y die Funktion $x \mapsto \langle y|x\rangle - \log \sum_{i=1}^n \exp(x_i)$ überhaupt ein endliches Maximum annehmen kann. Die Funktion ist jedenfalls konkav. Differenzieren nach x liefert die Bedingungen $y_i = \mathrm{e}^{x_i} / \sum_j \mathrm{e}^{x_j}$ für alle i. Diese Bedingungen sind nur lösbar, wenn $y > 0$ und $\sum_i y_i = 1$. Logarithmieren führt auf $x_i = f(x) + \log y_i$; Einsetzen auf $f^*(y) = \sum_i y_i \log y_i$ unter Ausnutzung von $\sum_i y_i = 1$. Dieser Ausdruck bleibt korrekt, wenn manche $y = 0$ sind und $y \geq 0$, $\sum_i y_i = 1$.

   Wir zeigen jetzt noch, dass andere y-Wahlen dazu führen, dass das Supremum unbeschränkt ist. Ist ein $y_k < 0$, dann sehen wir durch die Wahl $x_k = -t$ (mit $t \to \infty$) und $x_i = 0$ für $i \neq k$, dass $\langle y|x\rangle - f(x) = t|y_k| - (-t) \to \infty$. Ist $y \geq 0$, aber $Y := \sum_i y_i \neq 1$, wählen wir $x = (t, \ldots, t)$ für $t \in \mathbb{R}$ und sehen, dass $\langle y|x\rangle - f(x) = tY - t - \log n = t(Y-1) - \log n$. Dies ist für $Y < 1$ für $t \to -\infty$ und für $Y > 1$ für $t \to +\infty$ unbeschränkt.

   Insgesamt ergibt sich $f^*(y) = \sum_i y_i \log y_i$ mit $0 \log 0 := 0$, wobei $\operatorname{dom} f^*$ das Wahrscheinlichkeitssimplex $y \geq 0, \langle 1|y\rangle = 1$ ist.

♡

Wir betrachten jetzt noch die konjugierte Funktion einer beliebigen Norm $\|\cdot\|$ auf $\mathbb{R}^n$. Dazu ist folgende Definition nützlich.

6.18 Definition (duale Norm)**.** Zu einer Norm $\|\cdot\|$ auf $\mathbb{R}^n$ ist die *duale Norm* $\|\cdot\|_*$ definiert als

$$\|z\|_* := \sup \{ \langle z \mid x\rangle \mid \|x\| \leq 1 \} = \sup \{ \mid \langle z \mid x\rangle \mid \mid \|x\| \leq 1 \}.$$

(Dies ist die Operatornorm der durch z gegebenen linearen Funktion $x \mapsto \langle z|x\rangle$.)

Per Definition gilt stets die Ungleichung

$$\langle z|x\rangle \le \|z\|_*\|x\|.$$

6.19 Beispiel (duale Normen)**.** Die zur Euklidischen Norm duale Norm ist die Euklidische Norm selbst. Dies folgt mit der Cauchy-Schwarz-Ungleichung.

Die zur 1-Norm $\|x\|_1 = \sum_i |x_i|$ duale Norm ist die Maximums-Norm $\|z\|_\infty = \max_i |z_i|$, denn es ist $|\langle z|x\rangle| \le \max_i |z_i| \cdot \sum_i |x_i|$ mit Gleichheit, wenn alle Komponenten von z gleich groß sind. Umgekehrt ist die zur Maximums-Norm duale Norm die 1-Norm.

Allgemein kann man zeigen, dass die zur p-Norm duale Norm im $\mathbb{R}^n$ die q-Norm mit $1/p + 1/q = 1$ ist. Im $\mathbb{R}^n$ ist die duale Norm der dualen Norm die Norm selbst. In unendlichdimensionalen Vektorräumen muss dies nicht gelten. ♡

6.20 Beispiel (Konjugierte Funktion einer Norm)**.** Wir berechnen die konjugierte Funktion von $f(x) = \|x\|$. Sei $\|y\|_*$ die duale Norm.

Ist $\|y\|_* > 1$, dann gibt es nach Definition der dualen Norm ein z mit $\|z\| \le 1$ und $1 < \|y\|_* \le \langle y|z\rangle$. Setzen wir $x = tz$ mit $t \to \infty$, sehen wir, dass

$$\langle y|x\rangle - \|x\| = t(\langle y|z\rangle - \|z\|) \to \infty,$$

so dass $f^*(y) = \infty$ für $\|y\|_* > 1$.

Ist $\|y\|_* \le 1$, dann ist $\langle y|x\rangle \le \|x\|\,\|y\|_* \le \|x\|$ und daher

$$\langle y|x\rangle - \|x\| \le 0;$$

das Maximum 0 wird für $x = 0$ angenommen.

Es ergibt sich also $f^*(y) = 0$ mit $\operatorname{dom} f^* = \{\, y \mid \|y\|_* \le 1 \,\}$; dies ist die Indikatorfunktion der dualen Einheitskugel. ♡

6.21 Beispiel (quadrierte Norm)**.** Wir berechnen die konjugierte Funktion zu $f(x) = \frac{1}{2}\|x\|^2$. Sei $\|y\|_*$ die duale Norm. Nach Definition der dualen Norm ist

$$\langle y|x\rangle - \frac{1}{2}\|x\|^2 \;\le\; \|y\|_*\|x\| - \frac{1}{2}\|x\|^2;$$

die rechte Seite ist eine konkave quadratische Funktion in $\|x\|$; ihr Maximum wird für ein x mit $\langle y|x\rangle = \|y\|_*\|x\|$, so skaliert dass $\|x\| = \|y\|_*$, angenommen, und ist $\frac{1}{2}\|y\|_*^2$. Daher ist $f^*(y) = \frac{1}{2}\|y\|_*^2$. ♡

Abhängigkeit des dualen Problems von der Formulierung des primalen Problems. Wir zeigen nun, dass eine Umformulierung des primalen Problems durch einfache Äquivalenzumformungen starke Auswirkungen auf das duale Problem haben kann.

6.22 Beispiel (Einführen von Gleichungs-Bedingungen)**.** Das Problem

$$\text{Minimiere } f_0(x) := f(Ax + b)$$

mit einer geeigneten Funktion f hat keine Nebenbedingungen. Daher ist die Lagrange-Funktion $L(x) = f_0(x)$ und die duale Funktion ist die Konstante $g \equiv \inf_x L(x) = p^*$. Daher

besteht das duale Problem darin, die Konstante zu "maximieren", eine triviale Aufgabe (sofern man p^* kennt), die in keiner Weise nützlich ist.

Führen wir aber neue Variablen y ein und schreiben das Problem äquivalent als

$$\begin{aligned} &\text{Minimiere } f_0(x,y) := f(y), \\ &\text{so dass } Ax + b - y = 0, \end{aligned}$$

dann erhalten wir die Lagrange-Funktion

$$L((x,y),\nu) = f(y) + \langle \nu | Ax + b - y \rangle = f(y) + \langle x | A^T \nu \rangle + \langle b - y | \nu \rangle \,,$$

die über (x,y) minimiert werden muss. Die Funktion ist unbeschränkt nach unten, wenn nicht $A^T\nu = 0$. Ist $A^T\nu = 0$, bleibt

$$g(\nu) = \langle b|\nu\rangle + \inf_y (f(y) - \langle \nu | y \rangle) = \langle b|\nu\rangle - f^*(\nu)$$

unter Benutzung der konjugierten Funktion f^*. Das duale Problem lautet also

$$\begin{aligned} &\text{Maximiere } \langle b|\nu\rangle - f^*(\nu), \\ &\text{so dass } A^T\nu = 0. \end{aligned}$$

Dies ist eine sinnvolle und nützliche duale Formulierung (sofern sich f^* berechnen lässt).

Wir konkretisieren dies an einem unrestringierten geometrischen Programm in konvexer Form

$$\text{Minimiere } \log\Big(\sum_{i=1}^m \exp(\langle a_i | x\rangle + b_i)\Big).$$

Einführen von y und der Gleichung $Ax + b = y$ (die a_i sind die Zeilen von A) liefert das Problem

$$\begin{aligned} &\text{Minimiere } f(y) = \log\Big(\sum_{i=1}^m \exp(y_i)\Big), \\ &\text{so dass } Ax + b - y = 0, \end{aligned}$$

mit Variablen (x,y). Die konjugierte Funktion zu log-sum-exp haben wir als $f^*(\nu) = \sum_{i=1}^m \nu_i \log \nu_i$ für $\nu \geq 0$, $\langle 1|\nu\rangle = 1$ und $f^*(\nu) = \infty$ sonst kennengelernt. Das duale Problem lautet daher

$$\begin{aligned} &\text{Maximiere } \langle b|\nu\rangle - \sum_{i=1}^m \nu_i \log \nu_i, \\ &\text{so dass } A^T\nu = 0, \\ &\nu \geq 0, \\ &\langle 1|\nu\rangle = 1. \end{aligned}$$

Dies ist ein Entropiemaximierungsproblem mit Nebenbedingungen. ♡

6.23 Beispiel (Approximative Lösung eines LGS)**.** Wir betrachten das Problem "Minimiere $\|Ax - b\|$" mit einer beliebigen Norm. Dieses Problem ist zunächst unrestringiert und hat daher nur ein triviales duales Problem. Einführen von $y = Ax - b$ ergibt das Problem

$$\begin{aligned} &\text{Minimiere } \|y\|, \\ &\quad \text{so dass } Ax - b - y = 0. \end{aligned}$$

Dies führt auf das duale Problem (unter Benutzung der konjugierten Funktion für die Norm)

$$\begin{aligned} &\text{Maximiere } -\langle b|\nu\rangle\,, \\ &\quad \text{so dass } A^T\nu = 0, \\ &\qquad \|\nu\|_* \le 1. \end{aligned}$$

Formulieren wir das primale Problem äquivalent als

$$\begin{aligned} &\text{Minimiere } \frac{1}{2}\|y\|^2, \\ &\quad \text{so dass } Ax - b - y = 0, \end{aligned}$$

dann erhalten wir ein anderes duales Problem, nämlich

$$\begin{aligned} &\text{Maximiere } -\langle b|\nu\rangle - \frac{1}{2}\|\nu\|_*^2, \\ &\quad \text{so dass } A^T\nu = 0. \end{aligned}$$

Hierbei haben wir die konjugierte Funktion der quadrierten Norm benutzt. ♡

6.4 Starke Dualität und Slater's Bedingungen

Die Beziehung $d^* \le p^*$ nennt man schwache Dualität. Die Differenz $p^* - d^* \ge 0$ nennt man *optimale Dualitätslücke* (optimal duality gap). Interssant ist die Frage, wann starke Dualität, also $d^* = p^*$ gilt und die Lücke verschwindet.

Vorbereitende Definitionen. Um starke Dualität zu charakterisieren, benötigen wir einige weitere Begriffe im Zusammenhang mit konvexen Mengen: unterstützende und trennende Hyperebenen, sowie das relative Innere einer Menge.

6.24 Definition (Inneres, Abschluss, Rand)**.** Das *Innere* $\operatorname{int} C$ einer Menge $C \subset \mathbb{R}^n$ besteht aus den Punkten, für die gilt, dass eine ganze offene Umgebung des Punktes ebenfalls in C liegt.

Der *Abschluss* $\operatorname{cl} C$ von C besteht aus den Punkten, für die es eine Folge $(x_n) \subset C$ gibt, die gegen $x \in \mathbb{R}^n$ konvergiert. (Insbesondere ist $C \subset \operatorname{cl} C$.)

Der *Rand* $\operatorname{bd} C$ von C ist definiert als $\operatorname{bd} C := \operatorname{cl} C \setminus \operatorname{int} C$.

Liegt die Menge C ohnehin nur in einem echten affinen Unterraum von $\mathbb{R}^n$, gibt es zu keinem $x \in C$ eine offene Umgebung, die ganz in C liegt, und das Innere ist leer. Als Beispiel stelle man sich eine (flache) Kreissscheibe im dreidimensionalen Raum vor: Das Innere ist leer. Im zweidimensionalen Raum wäre das Innere jedoch die Kreisscheibe selbst (ohne Rand). Um vom Inneren einer Menge unabhängig vom umgebenden Raum sprechen zu können, führen wir den Begriff des relativen Inneren ein. Wir erinnern daran, dass $\operatorname{aff} C$ die affine Hülle von C ist, also der kleinste affine Raum, der C enthält. Verlangt wird, dass es zu x eine ε-Umgebung zumindest in $\operatorname{aff} C$ gibt, die ganz in C liegt.

6.25 Definition (relatives Inneres)**.** Sei C eine Menge, $\operatorname{aff} C$ ihre affine Hülle. Das *relative Innere* von C ist die Menge

$$\operatorname{relint} C := \{\, x \in C \mid \exists\, \varepsilon > 0 \text{ mit } B_\varepsilon(x) \cap \operatorname{aff} C \subset C \,\}\,.$$

Dabei ist $B_\varepsilon(x)$ die ε-Kugel um x (bezüglich irgendeiner Norm).

6.26 Definition (trennende Hyperebene)**.** Es seien $C, D \subset \mathbb{R}^n$ zwei beliebige Mengen. Wenn es eine Hyperebene $H = \{\, x \mid \langle a|x\rangle = b \,\}$ mit $a \neq 0$ gibt, so dass die Punkte $x \in C$ auf der einen Seite liegen ($\langle a|x\rangle \leq b$) und die Punkte $x \in D$ auf der anderen Seite ($\langle a|x\rangle \geq b$), dann heißt H eine *trennende Hyperebene* der Mengen C, D.

6.27 Satz (Existenz trennender Hyperebenen)**.** *Seien C, D konvexe nichtleere disjunkte Teilmengen des $\mathbb{R}^n$. Dann existiert eine C und D trennende Hyperebene.*

Wir verzichten auf einen Beweis.

6.28 Definition (unterstützende Hyperebene)**.** Sei $C \subset \mathbb{R}^n$ eine nichtleere beliebige Menge und $x_0 \in \operatorname{bd} C$ ein Randpunkt. Gibt es eine Hyperebene, die $D = \{x_0\}$ von C trennt, dann heißt diese eine *unterstützende Hyperebene* im Randpunkt x_0.

6.29 Satz (Existenz unterstützender Hyperebenen)**.** *Sei C eine nichtleere konvexe Teilmenge des $\mathbb{R}^n$. Dann existiert zu jedem Randpunkt $x_0 \in C$ eine unterstützende Hyperebene.*

Beweis. Ist $\operatorname{int} C$ nicht leer, dann folgt der Satz aus Satz 6.27 mit den disjunkten Mengen $\operatorname{int} C$ und x_0. Ist $\operatorname{int} C$ leer, dann liegen C und x_0 in einem affinen Unterraum von $\mathbb{R}^n$. Jede Hyperebene, die diesen affinen Unterraum enthält, ist eine unterstützende Hyperebene. □

Geometrische Charakterisierung starker Dualität. Da das duale Problem stets konvex ist, kann man nicht erwarten, dass starke Dualität immer möglich ist. (Sonst wären allgemeine Optimierungsprobleme stets auf konvexe Probleme vereinfachbar.) Wie steht es aber bei konvexen Problemen? Hier lässt sich "im Normalfall" starke Dualität beweisen. Wir zeigen dies mit einem geometrischen Argument, das auch klärt, wann der "Normalfall" verletzt wird.

Zu Problem (6.1) definieren wir die Punktmenge

$$\mathcal{A}_= := \{\, (u, v, t) \mid \exists x \in \mathcal{D} \text{ mit } f_i(x) = u_i\ (i = 1, \ldots, m);\ h_i(x) = v_i\ (i = 1, \ldots, p);\ f_0(x) = t \,\}\,.$$

Diese Menge enthält alle Werte, die von den Nebenbedingungsfunktionen und der Zielfunktion auf $\mathcal{D}$ angenommen werden. Weiter definieren wir die erweiterte Menge

$$\mathcal{A} := \{\, (u, v, t) \mid \exists x \in \mathcal{D} \text{ mit } f_i(x) \leq u_i\ (i = 1, \ldots, m);\ h_i(x) = v_i\ (i = 1, \ldots, p);\ f_0(x) \leq t \,\}\,.$$

TODO: Bild einbinden

Abbildung 6.1: Mengen $\mathcal{A}_=$ und $\mathcal{A}$ für eine Zielfunktion (Werte auf der t-Achse) mit einer Ungleichungs-Nebenbedingung (Werte auf der u-Achse). (a) nicht konvexe Menge, keine unterstützende Hyperebene in $(0, p^*)$; (b) nicht konvexe Menge mit unterstützender nichtvertikaler Hyperebene in $(0, p^*)$; (c) konvexe Menge mit nichtvertikaler unterstützender Hyperebene in $(0, p^*)$; (d) konvexe Menge, aber es existiert nur eine vertikale unterstützende Hyperebene in $(0, p^*)$.

Diese Menge enthält alle Werte, die von den Nebenbedingungsfunktionen und der Zielfunktion auf $\mathcal{D}$ angenommen werden, sowie schlechtere für die Ungleichungen und die Zielfunlktion. Die Menge $\mathcal{A}$ ist konvex, wenn das Optimierungsproblem 6.1 konvex ist.

Mit dieser Definition von $\mathcal{A}_=$ und $\mathcal{A}$ ist

$$p^* \;=\; \inf\{\, t \mid (u,v,t) \in \mathcal{A}_=, u \le 0, v = 0 \,\} \;=\; \inf\{\, t \mid (0,0,t) \in \mathcal{A} \,\}\,.$$

Ferner ist

$$L(x,\lambda,\nu) = \langle (\lambda,\nu,1) | (f(x), h(x), f_0(x)) \rangle$$

mit $f(x) = (f_1(x), \ldots, f_m(x))$ und $h(x) = (h_1(x), \ldots, h_p(x))$, so dass $(f(x), h(x), f_0(x)) \in \mathcal{A}_=$. Daraus folgt

$$g(\lambda,\nu) = \inf\{\, \langle (\lambda,\nu,1) \mid (u,v,t) \rangle \mid (u,v,t) \in \mathcal{A}_= \,\}\,.$$

Für $\lambda \ge 0$ ist auch

$$g(\lambda,\nu) = \inf\{\, \langle (\lambda,\nu,1) \mid (u,v,t) \rangle \mid (u,v,t) \in \mathcal{A} \,\}\,.$$

Die Punktmenge $H = \{\, (u,v,t) \mid \langle (\lambda,\nu,1) | (u,v,t) \rangle = g(\lambda,\nu) \,\}$ ist daher die "am tiefsten" gelegene Hyperebene der Form $t = g - \langle \lambda | u \rangle - \langle \nu | v \rangle$, die $\mathcal{A}$ berührt (g minimal). H definiert daher eine (nichtvertikale) unterstützende Hyperebene an $\mathcal{A}$. Für $u = v = 0$ ist $t = g(\lambda,\nu)$, d.h. die unterstützende Hyperebene geht durch den Punkt $(0, 0, g(\lambda,\mu))$; ihr Normalenvektor ist nach Konstruktion $(\lambda, \nu, 1)$.

Für alle Punkte (u,v,t) in $\mathcal{A}$ gilt $\langle (\lambda,\nu,1) | (u,v,t) \rangle \ge g(\lambda,\nu)$. Die Wahl des Randpunktes $(u,v,t) = (0,0,p^*)$ zeigt wieder $p^* \ge g(\lambda,\nu)$ für alle $\lambda \ge 0$ und alle ν.

Gleichheit (und damit starke Dualität) gilt genau dann, wenn die Menge $\mathcal{A}$ eine nichtvertikale unterstützende Hyperebene in ihrem Randpunkt $(0,0,p^*)$ besitzt, d.h. wenn $(0,0,g(\lambda,\mu))$ Randpunkt der Menge $\mathcal{A}$ mit nichtvertikaler unterstützender Hyperebene ist.

Eine konvexe Menge besitzt in *jedem* ihrer Randpunkte eine unterstützende Hyperebene; allerdings muss für die starke Dualität noch sichergestellt werden, dass es eine nicht vertikale Hyperebene gibt. Dazu sind zusätzliche technische Bedingungen erforderlich, die man an das Optimierungsproblem stellt. Diese heißen *constraint qualifications*. Natürlich kann auch für nicht konvexe Probleme im Einzelfall starke Dualität gelten. Abbildung 6.1 illustriert diese Sachverhalte für nicht konvexe und konvexe Mengen $\mathcal{A}$.

Wir zeigen zunächst an einem Beispiel, dass auch bei konvexen Problemen starke Dualität nicht gelten muss, wenn es keine nichtvertikale unterstützende Hyperebene an $\mathcal{A}$ gibt.

6.30 Beispiel (Konvexes Problem ohne starke Dualität)**.** Wir wollen auf $\mathcal{D} := \{ (x, y) \mid y > 0 \} \subset \mathbb{R}^2$ die Zielfunktion $f_0(x, y) := \exp(-x)$ unter der Nebenbedingung $f_1(x, y) := x^2/y \leq 0$ minimieren. Es handelt sich um ein konvexes Problem, denn f_1 wird über ein Perspektiv-Argument als konvex nachgewiesen.

Wir stellen fest, dass die zulässige Menge $\mathcal{F} = \{ (0, y) \mid y > 0 \}$ ist, dort ist $f_0 \equiv 1$ konstant. Der optimale Wert ist daher $p^* = 1$.

Die Lagrange-Funktion ist $L((x, y), \lambda) = \exp(-x) + \lambda x^2/y$; diese ist konvex in (x, y). Man sieht, dass das Infimum über (x, y) für $y > 0$ gleich Null ist (wenn man x und y geeignet gegen unendlich gehen lässt) und nicht erreicht wird.

Es ist daher die duale Funktion $g(\lambda) \equiv 0$ und $d^* = 0$. Die "Lücke" (optimal duality gap) beträgt $p^* - d^* = 1$.

Wie sieht für dieses Beispiel die Menge $\mathcal{A}_=$ aus? Sie besteht aus allen Punkten $(u, t) = (x^2/y, \exp(-x))$ für $x \in \mathbb{R}$ und $y > 0$. Dies ist genau die Menge $\{ (u, t) \mid u > 0, t > 0 \} \cup \{ (0, 1) \}$; zu gegebenem $(u, t) > 0$ wählt man $x = -\log t$ und $y = (\log t)^2/u$ und zu $(0, 1)$ wählt man $x = 0$ und $y > 0$ beliebig. Weiter ist $\mathcal{A} = \{ (u, t) \mid u > 0, t > 0 \} \cup \{ (0, t) \mid t \geq 1 \}$. Die Mengen $\mathcal{A}_=$ und $\mathcal{A}$ sind konvex.

Eine zweidimensionale Skizze im (u, t)-Raum veranschaulicht, dass es an $(0, p^*) = (0, 1)$ nur eine vertikale unterstützende Hyperebene gibt (Abbildung 6.1d). ♡

Dass sogar bei linearen Programmen starke Dualität nicht selbstverständlich ist, zeigt folgendes Beispiel.

6.31 Beispiel (Lineares Programm ohne starke Dualität)**.** Betrachte das eindimensionale Problem "Minimiere x, so dass $0x \leq -1$ und $1x \leq 1$". Dieses Problem ist offensichtlich nicht zulässig, also $p^* = \infty$. Die Lagrange-Funktion ist $L(x, \lambda_1, \lambda_2) = x + \lambda_1 + \lambda_2(x - 1) = \lambda_1 - \lambda_2 + (1 + \lambda_2)x$; diese ist unbeschränkt in x für $\lambda_2 \neq -1$. Für $\lambda \geq 0$ ist also $g(\lambda) \equiv -\infty$, und damit die dual zulässige Menge leer. Damit ergibt sich ein dualer optimaler Wert von $d^* = -\infty$ für das Maximierungsproblem. Also ist $d^* \neq p^*$, es gilt keine starke Dualität.

Hierbei fällt auf, dass sowohl primales als auch duales LP nicht zulässig sind. Es wird sich zeigen, dass starke Dualität bei linearen Programmen immer gilt, außer wenn das primale und duale Problem beide nicht zulässig sind. ♡

Eine hinreichende Bedingung, unter der starke Dualität gilt, also die Existenz eine nichtvertikalen unterstützenden Hyperebene an $(0, 0, p^*) \in \mathcal{A}$ garantiert werden kann, heißt *Slater's Bedingung.*

6.32 Satz (Slater's Satz zur starken Dualität)**.** *Das primale Problem* (6.1) *sei konvex mit konvexer Definitionsmenge $\mathcal{D}$. Von den m Ungleichungen $f_i(x) \leq 0$ seien die ersten $k \leq m$ affin. (Die p Gleichungen sind ohnehin affin bei einem konvexen Problem.)*

Angenommen, es gelten Slater's Bedingungen*, nämlich: Es gebe einen zulässigen Punkt, der die $m - k$ nicht-affinen Ungleichungen* strikt *erfüllt und die affinen Ungleichungen und Gleichungen erfüllt; ferner liege dieser Punkt im relativen Inneren des Definitionsbereichs* $\operatorname{relint} \mathcal{D}$*.*

Dann gilt starke Dualität $d^* = p^*$ und es gibt optimale primale und duale Punkte x^* und (λ^*, ν^*).

Beweis. Wir führen den Beweis nicht im Detail. Die Existenz eines x, das Slater's Bedingungen erfüllt, garantiert aber die Existenz eines $(u, v, t) \in \mathcal{A}$ mit $u < 0$, $v = 0$ und $t > p^*$, so dass eine nichtvertikale unterstützende Hyperebene durch $(0, 0, p^*)$ geht. □

6.33 Beispiel (Anwendung von Slater's Bedingungen)**.** Für Beispiel 6.9 (Lösung von $Ax = b$ mit minimaler Norm) sagen Slater's Bedingungen, dass starke Dualität gilt, wenn das Problem zulässig ist. Aber sogar, wenn es kein x mit $Ax = b$ gibt ($p^* = +\infty$), gilt hier starke Dualität: Dann gibt es nämlich (hier ohne Beweis) ein z mit $A^T z = 0$ und $\langle b|z\rangle \neq 0$, so dass die duale Funktion $-(1/4)\,\langle\nu|A^T A|\nu\rangle - \langle b|\nu\rangle$ entlang der Geraden $\{\, tz \mid t \in \mathbb{R} \,\}$ unbeschränkt ist, so dass auch $d^* = \infty$ gilt. Slater's Bedingung ist also hinreichend, aber nicht notwendig.

Für Beispiel 6.13 (Entropie-Maximierung unter Nebenbedingungen) sagen Slater's Bedingungen, dass starke Dualität gilt, wenn es ein $x > 0$ mit $Ax \leq b$ und $\sum_i x_i = 1$ gibt.

Für lineare Programme lautet Slater's Bedingung, dass das primale Problem zulässig ist. Angewendet auf das duale Problem lautet Slater's Bedingung, dass das duale Problem zulässig ist. Da das duale des dualen LPs äquivalent zum primalen LP ist, gilt bei LPs starke Dualität, sobald das primale oder das duale LP zulässig ist. Der Fall, dass beide LPs nicht zulässig sind, kann tatsächlich zur Verletzung der starken Dualität führen, siehe Beispiel 6.31.

Sind bei einem konvexen Optimierungsproblem alle Gleichungen und Ungleichungen affin und ist $\mathcal{D}$ offen, lautet Slater's Bedingung einfach, dass das primale Problem zulässig ist. ♡

Starke Dualität lässt sich auch mit der Lagrange-Funktion beschreiben.

6.34 Definition (Sattelpunkt)**.** Ein Punkt $(x^*, \lambda^*, \nu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ heißt *Sattelpunkt* der Lagrange-Funktion, wenn

$$L(x^*, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) \leq (x, \lambda^*, \nu^*)$$

für alle x, λ, ν gilt, wenn also

$$L(x^*, \lambda^*, \nu^*) = \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*) = \sup_{\lambda \geq 0; \nu} L(x^*, \lambda, \nu).$$

6.35 Lemma (Starke Dualität und Sattelpunkte)**.** *Sei (x^*, λ^*, ν^*) ein Sattelpunkt der Lagrange-Funktion. Dann gilt starke Dualität und x^* ist primal optimal und (λ^*, ν^*) ist dual optimal. Umgekehrt: Wenn starke Dualität gilt und x^* primal optimal und (λ^*, ν^*) dual optimal ist, dann ist (x^*, λ^*, ν^*) ein Sattelpunkt der Lagrange-Funktion.*

Daraus ergibt sich eine vereinfachte Formulierung des dualen Problems, bei der wir die Minimierung (zum Berechnen der dualen Funktion) dadurch ersetzen können, dass wir die Ableitung der Lagrange-Funktion nach x auf Null setzen.

6.5 Optimalitätsbedingungen (KKT)

Ein primal zulässiger Punkt x und ein dual zulässiger Punkt (λ, ν) schränken die optimalen primalen und dualen Werte ein und können (bei starker Dualität) als Abbruchbedingung für einen Optimierungsalgorithmus verwendet werden.

6.36 Satz (Zertifikat für ε-Optimalität)**.** *Sei in den zueinander dualen Optimierungsproblemen* (6.1), (6.2) *x primal zulässig und (λ, ν) dual zulässig. Dann ist*

$$p^* \in [g(\lambda, \nu), f_0(x)], \qquad d^* \in [g(\lambda, \nu), f_0(x)].$$

Ist $\varepsilon := f_0(x) - g(\lambda, \nu)$ die Dualitätslücke *zu (x, λ, ν), dann ist*

$$f_0(x) - p^* \leq \varepsilon.$$

Gilt starke Dualität, findet man zu jeder Dualitätslücke $\varepsilon > 0$ primal und dual zulässige Punkte $(x, (\lambda, \nu))$.

6.37 Satz (Komplementärer Schlupf)**.** *Es gelte starke Dualität und es sei (x^*, λ^*, ν^*) primal-dual optimal. Dann ist $\lambda_i^* f_i(x^*) = 0$ für alle $i = 1, \ldots, m$. (Der Lagrange'sche Multiplikator λ_i ist Null, es sei denn, die i-te Ungleichung ist im optimalen Punkt aktiv.)*

Beweis. Bei starker Dualität ist

$$\begin{aligned} f_0(x^*) = p^* = d^* &= g(\lambda^*, \nu^*) \\ &= \inf_x \Big(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu^* h_i(x) \Big) \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu^* h_i(x^*) \\ &\leq f_0(x^*). \end{aligned}$$

Es gilt also überall Gleichheit; daher muss $\sum_i \lambda_i^* f_i(x^*) = 0$ sein. Daraus folgt die Behauptung. □

Ab jetzt nehmen wir an, dass starke Dualität gilt und die Zielfunktion und Nebenbedingungs-Funktionen differenzierbar sind. Wir nehmen noch nicht an, dass das Problem konvex ist!

6.38 Satz (Notwendige KKT-Optimalitätsbedingungen bei starker Dualität)**.** *Sei* (6.1) *ein (nicht notwendig konvexes) Optimierungsproblem mit starker Dualität und differenzierbaren Ziel- und Nebenbedingungsfunktionen. Sei (x^*, λ^*, ν^*) primal-dual optimal. Dann gelten die Karush-Kuhn-Tucker-Bedingungen (KKT-Bedingungen):*

$$\begin{aligned} f_i(x^*) &\leq 0 && (i = 1, \ldots, m), \\ h_i(x^*) &= 0 && (i = 1, \ldots, p), \\ \lambda_i^* &\geq 0 && (i = 1, \ldots, m), \\ \lambda_i^* f_i(x^*) &= 0 && (i = 1, \ldots, m), \\ \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) &= 0. \end{aligned}$$

Beweis. Die ersten drei Zeilen sind primale und duale Zulässigkeit; die vierte Zeile gilt nach Satz 6.37. Die letzte Zeile sagt aus, dass der Gradient von $x \mapsto L(x, \lambda^*, \nu^*)$ in x^* verschwindet, denn x^* minimiert diese Funktion in x. □

6.39 Satz (Hinreichende KKT-Optimalitätsbedingungen bei konvexem Optimierungsproblem)**.** *Sei* (6.1) *ein konvexes Optierungsproblem. Sei* (x^*, λ^*, ν^*) *ein Punkt, der die KKT-Bedingungen erfüllt. Dann gilt starke Dualität und* (x^*, λ^*, ν^*) *ist primal-dual optimal.*

Beweis. Wir nutzen, dass $L(x, \lambda, \nu)$ für alle $\lambda \geq 0$ und alle ν konvex in x ist. Daher minimiert x^* die Funktion $x \mapsto L(x, \lambda^*, \nu^*)$, und es folgt $g(\lambda^*, \nu^*) = f_0(x^*)$ unter Ausnutzung der KKT-Bedingungen. □

6.40 Satz (Notwendige und hinreichende Optimalitätsbedingungen)**.** *Sei* (6.1) *ein konvexes Optierungsproblem, das Slater's Bedingungen erfüllt. Dann ist der Punkt* (x^*, λ^*, ν^*) *genau dann primal-dual optimal, wenn er die KKT-Bedingungen erfüllt.*

Beweis. Folgt aus den vorangehenden Sätzen und Slater's Satz. □

Die KKT-Bedingungen in den primalen und dualen Variablen liefern die Motivation für Optimierungsverfahren, die man als Verfahren zur Lösung der KKT-Bedingungen bzw. zum Findes eines Sattelpunktes der Lagrange-Funktion interpretieren kann. Wir betrachten Anwendungen der KKT-Bedingungen im nächsten Kapitel.

6.6 Sensitivitätsanalyse des primalen Problems

Statt des ursprünglichen primalen Problems (6.1) betrachten wir nun das *gestörte Problem*

$$\begin{aligned} &\text{Minimiere } f_0(x) && && (6.3)\\ &\text{so dass } f_i(x) \leq u_i && \text{für } i = 1, \ldots, m,\\ &\qquad h_i(x) = v_i && \text{für } i = 1, \ldots, p. \end{aligned}$$

mit $u \in \mathbb{R}^m$ und $v \in \mathbb{R}^p$.

Es sei $p^*(u, v)$ der optimale Wert des gestörten Problems; für $u = 0, v = 0$ ist $p^*(0, 0) = p^*$ der optimale Wert des Originalproblems.

6.41 Lemma (Konvexität von $p^*(u, v)$)**.** *Ist das Originalproblem konvex, dann ist* $p^*(u, v)$ *eine konvexe Funkton von* (u, v)*.*

Eine interessante Frage ist, wie stark sich $p^*(u, v)$ von p^* unterscheiden kann.

6.42 Satz (Globale Sensitivitätsanalyse)**.** *Es gelte starke Dualität und das duale Optimum werde für das nicht gestörte Problem in* (λ^*, ν^*) *angenommen. (Konvexität wird nicht vorausgesetzt.) Dann gilt für alle* u *und* v *die Ungleichung*

$$p^*(u, v) \geq p^* - \langle \lambda^* | u \rangle - \langle \nu^* | v \rangle .$$

Beweis. Sei x zulässig für das gestörte Problem. Dank starker Dualität gilt

$$\begin{aligned} p^* = g(\lambda^*, \nu^*) &\leq f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \\ &\leq f_0(x) + \langle \lambda^* | u \rangle + \langle \nu^* | v \rangle \,, \end{aligned}$$

da $\lambda \geq 0$. Daher ist für alle im gestörten Problem zulässigen x

$$f_0(x) \geq p^* - \langle \lambda^* | u \rangle - \langle \nu^* | v \rangle \,,$$

daher gilt diese Ungleichung auch für $p^*(u, v)$. □

Man beachte, dass das Resultat des vorangehendes Satzes nicht symmetrisch ist und nur eine untere Schranke (die aber global gültig ist) liefert.

Ein symmetrischeres Resultat erhalten wir, wenn wir die Differenzierbarkeit der Ziel- und Nebenbedingungsfunktionen voraussetzen und in der Nähe von $(u, v) = (0, 0)$ bleiben.

6.43 Satz (Lokale Sensitivitätsanalyse)**.** *Es gelte starke Dualität und das duale Optimum werde für das nicht gestörte Problem in (λ^*, ν^*) angenommen. (Konvexität wird nicht vorausgesetzt.) Ziel- und Nebenbedingungsfunktionen seien differenzierbar. Dann gilt*

$$\lambda_i^* = -\frac{\partial p^*(0,0)}{\partial u_i}, \qquad \nu_i^* = -\frac{\partial p^*(0,0)}{\partial v_i}.$$

Beweis. Wir betrachten für den i-ten Einheitsvektor e_i und $t > 0$ den Differenzenquotienten

$$\frac{p^*(te_i, 0) - p^*}{t} \geq -\lambda_i^*;$$

die Ungleichung gilt nach Satz 6.42. Für $t \to 0$ mit $t > 0$ erhalten wir $\frac{\partial p^*(0,0)}{\partial u_i} \geq -\lambda_i^*$. Entsprechendes mit $t < 0$ liefert die entgegengesetzte Ungleichung. Entsprechend zeigt man die Aussage für ν. □

Ändert man also beispielsweise u_i von 0 auf ein kleines ε, ändert sich der optimale Wert in etwa um $-\varepsilon\lambda_i$. Ist beispielsweise die i-te Ungleichung $f_i(x^*) \leq 0$ nicht aktiv, ist also $f_i(x^*) < 0$, dann gilt wegen des komplementären Schlupfes $\lambda_i^* = 0$ und man kann die Ungleichung ein wenig ändern, ohne dass sich der optimale Wert ändert. Ist die Ungleichung aber aktiv, so hängt das Ausmaß der Änderung von λ_i^* ab.

KAPITEL 7

Anwendungen

In diesem Kapitel betrachten wir (in unterschiedlicher Ausführlichkeit) verschiedene Anwendungsfragestellungen, die auf konvexe Optimierungsprobleme führen.

7.1 Klassifikation mit Support Vector Machines

Wir hatten in Abschnitt 1.2 das (lineare) Klassifikationsproblem betrachtet, die beste (maximum margin) trennende Hyperebene zwischen zwei Punktmengen, die wir die Klassen $+1$ und -1 genannt haben, zu berechnen. Gegeben waren n Punkte im d-Dimensionalen Raum $(x_i) \in \mathbb{R}^d$, $i = 1, \ldots, n$ und zugehörige Klassenbezeichnungen $y_i \in \{ \pm 1 \}$. Die Hyperebene $h = \{ x \mid \langle w|x \rangle = b \}$, die die Punkte am besten trennt, erhält man (siehe Abschnitt 1.2), wenn man das konvexe quadratische Programm

$$\begin{array}{lll} \text{Minimiere} & \dfrac{1}{2}\|w\|_2^2 = (1/2)\,\langle w|w\rangle\,, & \\ \text{so dass} & y_i \cdot (\langle w|x_i\rangle - b) - 1 \geq 0 & \text{für } i = 1, \ldots, n \end{array}$$

in den Variablen $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ und $b \in \mathbb{R}$ löst.

Wir wissen, dass die Zielfunktion in (w, b) (strikt) konvex ist und die Nebenbedingungen affin sind. Daher ist die Lagrange-Funktion für $\lambda \geq 0$ konvex in (w, b) und konkav in λ. Es gelten Slater's Bedingungen und daher starke Dualität. Es ist

$$L((w, b), \lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \lambda_i (1 - y_i(\langle w|x_i\rangle - b)).$$

Das Minimum über (w, b) finden wir (da L in (w, b) konvex ist) durch Nullsetzen der Ableitungen:

$$\partial L(w, b, \lambda)/\partial w = w - \sum_i \lambda_i y_i x_i \stackrel{!}{=} 0,$$
$$\partial L(w, b, \lambda)/\partial b = \sum_i \lambda_i y_i \stackrel{!}{=} 0.$$

Einsetzen von $w = \sum_i \lambda_i y_i x_i$ und der Bedingung $\sum_i \lambda_i y_i = 0$ in L liefert das duale Problem

$$\begin{array}{ll} \text{Maximiere} & -(1/2)\sum_{i,j} \lambda_i \lambda_j y_i y_j \langle x_i | x_j \rangle \ + \sum_i \lambda_i, \\ \text{so dass} & \lambda \geq 0, \\ & \langle \lambda | y \rangle = 0. \end{array}$$

Dies ist ein quadratisches konvexes Problem in λ. Wir bemerken, dass die Zielfunktion nur von den Skalarprodukten $\langle x_i | x_j \rangle$ abhängt.

Es ist selbstverständlich möglich, dass das primale Problem nicht zulässig ist, sich die Punkte in den Klassen $+1$ und -1 also nicht durch eine Hyperebene trennen lassen. Es gibt zwei Möglichkeiten, damit umzugehen. Wir werden zum Schluss beide gleichzeitig verwenden, beschreiben sie jetzt aber zunächst getrennt.

Kernel Trick. Die erste Möglichkeit besteht darin, mittels einer (nichtlinearen) Abbildung $\phi : \mathbb{R}^d \to \mathcal{Z}$ die Punkte zunächst in einen anderen Raum abzubilden und das Problem dann nur auf den $z_i = \phi(x_i)$ zu formulieren. Man stelle sich im $\mathbb{R}^2$ einige Punkte innerhalb der durch den Kreis $x_1^2 + x_2^2 = 1$ berandeten Kreisscheibe vor und andere Punkte außerhalb. Bildet man diese mittels $\phi(x) = (x_1, x_2, x_1^2 + x_2^2) = (z_1, z_2, z_3)$ ab, werden die Klassen durch die Hyperebene $z_3 = 1$ linear trennbar.

Die Nachteile liegen darin, dass man einerseits nicht weiß, welche Abbildungen ϕ für die gegebenen Punktmengen vermutlich "günstig" sein werden, und es andererseits sehr aufwändig sein kann, $z_i = \phi(x_i)$ für alle Datenpunkte explizit zu berechnen. Da aber die Zielfunktion des dualen Problems nur von $\langle x_i | x_j \rangle$ abhängt, lässt sich mit der (konzeptionellen) Definition $K(x_i, x_j) := \langle \phi(x_i) | \phi(x_j) \rangle = \langle z_i | z_j \rangle$ die explizite Berechnung von z_i und z_j umgehen, wenn man einen einfachen Ausdruck für das entsprechende Skalarprodukt finden kann.

Noch besser (und das ist der Trick!): Man muss weder den Raum $\mathcal{Z}$ noch explizit die Abbildung ϕ kennen, solange sich $K(\cdot, \cdot)$ so verhält wie ein Skalarprodukt (in irgendeinem Raum)!

7.1 Definition (Kernel)**.** Eine Funktion $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ heißt *Kernel* oder *Kernel-Funktion* oder *Kern* auf $\mathbb{R}^d$, wenn es einen Raum $\mathcal{Z}$, der mit einem Skalarprodukt $\langle \cdot | \cdot \rangle$ ausgestattet ist, und eine Abbildung $\phi : \mathbb{R}^d \to \mathcal{Z}$ gibt, so dass $K(x, y) = \langle \phi(x) | \phi(y) \rangle$ für alle $x, y \in \mathbb{R}^d$ gilt.

Wann ist das der Fall? Betrachten wir eine notwendige Bedingung im euklidischen Raum, etwa $\mathbb{R}^d$, denn ϕ könnte ja auch die identische Abbildung sein. Dann ist $K(x, y) = \langle x | y \rangle$. Sei nun $X = \{ x_1, \ldots, x_n \} \subset \mathbb{R}^d$ eine endliche Teilmenge des $\mathbb{R}^d$ und $K_X := (\langle x_i | x_j \rangle)_{i,j=1,\ldots,n}$ die (symmetrische) $n \times$n-Matrix der Werte des Skalarprodukts auf X (die *Kernel-Matrix* auf

der Menge X). Sei ferner $\hat{X}$ die $n \times d$-Matrix, deren Spalten die Punkte x_i bilden. Dann ist $K_X = \hat{X}\hat{X}^T$ positiv semidefinit, denn es ist $\langle c|K_X|c\rangle = \langle c|\hat{X}\hat{X}^T|c\rangle = \|c^T\hat{X}\|^2 \geq 0$ für alle $c \in \mathbb{R}^n$.

Der folgende Satz von Mercer zeigt, dass die Bedingung, dass die Kernelmatrix positiv semidefinit ist, auch hinreichend ist.

7.2 Satz (Satz von Mercer)**.** *Sei $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ eine symmetrische Funktion. Sei $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ eine endliche Teilmenge des $\mathbb{R}^d$. Sei $K_X := (K(x_i, x_j))_{i,j}$ die Matrix der Werte von K auf X. Wenn K_X für alle endlichen $X \subset \mathbb{R}^d$ positiv semidefinit ist, dann ist K ein Kernel.*

7.3 Beispiel (Kernel-Funktionen)**.** Die folgenden Funktionen sind Kernels auf $\mathbb{R}^d$.

1. $K(x, y) = \langle x|y\rangle^d$
2. $K(x, y) = (\langle x|y\rangle + 1)^d$
3. $K(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ für $\sigma > 0$
4. $K(x, y) = \tanh(k\,\langle x|y\rangle + c)$ für geeignete $k \geq 0$ und $c < 0$.

♡

Der Kernel-Trick besteht nun darin, diese Tatsache auszunutzen: Man kann Kernel-Funktionen konstruieren und angeben, *ohne* die Abbildung ϕ und den Zielraum $\mathcal{Z}$ explizit zu kennen. Der Trick wurde erstmals in der Arbeit [M. Aizerman, E. Braverman, and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control 25: 821–837] beschrieben und ist seitdem im maschinellen Lernen extrem wichtig geworden.

Wir betonen, dass sich zwar das duale Problem mit Hilfe von Kernels aufstellen lässt, ohne ϕ zu kennen, nicht aber das primale Problem: Dort müsste explizit jedes Vorkommen von x durch $\phi(x)$ ersetzt werden und der Gewichtsvektor w läge in dem unbekannten Raum $\mathcal{Z}$.

Schlupfvariablen. Besonders ärgerlich ist, wenn sich fast alle Punkte perfekt linear trennen lassen (ob mit oder ohne Kernel), es aber eine oder wenige Ausnahmen gibt, die die lineare Trennbarkeit zunichte machen. In diesem Fall möchte man zulassen, dass einige der Ungleichungen "ein wenig" verletzt sein können. Dasselbe Argument gilt für den Fall, wenn zwar alle Punkte linear trennbar sind, aber der Rand (margin) aufgrund weniger Ausreißer sehr klein ist und ohne diese Ausreißer sehr viel größer sein könnte.

Wir führen daher nichtnegative Schlupfvariablen $z_i \geq 0$ ein, um die wir die linke Seite der i-ten Ungleichung vergrößern. Damit die z_i nur dort positiv gewählt werden, wo es nötig oder lohnenswert ist, soll die Summe der z_i minimiert werden. Dies führt eigentlich auf ein mehrkriterielles Optimierungsproblem (Minimiere $\|w\|_2^2$ und $z := \sum_i z_i$). Um die Diskussion mehrkriterieller Optimierung hier zu vermeiden, kombinieren wir beide Ziele in eine gemeinsame Zielfunktion, in der wir z mit dem Faktor $C > 0$ gewichten.

Aufgabe 7.1. Was passiert, wenn $C \to 0$ oder $C \to \infty$?

Das primale Problem lautet dann insgesamt

$$\begin{aligned}
&\text{Minimiere} && \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} z_i, && \\
&\text{so dass} && y_i \cdot (\langle w|x_i\rangle - b) - 1 + z_i \geq 0 && \text{für } i = 1, \ldots, n, \\
& && z_i \geq 0. &&
\end{aligned}$$

Die Variablen sind $(w, b, z) \in \mathbb{R}^{d+1+n}$. Das Gewicht $C > 0$ ist eine vom Anwender sinnvoll zu wählende Konstante.

Der Vorteil dieser Formulierung ist, dass das Problem durch geeignete Wahlen von z_i immer lösbar ist. Wie groß die Summe der z_i im Optimum wird, hängt von der Wahl von C ab.

Auch dieses Problem lässt sich dualisieren, und auf dem dualen Problem lässt sich der Kernel-Trick anwenden. Es gelten wieder Slater's Bedingungen und daher starke Dualität. Es ist

$$L((w, b, z), (\lambda, \mu)) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} z_i + \sum_{i=1}^{n} \lambda_i(1 - y_i(\langle w|x_i\rangle - b) - z_i) - \sum_{i=1}^{n} \mu_i z_i$$

Das Minimum über (w, b, z) finden wir wieder (da L in (w, b, z) konvex ist) durch Nullsetzen der Ableitungen:

$$\begin{aligned}
\partial L(w, b, z, \lambda, \mu)/\partial w &= w - \sum_i \lambda_i y_i x_i \stackrel{!}{=} 0, \\
\partial L(w, b, z, \lambda, \mu)/\partial b &= \sum_i \lambda_i y_i \stackrel{!}{=} 0, \\
\partial L(w, b, z, \lambda, \mu)/\partial z_i &= C - \lambda_i - \mu_i \stackrel{!}{=} 0.
\end{aligned}$$

Einsetzen in L liefert das duale Problem

$$\begin{aligned}
&\text{Maximiere} && -(1/2)\sum_{i,j} \lambda_i \lambda_j y_i y_j \langle x_i|x_j\rangle + \sum_i \lambda_i, \\
&\text{so dass} && \lambda \geq 0, \\
& && \mu \geq 0, \\
& && \mu_i = C - \lambda_i \\
& && \langle \lambda|y\rangle = 0.
\end{aligned}$$

Wir kernelisieren die Zielfunktion und eliminieren μ, indem wir die Bedingung $\lambda \leq C$ hinzufügen. Das Problem lautet jetzt

$$\begin{aligned}
&\text{Maximiere} && -(1/2)\sum_{i,j} \lambda_i \lambda_j y_i y_j K(x_i, x_j) + \sum_i \lambda_i, \\
&\text{so dass} && 0 \leq \lambda \leq C, \\
& && \langle \lambda|y\rangle = 0.
\end{aligned}$$

Gegenüber dem Originalproblem ohne Schlupfvariablen ist also nur die Einschränkung $\lambda \leq C$ hinzugekommen.

Wir müssen noch klären, wie man aus der dualen Lösung λ^* den Klassifikator eines neuen Punktes x bekommt. Im primalen Problem ist das einfach das Vorzeichen von $\langle w|x\rangle - b$ (auf welcher Seite der Hyperebene liegt x?), aber da wir kernelisiert haben, müsste man $\langle w|\phi(x)\rangle$ in dem unbekannten Raum $\mathcal{Z} \ni w$ ausrechnen. Glücklichweise wissen wir aus der Optimalitätsbedingung für L, dass $w = \sum_i \lambda_i y_i \phi(x_i)$ gilt. Daher ist

$$\langle w|\phi(x)\rangle = \langle \sum_i \lambda_i y_i \phi(x_i)|\phi(x)\rangle = \sum_i \lambda_i y_i K(x_i, x),$$

und die Summe muss nur über die Datenpunkte x_i gebildet werden, für die $\lambda_i > 0$ ist. Diese nennt man *Stützvektoren* oder *support vectors*, daher der Name *support vector machine* für diese Klassifikationsmethode. Den Offset b bekommen wir mit Hilfe des komplementären Schlupfes: Ist x_j ein Stützvektor ($\lambda_i > 0$), dann muss $y_j(\langle w|\phi(x_j)\rangle - b) - 1 = 0$ gelten, also

$$b = \langle w|\phi(x_j)\rangle - y_j = \sum_i \lambda_i y_i K(x_i, x_j) - y_j.$$

Praxis. Normalisierung der Daten. Wahl von $C > 0$. Wahl des Kernels.

7.2 Optimierung der Kommunikationsrate

7.3 Basis-Pursuit-Probleme

KAPITEL 8

Innere-Punkte-Verfahren für konvexe Optimierungsprobleme

Abschließend kommen wir zu den Verfahren, allgemeine konvexe Optimierungsprobleme mit Nebenbedingungen zu lösen. Als Basis dient das in Kapitel 5 besprochenen Newton-Verfahren (alternativ das BFGS-Verfahren). Wir gehen so vor, dass wir zunächst nur Gleichungen $Ax = b$ als Nebenbedingungen betrachten und im Anschluss zu Methoden kommen, die auch Ungleichungen umgehen können. Es handelt sich bei den Verfahren um sogenannte *innere-Punkte-Verfahren*, da stets primale Punkte x betrachtet werden, die strikt zulässig sind, etwaige Ungleichungen also mit $f_i(x) < 0$ erfüllen.

Alle Probleme in diesem Kapitel werden als konvex vorausgesetzt, alle auftretenden Funktionen seien so glatt (differenzierbar) wie nötig (i.d.R. zwei mal stetig differenzierbar). Wir setzen weiter stillschweigend voraus, dass das primale Problem zulässig ist und (mindestens) eine optimale Lösung x^* mit Zielfunktionswert $f_0(x^*) = p^*$ existiert.

8.1 KKT-Bedingungen für konvexe Probleme mit Gleichungen

Wir behandeln das konvexe Problem

$$\begin{aligned} &\text{Minimiere } f(x), && (8.1)\\ &\quad \text{so dass } Ax = b && \text{mit } A \in \mathbb{R}^{p\times n},\ \text{Rang } A = p < n, \end{aligned}$$

wobei f zwei mal stetig differenzierbar sei. Die Voraussetzungen des vollen Rangs von A lässt sich durch Vorverarbeitung stets erreichen. Die Voraussetzung $p < n$ dient dazu, den Trivialfall, dass A invertierbar ist, zu vermeiden; in dem Fall wäre ohnehin nur $x^* = A^{-1}b$ zulässig und damit optimal.

Wenn wir voraussetzen, dass es überhaupt zulässige Punkte im relativen Inneren des Definitionsbereichs gibt, gilt auch starke Dualität.

Wir stellen die KKT-Bedingungen auf, die die Lösung (x^*, ν^*) charakterisieren. Da keine Ungleichungen vorkommen, fallen die KKT-Bedingungen weg, die λs enthalten. Es bleibt die primale Zulässigkeitsbedingung $Ax^* = b$ und die Bedingung, dass die x-Ableitung der Lagrangefunktion im Optimum verschwindet. Die Lagrange-Funktion lautet $L(x, \nu) = f(x) + \langle \nu | Ax - b \rangle$. Ableiten nach x liefert $\nabla f(x^*) + A^T \nu^* = 0$. Dies wird (nicht ganz korrekt) auch als duale Zulässigkeitsbedingung bezeichnet.

Insgesamt haben wir also die $n+p$ Variablen (x^*, ν^*), dazu das in der Regel nichtlineare Gleichungssystem zur dualen zulässigkeit mit n Gleichungen und das lineare Gleichungssystem zur primalen Zulässigkeit mit p Gleichungen $Ax^* = b$.

Quadratische Funktion. Es lohnt sich, den Spezialfall zu betrachten, dass f ein Polynom vom Grad höchstens 2 ist, also

$$f(x) = \frac{1}{2} \langle x | P | x \rangle + \langle q | x \rangle + r,$$

mit $P \in \S_+^n$, $q \in \mathbb{R}^n$ und $r \in \mathbb{R}$. Dadurch wird die Bedingung $\nabla f(x^*) + A^T \nu^* = 0$ linear und lautet $Px^* + q + A^T \nu^* = 0$. Insgesamt ergibt sich das lineare *KKT-System*

$$\begin{pmatrix} P & A^T \\ A & 0 \end{pmatrix} \cdot \begin{pmatrix} x^* \\ \nu^* \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}.$$

Die Matrix auf der linken Seite heißt auch *KKT-Matrix*. Wir können drei Fälle unterscheiden.

1. Ist die KKT-Matrix invertierbar, dann gibt es genau einen Lösung (x^*, ν^*) des KKT-Systems; dies ist der eindeutige primal-dual optimale Punkt.
2. Ist die KKT-Matrix singulär, das System aber lösbar, dann gibt es mehrere äquivalente Lösungen, die sich im Wert der primalen und dualen Funktion nicht unterscheiden.
3. Ist die KKT-Matrix singulär und das KKT-System nicht lösbar, dann ist das primale Problem unbeschränkt, denn den Fall, dass es nicht zulässig ist, haben wir ausgeschlossen.

8.2 Lösungsansätze für konvexe Probleme mit Gleichungen

Es stellt sich nun allgemein die Frage, wie man (auch im nichtquadratischen Fall) das primale Problem löst. Wir sprechen drei Möglichkeiten an, verfolgen im weiteren aber insbesondere die letzte.

1. Übergang zum dualen Problem mit Hilfe der konjugierten Funktion
2. Angabe einer Basis des Lösungsraums von $Ax = b$ und Neuformulierung des Problems als unrestringiertes Problem
3. Anpassung des Newton-Verfahrens an die Gleichungs-Bedingung $Ax = b$

Zu 1.: Das duale Problem lautet unter Benutzung der konjugierten Funktion f^* zu f wie folgt (Übung):

$$\text{Maximiere } -\langle b|\nu\rangle - f^*(-A^T\nu).$$

Dies ist formal ein unrestringiertes Problem, das aber nur dann wirklich unrestringiert ist, wenn f^* auf dem Bild von A^T wirklich definiert (und nicht unendlich) ist; ansonsten würden aus den Bedingungen, die den Definitionsbereich von f^* charakterisieren, neue Gleichungen oder Ungleichungen entspringen. Dieser Ansatz ist also nur dann sinnvoll, wenn sich f^* ausrechnen lässt oder bekannt ist und überall definiert ist.

Zu 2.: Allgemein lässt sich folgender Ansatz verwenden, der darin besteht, eine Basis des Lösungsraums für das System $Ax = b$ auszurechnen. Da $A \in \mathbb{R}^{p\times n}$ vollen Rang p hat, gibt es $n - p$ freie Variablen und der Lösungsraum hat Dimension $n - p$. Finden wir für diesen eine Basis $\{ q_1, \ldots, q_{n-p} \} \subset \mathbb{R}^n$, die wir als Spalten einer $n \times (n - p)$ Matrix Q schreiben und außerdem eine spezielle Lösung, dann lassen sich alle Lösungen schreiben als $\{ x \mid Ax = b \} = \{ x_0 + Qz \mid z \in \mathbb{R}^{n-p} \}$. Damit wird das Problem äquivalent formuliert als unrestringiertes Problem

$$\text{Minimiere } f_0(z) := f(x_0 + Qz).$$

Der Grund, warum diese Variante häufig nicht gewählt wird, ist, dass das "Lösen" von $Ax = b$ in Form von Q häufig die spezielle Struktur von A (insbesondere Dünnbesetztheit) zerstört und man diese im weiteren Verlauf nicht mehr ausnutzen kann.

Im weiteren behandeln wir daher eine Anpassung des unrestringierten Newton-Verfahrens an die zusätzliche Bedingung $Ax = b$.

8.3 Das Newton-Verfahren unter Gleichungs-Bedingungen

Das Newton-Verfahren zur Lösung des unrestringierten Problems (Kapitel 5) löst iterativ die Gleichung $r(x) := \nabla f(x) = 0$. (Wir schreiben hier einfach f für die Zielfunktion f_0.) Wir nennen den Wert $r(x)$ auch das zum Punkt x gehörende *Residuum*. Das Newton-Verfahren verwendet die Ableitungen von r, also die Hesse-Matrix $\nabla^2 f(x)$.

Unter Gleichungs-Bedingungen kommen duale Variablen $\nu \in \mathbb{R}^p$ hinzu, die Gleichungen werden komplizierter. Die KKT-Bedingungen in (x, ν) lauten ja

1. $\nabla f(x^*) + A^T\nu^* = 0$
2. $Ax^* - b = 0$

Wir definieren also das Residuum als

$$r(x,\nu) \equiv (r_{\text{dual}}(x,\nu), r_{\text{pri}}(x,\nu)) := (\nabla f(x) + A^T\nu, Ax - b)$$

und nennen die beiden Komponenten das *duale Residuum* und das *primale Residuum*. Mit dieser Definition ist die Aufgabe wieder, das (nichtlineare) Gleichungssystem $r(x, \nu) = 0$ zu lösen. Schreiben wir $y = (x, \nu)$ und $Dr(y)$ für die Jakobi-Matrix (Matrix der Ableitungen) von r, dann lautet die Newton-Gleichung für einen Newton-Schritt allgemein $Dr(y) \cdot \Delta y =$

$-r(y)$; dabei ist Δy die Newton-Richtung, also die Differenz zwischen alter und neuer Iteration.

Ausgeschrieben in (x, ν) und r_{dual}, r_{pri} lautet ein Newton-Schritt hier

$$\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \cdot \begin{pmatrix} \Delta x \\ \Delta \nu \end{pmatrix} = - \begin{pmatrix} \nabla f(x) + A^T \nu \\ Ax - b \end{pmatrix}. \tag{8.2}$$

Ist $\nabla^2 f(x)$ zumindest auf dem Nullraum von A stets positiv definit (was z.B. unter der Voraussetzung strikter oder starker Konvexitäts stets der Fall ist), ist dies System immer lösbar, und $(x + \Delta x, \nu + \Delta \nu)$ ist der Punkt, der die quadratische Taylor-Approximation von $r(x, \nu)$ minimiert.

Schrittweitensteuerung. Bekannt ist, dass das Newton-Verfahren lokal quadratisch konvergiert. Um globale Konvergenz für konvexe Funktionen zu erreichen, muss wie beim unrestringierten Newton-Verfahren eine Schrittweitensteuerung eingeführt werden, d.h. wird ggf. ein kürzerer Schritt $(t\Delta x, t\Delta \nu)$ mit $0 < t \leq 1$ statt einem vollen Newton-Schritt $(t = 1)$ ausgeführt.

Wir stellen zunächst fest, dass die Δx-Komponente der Newton-Richtung hier nicht notwendigerweise eine Abstiegsrichtung für f ist, dass aber stets $\Delta y := (\Delta x, \Delta \nu)$ eine Abstiegsrichtung für $\|r(y)\|_2^2$ ist, wobei $y = (x, \nu)$.

8.1 Lemma. *Die Newton-Richtung $\Delta y = (\Delta x, \Delta \nu)$ aus (8.2) ist eine Abstiegsrichtung für die (quadrierte) Norm des Residuums.*

Beweis. Sei $\phi_r(t) := \|r(y + t\Delta y)\|_2^2$. Dann ist $\phi_r'(0) = 2 \langle r(y) | Dr(y) | \Delta y \rangle = -2\|r(y)\|_2^2 < 0$, sofern $r(y) \neq 0$. Durch Übergang zur Wurzel folgt weiter $\frac{\mathrm{d}}{\mathrm{d}t}[\|r(y + t\Delta y)\|_2] \mid_{t=0} = -\|r(y)\|_2$.

Bemerkung: Ist hingegen $\phi_f(t) := f(x + t\Delta x)$, dann ist $\phi_f'(0) = ...$; dies ist nicht notwendig negativ! □

Daher ist eine Schrittweitensteuerung über die Residuums-Norm (aber nicht direkt über f!) möglich; wir stellen sicher, dass sie genügend stark fällt (im Vergleich zur Schrittweite). Wir erhalten das folgende Newton-Verfahren mit Armijo-Schrittweite.

- Gegeben: Startpunkt $(x, \nu) \in \mathcal{D} \times \mathbb{R}^p$, Genauigkeit $\varepsilon > 0$, Armijo-Schrittweitenparameter $0 < \alpha < 1/2$ und $0 < \beta < 1$.
- Wiederhole:
  1. Berechne den Newtonschritt $\Delta y = (\Delta x, \Delta \nu)$ aus 8.2.
  2. Berechne die Armijo-Schrittweite zu α und β:
     - $t := 1$
     - Solange $\|r(y + t\Delta y)\|_2 > (1 - \alpha t)\|r(y)\|_2$: $t := \beta t$.
  3. Aktualisiere $x := x + t\Delta x$ und $\nu := \nu + t\Delta \nu$.
- bis $Ax = b$ und $\|r(x, \nu)\|_2 \leq \varepsilon$.

Wir machen zu diesem Verfahren einige Aussagen:

1. In jedem Schritt reduziert sich das primale Residuum um den Faktor $1 - t$: Es ist $r^+_{\text{pri}} = A(x + t\Delta x) - b = (1 - t)(Ax - b) = (1 - t)r_{\text{pri}}$. Wird also einmal $t = 1$ gewählt, gilt danach stets primale Zulässigkeit $Ax = b$. Ab diesem Zeitpunkt kann statt $\|r\|_2$ auch wieder f zur Schrittweitenwahl und zur Terminierung herangezogen werden.
2. Die Wahl von $\|r\|_2$ zur Schrittweitenwahl garantiert, auch im Fall von Nichtzulässigkeit, dass die Schrittweitensuche nach endlich vielen Schritten abbricht, aufgrund von Lemma 8.1.
3. Statt der Hesse-Matrix $\nabla^2 f(x)$ kann auch wieder eine positiv definite Approximation verwendet werden (BFGS-Verfahren). Wir gehen auf die Details nicht ein. Zu beachten ist aber, dass nicht nur die Hesse-Matrix, sondern die gesamte KKT-Matrix invertiert werden muss.
4. Wir halten fest: Das Newton-Verfahren (oder ein Quasi-Newton-Verfahren wie BFGS) löst ein konvexes Problem mit Gleichungsbedingungen, indem es eine Folge von quadratischen Problemen mit Gleichungsbedingungen löst.

8.4 Das Logarithmische-Barriere-Verfahren

Die Hauptidee, um nun auch mit Ungleichungen umzugehen, besteht darin, um die zulässige Menge eine "Barriere" aufzubauen, so dass der Zielfunktionswert gegen unendlich geht, sobald man sich vom Inneren der zulässigen Menge dem Rand annähert. Ein konkretes Beispiel dafür ist die *logarithmische Barrierefunktion*, die wir im folgenden verwenden.

8.2 Definition (Logarithmische Barriere)**.** Zu den konvexen Ungleichungen $f_i(x) \leq 0$, $i = 1, \ldots, m$ ist die *logarithmische Barriere(funktion)* definiert als

$$\phi(x) := -\sum_{i=1}^{m} \log(-f_i(x)).$$

Ihr Definitionsbereich ist $\operatorname{dom}\phi = \{\, x \mid x \in \operatorname{dom} f_i \text{ und } f_i(x) < 0 \text{ für alle } i = 1, \ldots, m \,\}$.

8.3 Lemma (Konvexität der logarithmischen Barriere)**.** *Die logarithmische Barriere ϕ ist konvex, wenn die f_i konvex sind.*

Beweis. Die Funktion $u \mapsto -\log(-u)$ ist für $u < 0$ konvex, monoton wachsend (Skizze!) und differenzierbar. Die Komposition mit f_i ist daher wieder konvex. □

Die grundlegende Idee ist nun, die Zielfunktion f_0 mit der Barriere-Funktion zu kombinieren (gewichtete Addition), also mit $\gamma > 0$ die Funktion $f_0 + (1/\gamma)\phi$ zu bilden. Diese nimmt außerhalb von $\operatorname{dom}\phi$ den Wert unendlich an. Für große Werte von γ ist jedoch der Einfluss der Barriere innerhalb der zulässigen Menge zu vernachlässigen; daher darf man hoffen, dass das ursprüngliche Problem mit Ungleichungen und das Problem mit der modifizierten Zielfunktion gegen die gleiche Lösung konvergieren, wenn $\gamma \to \infty$. Es ist klar, dass man alternativ die Zielfunktion $\gamma f_0 + \phi$ betrachten kann. Die Gleichungs-Bedingungen werden nicht verändert.

8.4 Definition (Barriere-Problem)**.** Wir definieren zum konvexen Problem

$$\begin{aligned} &\text{Minimiere } f_0(x), \\ &\quad \text{so dass } f_i(x) \leq 0, && i = 1, \ldots, m, \\ &\qquad Ax = b && A \in \mathbb{R}^{p \times n}, \operatorname{Rang} A = p < n, \end{aligned} \tag{8.3}$$

das konvexe *Barriere-Problem* mit *Zentralitätsparameter* $\gamma > 0$ als

$$\begin{aligned} &\text{Minimiere } \gamma f_0(x) + \phi(x), \\ &\quad \text{so dass } Ax = b. \end{aligned} \tag{8.4}$$

Dabei ist $\phi(x)$ die Barriere-Funktion aus Definition 8.2.

Wir nehmen an, dass das Barriere-Problem für alle $\gamma > 0$ eine eindeutige Lösung $x^*(\gamma)$ besitzt. Dies kann zum Beispiel dadurch erreicht werden, dass die Ungleichung $\|x\|_2^2 \leq R^2$ für ein genügend großes $R > 0$ zum Problem hinzugefügt wird, so dass ϕ den Term $\|x\|_2^2 - R^2$ enthält, der ϕ strikt konvex macht.

8.5 Definition (Zentraler Pfad)**.** Die Lösungen $x^*(\gamma)$ des Barriere-Problems heißen *zentrale Punkte*. Die Menge $\{\, x^*(\gamma) \mid \gamma > 0 \,\}$ heißt zentraler Pfad.

Das Barriere-Problem ist ein konvexes Problem mit ausschließlich Gleichungsbedingungen und kann daher mit den Methoden aus den Abschinitten 8.2 und 8.3 gelöst werden. Entscheidend und möglicherweise problematisch ist, dass man zum Start überhaupt erst ein x_0 braucht, dass alle Ungleichungen strikt erfüllt. Im Moment nehmen wir an, dass dieses x_0 leicht zu beschaffen ist; wir kommen auf die Problematik in Abschnitt 8.5 zurück.

Mit der Zielfunktion $f(x) := \gamma f_0(x) + \phi(x)$ und Nebenbedingungen $Ax = b$ lauten Gradient und Hessematrix für das Newton-Verfahren wie folgt:

$$\begin{aligned} \nabla f(x) &= \gamma \nabla f_0(x) + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x) \\ \nabla^2 f(x) &= \gamma \nabla^2 f_0(x) + \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \left|\nabla f_i(x)\right\rangle \left\langle \nabla f_i(x)\right| + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x) \end{aligned}$$

Es gibt eine mechanische Interpretation der zentralen Punkte; dabei lassen wir Gleichungsbedingungen der Einfachheit halber weg. Wir stellen uns vor, dass auf jeden (strikt zulässigen) Punkt Kräfte wirken, und zwar zunächst die Kraft $F_0(x) := -\gamma \nabla f_0(x)$ in Richtung zum Minimum, andererseits weg vom Rand mit $F_i(x) := \nabla \log(-f_i(x)) = 1/f_i(x) \cdot \nabla f_i(x)$ für $i = 1, \ldots, m$. Der zentrale Punkt $x^*(\gamma)$ ist derjenige, in dem insgesamt keine Kraft wirkt.

Woher weiß man, welches $\gamma > 0$ eine sinnvolle Parameterwahl zum Lösen des Problems (8.3) ist? Interessanterweise liefern Punkte auf dem zentralen Pfad auch dual zulässige Punkte, mit denen die Dualitätslücke zu jedem $\gamma > 0$ bestimmt werden kann.

8.6 Satz (Duale Punkte und Dualitätslücken auf dem zentralen Pfad)**.** *Sei* $(x^*(\gamma), \tilde{\nu}_\gamma)$ *die primal-duale Lösung des Barriere-Problems zu* $\gamma > 0$*. Dann sind*

$$\lambda_i^*(\gamma) := -\frac{1}{\gamma f_i(x^*(\gamma))} > 0, \qquad \nu^*(\gamma) := \tilde{\nu}_\gamma / \gamma$$

dual zulässig für das primale Problem (8.3)*, und die Dualitätslücke beträgt* $f_0(x^*(\gamma)) - g(\lambda^*(\gamma), \nu^*(\gamma)) = m/\gamma$ *(dabei war* m *die Anzahl der Ungleichungen). Insbesondere gilt*

$$f_0(x^*(\gamma)) - p^* \leq m/\gamma.$$

Beweis. Die Optimalitätsbedingung aus dem KKT-System für das Barriereproblem lautet $\nabla f(x^*(\gamma)) + A^T \tilde{\nu}_\gamma = 0$. Nach Division durch γ und Einsetzen des Gradienten heißt das

$$\nabla f_0(x^*(\gamma)) \, + \, \sum_{i=1}^{m} \frac{1}{-\gamma f_i(x^*(\gamma))} \nabla f_i(x^*(\gamma)) \, + \, A^T \tilde{\nu}_\gamma / \gamma \, = \, 0.$$

Vergleich mit der KKT-Optimalitätsbedingung für das primale Problem

$$\nabla f_0(x^*(\gamma)) + \sum_{i=1}^{m} \lambda_i^*(\gamma) \, \nabla f_i(x^*(\gamma)) + A^T \, \nu^*(\gamma) = 0$$

zeigt, dass $x^*(\gamma)$ die Lagrange-Funktion $L(x, \lambda^*(\gamma), \nu^*(\gamma))$ minimiert; daher und wegen $\lambda^*(\gamma) > 0$ ist $(\lambda^*(\gamma), \nu^*(\gamma))$ dual zulässig.

Der Wert der dualen Funktion in $(\lambda^*(\gamma), \nu^*(\gamma))$ ist daher

$$\begin{aligned} g(\lambda^*(\gamma), \nu^*(\gamma)) &= f_0(x^*(\gamma)) + \sum_{i=1}^{m} \lambda_i^*(\gamma) \, f_i(x^*(\gamma)) + \langle \nu^*(\gamma) | A x^*(\gamma) - b \rangle \\ &= f_0(x^*(\gamma)) - m/\gamma; \end{aligned}$$

dies beweist die Aussage über die Dualitätslücke. □

Man kann den Satz noch etwas anders formulieren und die Bedingungen als *deformierte KKT-Bedingungen* interpretieren.

8.7 Satz. *Ein Punkt* $x \in \mathbb{R}^n$ *ist genau dann gleich* $x^*(\gamma)$*, wenn es* $0 < \lambda \in \mathbb{R}^m$ *und* $\nu \in \mathbb{R}^p$ *gibt, so dass die folgenden deformierten KKT-Bedingungen gelten:*

$$\begin{aligned} Ax &= b, & \\ f_i(x) &< 0, & i = 1, \ldots, m, \\ \lambda &> 0, & \\ \lambda_i f_i(x) &= -1/\gamma & i = 1, \ldots, m, \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + A^T \nu &= 0. & \end{aligned}$$

(Der einzige Unterschied zu den "normalen" KKT-Bedingungen liegt in den komplementären Schlupf-Bedingungen und besteht im Auftreten des Parameters γ*. Für* $\gamma \to \infty$ *ergeben sich die normalen KKT-Bedingungen.)*

Beweis. Der Beweis ergibt sich unmittelbar aus dem vorigen Satz. □

Es scheint nun so, als müsste man nur das Barriere-Problem für $\gamma := 2m/\varepsilon$ lösen, um eine Lösung des primalen Problems mit Genauigkeit $\varepsilon/2$ zu erhalten. Wenn man dieses Barriere-Problem auch mit Genauigkeit $\varepsilon/2$ löst, ist der Fehler höchstens ε.

Die direkte Lösung funktioniert leider nicht besonders gut. Das Problem ist, dass die Zielfunktion f für große Werte von γ am Rand der zulässigen Menge beinahe unstetig wird ("Sprung" auf unendlich am Rand). Das Newton-Verfahren konvergiert daher nur sehr langsam, wenn man mit einem weit entfernten Startwert startet. Eine bessere Idee ist es, den Parameter γ zunächst relativ klein zu wählen, die Lösung $x^*(\gamma)$ zu berechnen (mit vergleichsweise wenig Newton-Iterationen), dann γ zu erhöhen und von der alten Lösung aus die neue Lösung zu berechnen. Auf diese Art und Weise "folgt" man dem zentralen Pfad bis zum Optimum.

Algorithmus. Es ergibt sich folgender Algorithmus, das *Barriere-Verfahren*:

- Gegeben: primales Problem (8.3), Startpunkt x strikt zulässig in den Ungleichungen, Toleranz $\varepsilon > 0$, Startparameter $\gamma > 0$, Multiplikator $\mu > 1$.
- Wiederhole:
  1. Zentrierungsschritt: Bilde und löse das Barriere-Problem (8.4) zum Parameter γ mit hinreichender Genauigkeit (z.B. $\varepsilon/2$, insbesondere wenn $\gamma \geq 2m/\varepsilon$); die Lösung sei $x^*(\gamma)$.
  2. Setze $x := x^*(\gamma)$.
  3. Abbruch mit Lösung x, wenn $m/\gamma \leq \varepsilon/2$.
  4. Erhöhe $\gamma := \mu \cdot \gamma$.
- Mit dem gefundenen x gilt $f_0(x) - p^* \leq \varepsilon$.
  Ohne wesentlichen Mehraufwand kann auch ein passender dualer Punkt zurückgeliefert werden.

Höchstens im ersten Zentrierungsschritt wird mit einem unzulässigen x (das nicht $Ax = b$ erfüllt) gestartet; das erste Barriere-Problem wird mindestens so genau gelöst, dass danach $Ax = b$ gilt. Um die jeweiligen Barriere-Probleme zu lösen, müssen noch entsprechende Schrittweiten-Parameter α, β festgelegt werden.

Parameterwahl. Es stellt sich die Frage nach der Wahl des Startparameters $\gamma > 0$. Hierzu gibt es verschiedene Stragegien. Am einfachsten ist es, wenn die duale Funktion und ein dual zulässiger Punkt (λ, μ) für das Problem bekannt sind. Dann lässt sich mit dem Startpunkt $x = x_0$ die Dualitätslücke $\delta := f(x_0) - g(\lambda, \mu)$ berechnen $\gamma := m/\delta$ setzen; damit finden wir einen Punkt auf dem zentralen Pfad mit derselben Dualitätslücke. Alternativ können wir zur KKT-Bedingung des Barriere-Problem mit festem x_0,

$$\gamma \nabla f_0(x_0) + \nabla \phi(x_0) + A^T \nu = 0$$

optimale (γ, ν) finden, indem wir das unrestringierte kleinste-Quadrate-Problem

$$\min \|\gamma \nabla f_0(x_0) + \nabla \phi(x_0) + A^T \nu\|_2^2$$

lösen. Es gibt viele andere Möglichkeiten, zum Beispiel auch einfach $\gamma = 1$.

Es stellt weiter sich die Frage, wie hoch der Multiplikator $\mu > 1$ gewählt werden sollte und wie genau die einzelnen Barriere-Probleme (außer dem letzten) gelöst werden müssen und wie viele Newton-Schritte dies jeweils benötigt. Intuitiv erwarten wir folgendes:

- Ist $\mu \approx 1$, benötigt das Barriere-Verfahren viele äußere Iterationen, aber jedes einzelne Barriere-Problem kann in wenigen (inneren) Iterationen (z.B. Newton- oder BFGS-Schritten) gelöst werden; die einzelnen berechneten Zwischenlösungen $x^*(\gamma)$ liegen dicht beieinander auf dem zentralen Pfad.
- Ist μ sehr groß, benötigt das Verfahren im Extremfall nur zwei äußere Iterationen, die aber beide relativ viele innere Iterationen erfordern können.

In der Praxis zeigt sich, dass eine Wahl $\mu \in [3, 200]$ fast keinen Unterschied macht, da sich die Effekte nahezu gegenseitig aufheben. Ungünstig sind zu kleine und zu große Wahlen.

8.5 Finden eines Startpunktes: Phase I

Die Barriere-Methode setzt voraus, dass ein Startpunkt bekannt ist, der alle Ungleichungen strikt erfüllt. Oft ist aber gar nicht bekannt, ob ein solcher überhaupt existiert. Um dies festzustellen, kann man ein Machbarkeits- oder auch Zulässigkeitsproblem zum Problem (8.3) formulieren und lösen.

Hierzu gibt es mehrere Möglichkeiten. Die grundlegende Idee ist stets, die Schranke 0 bei den Ungleichungen variabel zu machen, die Ungleichung also durch $f_i(x) \leq s_i$ zu ersetzen und dann zu fragen, ob es zu negativen s_i Punkte x gibt, die die Ungleichungen erfüllen.

8.8 Problem (Unzulässigste Ungleichung)**.** Wir betrachten zum Problem (8.3) das folgende Problem:

$$\begin{aligned} &\text{Minimiere } s, \\ &\quad \text{so dass } f_i(x) \leq s, && i = 1, \dots, m, \\ &\qquad\quad Ax = b. \end{aligned}$$

Wir suchen also das kleinste s, für das ein zulässiges x existiert, mit dem *alle* s-Ungleichungen erfüllt sind. •

Ist in Problem 8.8 das optimale s^* negativ, sind alle Ungleichungen strikt erfüllbar, und das optimale x^* erfüllt, wenn es angenommen wird, alle Ungleichungen strikt, kann also als Startwert für das Hauptproblem (8.3) herangezogen werden. In jedem Fall gibt es strikt zulässige x. Ist $s^* = 0$, dann ist die zulässige Menge nur dann nicht leer, wenn das Optimum in einem x^* angenommen wird. In keinem Fall gibt es strikt zulässige x. Ist $s^* > 0$, dann ist die zulässige Menge leer; insbesondere gibt es keine strikt zulässigen x. Aufgrund numerischer Genauigkeit lässt sich 0 nicht von kleinen $\pm\varepsilon$ unterscheiden. Der Einsatz der Barriere-Methode ist nur dann sinnvoll möglich, wenn s^* hinreichend negativ ist.

8.9 Problem (Kleinste Summe der Unzulässigkeiten)**.** Wir betrachten zum Problem (8.3) das folgende Problem:

$$\begin{aligned} &\text{Minimiere } \sum_i s_i, \\ &\text{so dass } f_i(x) \le s_i, && i = 1, \dots, m, \\ &\qquad Ax = b, \\ &\qquad s \ge 0. \end{aligned}$$

Wir suchen also nichtnegative s_i, so deren Summe minimal ist, und es ein $x \in \mathcal{D}$ gibt, für das $f_i(x) \le s_i$ für alle i gilt. •

Ein Zielfunktionswert von 0 zeigt an, dass das Problem zulässig ist und die Lösung liefert ein (eventuell strikt) zulässiges x. Im Gegensatz zum ersten Problem erhält man bei Nichtzulässigkeit hier häufig Lösungen, bei der mehr Ungleichungen mit $s_i = 0$ erfüllt sind. Dies kann dann nützlich sein, wenn man bereit ist, auf einige der Ungleichungen (aber nicht zu viele) zu verzichten.

In beiden Fällen kann man als Startpunkt irgendein x im Schnitt der Definitionsbereiche der f_i nehmen, $s_i := f_i(x)$ ausrechnen und für das erste Problem $s := \max_i s_i$ als Startwert wählen.

8.6 Implementierungsprojekt

KAPITEL 9

Verallgemeinerungen und Spezialisierungen der Konvexität

9.1 Quasikonvexe Funktionen

Wir haben gesehen, dass die Niveaumengen einer konvexen Funktion stets konvex sind, aber dass die Umkehrung im allgemeinen nicht gilt. Hierfür füren wir einen eigenen Begriff ein, die *Quasikonvexität*. Es ist dann klar, dass jede konvexe Funktion auch quasikonvex ist, aber nicht umgekehrt.

9.1 Definition (quasikonvex, quasikonkav, quasilinear)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit $\operatorname{dom} f = C$ heißt *quasikonvex*, wenn C konvex ist und für alle $t \in \mathbb{R}$ die Niveaumenge $S_t = \{ x \in \operatorname{dom} f \mid f(x) \leq t \}$ konvex ist. Sie heißt *quasikonkav*, wenn $-f$ quasikonvex ist. Sei heißt *quasilinear*, wenn sie quasikonvex und quasikonkav ist.

9.2 Beispiel (Quasikonvexe, quasikonkave und quasilineare Funktionen)**.**

- $x \mapsto \log x$ ist auf $\mathbb{R}_{++}$ quasikonvex und konkav (und damit auch quasikonkav), und daher quasilinear.
- $x \mapsto \lceil x \rceil = \inf \{ z \in \mathbb{Z} \mid z \geq x \}$ ("Aufrunden") ist quasikonvex und quasikonkav, daher quasilinear.
- Wir definieren die Länge eines Vektors $x \in \mathbb{R}^n$ als den größten Index i mit $x_i \neq 0$ und die Länge des Nullvektors als 0. Die Länge ist quasikonvex.
- Die Funktion $f(x_1, x_2) = x_1 x_2$ ist quasikonkav auf $\mathbb{R}^2_+$, aber nicht auf ganz $\mathbb{R}^2$.
- Linear-fraktionale Funktionen der Form $f(x) = \frac{\langle a|x\rangle + b}{\langle c|x\rangle + d}$ sind auf ihrem Definitionsbereich $\operatorname{dom} f = \{ x \mid \langle c|x\rangle + d > 0 \}$ quasikonvex, quasikonkav, und daher quasilinear.

♡

9.3 Lemma (Charakterisierung quasikonvexer Funktionen)**.** *Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit $\operatorname{dom} f = C$ konvex ist genau dann quasikonvex, wenn für alle $x, y \in C$ und alle $\lambda \in [0,1]$ gilt, dass*

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}.$$

Beweis. Sei f quasikonvex, seien $x, y \in C$, sei $\lambda \in [0,1]$. Setze $M := \max\{f(x), f(y)\}$; dann liegen $x, y \in S_M$. Da die Niveaumenge S_M konvex ist, liegen auch alle Konvexkombinationen von x und y in S_M, und die Ungleichung gilt.

Umgekehrt gelte nun die Ungleichung für alle $x, y \in C$ und $\lambda \in [0,1]$. Es seien $x, y \in S_t$; aufgrund der Ungleichung liegen dann auch alle Konvexkombinationen von x, y in S_t; damit ist S_t konvex. □

Aufgabe 9.1 (Stetige quasikonvexe Funktionen auf $\mathbb{R}$)**.** Eine stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ ist genau dann quasikonvex, wenn mindestens eine der folgenden Bedingungen erfüllt ist.

1. f ist wachsend.
2. f ist fallend.
3. Es gibt ein $c \in \operatorname{dom} f$, so dass f links von c fallend und rechts von c wachsend ist.

9.4 Satz. *Quasikonvexitätskriterien erster Ordnung Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar mit offenem konvexem $\operatorname{dom} f = C$. Dann ist f quasikonvex genau dann wenn für alle $x, y \in C$ gilt: Aus $f(y) \leq f(x)$ folgt $\langle f(x) | y - x \rangle \leq 0$.*

Beweis. Es genügt, den Satz für $f : \mathbb{R} \to \mathbb{R}$ zu beweisen (Restriktion von allgemeinem f auf die Strecke zwischen x und y). **TODO: schreiben** □

9.5 Satz. *Quasikonvexitätskriterien zweiter Ordnung Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit offenem konvexem $\operatorname{dom} f = C$. Ist f quasikonvex, dann gilt für alle $x \in C$ und $p \in \mathbb{R}^n$: Aus $\langle \nabla f(x) | p \rangle = 0$ folgt $\langle p | \nabla^2 f(x) | p \rangle \geq 0$.*

Umgekehrt, wenn für alle $x \in C$ und alle $p \in \mathbb{R}^n$ gilt: Aus $\langle \nabla f(x) | p \rangle = 0$ folgt $\langle p | \nabla^2 f(x) | p \rangle > 0$, dann ist f quasikonvex.

Beweis. Es genügt wieder, den Satz für $f : \mathbb{R} \to \mathbb{R}$ zu beweisen.

Sei f also quasikonvex auf einem offenen Intervall (a, b); zu zeigen ist, dass für $c \in (a, b)$ aus $f'(c) = 0$ stets $f''(c) \geq 0$ folgt. Angenommen, es sei $f'(c) = 0$ und $f''(c) < 0$. Dann gibt es $\varepsilon > 0$, so dass sowohl $f(c - \varepsilon) < f(c)$ als auch $f(c + \varepsilon) < f(c)$; also ist für kleines $\delta > 0$ die Niveaumenge $S_{f(c)-\delta}$ nicht zusammenhängend und damit nicht konvex, im Widerspruch zur Quasikonvexität von f.

Zur Umkehrung: Für jedes $c \in (a, b)$ mit $f'(c) = 0$ sei $f''(c) > 0$. Daher kann es höchstens ein c mit $f'(c) = 0$ geben. Wenn es kein solches c gibt, ist stets $f' < 0$ oder $f' > 0$, also f strikt fallend oder strikt wachsend und damit quasikonvex. Wenn es genau ein solches c ist, ist f links von c fallend und rechts wachsend und damit quasikonvex. □

9.6 Lemma (Quasikonvexitätserhaltende Operationen)**.**

Positiv gewichtete Maxima *Sind $f_1, \dots, f_m$ quasikonvex und $w_1, \dots, w_m \geq 0$, dann ist auch $\max\{w_1 f_1, \dots, w_m f_m\}$ quasikonvex. (Beweis mit Niveaumengen.)*

Komposition mit wachsender Funktion links *Ist $g : \mathbb{R}^n \to \mathbb{R}$ quasikonvex und $h : \mathbb{R} \to \mathbb{R}$ wachsend, dann ist $f := h \circ g$ quasikonvex. (Beweis mit Niveaumengen).*

Komposition mit affiner oder linear-Fraktionaler Funktion rechts *Ist $g : \mathbb{R}^n \to \mathbb{R}$ quasikonvex, dann sind auch $f(x) = g(Ax + b)$ und $f(x) = g((Ax + b)/(\langle c|x\rangle + d))$ quasikonvex.*

Minimierung *Ist $g(x, y)$ quasikonvex in (x, y) und Q eine konvexe Menge, dann ist $f(x) := \inf_{y \in Q} g(x, y)$ quasikonvex. (Beweis wie bei konvexen Funktionen, Satz 3.38.)*

Beweis. TODO: Elementare Übungsaufgabe. □

Repräsentierung einer quasikonvexen Funktion durch konvexe Funktionen. Für eine quasikonvexe Funktion f ist die Niveaumenge $S_t = \{ x \mid f(x) \leq t \}$ konvex, d.h. es gibt eine konvexe Funktion f_t, so dass S_t die Null-Niveaumenge von f_t ist. Zum Beispiel kann man $f_t(x) := 0$, wenn $f(x) \leq t$, und $f_t(x) := \infty$ sonst wählen. Interessanter sind aber Familien (f_t) mit Eigenschaften wie Differenzierbarkeit, etc.

Wir betrachten ein wichtiges Beispiel: Es sei $p \geq 0$ konvex, $q > 0$ konkav. Dann ist $f = p/q$ quasikonkav, wie man sich leicht überlegt. Es ist $f(x) \leq t \iff p(x)/q(x) \leq t \iff p(x) - tq(x) \leq 0$, und $f_t := p - tq$ ist eine konvexe Funktion. Für festes x ist $f_t(x)$ fallend in t.

9.2 Logarithmische Konvexität

Während Quasikonvexität eine Verallgemeinerung (Lockerung) des Konvexitätsbegriffs ist, führen wir jetzt noch eine Spezialisierung ein: die logarithmische Konvexität oder log-Konvexität.

9.7 Definition (log-konvex, log-konkav)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *logarithmisch konvex* oder *log-konvex*, wenn $f \geq 0$ und $\log f$ konvex ist (dabei ist $\log 0 := -\infty$). Sie heißt *logarithmisch konkav* oder *log-konkav*, wenn $f \geq 0$ und $\log f$ konkav ist.

9.8 Lemma (Logarithmenfreie Charakterisierung der log-Konvexität)**.** *Es ist $f > 0$ genau dann log-konvex, wenn*

$$f(\lambda x + (1 - \lambda)y) \leq f(x)^\lambda f(y)^{1-\lambda}$$

für alle $x, y \in \operatorname{dom} f$ und alle $\lambda \in [0, 1]$.

Beweis. Übungsaufgabe □

9.9 Lemma (Implikationen)**.** *Ist f log-konvex, dann ist f auch konvex und quasikonvex. Ist f konkav, dann ist f auch log-konkav. Ist f log-konkav, dann ist f auch quasikonkav.*

Beweis. Nach den Kompositionsregeln ist mit $g := \log f$ auch $\exp(g) = f$ konvex. Ebenso ist mit mit f auch $\log f$ konkav. Ist f log-konkav, dann ist $\log f$ konkav und damit quasikonkav, aber log ist streng monoton; daher ist auch f quasikonkav. □

Während also log-konvex "mehr" als konvex ist, ist log-konkav "weniger" als konkav, aber noch "mehr" als quasikonkav. Wir betrachten einige Beispiele.

9.10 Beispiel (Einfache Beispiele)**. TODO: Buch S.104** ♡

9.11 Beispiel (Wahrscheinlichkeitsdichten)**. TODO: Buch S.105** ♡

9.12 Lemma (Verhalten unter Operationen)**.**

1. *Sind f, g log-konvex (log-konkav) und $c > 0$, dann auch $f \cdot g$ und cf.*
2. *Sind f, g log-konvex, dann auch $f + g$. Allgemeiner: Ist $f(x, y)$ log-konvex in x für alle $y \in I$ (eine Indexmenge), dann ist auch $g(x) := \int_Q f(x, y)\mathrm{d}y$ log-konvex. (Abgeschlossenheit unter Addition gilt im allgemeinen* nicht *für log-konkave Funktionen.)*

Beweis. Zu 1.: Konvexität und Konkavität sind unter Addition (auch von Konstanten) abgeschlossen. Zu 2.: Es sind $F := \log f$ und $G := \log g$ konvex. Nach den Kompositionsregeln für konvexe Funktionen (log-sum-exp) ist damit auch $\log(\exp F + \exp G) = \log(f + g)$ konvex, also $f + g$ log-konvex. □

9.13 Beispiel (Laplace-Transformation)**.** Sei $p : \mathbb{R}^n \to \mathbb{R}$ nichtnegativ. Dann ist die Laplace-Transformierte $P(z)$ von $p(x)$ log-convex,

$$P(z) = \int_{\mathbb{R}^n} p(x)\, \exp(-\langle z|x\rangle)\, \mathrm{d}x.$$

Ist p eine Wahrscheinlichkeitsdichte, ist insbesondere die Moment-erzeugende Funktion $M(z) = P(-z)$ log-konvex und die Kumulanten-erzeugende Funktion $C(z) := \log M(z)$ konvex. ♡

9.14 Lemma (Integration log-konkaver Funktionen)**.** *Sei $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ log-konkav in (x, y). Dann ist $g(y) := \int_{set R^m} f(x, y)\, \mathrm{d}y$ log-konkav in x.*

Beweis. TODO: nicht einfach, siehe Prékopa (1971,1973), weggelassen. □

Das vorige Lemma hat wichtige Konsequenzen für Wahrscheinlichkeitsverteilungen, zum Beispiel:

- Marginaldichten von log-konkaven Wahrscheinlichkeitsdichten sind log-konkav.
- Faltungen von log-konkaven Dichten sind log-konkav.
- Die kumulative Verteilungsfunktion einer log-konkaven Dichte ist log-konkav (z.B. die der Gaussverteilung).
- Ist $C \subset \mathbb{R}^n$ konvex, $x \in \mathbb{R}^n$ und W ein zufälliger Vektor mit Dichte p auf $\mathbb{R}^n$. Dann ist $f(x) := \mathbb{P}(x + W \in C)$ log-konkav in x.

Literaturverzeichnis