Oliver Stein

Globale Optimierung I und II

Skript zur Vorlesung am

Karlsruher Institut für Technologie

im Sommersemester 2017

Inhaltsverzeichnis

Vorwort

Dieses Skript ist aus den Lecture Notes zu meiner Vorlesung „Globale Optimierung I und II“ entstanden, die ich am Karlsruher Institut für Technologie seit 2008 jährlich halte. Gegenstand ist die globale Minimierung nichtlinearer Funktionen unter nichtlinearen Nebenbedingungen, wie sie in den Natur-, Ingenieur- und Wirtschaftswissenschaften sehr oft auftreten. Dabei besteht häufig das Problem, dass numerische Lösungsverfahren zwar effizient lokale Optimalpunkte finden können, während globale Optimalpunkte viel schwerer zu identifizieren sind. Dies entspricht der Tatsache, dass man mit lokalen Suchverfahren zwar gut den Gipfel des nächstgelegenen Berges finden kann, während die Suche nach dem Gipfel des Mount Everest eher aufwändig ist.

Der Inhalt der Kapitel 1 und 2 dieses Skriptes, also im Wesentlichen Existenzaussagen für globale Minimalpunkte und die konvexe Optimierung, entsprechen dem Vorlesungsteil „Globale Optimierung I“. Die in Kapitel 3 besprochenen Techniken zur globalen Optimierung nichtkonvexer Probleme sind Inhalt des Teils „Globale Optimierung II“.

Die Vorlesung stützt sich teilweise auf Darstellungen aus den Büchern [2] von M.S. Bazaraa, H.D. Sherali und C.M. Shetty, [3] von C.A. Floudas, [4] von O. Güler, [6] von R. Horst und H. Tuy, [5] von J.-B. Hiriart-Urruty und C. Lemaréchal sowie [7] von H.Th. Jongen, K. Meer und E. Triesch, die auch viele über die Vorlesung hinausgehende Fragestellungen behandeln. Zu Grundlagen der (lokalen) Nichtlinearen Optimierung sei auf mein Vorlesungsskript [13] verwiesen, und zu allgemeinen Grundlagen der Optimierung auf [10].

In kleinerem Schrifttyp gesetzter Text bezeichnet Material, das zur Vollständigkeit angegeben ist, in der Vorlesung aber höchstens kurz angerissen wird.

Dieses Skript wurde in LaTeX2e gesetzt. Die Abbildungen stammen aus *Xfig* oder wurden als Ausgabe von *Matlab* erzeugt.

Für Kommentare jeglicher Art zu diesem Skript bin ich unter `stein@kit.edu` erreichbar.

Karlsruhe, im März 2017 Oliver Stein

Kapitel 1

Einführung

Die endlich-dimensionale kontinuierliche Optimierung behandelt die Minimierung oder Maximierung einer Zielfunktion in einer endlichen Anzahl kontinuierlicher Entscheidungsvariablen. Anwendungen im Operations Research finden sich nicht nur bei linearen Modellen (wie zur Gewinnmaximierung oder bei Transportproblemen, [10]), sondern auch bei wichtigen nichtlinearen Modellen aus verschiedenen Anwendungen. Dazu gehören geometrische Probleme, mechanische Probleme, Parameter-Fitting-Probleme, Schätzprobleme, Datenklassifikation und Sensitivitätsanalyse. Als Lösungswerkzeug benutzt man sie außerdem bei nichtkooperativen Spielen ([14]), in der robusten Optimierung ([14]) oder bei der Relaxierung diskreter Optimierungsprobleme ([11]).

1.1 Beispiele und Begriffe

Zwei grundlegende Begriffe in der Optimierung sind der optimale *Punkt* und der optimale *Wert*. Als Beispiel kann man fragen, wer in einer Gruppe von Personen die meisten Münzen bei sich trägt. Der optimale Wert ist die gefundene größtmögliche Anzahl von Münzen. Er ist eindeutig bestimmt.

Im Gegensatz dazu können durchaus mehrere Personen diese Anzahl von Münzen bei sich tragen. Demnach ist der *Punkt* (hier: die Person), an dem der optimale *Wert* angenommen wird, nicht notwendigerweise eindeutig. Dies gilt analog bei jedem Optimierungsproblem.

1.1.1 Beispiel (Projektion auf eine Menge)
Für eine Menge $M \subseteq \mathbb{R}^n$ und einen Punkt $z \in \mathbb{R}^n$ sei der (Euklidische) Abstand von z zu M gesucht, sowie ein Punkt $\bar{x}$ in M, der am nächsten an z liegt (s. Abb. 1.1).

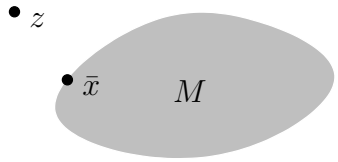


Abbildung 1.1: Projektion auf eine Menge

Die Formulierung als Optimierungsproblem lautet

$$P: \quad \min_{x \in \mathbb{R}^n} \|x - z\|_2 \quad s.t. \quad x \in M,$$

wobei in Abbildung 1.1 der Fall $n = 2$ illustriert ist.

Jeden optimalen Punkt $\bar{x}$ von P nennen wir Projektion von z auf M. *Wenn ein optimaler Punkt $\bar{x}$ gefunden ist, berechnet sich der optimale Wert (hier also der Abstand von z zu M) einfach durch Einsetzen von $\bar{x}$ in die Zielfunktion $f(x) = \|x - z\|_2$ zu $v = \|\bar{x} - z\|_2$.*

1.1.2 Bemerkung *Anstelle der in Anwendungen typischen Wahl der Euklidischen Norm kann man im Projektionsproblem auch jede andere Norm zur Abstandsmessung benutzen. Dies kann z.B. geometrische Gründe haben und führt im Allgemeinen zu anderen Ergebnissen für die Projektion. Zur Abgrenzung spricht man bei der Wahl der Euklidischen Norm auch von* orthogonaler *Projektion.*

Falls M eine Hyperebene ist, lässt sich das Problem P aus Beispiel 1.1.1 explizit lösen (s. Abb. 1.2). Dazu sei $M = \{x \in \mathbb{R}^n | \, a^\intercal x = b\}$, d.h. M sei durch eine lineare Gleichung beschrieben. Dann wird das Optimierungsproblem P zu

$$P: \quad \min_{x \in \mathbb{R}^n} \|x - z\|_2 \quad \text{s.t.} \quad a^\intercal x = b\,.$$

Der (eindeutige) optimale Punkt berechnet sich zu (s.u., Bsp. 2.6.8)

$$\bar{x} \;=\; z - \frac{a^\intercal z - b}{a^\intercal a} \cdot a\,,$$

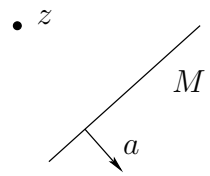


Abbildung 1.2: Projektion auf eine Gerade

und der optimale Wert ist

$$v \;=\; \left\| \frac{a^\intercal z - b}{a^\intercal a} \cdot a \right\|_2 \;=\; \frac{|a^\intercal z - b|}{\|a\|_2}\,.$$

Ein Beispiel für einen nicht eindeutigen optimalen Punkt im Projektionsproblem ist in Abbildung 1.3 gegeben. Hier besitzen sowohl $\bar{x}$ als auch $\widetilde{x}$ unter allen Punkten der Menge M minimalen Abstand von z, es gibt also zwei verschiedene Projektionen von z auf M !

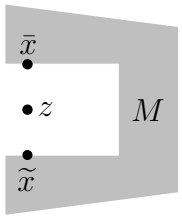


Abbildung 1.3: Nichteindeutiger optimaler Punkt

Verschiebt man in Abbildung 1.3 den Punkt z in Richtung $\bar{x}$, so wird $\bar{x}$ eindeutiger Minimalpunkt. Der Punkt $\widetilde{x}$ behält allerdings die Eigenschaft, dass es in einer hinreichend kleinen Umgebung U um $\widetilde{x}$ keine Punkte in M gibt, die näher an z liegen als $\widetilde{x}$ (s. Abb. 1.4). Man unterscheidet diese beiden Situationen, indem man $\bar{x}$ *globalen* Minimalpunkt und $\widetilde{x}$ *lokalen* Minimalpunkt nennt.

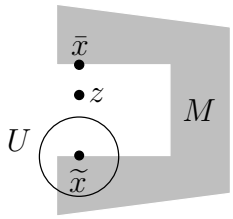


Abbildung 1.4: Lokaler und globaler Minimalpunkt

Diese wichtige Unterscheidung wird in der folgenden Definition formal festgehalten.

1.1.3 Definition (Minimalpunkte und Minimalwerte)
Gegeben seien eine Menge von zulässigen Punkten $M \subseteq \mathbb{R}^n$ und eine Zielfunktion $f : M \to \mathbb{R}$.

a) $\bar{x} \in M$ heißt lokaler Minimalpunkt *von f auf M, falls eine Umgebung U von $\bar{x}$ existiert mit*

$$\forall\, x \in U \cap M : \quad f(x) \geq f(\bar{x}).$$

b) $\bar{x} \in M$ heißt globaler Minimalpunkt *von f auf M, falls man in a) $U = \mathbb{R}^n$ wählen kann.*

c) Ein lokaler oder globaler Minimalpunkt heißt strikt*, falls in a) bzw. b) für $x \neq \bar{x}$ sogar die strikte Ungleichung „$>$“ gilt.*

d) Zu jedem globalen Minimalpunkt $\bar{x}$ heißt $f(\bar{x})$ $(= v = \min_{x \in M} f(x))$ globaler Minimalwert*, und zu jedem lokalen Minimalpunkt $\bar{x}$ heißt $f(\bar{x})$* lokaler Minimalwert*.*

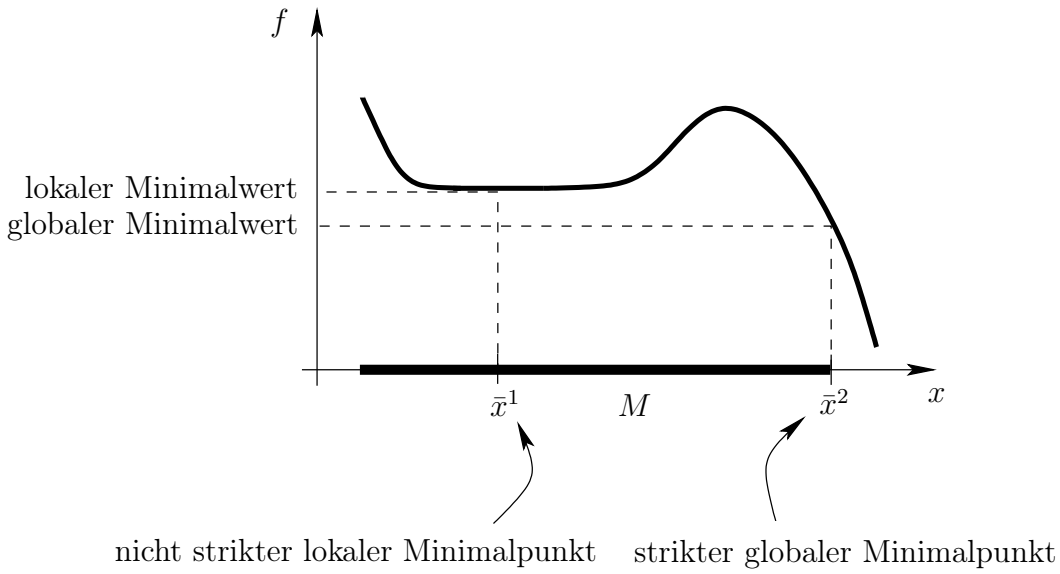


Abbildung 1.5: Lokale und globale Minima

Bemerkungen:

- Jeder globale Minimalpunkt ist auch lokaler Minimalpunkt.
- Damit die Forderung „$f(x) \geq f(\bar{x})$“ sinnvoll ist, muss der Bildbereich von f geordnet sein. Zum Beispiel ist die Minimierung von $f : \mathbb{R}^n \to \mathbb{R}^2$ nicht sinnvoll (allerdings befasst sich das Gebiet der *Mehrzieloptimierung* damit, wie man solche Probleme trotzdem behandeln kann).
- Strikte globale Minimalpunkte sind eindeutig, und strikte lokale Minimalpunkte sind lokal eindeutig.
- Lokale und globale *Maximalpunkte* sind analog definiert. Da die Maximalpunkte von f genau die Minimalpunkte von $-f$ sind, reicht es, Minimierungsprobleme zu betrachten (Achtung: dabei ändert sich allerdings das Vorzeichen des Optimalwertes: $\max f(x) = -\min(-f(x))$, vgl. Abb. 1.6 und Ü. 1.1.4 sowie für etwas allgemeinere Aussagen Ü. 1.3.1).
- Wegen der ähnlichen Notation besteht eine Verwechslungsgefahr zwischen dem Minimalwert $\min_{x \in M} f(x)$ und seinem zugrundeliegenden Problem P!

1.1.4 Übung *Gegeben seien eine Menge von zulässigen Punkten $M \subseteq \mathbb{R}^n$ und eine Zielfunktion $f : M \to \mathbb{R}$. Zeigen Sie:*

a) Die globalen Maximalpunkte von f auf M sind genau die globalen Minimalpunkte von $-f$ auf M.

b) Sofern f globale Maximalpunkte besitzt, gilt für den globalen Maximalwert

$$\max_{x \in M} f(x) \;=\; -\min_{x \in M} (-f(x)).$$

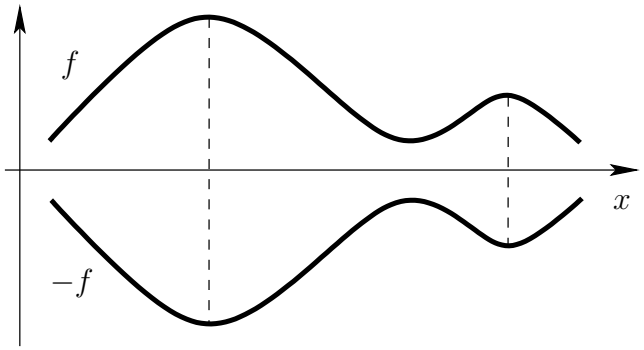


Abbildung 1.6: Maximierung von f durch Minimierung von $-f$

1.1.5 Beispiel (Zentrum einer Punktewolke)

Gegeben seien Punkte $x^1, x^2, ..., x^m \in \mathbb{R}^n$ (d.h. $x^i = \begin{pmatrix} x_1^i \\ \vdots \\ x_n^i \end{pmatrix}$, $i = 1, ..., m$).

Gesucht ist ein Punkt $z \in \mathbb{R}^n$ „im Zentrum der $x^1, x^2, \ldots, x^m$“.

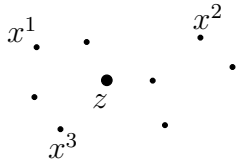


Abbildung 1.7: Punktewolke

Dazu ist zunächst zu klären, wie man „Zentrum“ definieren soll. Die auftretenden Abstände $\|z - x^1\|_2, \|z - x^2\|_2, \ldots, \|z - x^m\|_2$ sind alle gleichzeitig

nahe bei null, falls die Norm ihres Vektors, $\left\| \begin{pmatrix} \|z-x^1\|_2 \\ \vdots \\ \|z-x^m\|_2 \end{pmatrix} \right\|_2$, *möglichst klein ist. Dies führt auf das Optimierungsproblem*

$$P: \quad \min_{z\in\mathbb{R}^n} \left\| \begin{pmatrix} \|z-x^1\|_2 \\ \vdots \\ \|z-x^m\|_2 \end{pmatrix} \right\|_2 .$$

P besitzt offenbar keine Nebenbedingungen. Als Optimalpunkt (zur Technik dafür vgl. Kap. 2.4) berechnet man das arithmetische Mittel der Punkte,

$$\bar{z} \;=\; \frac{1}{m}\sum_{i=1}^{m} x^i.$$

Ein Problem dieses Ansatzes besteht darin, dass Ausreißer bei wachsender Punkteanzahl m zunehmend vernachlässigt werden. Dies kann erwünscht sein (z.B. wenn man von Messfehlern in den Punkten ausgehen muss) oder auch unerwünscht (z.B. wenn die Punkte geographisch Orten entsprechen und abgelegene Orte gleichberechtigt behandelt werden sollen).

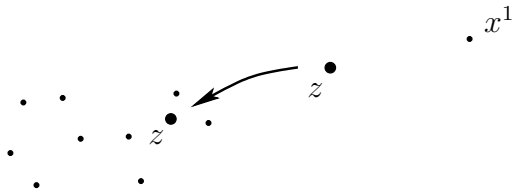


Abbildung 1.8: Ausreißer x^1

Dies lässt sich durch Wahl einer anderen (äußeren) Norm verhindern, zum Beispiel der Maximumsnorm

$$\|x\|_\infty = \max_{j=1,..,n} |x_j| .$$

Durch Wahl dieser Norm kann man z als Mittelpunkt einer möglichst kleinen Kugel auffassen, die alle x^i enthält. Um dies einzusehen, führen wir den sogenannten Epigraphen einer Funktion ein.

Für $M \subseteq \mathbb{R}^n$ und $f: M \to \mathbb{R}$ heißt die Menge

$$\mathrm{epi}\,(f,M) \;=\; \{(x,\alpha) \in M\times\mathbb{R} |\; f(x) \le \alpha\}$$

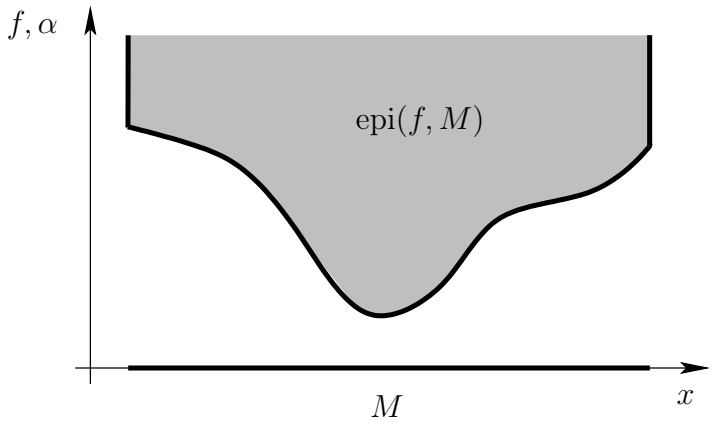


Abbildung 1.9: Epigraph von f

Epigraph von f auf M. Der Epigraph besteht aus dem Graphen von f auf M sowie allen darüberliegenden Punkten (s. Abb. 1.9). Man sieht leicht ein (vgl. Ü. 1.3.7), dass man einen Minimalpunkt von f auf M dadurch finden kann, dass man im Epigraphen von f einen Punkt (x, α) mit minimaler α-Komponente sucht. Das Problem

$$P: \quad \min_{x} f(x) \quad \text{s.t.} \quad x \in M$$

ist in diesem Sinne also äquivalent zu

$$P_{\text{epi}}: \quad \min_{(x,\alpha)} \alpha \quad \text{s.t.} \quad f(x) \leq \alpha, \quad x \in M.$$

Diese sogenannte *Epigraph-Umformulierung* von P kann beispielsweise dann Vorteile haben, wenn die Funktion f selbst das Maximum anderer Funktionen ist, oder wenn man gezwungen ist, eine lineare Zielfunktion zu benutzen.

Im vorliegenden Fall des Zentrums einer Punktewolke mit Maximumsnorm als äußerer Norm gilt

$$P: \quad \min_{z \in \mathbb{R}^n} \left\| \begin{pmatrix} \|z - x^1\|_2 \\ \vdots \\ \|z - x^m\|_2 \end{pmatrix} \right\|_\infty ,$$

also $M = \mathbb{R}^n$ und

$$f(z) = \max_{i=1,\ldots,m} \|z - x^i\|_2 .$$

Als Epigraph-Umformulierung erhält man das äquivalente Problem

$$P_{\text{epi}}: \quad \min_{(z,\alpha)} \alpha \quad \text{s.t.} \quad f(z) \leq \alpha\,,$$

wobei die Nebenbedingung von P_{epi} ausgeschrieben lautet:

$$\max_{i=1,...,m} \|z - x^i\|_2 \;\leq\; \alpha.$$

Da das Maximum von m Zahlen genau dann unter der Schranke α liegt, wenn alle m Zahlen unter α liegen, lässt sich diese Restriktion äquivalent umformulieren zu

$$\|z - x^i\|_2 \;\leq\; \alpha, \quad i = 1, ..., m,$$

und wir erhalten die Äquivalenz von P zu

$$P_{\text{epi}}: \quad \min_{(z,\alpha)} \alpha \quad \text{s.t.} \quad \|z - x^i\|_2 \;\leq\; \alpha, \quad i = 1, ..., m.$$

Fasst man α als Radius auf, so versucht man also, den Kreis mit Mittelpunkt z und minimalem Radius α zu finden, der alle Punkte $x^1, ..., x^m$ enthält.

1.1.6 Beispiel (Clusteranalyse)

Falls in Beispiel 1.1.5 zu vermuten ist, dass sich die Punkte $x^1, x^2, ..., x^m$ an k Stellen „häufen", kann man versuchen, „k Zentren gleichzeitig" zu bestimmen.

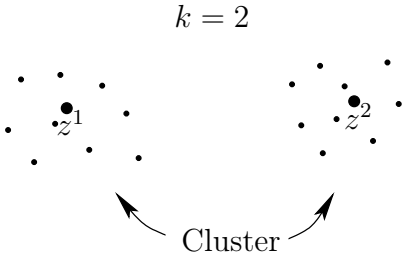


Abbildung 1.10: Cluster

Nennt man die Zentren $z^1, ..., z^k$, dann gibt es zu jedem z^ℓ ein Cluster C^ℓ, $\ell = 1, ..., k$, das aus den nächstliegenden Punkten besteht. Eine entscheidende Frage ist, zu welchem Cluster ein Punkt x^i gehört.

Der Index dieses Clusters sei $\ell(i)$. Dann gilt $x^i \in C^{\ell(i)}$ offenbar genau dann, wenn x^i von $z^{\ell(i)}$ kleineren Abstand hat als von allen anderen z^ℓ, d.h. wenn

$$\|z^{\ell(i)} - x^i\| = \min_{\ell=1,...,k} \|z^\ell - x^i\|$$

gilt.

Die im Problem der Clusteranalyse auftretende Abstände sind also $\|z^{\ell(1)} - x^1\|$, ..., $\|z^{\ell(m)} - x^m\|$, *und das zugehörige Optimierungsproblem lautet*

$$P: \quad \min_{z^1,\ldots,z^k \in \mathbb{R}^n} \left\| \begin{pmatrix} \|z^{\ell(1)} - x^1\| \\ \vdots \\ \|z^{\ell(m)} - x^m\| \end{pmatrix} \right\| = \left\| \begin{pmatrix} \min_{\ell=1,\ldots,k} \|z^\ell - x^1\| \\ \vdots \\ \min_{\ell=1,\ldots,k} \|z^\ell - x^m\| \end{pmatrix} \right\|.$$

Als „äußere" Norm benutzt man häufig die ℓ_1*-Norm (*$\|x\|_1 = \sum_{i=1}^n |x^i|$*), so dass* P *die Form*

$$P_1: \quad \min_{z^1,\ldots,z^k \in \mathbb{R}^n} \sum_{i=1}^m \min_{\ell=1,\ldots,k} \|z^\ell - x^i\|$$

erhält, oder auch die ℓ_∞*-Norm mit*

$$P_\infty: \quad \min_{z^1,\ldots,z^k \in \mathbb{R}^n} \max_{i=1,\ldots,m} \min_{\ell=1,\ldots,k} \|z^\ell - x^i\|.$$

In jedem Fall besitzt der Vektor der Entscheidungsvariablen im Problem der Clusteranalyse die Dimension $n \cdot k$, so dass in praktisch relevanten Anwendungen häufig hochdimensionale Probleme auftreten.

Die numerische Lösung solcher Probleme zu globaler Optimalität gilt nicht nur wegen der Hochdimensionalität als sehr schwer, sondern insbesondere aufgrund der „Minimumsstruktur" in der Zielfunktion.

1.2 Lösbarkeit

Ob ein Optimierungsproblem überhaupt optimale Punkte besitzt, liegt nicht immer auf der Hand und muss bei vielen Lösungsverfahren vorab vom Anwender selbst geprüft werden. Damit befasst sich der folgende Abschnitt.

Ohne irgendwelche Voraussetzungen an die Menge $M \subseteq \mathbb{R}^n$ und die Funktion $f : M \to \mathbb{R}$ lässt sich jedenfalls jedem Minimierungsproblem

$$P: \quad \min f(x) \quad \text{s.t.} \quad x \in M$$

ein „verallgemeinerter Minimalwert“ zuordnen, nämlich das sogenannte *Infimum* von f auf M. Um es formal einzuführen, bezeichnen wir $\alpha \in \mathbb{R}$ als *untere Schranke* für f auf M, falls

$$\forall\, x \in M: \quad \alpha \leq f(x)$$

gilt. Das Infimum von f auf M ist die *größte* untere Schranke von f auf M, es gilt also $v = \inf_{x\in M} f(x)$ falls

- $\forall\, x \in M: \quad v \leq f(x)$ (d.h. v ist selbst untere Schranke) und
- für alle unteren Schranken α von f auf M gilt $\alpha \leq v$.

Analog wird das *Supremum* $\sup_{x\in M} f(x)$ von f auf M als kleinste obere Schranke definiert.

1.2.1 Beispiel *Es gilt* $\inf_{x\in\mathbb{R}} (x-5)^2 = 0$ *und* $\inf_{x\in\mathbb{R}} e^x = 0$.

Falls f auf M nicht nach unten beschränkt ist, setzt man formal

$$\inf_{x\in M} f(x) \;=\; -\infty,$$

und für das Infimum über die leere Menge definiert man formal

$$\inf_{x\in\emptyset} f(x) \;=\; +\infty$$

(wobei die Gestalt von f dann keine Rolle spielt). Warum diese formalen Setzungen sinnvoll sind, werden wir nach Satz 1.2.8 sehen.

1.2.2 Beispiel *Es gilt* $\inf_{x\in\mathbb{R}} (x-5) = -\infty$ *und* $\inf_{x\in\emptyset} (x-5) = +\infty$.

Der „verallgemeinerte Minimalwert“ $\inf_{x\in M} f(x)$ von P ist also stets ein Element der *erweiterten reellen Zahlen* $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. In der Analysis wird gezeigt, dass das so definierte Infimum ohne Voraussetzungen an f und M stets existiert und eindeutig bestimmt ist.

1.2.3 Definition (Lösbarkeit)
Das Minimierungsproblem P heißt lösbar*, falls ein $\bar{x} \in M$ mit $\inf_{x\in M} f(x) = f(\bar{x})$ existiert.*

Lösbarkeit von P bedeutet also, dass das Infimum von f auf M als Zielfunktionswert eines zulässigen Punktes realisiert werden kann, dass also das Infimum *angenommen* wird. Um anzudeuten, dass das Infimum angenommen wird schreiben wir $\min_{x\in M} f(x)$ anstelle von $\inf_{x\in M} f(x)$.

1.2.4 Beispiel *Es gilt $0 = \min_{x\in\mathbb{R}} (x-5)^2 = (\bar{x}-5)^2$ mit $\bar{x} = 5$, aber es gibt kein $\bar{x} \in \mathbb{R}$ mit $0 = \inf_{x\in\mathbb{R}} e^x = e^{\bar{x}}$.*

Der folgende Satz besagt, dass man zur Lösbarkeit genauso gut die Existenz eines globalen Minimalpunktes fordern kann.

1.2.5 Satz *Das Minimierungsproblem P ist genau dann lösbar, wenn es einen globalen Minimalpunkt besitzt.*

Beweis. Zunächst sei P lösbar. Dann gibt es ein $\bar{x} \in M$ mit $\min_{x\in M} f(x) = f(\bar{x})$. Als Infimum ist $f(\bar{x})$ eine untere Schranke für f auf M, es gilt also $f(\bar{x}) \le f(x)$ für alle $x \in M$. Nach Definition 1.1.3 ist $\bar{x}$ demnach globaler Minimalpunkt von f auf M.

Andererseits sei $\bar{x}$ ein globaler Minimalpunkt von f auf M. Dann gilt $\bar{x} \in M$, und $f(\bar{x})$ ist eine untere Schranke für f auf M. Würde es eine größere untere Schranke α für f auf M geben, so hätten wir

$$\forall\, x \in M: \quad f(\bar{x}) < \alpha \le f(x),$$

was für $x = \bar{x}$ zu einem Widerspruch führt. Daher ist $f(\bar{x})$ die größte untere Schranke für f auf M, es gilt also $\inf_{x\in M} f(x) = f(\bar{x})$. •

Arten von Unlösbarkeit

Bevor wir uns hinreichenden Kriterien für die *Lösbarkeit* von P zuwenden, widmen wir uns zunächst der Frage, welche *Arten von Unlösbarkeit* möglich sind. Dies ist für numerische Algorithmen interessant, die nicht nur P lösen, sondern im Falle der Unlösbarkeit auch eine entsprechende Meldung liefern (etwa der Simplex-Algorithmus der linearen Optimierung, s. [10]).

Dazu betrachten wir die sogenannte *Parallelprojektion* auf die „α-Achse“ des Epigraphen

$$\operatorname{epi}(f, M) \;=\; \{(x, \alpha) \in M \times \mathbb{R} |\; f(x) \leq \alpha\}$$

von f auf M. Sie hängt wie folgt mit der orthogonalen Projektion eines Punktes z auf eine Menge M aus Beispiel 1.1.1 zusammen.

1.2.6 Übung *Wir bezeichnen mit* $\operatorname{pr}(z, M)$ *die Menge der orthogonalen Projektionen eines Punktes* $z \in \mathbb{R}^n$ *auf* $M \subseteq \mathbb{R}^n$ *und mit*

$$\operatorname{pr}(Z, M) \;=\; \{\operatorname{pr}(z, M) |\; z \in Z\}$$

die orthogonale Projektion einer Menge $Z \subseteq \mathbb{R}^n$ *auf* M.

Wir betrachten nun den speziellen Fall $M = \mathbb{R}^k \times \{0_\ell\}$ *mit* $0_\ell \in \mathbb{R}^\ell$ *und* $k + \ell = n$. *Dazu spalten wir den Vektor* x *auf in* $x = (a, b)$ *mit* $a \in \mathbb{R}^k$ *und* $b \in \mathbb{R}^\ell$. *Zeigen Sie für jede Menge* $Z \subseteq \mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^\ell$ *die Identität*

$$\operatorname{pr}(Z, \mathbb{R}^k \times \{0_\ell\}) \;=\; \{(a, 0_\ell) \in \mathbb{R}^k \times \{0_\ell\} |\; \exists\, b \in \mathbb{R}^\ell :\; (a, b) \in Z\}.$$

Das Unterschlagen der Menge $\{0_\ell\}$ in der Identität für $\operatorname{pr}(Z, \mathbb{R}^k \times \{0_\ell\})$ aus Übung 1.2.6 motiviert die folgende Definition.

1.2.7 Definition (Parallelprojektion)
Für $Z \subseteq \mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^\ell$ *heißt*

$$\operatorname{pr}_a(Z) \;=\; \{a \in \mathbb{R}^k |\; \exists\, b \in \mathbb{R}^\ell :\; (a, b) \in Z\}.$$

Parallelprojektion *von* Z *in die (etwas lax als „a-Raum“ bezeichnete) Menge* $\mathbb{R}^k$.

Die Parallelprojektion von $\operatorname{epi}(f, M)$ auf den „α-Raum“ $\mathbb{R}$ lautet also

$$\operatorname{pr}_\alpha \operatorname{epi}(f, M) \;=\; \{\alpha \in \mathbb{R} |\; \exists\, x \in M :\; f(x) \leq \alpha\}.$$

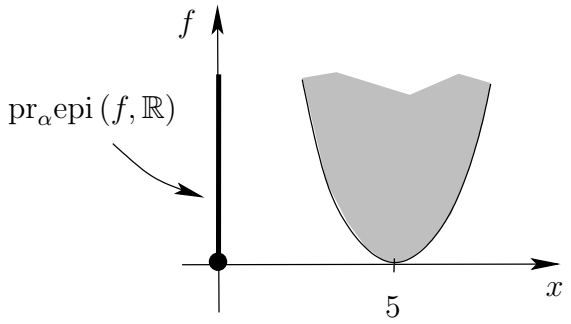


Abbildung 1.11: Die Menge $\mathrm{pr}_\alpha \mathrm{epi}\,(f, \mathbb{R})$ für $f(x) = (x-5)^2$

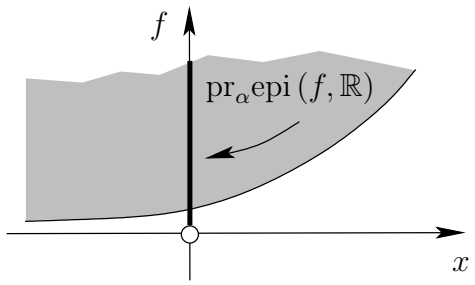


Abbildung 1.12: Die Menge $\mathrm{pr}_\alpha \mathrm{epi}\,(f, \mathbb{R})$ für $f(x) = e^x$

Die Abbildungen 1.11 und 1.12 zeigen die Mengen $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ für die beiden Optimierungsprobleme aus Beispiel 1.2.1. In Beispiel 1.2.2 haben wir $\mathrm{pr}_\alpha \mathrm{epi}\,(x-5, \mathbb{R}) = \mathbb{R}$ und $\mathrm{pr}_\alpha \mathrm{epi}\,(x-5, \emptyset) = \emptyset$.

Der folgende Satz besagt, dass $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ auch allgemein nur eine von genau vier Gestalten besitzen kann, von denen außerdem genau eine der Lösbarkeit von P entspricht. Hier und im Folgenden schreiben wir kurz (x^ν) für eine Folge $(x^\nu)_{\nu \in \mathbb{N}}$ sowie $\lim_\nu$ für $\lim_{\nu \to \infty}$.

1.2.8 Satz *Für jedes Optimierungsproblem P hat die Menge $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ genau eine der vier Gestalten*

a) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = \emptyset$,

b) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = \mathbb{R}$,

c) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = (v, +\infty)$ *mit* $v \in \mathbb{R}$,

d) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = [v, +\infty)$ *mit* $v \in \mathbb{R}$.

Ferner ist der Fall a) zu $M = \emptyset$ äquivalent, Fall b) zur Unbeschränktheit nach unten von f auf M, Fall c) dazu, dass das endliche Infimum $v = \inf_{x\in M} f(x)$ nicht angenommen wird, und Fall d) zur Lösbarkeit von P mit $v = \min_{x\in M} f(x)$.

Beweis. Zur Abkürzung setzen wir $V := \mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ und unterscheiden zunächst, ob in P die zulässige Menge M leer ist oder nicht. Für $M = \emptyset$ folgt sofort $V = \emptyset$, so dass gezeigt ist, dass Fall a) auftreten kann. Um zu sehen, dass Fall a) *nur* für $M = \emptyset$ eintritt, wähle im Falle $M \neq \emptyset$ ein $\bar{x} \in M$ und setze $\bar{\alpha} = f(\bar{x})$. Dann gilt $\bar{\alpha} \in V$ und damit $V \neq \emptyset$.

Im Folgenden sei $M \neq \emptyset$ und damit $V \neq \emptyset$. Wähle ein beliebiges $\alpha \in V$. Dann liegen auch alle $\widetilde{\alpha} \geq \alpha$ in V, denn es gibt ein $x \in M$ mit $f(x) \leq \alpha \leq \widetilde{\alpha}$. Daraus folgt

$$V \;\supseteq\; \bigcup_{\alpha\in V} [\alpha, +\infty).$$

Nun sei f auf M nach unten unbeschränkt. Dann gibt es eine Folge $(x^\nu) \subseteq M$ mit $f(x^\nu) \leq -\nu$ für alle $\nu \in \mathbb{N}$. Insbesondere gilt $-\nu \in V$ für alle $\nu \in \mathbb{N}$ und damit

$$V \;\supseteq\; \bigcup_{\alpha\in V} [\alpha, +\infty) \;\supseteq\; \bigcup_{\nu\in\mathbb{N}} [-\nu, +\infty) \;=\; \mathbb{R},$$

also Fall b). Um zu sehen, dass dieser Fall *nur* für auf M nach unten unbeschränktes f auftritt, wähle im Fall einer auf M nach unten beschränkten Funktion f eine Unterschranke $\bar{\alpha} \in \mathbb{R}$ mit $\bar{\alpha} \leq f(x)$ für alle $x \in M$. Wähle nun zu einem beliebigen $\alpha \in V$ ein beliebiges $x \in M$ mit $f(x) \leq \alpha$. Dann gilt $\bar{\alpha} \leq f(x) \leq \alpha$ und demnach $V \subseteq [\bar{\alpha}, +\infty)$. Wegen $\bar{\alpha} \in \mathbb{R}$ schließt dies den Fall b) aus.

Im Folgenden sei f auf $M \neq \emptyset$ mit $\bar{\alpha} \in \mathbb{R}$ nach unten beschränkt. Dann ist auch die größte Unterschranke von f auf M eine reelle Zahl, also $v := \inf_{x\in M} f(x) \in \mathbb{R}$. In der Analysis wird gezeigt, dass in diesem Fall das Infimum einer Menge genau diejenige Unterschranke ist, die sich durch Elemente der Menge beliebig genau approximieren lässt. Hier bedeutet dies, dass eine Folge $(x^\nu) \subseteq M$ mit $\lim_\nu f(x^\nu) = v$ existiert, wobei die Folge $(f(x^\nu))$ ohne Beschränkung der Allgemeinheit als streng monoton fallend angenommen werden kann. Wegen $(f(x^\nu)) \subseteq V$ folgt

$$[v, +\infty) \;\supseteq\; V \;\supseteq\; \bigcup_{\alpha\in V} [\alpha, +\infty) \;\supseteq\; \bigcup_{\nu\in\mathbb{N}} [f(x^\nu), +\infty) \;=\; (v, +\infty).$$

Für die Gestalt der Menge V folgen daraus die Fälle c) und d). Falls v als Infimum angenommen wird, gibt es ein $\bar{x} \in M$ mit $f(\bar{x}) = v$, so dass $v \in V$ und damit $V = [v, +\infty)$ gilt, also Fall d). Falls v als Infimum nicht angenommen wird, resultiert analog $v \notin V$ und damit Fall c). •

Nach Satz 1.2.8 gibt es genau drei Gründe für die Unlösbarkeit von P, nämlich *Inkonsistenz* von P (genauer: von M) in Fall a), die *Unbeschränktheit* von P (genauer: von f auf M) in Fall b) und die Tatsache, dass ein *endliches Infimum* (d.h. $v \notin \{\pm\infty\}$) *nicht angenommen* wird, in Fall c).

Insbesondere ist P genau dann unlösbar, wenn $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ eine offene Menge ist. Da die Mengen in Fall a) und b) gleichzeitig auch abgeschlossen sind, entfällt die Alternative c) für alle Klassen von Optimierungsproblemen, in denen $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M)$ stets eine abgeschlossene Menge ist. Dies ist beispielsweise in der linearen Optimierung der Fall (da dann $\mathrm{epi}\,(f, M)$ ein konvexes Polyeder ist, und Parallelprojektionen konvexer Polyeder wieder konvexe Polyeder und damit insbesondere abgeschlossen sind). In der Tat liefert der Simplex-Algorithmus der linearen Optimierung bei Unlösbarkeit entweder die Meldung der Inkonsistenz oder die der Unbeschränktheit von P.

Macht man davon Gebrauch, dass das Infimum $v = \inf_{x \in M} f(x)$ ein Element der erweiterten reellen Zahlen ist, wird die Klassifikation in Satz 1.2.8 noch übersichtlicher, und außerdem wird der Sinn der formalen Setzungen des Infimums in den Fällen von Inkonsistenz und Unbeschränktheit klar. In den vier Fällen aus Satz 1.2.8 gilt nämlich:

a) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = (v, +\infty)$ mit $v = +\infty$,

b) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = (v, +\infty)$ mit $v = -\infty$,

c) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = (v, +\infty)$ mit $v \in \mathbb{R}$,

d) $\mathrm{pr}_\alpha \mathrm{epi}\,(f, M) = [v, +\infty)$ mit $v \in \mathbb{R}$.

Nach Satz 1.2.8 stimmt v in den Fällen c) und d) mit dem Infimum von f auf M überein. Die obige Darstellung erklärt nun, warum man analog in Fall a) $\inf_{x \in \emptyset} f(x) = +\infty$ und in Fall b) $\inf_{x \in M} f(x) = -\infty$ setzt.

Ob Unlösbarkeit tatsächlich problematisch ist, hängt von der Anwendung ab, wie das folgende Beispiel zeigt.

1.2.9 Beispiel (Distanz eines Punktes von einer Menge)
Für eine Menge $M \subseteq \mathbb{R}^n$ und einen Punkt $z \in \mathbb{R}^n$ heißt

$$\operatorname{dist}(z, M) := \inf_{x \in M} \|x - z\|_2$$

Distanz *von z zu M. Wie in Beispiel 1.1.1 gesehen, heißt ein Punkt $\bar{x} \in M$, dessen Zielfunktionswert $\|\bar{x} - z\|_2$ die Distanz realisiert, Projektion von z auf M. Die Distanz ist aber auch dann sinnvoll definiert, wenn eine solche Projektion nicht existiert, das zugrunde liegende Optimierungsproblem also unlösbar ist. Dies illustriert Abbildung 1.13 für eine nicht abgeschlossene Menge M, d.h. es gibt eine Folge $(x^\nu) \subseteq M$, die einen Grenzwert $x^\star \notin M$ besitzt. Ein optimaler Punkt existiert hier deswegen nicht, weil sich jeder zulässige Punkt $x \in M$ verbessern lässt.*

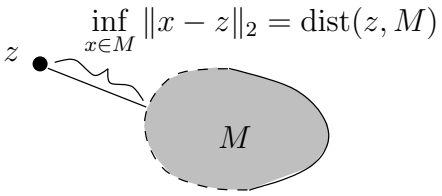


Abbildung 1.13: Distanz zu einer nicht abgeschlossenen Menge

Wegen der Nichtnegativität der Norm ist Unlösbarkeit wegen Unbeschränktheit für dieses Problem nicht möglich. Ferner gilt $\operatorname{dist}(z, \emptyset) = +\infty$ *für jedes* $z \in \mathbb{R}^n$.

Der Satz von Weierstraß

Wir wenden uns nun hinreichenden Bedingungen für die Lösbarkeit von P zu. Eine Minimalvoraussetzung dafür ist offensichtlich die Konsistenz von M, also $M \neq \emptyset$. In Beispiel 1.2.9 haben wir außerdem gesehen, dass fehlende Abgeschlossenheit von M zu Unlösbarkeit führen kann.

Die Unlösbarkeit der Minimierung von e^x auf $\mathbb{R}$ liegt an einem asymptotischen Effekt für $x \to -\infty$, der sich sicher dann ausschließen lässt, wenn man zusätzlich die Beschränktheit von M fordert, also die Existenz eines $R > 0$ mit $M \subseteq \{x \in \mathbb{R}^n | \, \|x\| \leq R\}$ (d.h. M liegt in einer hinreichend großen Kugel um den Nullpunkt – die Wahl der Norm ist dabei gleichgültig). Eine gleichzeitig abgeschlossene und beschränkte Menge $M \subseteq \mathbb{R}^n$ heißt auch *kompakt*.

Schließlich kann Unlösbarkeit auch durch Sprungstellen der Zielfunktion f ausgelöst werden. Beispielsweise besitzt

$$f(x) \;=\; \begin{cases} 1, & x \leq 0 \\ x, & x > 0 \end{cases}$$

keinen globalen Minimalpunkt auf $\mathbb{R}$, denn wieder gibt es zu jedem Lösungskandidaten eine Verbesserungsmöglichkeit (s. Abb. 1.14). Ausgeschlos-

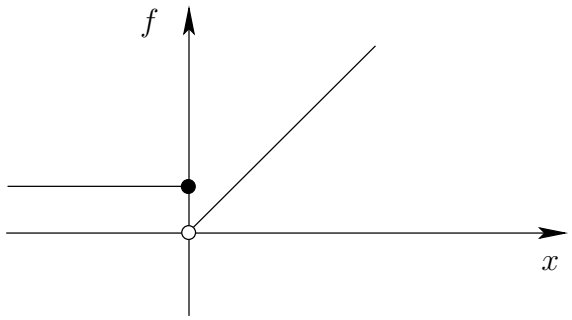


Abbildung 1.14: Unlösbarkeit wegen Sprungstelle von f

sen wird diese Situation sicher durch die Forderung der Stetigkeit von f.

Der folgende zentrale Satz zur Existenz von Minimal- und Maximalpunkten zeigt, dass die aufgeführten Forderungen an f und M ausreichen, um tatsächlich die Lösbarkeit von P zu garantieren.

1.2.10 Satz (Satz von Weierstraß)
Die Menge $M \subseteq \mathbb{R}^n$ sei nicht-leer und kompakt, und die Funktion $f : M \to \mathbb{R}$ sei stetig. Dann besitzt f auf M einen globalen Minimalpunkt und einen globalen Maximalpunkt.

Beweis. Es sei $v = \inf_{x \in M} f(x)$. Wegen $M \neq \emptyset$ gilt $v < +\infty$. Zu zeigen ist die Existenz eines $\bar{x}$ in M mit $v = f(\bar{x})$. Da v Infimum ist, existiert eine Folge $(x^\nu) \subseteq M$ mit $\lim_\nu f(x^\nu) = v$. In der Analysis wird bewiesen, dass jede in einer kompakten Menge M liegende Folge (x^ν) eine in M konvergente Teilfolge besitzt (Satz von Bolzano-Weierstraß). OBdA sei (x^ν) bereits selbst konvergent, es existiert also ein $x^\star \in M$ mit $\lim_\nu x^\nu = x^\star$. Aufgrund der Stetigkeit von f auf M gilt nun

$$f(x^\star) \;=\; f\left(\lim_\nu x^\nu\right) \;=\; \lim_\nu f(x^\nu) \;=\; v,$$

man kann also $\bar{x} := x^\star$ wählen. Der Beweis zur Existenz eines globales Maximalpunkts verläuft analog. •

1.2.11 Beispiel (Projektion auf eine Menge - Fortsetzung 1)
Nach Satz 1.2.10 ist das Projektionsproblem aus Beispiel 1.1.1 für jede nichtleere und kompakte Menge $M \subseteq \mathbb{R}^n$ lösbar, denn $f(x) = \|x - z\|_2$ ist eine stetige Funktion.

Während der Satz von Weierstraß die zentralen *hinreichenden* Bedingungen für die Lösbarkeit eines Optimierungsproblems liefert, wird für das Folgende entscheidend sein, dass außer der Konsistenz von M keine dieser Bedingungen auch *notwendig* für Lösbarkeit ist. Dass die Bedingungen stärker als nötig sind, sieht man alleine schon daran, dass der Satz von Weierstraß auch die Existenz eines globalen *Maximal*punkts garantiert, an der wir aber gar nicht interessiert waren.

Beispielsweise ist die Existenz von Minimalpunkten (aber nicht von Maximalpunkten) auch noch für gewisse unstetige Funktionen f garantiert. Eine Diskussion dieser sogenannten *Unterhalbstetigkeit* findet sich beispielsweise in [14].

Im Folgenden werden wir untersuchen, wie sich die Beschränktheit und die Abgeschlossenheit von M abschwächen lassen, da diese Voraussetzungen in Anwendungsproblemen häufig verletzt sind. Insbesondere für Probleme ohne Nebenbedingungen, sogenannte *unrestringierte Probleme*, gilt $M = \mathbb{R}^n$ (z.B. beim Zentrum einer Punktewolke, Bsp. 1.1.5, und in der Clusteranalyse, Bsp. 1.1.6). Zwar ist M dann nicht-leer und abgeschlossen, aber *nicht* beschränkt. Daher ist Satz 1.2.10 auf unrestringierte Probleme nicht anwendbar.

Unbeschränkte zulässige Mengen

Um den Satz von Weierstraß für Probleme mit unbeschränkter Menge M anwendbar zu machen, bedient man sich eines Tricks und betrachtet sogenannte untere Niveaumengen von f: Für $M \subseteq \mathbb{R}^n$, $f : M \to \mathbb{R}$ und $\alpha \in \overline{\mathbb{R}}$ heißt

$$\mathrm{lev}^{\alpha}_{\leq}(f, M) \;=\; \{x \in M |\; f(x) \leq \alpha\}$$

untere Niveaumenge von f auf M zum Niveau α. Im Falle $M = \mathbb{R}^n$ schreiben wir auch kurz

$$f^{\alpha}_{\leq} \;:=\; \mathrm{lev}^{\alpha}_{\leq}(f, \mathbb{R}^n) \quad (=\; \{x \in \mathbb{R}^n |\; f(x) \leq \alpha\}\,).$$

Für jede reellwertige Funktion f gilt $\mathrm{lev}^{+\infty}_{\leq}(f, M) = M$ und $\mathrm{lev}^{-\infty}_{\leq}(f, M) = \emptyset$.

Achtung: Die Menge $\text{lev}_{\leq}^{\alpha}(f, M)$ ist nicht mit dem Epigraphen von f auf M,

$$\text{epi}\,(f, M) \;=\; \{(x, \alpha) \in M \times \mathbb{R} |\; f(x) \leq \alpha\}\,,$$

zu verwechseln!

1.2.12 Beispiel *Für* $f(x) = x^2$ *gilt* $f_{\leq}^{1} = [-1, 1]\,,\; f_{\leq}^{0} = \{0\}\,,\; f_{\leq}^{-1} = \emptyset$ *(vgl. Abb. 1.15), und für* $f(x) = x_1^2 + x_2^2$ *gilt* $f_{\leq}^{1} = \{x \in \mathbb{R}^2 |\; x_1^2 + x_2^2 \leq 1\}$ *(die*

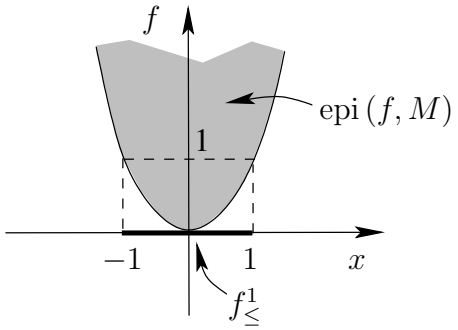


Abbildung 1.15: Untere Niveaumenge $f_{\leq}^{1}$ und Epigraph von $f(x) = x^2$ auf $\mathbb{R}$

Einheitskreisscheibe), $f_{\leq}^{0} = \{0\}\,,\; f_{\leq}^{-1} = \emptyset$ *(vgl. Abb. 1.16).*

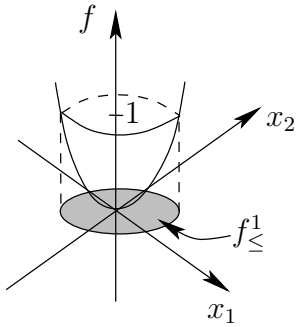


Abbildung 1.16: Untere Niveaumenge $f_{\leq}^{1}$ von $f(x) = x_1^2 + x_2^2$ auf $\mathbb{R}^2$

In der Analysis (und in [13]) wird gezeigt, dass für stetiges f die Mengen $f_{\leq}^{\alpha}$ für alle $\alpha \in \mathbb{R}$ abgeschlossen sind.

Für die folgenden Ergebnisse führen wir die Menge der globalen Minimalpunkte

$$S \;=\; \{\bar{x} \in M |\; \forall\, x \in M :\; f(x) \geq f(\bar{x})\}$$

von P ein. Die Lösbarkeit von P ist dann zu $S \neq \emptyset$ äquivalent, wir werden aber noch weitere Eigenschaften der Menge S zeigen können.

1.2.13 Lemma *Es sei $v = \inf_{x\in M} f(x)$. Dann gilt $S = \mathrm{lev}^v_{\leq}(f, M)$.*

Beweis. Es gilt

$$\begin{array}{rcl} \bar{x} \in S & \Leftrightarrow & \bar{x} \text{ globaler Minimalpunkt} \\ & \Leftrightarrow & \bar{x} \in M \text{ und } f(\bar{x}) = v \\ & \overset{\{x\in M \mid f(x)<v\} = \emptyset}{\Leftrightarrow} & \bar{x} \in M \text{ und } f(\bar{x}) \leq v \\ & \Leftrightarrow & \bar{x} \in \mathrm{lev}^v_{\leq}(f, M). \end{array}$$

•

1.2.14 Übung *Der Beweis von Lemma 1.2.13 deckt formal auch den Fall unlösbarer Probleme P ab, also $S = \emptyset$. Wie sieht man für die verschiedenen Fälle von Unlösbarkeit unabhängig vom Beweis von Lemma 1.2.13, dass die Menge $\mathrm{lev}^v_{\leq}(f, M)$ leer ist?*

1.2.15 Lemma *Für ein $\alpha \in \mathbb{R}$ sei $\mathrm{lev}^\alpha_{\leq}(f, M) \neq \emptyset$. Dann gilt $S \subseteq \mathrm{lev}^\alpha_{\leq}(f, M)$.*

Beweis. Wegen $\mathrm{lev}^\alpha_{\leq}(f, M) \neq \emptyset$ gibt es einen Punkt $\widetilde{x}$ in M mit $f(\widetilde{x}) \leq \alpha$. Nun sei $\bar{x}$ ein beliebiger globaler Minimalpunkt von P. Dann gilt $\bar{x} \in M$ und $f(\bar{x}) \leq f(\widetilde{x}) \leq \alpha$, also $\bar{x} \in \mathrm{lev}^\alpha_{\leq}(f, M)$. •

1.2.16 Satz (Verschärfter Satz von Weierstraß)
Für eine Menge $M \subseteq \mathbb{R}^n$ sei $f : M \to \mathbb{R}$ stetig, und mit einem $\alpha \in \mathbb{R}$ sei $\mathrm{lev}^\alpha_{\leq}(f, M)$ nicht-leer und kompakt. Dann ist auch S nicht-leer und kompakt.

Beweis. Wegen Lemma 1.2.15 kann man P äquivalent durch

$$\widetilde{P} : \ \min f(x) \quad \text{s.t.} \quad x \in \mathrm{lev}^\alpha_{\leq}(f, M)$$

ersetzen. $\widetilde{P}$ erfüllt die Voraussetzungen von Satz 1.2.10, woraus die Behauptung $S \neq \emptyset$ folgt. Außerdem impliziert $S = \mathrm{lev}^v_{\leq}(f, M) \subseteq \mathrm{lev}^\alpha_{\leq}(f, M)$ die Beschränktheit von S. Zum Nachweis der Abgeschlossenheit von S betrachte eine konvergente Folge $(x^\nu) \subseteq S$ mit Limes $x^\star$. Da S in der abgeschlossenen Menge $\mathrm{lev}^\alpha_{\leq}(f, M)$ enthalten ist, folgt $x^\star \in \mathrm{lev}^\alpha_{\leq}(f, M) \subseteq M$, und die Stetigkeit von f auf M garantiert, dass aus $f(x^\nu) \leq v$, $\nu \in \mathbb{N}$, auch $f(x^\star) \leq v$ folgt, insgesamt also $x^\star \in \mathrm{lev}^v_{\leq}(f, M) = S$. Damit ist S kompakt. •

Die *Verschärfung* von Satz 1.2.16 gegenüber Satz 1.2.10 bezieht sich darauf, dass die uns interessierende Aussage des Satzes von Weierstraß, nämlich die

Existenz eines globalen *Minimal*punkts, auch unter der schwächeren Voraussetzung von Satz 1.2.16 folgt. Da nun allerdings keine Aussage mehr zur Existenz eines globalen *Maximal*punkts gemacht werden kann, sind die beiden Sätze genau genommen unabhängig voneinander.

1.2.17 Beispiel *Betrachtet werde das Problem*

$$P: \quad \min e^x \quad s.t. \quad x \geq 0\,.$$

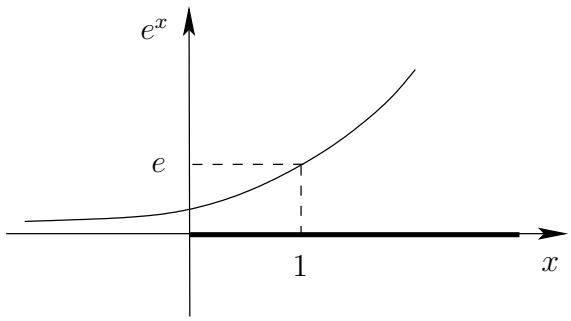


Abbildung 1.17: e^x mit $x \geq 0$

Hier ist $M = \{x \in \mathbb{R} | \, x \geq 0\}$ *unbeschränkt, Satz 1.2.10 also nicht anwendbar. Aber beispielsweise mit* $\alpha = e$ *ist*

$$\mathrm{lev}^e_{\leq}(f, M) \;=\; \{x \in M | \; e^x \leq e\} \;=\; \{x \geq 0 | \; x \leq 1\} \;=\; [0, 1]$$

nicht-leer und kompakt. Folglich ist Satz 1.2.16 anwendbar und P *daher lösbar.*

Der verschärfte Satz von Weierstraß zeigt auch, dass zur Lösbarkeit des Projektionsproblems aus Beispiel 1.1.1 die in Beispiel 1.2.11 getroffene Voraussetzung der Beschränktheit von M unnötig ist:

1.2.18 Beispiel (Projektion auf eine Menge - Fortsetzung 2)
Das Projektionsproblem aus Beispiel 1.1.1 für jede nicht-leere und abgeschlossene Menge $M \subseteq \mathbb{R}^n$ *lösbar. In der Tat bildet für jedes* $\alpha \geq 0$ *die Menge*

$$f^\alpha_{\leq} \;=\; \{x \in \mathbb{R}^n | \; \|x - z\|_2 \leq \alpha\}$$

eine Kugel mit Mittelpunkt z *und Radius* α*. Für einen beliebigen Punkt* $\widetilde{x} \in M$ *wählen wir* α *so groß, dass* $\widetilde{x} \in f^\alpha_{\leq}$ *gilt, etwa mit der Wahl* $\alpha = \|\widetilde{x} - z\|_2$*. Damit gilt* $\widetilde{x} \in f^\alpha_{\leq} \cap M = \mathrm{lev}^\alpha_{\leq}(f, M)$*, so dass* $\mathrm{lev}^\alpha_{\leq}(f, M)$ *nicht-leer ist, und außerdem ist* $\mathrm{lev}^\alpha_{\leq}(f, M)$ *als Schnitt der kompakten Menge* $f^\alpha_{\leq}$ *mit der abgeschlossenen Menge* M *kompakt. Satz 1.2.16 liefert nun die Behauptung.*

1.2.19 Korollar (Verschärfter S. v. Weierstraß für unrest. Probleme)
$f : \mathbb{R}^n \to \mathbb{R}$ sei stetig, und mit einem $\alpha \in \mathbb{R}$ sei $f^\alpha_\leq$ nicht-leer und kompakt. Dann ist auch S nicht-leer und kompakt.

Beweis. Satz 1.2.16 mit $M = \mathbb{R}^n$. •

1.2.20 Beispiel *Für $f(x) = (x-5)^2$ ist $f^1_\leq = [4, 6]$ nicht-leer und kompakt, also besitzt f nach Korollar 1.2.19 einen globalen Minimalpunkt auf $\mathbb{R}$.*

1.2.21 Beispiel *Mit $f(x) = e^x$ gilt $f^\alpha_\leq = \emptyset$ für alle $\alpha \leq 0$ sowie $f^\alpha_\leq = (-\infty, \log(\alpha)]$ für alle $\alpha > 0$. Daher ist $f^\alpha_\leq$ für kein α nicht-leer und kompakt, Korollar 1.2.19 also nicht anwendbar. f besitzt auch tatsächlich keinen globalen Minimalpunkt auf $\mathbb{R}$.*

1.2.22 Beispiel *Für $f(x) = \sin(x)$ ist Korollar 1.2.19 ebenfalls nicht anwendbar, da alle $f^\alpha_\leq$ unbeschränkt oder leer sind. f besitzt aber trotzdem globale Minimalpunkte auf $\mathbb{R}$ (wobei S allerdings nicht kompakt ist).*

Im Folgenden leiten wir ein einfaches Kriterium her, aus dem die Kompaktheit von $\text{lev}^\alpha_\leq(f, M)$ mit *jedem* $\alpha \in \mathbb{R}$ folgt. Dadurch kann man die Voraussetzungen von Satz 1.2.16 und Korollar 1.2.19 garantieren, ohne ein explizites Niveau α angeben zu müssen.

1.2.23 Definition (Koerzivität bei ∞)
Gegeben sei eine Funktion $f : M \to \mathbb{R}$ mit $M \subseteq \mathbb{R}^n$. Falls für alle Folgen $(x^\nu) \subseteq M$ mit $\lim_\nu \|x^\nu\| \to +\infty$ auch

$$\lim_\nu f(x^\nu) \;=\; +\infty$$

gilt, dann heißt f koerziv bei ∞ *auf M. Falls M abgeschlossen ist, heißt f kurz* koerziv *auf M.*

1.2.24 Beispiel *$f(x) = (x-5)^2$ ist koerziv auf $\mathbb{R}$.*

1.2.25 Beispiel *$f(x) = e^x$ ist nicht koerziv auf $M = \mathbb{R}$, wohl aber auf der Menge $M = \{x \in \mathbb{R} |\; x \geq 0\}$.*

1.2.26 Beispiel (Beispiel 1.1.5 - Fortsetzung 1)
Die Zielfunktion

$$f(z) \;=\; \left\| \begin{pmatrix} \|z - x^1\|_2 \\ \vdots \\ \|z - x^m\|_2 \end{pmatrix} \right\|_2$$

aus dem Problem, das Zentrum einer Punktewolke zu finden (Bsp. 1.1.5), ist koerziv auf $\mathbb{R}^n$, denn es gilt

$$\begin{aligned} f(z) \;&=\; \sqrt{\sum_{i=1}^{m} \|z - x^i\|_2^2} \;\geq\; \sqrt{\|z - x^1\|_2^2} \;=\; \|z - x^1\|_2 \\ &\geq\; \Big| \|z\|_2 - \|x^1\|_2 \Big| \;\overset{\|z\| \to \infty}{\longrightarrow}\; +\infty. \end{aligned}$$

1.2.27 Beispiel (Beispiel 1.1.6 - Fortsetzung 1)
Die Zielfunktion

$$f(z^1, \ldots, z^k) \;=\; \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \|z^\ell - x^i\|$$

mit $k \geq 2$ aus dem Problem P_1 der Clusteranalyse (Bsp. 1.1.6) ist zwar stetig (s. Analysis), aber nicht koerziv auf $\mathbb{R}^{nk}$:

Wähle $z^{1,\nu} = \ldots = z^{k-1,\nu} = 0, z^{k,\nu} = \nu e_n$ (mit dem n-ten Einheitsvektor e_n) für alle $\nu \in \mathbb{N}$. Dann gilt zwar

$$\left\| \begin{pmatrix} z^{1,\nu} \\ \vdots \\ z^{k,\nu} \end{pmatrix} \right\| \;\overset{\nu \to \infty}{\longrightarrow}\; \infty,$$

aber

$$\begin{aligned} f(z^{1,\nu}, \ldots, z^{k,\nu}) \;&=\; \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \|z^{\ell,\nu} - x^i\| \\ &=\; \sum_{i=1}^{m} \min \Big\{ \underbrace{\min_{\ell=1,\ldots,k-1} \|x^i\|}_{\|x^i\|} \;,\; \underbrace{\|z^{k,\nu} - x^i\|}_{\geq \left| \|z^{k,\nu}\| - \|x^i\| \right|} \Big\} \\ &\overset{\nu \geq \nu_0}{=}\; \sum_{i=1}^{m} \|x^i\|. \end{aligned}$$

Da dieser Ausdruck von $(z^{1,\nu}, ..., z^{k,\nu})$ unabhängig ist und somit für $\nu \to \infty$ nicht gegen unendlich geht, ist f auf keiner Menge koerziv, die die Folge der Punkte $(z^{1,\nu}, ..., z^{k,\nu})$ enthält, und damit insbesondere nicht auf $\mathbb{R}^{nk}$.

1.2.28 Beispiel *Auf kompakten Mengen M ist jede Funktion f trivialerweise koerziv.*

Für das folgende Ergebnis erinnern wir daran, dass die leere Menge trivialerweise beschränkt ist.

1.2.29 Lemma *Die Funktion $f : M \to \mathbb{R}$ sei koerziv bei ∞ auf der Menge $M \subseteq \mathbb{R}^n$. Dann sind die Mengen $\mathrm{lev}^\alpha_\leq(f, M)$ für jedes Niveau $\alpha \in \mathbb{R}$ beschränkt.*

Beweis. Wir wählen ein beliebiges $\alpha \in \mathbb{R}$ und nehmen an, $\mathrm{lev}^\alpha_\leq(f, M)$ sei unbeschränkt. Dann existiert eine Folge $(x^\nu) \subseteq \mathrm{lev}^\alpha_\leq(f, M)$ mit $\lim_\nu \|x^\nu\| = +\infty$. Aufgrund der Koerzivität bei ∞ von f auf M folgt hieraus $\lim_\nu f(x^\nu) = +\infty$. Dies steht im Widerspruch zu $f(x^\nu) \leq \alpha$ für alle $\nu \in \mathbb{N}$. Also ist $\mathrm{lev}^\alpha_\leq(f, M)$ beschränkt. •

1.2.30 Korollar *Es sei M nicht-leer und abgeschlossen, und $f : M \to \mathbb{R}$ sei stetig und koerziv auf M. Dann ist S nicht-leer und kompakt.*

Beweis. Wegen $M \neq \emptyset$ können wir zunächst ein $\bar{x} \in M$ wählen und $\alpha = f(\bar{x})$ setzen. Dann enthält die Menge $\mathrm{lev}^\alpha_\leq(f, M)$ offenbar den Punkt $\bar{x}$ und ist insbesondere nicht leer.

Aufgrund der Stetigkeit von f auf M und der Abgeschlossenheit von M ist $\mathrm{lev}^\alpha_\leq(f, M)$ außerdem abgeschlossen, und aufgrund von Lemma 1.2.29 auch beschränkt.

Insgesamt haben wir ein $\alpha \in \mathbb{R}$ gefunden, so dass $\mathrm{lev}^\alpha_\leq(f, M)$ nicht-leer und kompakt ist. Damit liefert Satz 1.2.16 die Behauptung. •

1.2.31 Beispiel (Beispiel 1.1.5 - Fortsetzung 2)
Das Problem, das Zentrum einer Punktewolke zu finden (Bsp. 1.1.5), ist nach Beispiel 1.2.26 und Korollar 1.2.30 lösbar.

Wegen Beispiel 1.2.27 lässt sich Korollar 1.2.30 nicht benutzen, um die Lösbarkeit des Problems P_1 der Clusteranalyse zu zeigen. Dass sie aber trotzdem gilt, sehen wir im Folgenden anhand einer weiteren Möglichkeit, den Satz von Weierstraß (S. 1.2.10) auf unrestringierte Probleme anwendbar zu machen.

1.2.32 Beispiel (Beispiel 1.1.6 - Fortsetzung 2)
Um zu zeigen, dass das Problem P_1 der Clusteranalyse (Bsp. 1.1.6) mit Euklidischer Norm als innerer Norm, also die unrestringierte Minimierung der Zielfunktion

$$f(z^1, \ldots, z^k) = \left\| \begin{pmatrix} \min_{\ell=1,\ldots,k} \|z^\ell - x^1\|_2 \\ \vdots \\ \min_{\ell=1,\ldots,k} \|z^\ell - x^m\|_2 \end{pmatrix} \right\|_1 ,$$

für $k \geq 2$ lösbar ist, werden wir eine nicht-leere und kompakte Menge $M \subseteq \mathbb{R}^{nk}$ konstruieren, außerhalb der man nicht nach globalen Minimalpunkten zu suchen braucht. Die unrestringierte Minimierung von f ist demnach äquivalent zur Minimierung von f über M, und Satz 1.2.10 liefert die Behauptung.

Dazu betrachten wir die Box (s. auch Kap. 3.3) $X = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_n, \overline{x}_n]$ mit

$$\underline{x}_j \;:=\; \min_{i=1,\ldots,m} x_j^i\,, \quad \overline{x}_j \;:=\; \max_{i=1,\ldots,m} x_j^i\,, \quad j = 1, \ldots, n,$$

also die kleinste Box in $\mathbb{R}^n$, die alle Datenpunkte enthält, und setzen $M := \prod_{\ell=1}^{k} X$.

Im Folgenden werden wir zeigen, dass zu jedem Punkt $(z^1, \ldots, z^k)$ in $\mathbb{R}^{nk}$ ein Punkt $(\tilde{z}^1, \ldots, \tilde{z}^k)$ in M mit mindestens ebenso gutem Zielfunktionswert existiert. Selbst wenn f also globale Minimalpunkte außerhalb von M besitzen sollte, gäbe es andere globale Minimalpunkte in M, so dass die restringierte Minimierung von f über M anstelle der unrestringierten Minimierung zulässig ist.

Es sei also $(z^1, \ldots, z^k) \in \mathbb{R}^{nk}$. Wir setzen für jedes $\ell \in \{1, \ldots, k\}$ und jedes $j \in \{1, \ldots, n\}$

$$\tilde{z}_j^\ell \;=\; \begin{cases} z_j^\ell, & \text{falls } z_j^\ell \in [\underline{x}_j, \overline{x}_j] \\ \underline{x}_j, & \text{falls } z_j^\ell < \underline{x}_j \\ \overline{x}_j, & \text{falls } z_j^\ell > \overline{x}_j. \end{cases}$$

Dann liegt der Punkt $(\tilde{z}^1, \ldots, \tilde{z}^k)$ offensichtlich in M. Ferner gilt für jedes $\ell \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n\}$ und $i \in \{1, \ldots, m\}$ im ersten obigen Fall

$$|\tilde{z}_j^\ell - x_j^i| \;=\; |z_j^\ell - x_j^i|,$$

im zweiten Fall

$$|\tilde{z}_j^\ell - x_j^i| \;=\; |\underline{x}_j - x_j^i| \;=\; x_j^i - \underline{x}_j \;<\; x_j^i - z_j^\ell \;=\; |z_j^\ell - x_j^i|,$$

und im dritten Fall

$$|\tilde{z}_j^\ell - x_j^i| \;=\; |\overline{x}_j - x_j^i| \;=\; \overline{x}_j - x_j^i \;<\; z_j^\ell - x_j^i \;=\; |z_j^\ell - x_j^i|,$$

insgesamt also in jedem der drei Fälle $|\tilde{z}_j^\ell - x_j^i| \leq |z_j^\ell - x_j^i|$. Mit der Definition der Euklidischen Norm macht man sich leicht klar, dass dann auch

$$\|\tilde{z}^\ell - x^i\|_2 \;\leq\; \|z^\ell - x^i\|_2 \tag{1.2.1}$$

gilt. Hieraus folgt

$$\min_{\ell=1,\ldots,k} \|\tilde{z}^\ell - x^i\|_2 \;\leq\; \min_{\ell=1,\ldots,k} \|z^\ell - x^i\|_2$$

sowie per Definition der ℓ_1-Norm,

$$f(\tilde{z}^1, \ldots, \tilde{z}^k) \;\leq\; f(z^1, \ldots, z^k). \tag{1.2.2}$$

Dies ist die Behauptung.

Wir merken an, dass wir im obigen Beispiel zum Nachweis sowohl von (1.2.1) also auch von (1.2.2) die Eigenschaft der beteiligten Normen benutzt haben, dass aus $|a_j| \leq |b_j|$, $j = 1, \ldots, n$, stets $\|a\| \leq \|b\|$ folgt. Jede Norm mit dieser Eigenschaft heißt *absolut-monoton* auf $\mathbb{R}^n$ (diese Eigenschaft wird in der Literatur auch häufig als Monotonie bezeichnet, was aber nicht konsistent mit der Definition von Monotonie für Funktionale ist, vgl. Def. 1.3.8).

1.2.33 Übung *Zeigen Sie, dass alle ℓ_p-Normen mit $p \in [1, \infty]$ absolut-monoton auf $\mathbb{R}^n$ sind.*

1.2.34 Übung *Geben Sie ein Beispiel einer auf $\mathbb{R}^2$ nicht absolut-monotonen Norm.*

Da außer der Absolut-Monotonie der beteiligten Normen keine weiteren ihrer speziellen Eigenschaften benutzt wurden, übertragen sich die Argumente aus Beispiel 1.2.32 ohne weiteres auf die Lösbarkeit von Problemen der Clusteranalyse mit beliebigen inneren und äußeren Normen, solange beide absolut-monoton sind (z.B. als ℓ_p-Normen, vgl. Ü. 1.2.33).

Ferner lässt sich mit derselben Idee ein zu Beispiel 1.2.31 alternativer Beweis der Lösbarkeit des Problems zur Bestimmung des Zentrums einer Punktewolke angeben, der dann ebenfalls auf beliebige absolut-monotone innere und äußere Normen übertragbar ist.

1.2.35 Übung *Auf der abgeschlossenen Menge $X \subseteq \mathbb{R}^n$ seien die Funktionen f, g_i, $i \in I$, stetig, die Menge $\{x \in X \,|\, g_i(x) \leq 0,\, i \in I\}$ sei nicht-leer, und mindestens eine der Funktionen f, g_i, $i \in I$, sei koerziv auf X. Zeigen Sie, dass die Menge S der Optimalpunkte von f auf M dann nicht-leer und kompakt ist.*

Nicht-abgeschlossene zulässige Mengen

Das folgende Beispiel zeigt, dass in Anwendungen auch Optimierungsprobleme mit nicht-abgeschlossenen zulässigen Mengen auftreten können. Deren Lösbarkeit lässt sich mit den bislang hergeleiteten Resultaten nicht garantieren.

1.2.36 Beispiel (Maximum-Likelihood-Schätzer)
Gegeben seien N Beobachtungen $\widehat{x}_1, \ldots, \widehat{x}_N \geq 0$ mit $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \widehat{x}_i > 0$, die als Realisierungen stochastisch unabhängiger und mit Parameter $\lambda > 0$ exponentialverteilter Zufallsvariablen $X_1, \ldots, X_N$ aufgefasst werden. Gesucht ist der Parameter λ, der zu den Beobachtungen „am besten passt“. Dazu kann man mit Hilfe der Dichtefunktionen der einzelnen X_i,

$$f(\lambda, x_i) \;=\; \begin{cases} \lambda e^{-\lambda x_i}, & x_i \geq 0 \\ 0, & x_i < 0, \end{cases}$$

zunächst die gemeinsame Dichte aller Zufallsvariablen

$$L(\lambda, x) \;=\; \prod_{i=1}^{N} f(\lambda, x_i)$$

betrachten. Der Maximum-Likelihood-Schätzer bestimmt dann λ als optimalen Punkt des Problems

$$ML: \qquad \max_{\lambda} L(\lambda, \widehat{x}) \quad s.t. \quad \lambda > 0.$$

Die zulässige Menge $M = (0, +\infty)$ dieses Problems ist offensichtlich nicht abgeschlossen. Es ist auch sinnlos, den Parameterwert $\lambda = 0$ künstlich hinzuzufügen, da $f(0, x)$ keine Wahrscheinlichkeitsdichte ist.

Wir werden im Folgenden sehen, wie die Lösbarkeit dieses Problems trotzdem garantiert werden kann und später auch einen globalen Maximalpunkt bestimmen. Dazu berechnen wir zunächst

$$L(\lambda, \widehat{x}) \;=\; \prod_{i=1}^{N} \lambda e^{-\lambda \widehat{x}_i} \;=\; \lambda^N e^{-\lambda N \bar{x}}.$$

Abbildung 1.18 zeigt den Verlauf dieser Funktion im Fall $N = 2$ und $\bar{x} = 1$.

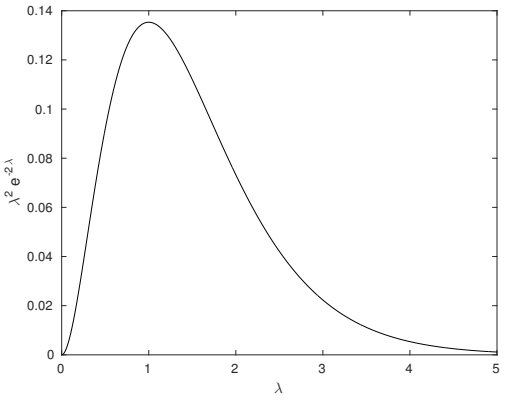


Abbildung 1.18: Graph der Likelihood-Funktion L

Da die Likelihood-Funktion L eine ausgeprägte Produktstruktur sowie ausschließlich positive Werte besitzt, bietet es sich an, stattdessen die sogenannte Log-Likelihood-Funktion

$$\ell(\lambda, \widehat{x}) \;:=\; \log\left(L(\lambda, \widehat{x})\right) \;=\; N \log(\lambda) - \lambda N \bar{x}$$

zu betrachten. Da die Funktion log *streng monoton wachsend auf dem Bildbereich $(0, +\infty)$ von L ist, kann man mit Hilfe von Übung 1.3.5 zeigen, dass das Problem*

$$ML_{\log}: \qquad \max_{\lambda} \ell(\lambda, \widehat{x}) \quad s.t. \quad \lambda > 0$$

dieselben Optimalpunkte wie ML besitzt. Schließlich streichen wir mittels Übung 1.3.1a) die Konstante $N > 0$ aus der Zielfunktion und gehen zum äquivalenten Minimierungsproblem

$$P_{ML}: \qquad \min_{\lambda} \lambda\bar{x} - \log(\lambda) \qquad s.t. \qquad \lambda > 0$$

über, dessen Zielfunktion für $\bar{x} = 1$ in Abbildung 1.19 geplottet ist.

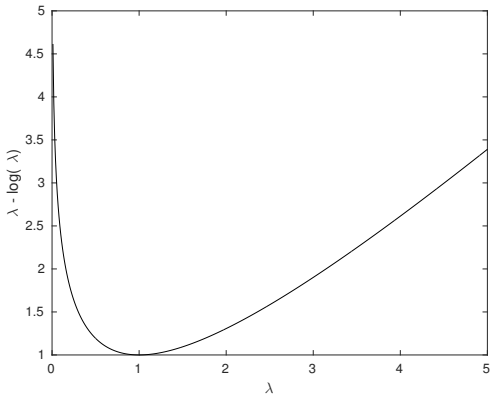


Abbildung 1.19: Zielfunktion des Problems P_{ML}

Beispiel 1.2.36 motiviert, dass man Koerzivität einer Funktion f auf einer *nicht notwendigerweise abgeschlossenen* Menge M nicht wie in Definition 1.2.23 nur „bei ∞“ fordern sollte, sondern auch an gewissen Punkten „im Endlichen“:

1.2.37 Definition (Koerzivität)
Gegeben seien eine (nicht notwendigerweise abgeschlossene) Menge $M \subseteq \mathbb{R}^n$ und eine Funktion $f : M \to \mathbb{R}$. Falls für alle Folgen $(x^\nu) \subseteq M$ mit $\lim_\nu \|x^\nu\| \to \infty$ und alle konvergenten Folgen $(x^\nu) \subseteq M$ mit $\lim_\nu x^\nu \notin M$ die Bedingung

$$\lim_{\nu} f(x^\nu) \;=\; +\infty$$

gilt, dann heißt f koerziv *auf M.*

Für abgeschlossene Mengen M stimmt der Koerzivitätsbegriff aus Definition 1.2.37 offenbar mit der Koerzivität bei ∞ aus Definition 1.2.23 überein, was die dort eingeführte abkürzende Sprechweise für Koerzivität bei ∞ auf abgeschlossenen Mengen begründet.

1.2.38 Beispiel (Beispiel 1.2.36 - Fortsetzung 1)
Für $\bar{x} > 0$ ist die Zielfunktion $f(\lambda) = \lambda\bar{x} - \log(\lambda)$ des Problems P_{ML} aus Beispiel 1.2.36 koerziv auf $M = (0, +\infty)$. Wir merken an, dass nur die Log-Likelihood-Funktion ℓ zu Koerzivität führt, die Likelihood-Funktion L selbst aber nicht.

1.2.39 Lemma *Die Funktion $f : M \to \mathbb{R}$ sei stetig und koerziv auf der Menge $M \subseteq \mathbb{R}^n$. Dann sind die Mengen $\mathrm{lev}^{\alpha}_{\leq}(f, M)$ für jedes Niveau $\alpha \in \mathbb{R}$ kompakt.*

Beweis. Wir wählen ein beliebiges $\alpha \in \mathbb{R}$. Nach Lemma 1.2.29 ist die Menge $\mathrm{lev}^{\alpha}_{\leq}(f, M)$ beschränkt. Ihre Abgeschlossenheit ist im Falle $\mathrm{lev}^{\alpha}_{\leq}(f, M) = \emptyset$ klar. Ansonsten wählen wir eine konvergente Folge $(x^\nu) \subseteq \mathrm{lev}^{\alpha}_{\leq}(f, M)$, deren Grenzwert wir mit $x^\star$ bezeichnen. Angenommen, $x^\star$ läge nicht in M. Wegen der Koerzivität von f auf M folgte dann $\lim_\nu f(x^\nu) = +\infty$, im Widerspruch zu $f(x^\nu) \leq \alpha$ für alle $\nu \in \mathbb{N}$. Daher gilt $x^\star \in M$, und die Stetigkeit von f an $x^\star$ liefert schließlich $f(x^\star) \leq \alpha$, also insgesamt $x^\star \in \mathrm{lev}^{\alpha}_{\leq}(f, M)$. Damit ist $\mathrm{lev}^{\alpha}_{\leq}(f, M)$ auch abgeschlossen, insgesamt also kompakt. •

1.2.40 Korollar *Es sei M nicht-leer, und $f : M \to \mathbb{R}$ sei stetig und koerziv auf M. Dann ist S nicht-leer und kompakt.*

Beweis. Wie im Beweis zu Korollar 1.2.30 setzen wir $\alpha = f(\bar{x})$ mit einem $\bar{x} \in M$ und erhalten wie dort sofort die Konsistenz der Menge $\mathrm{lev}^{\alpha}_{\leq}(f, M)$. Nach Lemma 1.2.39 ist diese Menge außerdem kompakt, also haben wir wieder ein $\alpha \in \mathbb{R}$ gefunden, so dass $\mathrm{lev}^{\alpha}_{\leq}(f, M)$ nicht-leer und kompakt ist. Damit liefert Satz 1.2.16 die Behauptung. •

1.2.41 Beispiel (Beispiel 1.2.36 - Fortsetzung 2)
Das Problem P_{ML} und damit auch das Problem ML aus Beispiel 1.2.36 sind nach Beispiel 1.2.38 und Korollar 1.2.40 lösbar.

1.3 Rechenregeln und Umformungen

Dieser Abschnitt führt eine Reihe von Rechenregeln und äquivalenten Umformungen von Optimierungsproblemen auf, die im Rahmen der Vorlesung von Interesse sind. Die Existenz aller auftretenden Optimalpunkte und -werte wird in diesem Abschnitt ohne weitere Erwähnung vorausgesetzt und muss bei Anwendung der Resultate zunächst zum Beispiel mit den Techniken aus Kapitel 1.2 garantiert werden. Die Übertragung der Resultate zu Optimalwerten auf Fälle von nicht angenommenen Infima und Suprema ist dem Leser als weitere Übung überlassen.

1.3.1 Übung (Skalare Vielfache und Summen)
Gegeben seien $M \subseteq \mathbb{R}^n$ und $f, g : M \to \mathbb{R}$. Dann gilt:

a) $\forall\, \alpha \geq 0,\ \beta \in \mathbb{R}: \ \min\limits_{x \in M} (\alpha f(x) + \beta) \ = \ \alpha \, (\min\limits_{x \in M} f(x)) + \beta.$

b) $\forall\, \alpha < 0,\ \beta \in \mathbb{R}: \ \min\limits_{x \in M} (\alpha f(x) + \beta) \ = \ \alpha \, (\max\limits_{x \in M} f(x)) + \beta.$

c) $\min\limits_{x \in M} (f(x) + g(x)) \ \geq \ \min\limits_{x \in M} f(x) + \min\limits_{x \in M} g(x).$

d) In c) kann „>“ auftreten.

In a) und b) stimmen außerdem jeweils die lokalen bzw. globalen Optimalpunkte der Optimierungsprobleme überein.

1.3.2 Übung (Separable Zielfunktion auf kartesischem Produkt)
Es seien $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$, $f : X \to \mathbb{R}$ und $g : Y \to \mathbb{R}$. Dann gilt

$$\min_{(x,y) \in X \times Y} (f(x) + g(y)) \ = \ \min_{x \in X} f(x) + \min_{y \in Y} g(y).$$

1.3.3 Übung (Vertauschung von Minima und Maxima)
Es seien $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$, $M = X \times Y$ und $f : M \to \mathbb{R}$ gegeben. Dann gilt

a) $\min\limits_{(x,y) \in M} f(x,y) \ = \ \min\limits_{x \in X} \min\limits_{y \in Y} f(x,y) \ = \ \min\limits_{y \in Y} \min\limits_{x \in X} f(x,y).$

b) $\max\limits_{(x,y) \in M} f(x,y) \ = \ \max\limits_{x \in X} \max\limits_{y \in Y} f(x,y) \ = \ \max\limits_{y \in Y} \max\limits_{x \in X} f(x,y).$

c) $\min\limits_{x \in X} \max\limits_{y \in Y} f(x,y) \ \geq \ \max\limits_{y \in Y} \min\limits_{x \in X} f(x,y).$

d) In c) kann „>“ auftreten.

1.3.4 Übung (Vereinigung)
Es seien I eine beliebige Indexmenge, $M_i \subseteq \mathbb{R}^n$, $i \in I$, und $f : \bigcup_{i\in I} M_i \to \mathbb{R}$ gegeben. Dann gilt

$$\min_{x\in\bigcup_{i\in I} M_i} f(x) \;=\; \min_{i\in I} \min_{x\in M_i} f(x).$$

1.3.5 Übung (Monotone Transformation)
Zu $M \subseteq \mathbb{R}^n$ und einer Funktion $f : M \to Y$ mit $Y \subseteq \mathbb{R}$ sei $\psi : Y \to \mathbb{R}$ eine streng monoton wachsende Funktion. Dann gilt

$$\min_{x\in M} \psi(f(x)) \;=\; \psi\,(\min_{x\in M} f(x)),$$

und die lokalen bzw. globalen Minimalpunkte stimmen überein.

Für die folgende Übung erinnern wir an die Definition der Parallelprojektion aus Definition 1.2.7.

1.3.6 Übung (Projektions-Umformulierung)
Gegeben seien $M \subseteq \mathbb{R}^n \times \mathbb{R}^m$ und eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$, die nicht von den Variablen aus $\mathbb{R}^m$ abhängt. Dann sind die Probleme

$$P: \qquad \min_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m} f(x) \quad s.t. \quad (x,y) \in M$$

und

$$P_{\text{proj}}: \qquad \min_{x\in\mathbb{R}^n} f(x) \quad s.t. \quad x \in \text{pr}_x\, M$$

in folgendem Sinne äquivalent:

a) Für jeden lokalen bzw. globalen Minimalpunkt $(x^\star, y^\star)$ von P ist $x^\star$ lokaler bzw. globaler Minimalpunkt von P_{proj}.

b) Für jeden lokalen bzw. globalen Minimalpunkt $x^\star$ von P_{proj} existiert ein $y^\star \in \mathbb{R}^m$, so dass $(x^\star, y^\star)$ lokaler bzw. globaler Minimalpunkt von P ist.

c) Die Minimalwerte von P und P_{proj} stimmen überein.

1.3.7 Übung (Epigraph-Umformulierung)
Gegeben seien $M \subseteq \mathbb{R}^n$ und eine Funktion $f : M \to \mathbb{R}$. Dann sind die Probleme

$$P: \qquad \min_{x \in \mathbb{R}^n} f(x) \quad s.t. \quad x \in M$$

und

$$P_{\text{epi}}: \qquad \min_{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}} \alpha \quad s.t. \quad f(x) \leq \alpha, \quad x \in M$$

in folgendem Sinne äquivalent:

a) *Für jeden lokalen bzw. globalen Minimalpunkt $x^\star$ von P ist $(x^\star, f(x^\star))$ lokaler bzw. globaler Minimalpunkt von P_{epi}.*
b) *Für jeden lokalen bzw. globalen Minimalpunkt $(x^\star, \alpha^\star)$ von P_{epi} ist $x^\star$ lokaler bzw. globaler Minimalpunkt von P.*
c) *Die Minimalwerte von P und P_{epi} stimmen überein.*

1.3.8 Definition (Monotones Funktional)
Wir nennen $F : \mathbb{R}^k \to \mathbb{R}$ monoton *(auf $\mathbb{R}^k$), falls*

$$\forall\, x, y \in \mathbb{R}^k: \quad x \leq y \quad \Rightarrow \quad F(x) \leq F(y)$$

gilt.

1.3.9 Übung (Verallgemeinerte Epigraph-Umformulierung)
Gegeben seien $X \subseteq \mathbb{R}^n$, Funktionen $f : X \to \mathbb{R}^k$ und $g : X \to \mathbb{R}^\ell$ sowie monotone Funktionen $F : \mathbb{R}^k \to \mathbb{R}$ und $G : \mathbb{R}^\ell \to \mathbb{R}$. Dann sind die Probleme

$$P: \qquad \min_{x \in \mathbb{R}^n} F(f(x)) \quad s.t. \quad G(g(x)) \leq 0,\ x \in X$$

und

$$P_{\text{epi}}: \qquad \min_{(x,\alpha,\beta) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^\ell} F(\alpha) \quad s.t. \quad \begin{aligned} G(\beta) &\leq 0 \\ f(x) &\leq \alpha \\ g(x) &\leq \beta \\ x &\in X \end{aligned}$$

in folgendem Sinne äquivalent:

a) *Für jeden lokalen bzw. globalen Minimalpunkt $x^\star$ von P ist $(x^\star, f(x^\star), g(x^\star))$ lokaler bzw. globaler Minimalpunkt von P_{epi}.*
b) *Für jeden lokalen bzw. globalen Minimalpunkt $(x^\star, \alpha^\star, \beta^\star)$ von P_{epi} ist $x^\star$ lokaler bzw. globaler Minimalpunkt von P.*
c) *Die Minimalwerte von P und P_{epi} stimmen überein.*

1.3.10 Übung *Stellen Sie ein zu dem nicht-glatten Optimierungsproblem*

$$\begin{aligned} P: \quad \min_{x \in \mathbb{R}^2} \ & \left(\max\{x_1 + 3x_2, -x_1 + x_2\} + 2\max\{5x_1 - x_2, -3x_1 + x_2, x_1\}\right) \\ s.t. \quad & x_1 - x_2 + \max\{x_1 + 7x_2, 2x_1 - x_2\} + \max\{-x_1 - x_2, x_1 + 4x_2\} \leq 0 \end{aligned}$$

äquivalentes lineares Optimierungsproblem auf.

Kapitel 2

Konvexe Optimierungsprobleme

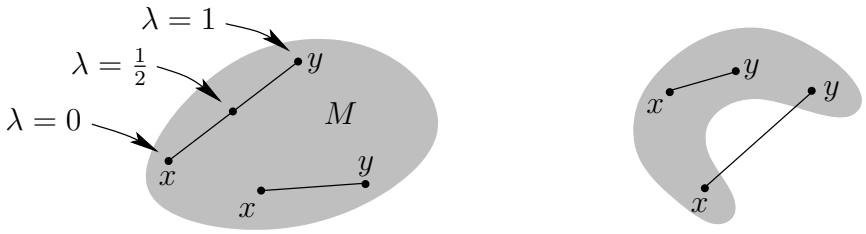
2.1 Konvexität

2.1.1 Definition (Konvexe Mengen und Funktionen)

a) Eine Menge $M \subseteq \mathbb{R}^n$ heißt konvex, *falls*

$$\forall\ x, y \in M,\ \lambda \in (0,1): \quad (1-\lambda)x + \lambda y \in M$$

gilt (d.h. die Verbindungsstrecke von je zwei beliebigen Punkten in M gehört komplett zu M).



M konvex $\qquad$ M nicht konvex

Abbildung 2.1: Konvexität von Mengen

b) Für eine konvexe Menge $M \subseteq \mathbb{R}^n$ heißt eine Funktion $f : M \to \mathbb{R}$ konvex (auf M), *falls*

$$\forall\ x, y \in M,\ \lambda \in (0,1): \quad f(\,(1-\lambda)x + \lambda y\,) \ \leq\ (1-\lambda)f(x) + \lambda f(y)$$

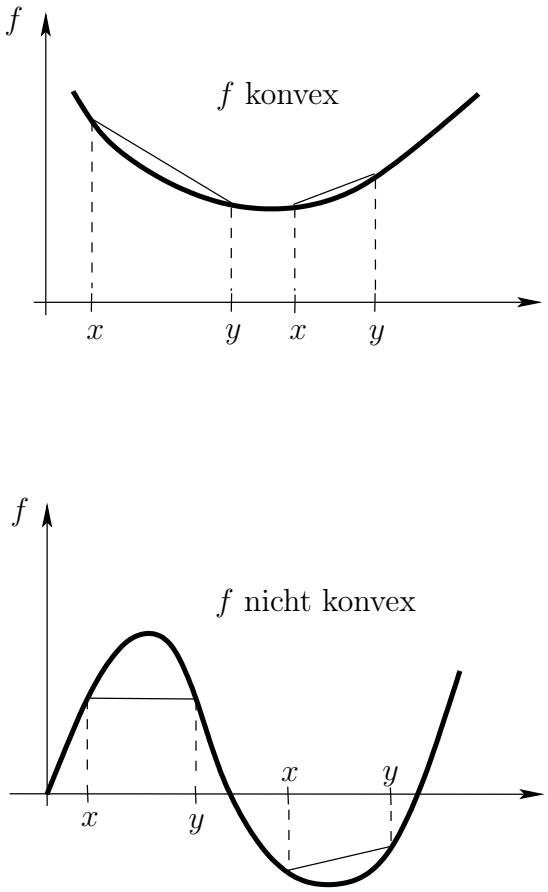


Abbildung 2.2: Konvexität von Funktionen auf $\mathbb{R}$

gilt (d.h. der Funktionsgraph von f verläuft unter *jeder seiner Sekanten).*

c) *Für eine konvexe Menge $M \subseteq \mathbb{R}^n$ heißt eine Funktion $f : M \to \mathbb{R}$* strikt konvex (auf M)*, falls in b) für $x \neq y$ sogar die strikte Ungleichung „$<$“ gilt (d.h. der Funktionsgraph von f verläuft echt unter jeder seiner Sekanten).*

d) *Für eine konvexe Menge $M \subseteq \mathbb{R}^n$ heißt eine Funktion $f : M \to \mathbb{R}$* gleichmäßig konvex (auf M)*, falls mit einer Konstanten $c > 0$ die Funktion $f(x) - \frac{c}{2}\|x\|_2^2$ konvex auf M ist.*

e) *Für eine konvexe Menge $M \subseteq \mathbb{R}^n$ heißt eine Funktion $f : M \to \mathbb{R}$* konkav, strikt konkav *oder* gleichmäßig konkav (auf M) *falls $-f$ konvex, strikt konvex bzw. gleichmäßig konvex auf M ist.*

Beispiele:

- Die Mengen $\emptyset$ und $\mathbb{R}^n$ sind konvex.
- $\{x \in \mathbb{R}^2 |\ x_1 \geq 0\}$ ist konvex, und
- $\{x \in \mathbb{R}^2 |\ x_1^2 + x_2^2 < 1\}$ ist konvex (d.h. konvexe Mengen brauchen weder beschränkt noch abgeschlossen zu sein).
- $f(x) = \sin(x)$ ist konkav auf $M_1 = [0, \pi]$, konvex auf $M_2 = [\pi, 2\pi]$ und weder konvex noch konkav auf $M_3 = [0, 2\pi]$.
- $f(x) = |x|$ ist konvex auf $\mathbb{R}$, und $f(x) = -\sqrt{1 - x^2}$ ist konvex auf $[-1, 1]$ (d.h. konvexe Funktionen brauchen nicht differenzierbar zu sein).
- Jede lineare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist konvex, denn für alle $x, y \in \mathbb{R}^n$ und $\sigma, \mu \in \mathbb{R}$ gilt $f(\sigma x + \mu y) = \sigma f(x) + \mu f(y)$, so dass die spezielle Wahl $\sigma = 1 - \lambda$ und $\mu = \lambda$ die Behauptung sogar mit „=“ statt „≤“ ergibt. Lineare Funktionen sind aber nicht strikt konvex.
- $f(x) = e^x$ ist strikt konvex, aber nicht gleichmäßig konvex auf $\mathbb{R}$ (s.u.).
- $f(x) = (x - 5)^2$ ist gleichmäßig und damit auch strikt konvex $\mathbb{R}$ (s.u.).
- $f(z) = \left\| \begin{pmatrix} \|z - x^1\|_2 \\ \vdots \\ \|z - x^m\|_2 \end{pmatrix} \right\|_2$ ist konvex auf $\mathbb{R}^n$ (s.u.).
- $f(z^1, .., z^k) = \sum_{i=1}^{m} \min_{\ell=1,..,k} \|z^\ell - x^i\|$ mit $k \geq 2$ ist nicht konvex auf $\mathbb{R}^{n \cdot k}$.

2.1.2 Übung *Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ ist die Funktion $f : M \to \mathbb{R}$ genau dann konvex, wenn die Menge* $\operatorname{epi}(f, M)$ *konvex ist.*

2.1.3 Definition (Konvexes Optimierungsproblem)
Das Optimierungsproblem

$$P: \quad \min f(x) \quad s.t. \quad x \in M$$

heißt konvex*, falls M und $f : M \to \mathbb{R}$ konvex sind.*

Da $M = \mathbb{R}^n$ eine konvexe Menge ist, sind unrestringierte Probleme genau dann konvex, wenn f konvex auf $\mathbb{R}^n$ ist.

Der folgende Satz ist von zentraler Bedeutung für konvexe Optimierungsprobleme:

2.1.4 Satz *P sei konvex. Dann ist jeder lokale Minimalpunkt von P auch globaler Minimalpunkt von P.*

Beweis. Der Punkt $\bar{x} \in M$ sei ein lokaler Minimalpunkt von P. Angenom-

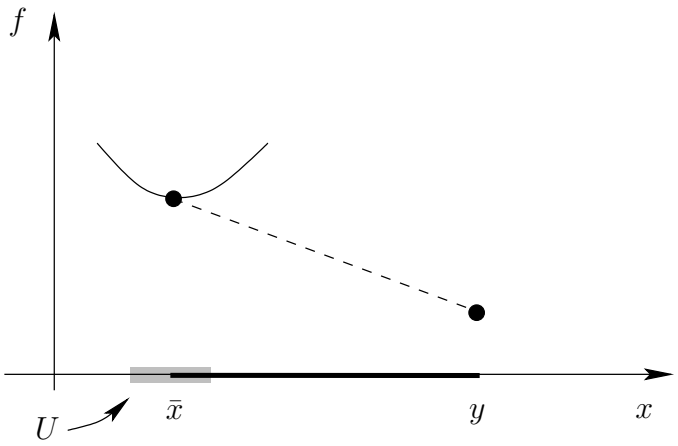


Abbildung 2.3: Beweisidee zu Satz 2.1.4

men, $\bar{x}$ sei nicht globaler Minimalpunkt von P. Dann existiert ein $y \in M$ mit $f(y) < f(\bar{x})$. Die Punkte auf der Verbindungsstrecke von $\bar{x}$ und y, also $x(\lambda) = (1-\lambda)\bar{x} + \lambda y$ mit $\lambda \in (0,1)$, liegen wegen der Konvexität von M sämtlich in M, und wegen der Konvexität von f auf M gilt für alle $\lambda \in (0,1)$

$$f(x(\lambda)) \;\leq\; (1-\lambda)f(\bar{x}) + \lambda \underbrace{f(y)}_{< f(\bar{x})} \;<\; f(\bar{x}).$$

Folglich existiert für jede Umgebung U von $\bar{x}$ ein $\lambda \in (0,1)$ mit $x(\lambda) \in U \cap M$ und $f(x(\lambda)) < f(\bar{x})$. Dies steht aber im Widerspruch dazu, dass $\bar{x}$ ein lokaler Minimalpunkt von P ist. Folglich ist die Annahme falsch, $\bar{x}$ sei kein globaler Minimalpunkt. •

Bei konvexen Optimierungsproblemen genügt es also, nach lokalen Minimalpunkten zu suchen, um globale Minimalpunkte zu finden! Der Grund für diesen Effekt liegt darin, dass Konvexität eine *globale* Voraussetzung an P ist.

In Anwendungen ist die zulässige Menge M häufig nicht abstrakt gegeben, sondern wird durch Gleichungen und Ungleichungen beschrieben. Im Folgenden leiten wir Eigenschaften der beteiligten Funktionen her, die dann die Konvexität von M garantieren.

2.1.5 Übung *Die Menge $M \subseteq \mathbb{R}^n$ und die Funktion $f : M \to \mathbb{R}$ seien konvex. Dann ist $\mathrm{lev}^{\alpha}_{\leq}(f, M)$ für jedes $\alpha \in \overline{\mathbb{R}}$ konvex. Die Umkehrung dieser Aussage ist falsch.*

2.1.6 Übung *Der Schnitt beliebig vieler konvexer Mengen ist konvex.*

2.1.7 Korollar *Die Funktionen $g_i : \mathbb{R}^n \to \mathbb{R},\ i \in I$, mit beliebiger Indexmenge I seien konvex. Dann ist die Menge $M = \{x \in \mathbb{R}^n |\, g_i(x) \leq 0,\, i \in I\}$ konvex.*

Beweis. Aus der Darstellung

$$M \ = \ \bigcap_{i \in I} \ (g_i)^0_{\leq}$$

folgt mit den Übungen 2.1.5 und 2.1.6 die Behauptung. •

Wir klären nun, wann für $h : \mathbb{R}^n \to \mathbb{R}$ die durch eine *Gleichung* beschriebene Menge $H = \{x \in \mathbb{R}^n |\, h(x) = 0\}$ konvex ist. Wegen

$$h(x) = 0 \ \Leftrightarrow \ h(x) \leq 0 \ \text{ und } \ -h(x) \leq 0$$

folgt mit den Übungen 2.1.5 und 2.1.6, dass H konvex ist, falls h und $-h$ konvex sind. Dies ist gleichbedeutend damit, dass h gleichzeitig konvex und konkav ist, also

$$\forall\ x, y \in \mathbb{R}^n,\ \lambda \in (0,1): \quad h(\,(1-\lambda)x + \lambda y\,) \ = \ (1-\lambda)h(x) + \lambda h(y).$$

Folglich ist für jede lineare Funktion h die Menge H konvex.

2.1.8 Korollar *Mit beliebigen Indexmengen I, J seien die Funktionen $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in I$, konvex, und die Funktionen $h_j : \mathbb{R}^n \to \mathbb{R}$, $j \in J$, linear. Dann ist die Menge $M = \{x \in \mathbb{R}^n |\, g_i(x) \leq 0,\, i \in I,\, h_j(x) = 0,\, j \in J\}$ konvex.*

2.1.9 Beispiel *Falls f, g_i, $i \in I$, auf $\mathbb{R}^n$ konvexe und h_j, $j \in J$, lineare Funktionen sind, dann ist*

$$P: \quad \min\ f(x) \quad \ s.t. \quad \ g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J$$

ein konvexes Optimierungsproblem.

2.1.10 Beispiel (Lineare Optimierung)

Mit $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ *und einer* (m,n)*-Matrix* $A = \begin{pmatrix} a_1^\intercal \\ \vdots \\ a_m^\intercal \end{pmatrix}$ *mit* $a_i \in \mathbb{R}^n, i = 1, ..., m$ *ist*

$$P: \ \min c^\intercal x \quad s.t. \quad Ax \leq b$$

ein lineares Optimierungsproblem (die Ungleichungsnebenbedingungen können sowohl eine Nichtnegativitätsbedingung $x \geq 0$ *enthalten als auch Gleichungen modellieren).*

P ist auch ein konvexes *Optimierungsproblem, denn mit den Setzungen* $f(x) = c^\intercal x$, $I = \{1, ..., m\}$ *und* $g_i(x) = a_i^\intercal x - b_i$, $i \in I$, *sind* $f, g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in I$, *linear und damit konvex auf* $\mathbb{R}^n$.

Zum Beispiel setzt man für das lineare Optimierungsproblem

$$\min x_1 + x_2 \quad s.t. \quad x \geq 0$$

$f(x) = x_1 + x_2$, $g_1(x) = -x_1$ *und* $g_2(x) = -x_2$.

2.2 C^1-Charakterisierung von Konvexität

Mehrdimensionale erste Ableitungen

Für eine nicht-leere offene Menge $U \subseteq \mathbb{R}^n$ und $f : U \to \mathbb{R},\ x \mapsto f(x)$ bezeichne $\partial_{x_i} f(\bar{x})$ die partielle Ableitung von f nach x_i an der Stelle $\bar{x} \in U$ (sofern sie existiert).

Beispiel:

Für $U = \mathbb{R}^2$, $f(x) = x_1^2 + x_2$ und $\bar{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ gilt $\partial_{x_1} f \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2$ und $\partial_{x_2} f \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1$.

Als erste Ableitung von f an $\bar{x}$ betrachtet man den *Zeilenvektor*

$$Df(\bar{x}) \ := \ (\, \partial_{x_1} f(\bar{x}), \ldots, \partial_{x_n} f(\bar{x})\,).$$

Sein Transponiertes $\nabla f(\bar{x}) := (Df(\bar{x}))^\intercal$ (Spaltenvektor) heißt auch *Gradient* von f an $\bar{x}$. Im Falle $n = 1$ gilt $Df(\bar{x}) = \nabla f(\bar{x}) = f'(\bar{x})$.

Beispiel:

$$f(x) = x_1^2 + x_2 \ \Rightarrow \ Df \begin{pmatrix} 1 \\ -1 \end{pmatrix} = (2, 1),\ \nabla f \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

f heißt auf U *stetig differenzierbar*, falls ∇f auf U existiert und eine stetige Funktion von x ist. Man schreibt dann kurz $f \in C^1(U, \mathbb{R})$. Für eine nicht notwendigerweise offene Menge $M \subseteq \mathbb{R}^n$ bedeutet die Forderung $f \in C^1(M, \mathbb{R})$, dass es eine offene Menge $U \supseteq M$ mit $f \in C^1(U, \mathbb{R})$ gibt.

Für eine *vektorwertige* Funktion $f : \mathbb{R}^n \to \mathbb{R}^m,\ x \to \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$ definiert man die erste Ableitung an $\bar{x}$ als

$$Df(\bar{x}) \ := \ \begin{pmatrix} Df_1(\bar{x}) \\ \vdots \\ Df_m(\bar{x}) \end{pmatrix}.$$

Dies ist eine (m, n)-Matrix, die *Jacobimatrix* oder *Funktionalmatrix* von f an $\bar{x}$ heißt.

Beispiel:

$$f(x) = \begin{pmatrix} x_1^2 + x_2 \\ x_1 - x_2^2 \\ x_1 x_2 \end{pmatrix} \Rightarrow Df(x) = \begin{pmatrix} 2x_1 & 1 \\ 1 & -2x_2 \\ x_2 & x_1 \end{pmatrix} \Rightarrow Df \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \end{pmatrix}.$$

Eine wichtige Rechenregel für differenzierbare Funktionen ist die *Kettenregel*:

Es seien $g : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar in $\bar{x} \in \mathbb{R}^n$ und $f : \mathbb{R}^m \to \mathbb{R}^k$ differenzierbar in $g(\bar{x}) \in \mathbb{R}^m$. Dann ist $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ differenzierbar in $\bar{x}$ mit

$$D(f \circ g)(\bar{x}) \;=\; Df(g(\bar{x})) \cdot Dg(\bar{x}).$$

Ein wesentlicher Grund dafür, die Jacobimatrix einer Funktion wie oben zu definieren, besteht darin, dass die Kettenregel dann völlig analog zum eindimensionalen Fall ($n = m = k = 1$) formuliert werden kann, obwohl das auftretende Produkt ein Matrixprodukt ist.

Der folgende Satz wird in den Grundlagen der Analysis bewiesen.

2.2.1 Satz (Linearisierung per Satz von Taylor im $\mathbb{R}^n$)
Für eine nicht-leere, offene und konvexe Menge $U \subseteq \mathbb{R}^n$ sei die Funktion $f : U \to \mathbb{R}$ differenzierbar an $x \in U$. Dann gilt für alle $y \in U$

$$f(y) \;=\; f(x) + \langle \nabla f(x), y - x \rangle + o(\|y - x\|),$$

wobei $o(\|y-x\|)$ einen Ausdruck der Form $\omega(y)\|y-x\|$ mit einer an x stetigen Funktion ω und $\omega(x) = 0$ bezeichnet.

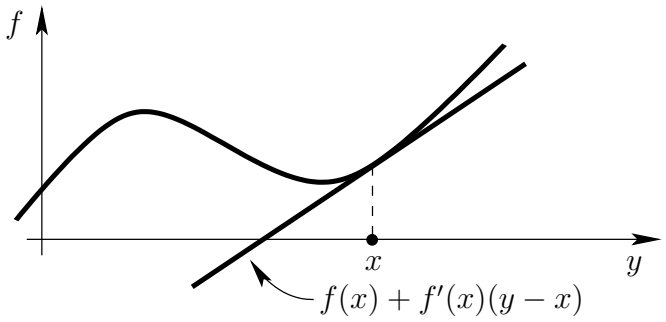


Abbildung 2.4: Lineare Approximation von f um x für $n = 1$

C^1-Charakterisierung

2.2.2 Satz (C^1-Charakterisierung von Konvexität)
Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ ist eine Funktion $f \in C^1(M, \mathbb{R})$ genau dann konvex, wenn

$$\forall\, x, y \in M: \quad f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle$$

gilt.

Beweis. f sei konvex auf M, und U sei eine konvexe offene Obermenge von von M, auf der f stetig differenzierbar ist. Dann gilt nach Satz 2.2.1 (mit $x + \lambda(y - x)$ in der Rolle von y) für alle $x, y \in U$ und $\lambda \in (0, 1)$

$$\begin{aligned}(1-\lambda)f(x) + \lambda f(y) \;&\geq\; f((1-\lambda)x + \lambda y) \;=\; f(x + \lambda(y - x)) \\ &=\; f(x) + \lambda \langle \nabla f(x), y - x \rangle + o(\lambda \|y - x\|),\end{aligned}$$

wobei $o(\lambda\|y - x\|)$ einen Ausdruck der Form $\omega(x + \lambda(y - x))\lambda\|y - x\|$ mit einer an x stetigen Funktion ω und $\omega(x) = 0$ bezeichnet. Nach Umstellung und Division durch λ folgt daraus

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle + \omega(x + \lambda(y - x))\|y - x\|.$$

Der Grenzübergang $\lambda \to 0$ liefert wegen der Stetigkeit von ω an x und $\omega(x) = 0$ die gewünschte Ungleichung für alle $x, y \in U$ und damit auch für alle $x, y \in M$.

Es seien andererseits $x, y \in M$, $\lambda \in (0, 1)$ und $z := (1 - \lambda)x + \lambda y$. Dann gelten die beiden Ungleichungen

$$\begin{aligned}f(x) \;&\geq\; f(z) + \langle \nabla f(z), x - z \rangle\,, \\ f(y) \;&\geq\; f(z) + \langle \nabla f(z), y - z \rangle\,.\end{aligned}$$

Deren Konvexkombination liefert

$$\begin{aligned}(1-\lambda)f(x) + \lambda f(y) \;&\geq\; f(z) + \langle \nabla f(z), \underbrace{(1-\lambda)(x - z) + \lambda(y - z)}_{=\,(1-\lambda)x + \lambda y - z \,=\, 0} \rangle \\ &=\; f(z) \;=\; f((1-\lambda)x + \lambda y)\end{aligned}$$

und damit die Konvexität von f. •

Die C^1-Charakterisierung besagt, dass eine C^1-Funktion genau dann konvex auf M ist, wenn ihr Graph *über* jeder seiner Tangentialebenen verläuft (s. Abb. 2.5).

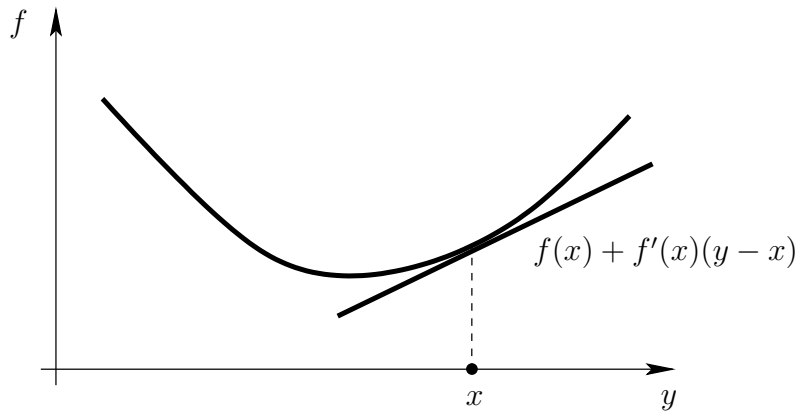


Abbildung 2.5: C^1-Charakterisierung von Konvexität für $n = 1$

Eine gründliche Durchsicht des Beweises zu Satz 2.2.2 zeigt, dass die *Stetigkeit* der ersten Ableitung von f nirgends erforderlich ist. Es reicht also auch, eine auf M differenzierbare Funktion f vorauszusetzen. Die meisten in Anwendungen auftretenden differenzierbaren Funktionen sind allerdings auch gleichzeitig stetig differenzierbar, so dass sich die C^1-Voraussetzung eingebürgert hat.

C^1-Charakterisierungen von *strikter* und *gleichmäßiger* Konvexität sind ebenfalls bekannt (s. z.B. [5, Theorem 4.1.1]).

2.3 Lösbarkeit

Für die Lösbarkeit von P ist Konvexität alleine kein wesentlicher Vorteil: sowohl $f_1(x) = (x-5)^2$ als auch $f_2(x) = e^x$ sind sogar strikt konvex, aber nur f_1 besitzt einen globalen Minimalpunkt auf $\mathbb{R}$. Im Folgenden werden wir sehen, dass der entscheidende Vorteil von f_1 gegenüber f_2 die *gleichmäßige* Konvexität ist.

2.3.1 Übung *Für eine konvexe Menge $M \subseteq \mathbb{R}^n$ sei $f : M \to \mathbb{R}$ gleichmäßig konvex. Dann ist f auch strikt konvex auf M.*

2.3.2 Lemma *Für eine abgeschlossene und konvexe Menge $M \subseteq \mathbb{R}^n$ sei $f : M \to \mathbb{R}$ gleichmäßig konvex. Dann ist f auch*

a) koerziv auf M und

b) stetig auf dem Inneren von M.

Beweis. Wir geben die grundlegende Beweisidee von Teil a) unter der zusätzlichen Annahme der stetigen Differenzierbarkeit von f auf M an.

Für $M = \emptyset$ ist nichts zu zeigen. Ansonsten setzen wir im Folgenden ohne Beschränkung der Allgemeinheit $0 \in M$ voraus (ansonsten ersetze mit einem $\bar{x} \in M$ die Variable x durch $y = x - \bar{x}$ und argumentiere, dass die Behauptung unabhängig von dieser Koordinatentransformation gilt).

Aufgrund der gleichmäßigen Konvexität von f auf M gilt zunächst für ein $c > 0$

$$f(x) \;=\; F(x) + \frac{c}{2}\|x\|_2^2$$

mit der auf M konvexen Funktion $F(x) = f(x) - \frac{c}{2}\|x\|_2^2$. Damit hieraus $\lim_{\|x\|_2 \to \infty} f(x) = +\infty$ folgen kann, darf $F(x)$ für $\|x\|_2 \to \infty$ nicht „zu schnell“ gegen $-\infty$ streben. Um dies zu sehen, dürfen wir die C^1-Charakterisierung von Konvexität aus Satz 2.2.2 benutzen, da mit f auch F stetig differenzierbar auf M ist:

$$\forall\, x \in M: \quad F(x) \;\geq\; F(0) + \langle \nabla F(0), x\rangle \;\geq\; f(0) - \|\nabla f(0)\|_2 \cdot \|x\|_2\,,$$

wobei die zweite Abschätzung aus der Cauchy-Schwarz-Ungleichung und der Definition von F folgt. Sie bedeutet, dass F für $\|x\|_2 \to \infty$ „höchstens mit linearer Geschwindigkeit“ gegen $-\infty$ fallen kann.

Insgesamt erhalten wir

$$\forall\, x \in M: \quad f(x) \;=\; F(x) + \frac{c}{2}\|x\|_2^2 \;\geq\; f(0) - \|\nabla f(0)\|_2 \cdot \|x\|_2 + \frac{c}{2}\|x\|_2^2\,,$$

woraus die Koerzivität von f auf M folgt.

Falls die stetige Differenzierbarkeit von f nicht vorliegt, verläuft der Beweis völlig analog durch Betrachtung eines Elements des sogenannten *Subdifferentials* $\partial F(0)$ anstelle der C^1-Charakterisierung von F, auf dessen Einführung wir im Rahmen dieser Vorlesung allerdings verzichten und stattdessen auf [1,5,12] verweisen.

Ebendort wird auch gezeigt, dass jede auf M konvexe Funktion auf dem Inneren von M stetig ist. Nach Übung 2.3.1 ist die gleichmäßig konvexe Funktion f auch strikt konvex und damit insbesondere konvex auf M, so dass die Behauptung b) folgt. •

2.3.3 Satz *P sei konvex. Dann gelten die folgenden Aussagen:*

a) Die Menge der Minimalpunkte S ist konvex.

b) Falls f strikt konvex auf M ist, dann besitzt S höchstens ein Element.

c) Es sei M nicht-leer und abgeschlossen, und f sei gleichmäßig konvex und stetig auf M. Dann besitzt S genau ein Element (d.h. P ist eindeutig lösbar).

Beweis.

1. Fall: $S = \emptyset$ (z.B. $f(x) = e^x, M = \mathbb{R}$).

Dann gelten die Aussagen von Teil a) und b) trivialerweise.

2. Fall: $S \neq \emptyset$

Dann gibt es ein $\bar{x} \in S$, und mit $v = f(\bar{x})$ gilt $S = \mathrm{lev}^v_{\leq}(f, M)$. Mit den Übungen 2.1.5 und 2.1.6 folgt daraus die Behauptung von Teil a).

Um Teil b) zu beweisen, nehmen wir an, es existiere ein Punkt $\widetilde{x} \in S \setminus \{\bar{x}\}$. Dann gilt $f(\bar{x}) = f(\widetilde{x}) = v$, und nach Teil a) auch $f(\frac{1}{2}\bar{x} + \frac{1}{2}\widetilde{x}) = v$. Aus der strikten Konvexität von f auf M folgt damit der Widerspruch

$$v \;=\; f\left(\tfrac{1}{2}\bar{x} + \tfrac{1}{2}\widetilde{x}\right) \;<\; \tfrac{1}{2}f(\bar{x}) + \tfrac{1}{2}f(\widetilde{x}) \;=\; v.$$

Für den Beweis von Teil c) folgt aus Übung 2.3.1 und Teil b) sofort, dass S *höchstens* ein Element enthält. Nach Lemma 2.3.2a) ist f außerdem koerziv auf M. Nach Korollar 1.2.30 enthält S also auch *mindestens* ein Element. •

Angemerkt sei, dass die Stetigkeitsforderung an f in Satz 2.3.3c) nach Lemma 2.3.2b) unnötig ist, falls die Menge M mit ihrem Inneren übereinstimmt (z.B. für $M = \mathbb{R}^n$).

2.4 Optimalitätsbedingungen für unrestringierte Probleme

Wir betrachten im Folgenden das unrestringierte Problem

$$P: \quad \min f(x).$$

Allgemeiner bezeichnet man auch Probleme mit *offener* zulässiger Menge M als unrestringiert. In der Tat lassen sich die im Folgenden besprochenen Optimalitätsbedingungen ohne weiteres auf diesen Fall übertragen, was wir aus Gründen der Übersichtlichkeit aber nicht explizit angeben werden.

2.4.1 Definition (Kritischer Punkt)
Ein Punkt $\bar{x} \in \mathbb{R}^n$ heißt kritisch *für $f \in C^1(\mathbb{R}^n, \mathbb{R})$, falls $\nabla f(\bar{x}) = 0$ gilt.*

Der folgende grundlegende Satz gilt ohne Konvexitätsannahme und wird z.B. in [13] bewiesen.

2.4.2 Satz (Fermat'sche Regel, notwendige Optimalitätsbedingung)
Der Punkt $\bar{x} \in \mathbb{R}^n$ sei lokal minimal für $f \in C^1(\mathbb{R}^n, \mathbb{R})$. Dann ist $\bar{x}$ kritischer Punkt von f.

Beispiele:

$\bar{x} = (0,0)$ ist lokaler Minimalpunkt und kritischer Punkt von $f_1(x) = x_1^2 + x_2^2$.

Allerdings ist $\bar{x} = (0,0)$ kein lokaler Minimalpunkt, aber trotzdem kritischer Punkt von $f_2(x) = -x_1^2 - x_2^2$ und auch von $f_3(x) = x_1^2 - x_2^2$. Folglich ist nicht jeder kritische Punkt notwendigerweise lokaler Minimalpunkt von f. Da analog auch nicht jeder lokale Minimalpunkt von f gleichzeitig *globaler* Minimalpunkt ist, liefert Satz 2.4.2 das Mengendiagramm in Abbildung 2.6.

In den folgenden beiden Beispielen genügen diese Zusammenhänge bereits, um globale Minimalpunkte zu bestimmen.

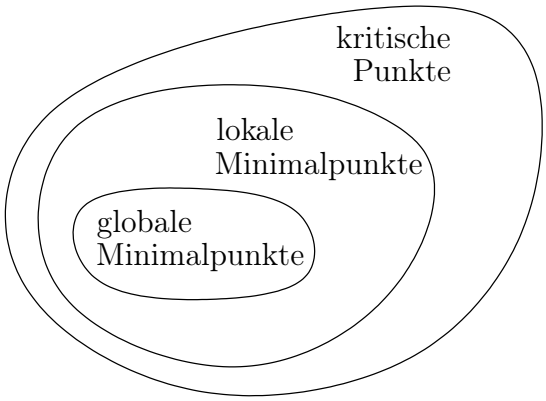


Abbildung 2.6: Notwendige Optimalitätsbedingung

2.4.3 Beispiel (Beispiel 1.1.5 - Fortsetzung 3)
Für die Funktion f aus Beispiel 1.1.5 gilt

$$f(z) = \left\| \begin{pmatrix} \|z-x^1\|_2 \\ \vdots \\ \|z-x^m\|_2 \end{pmatrix} \right\|_2 = \sqrt{\sum_{i=1}^m \|z-x^i\|_2^2}$$
$$= \sqrt{\sum_{i=1}^m \underbrace{(z-x^i)^\intercal (z-x^i)}_{z^\intercal z - 2z^\intercal x^i + (x^i)^\intercal x^i}} = \sqrt{mz^\intercal z - 2z^\intercal \sum_{i=1}^m x^i + \sum_{i=1}^m \|x^i\|_2^2},$$

so dass f nicht differenzierbar ist und Satz 2.4.2 nicht angewendet werden kann. Nach Übung 1.3.5 mit $\psi(y) = y^2$ und $Y = \{y \in \mathbb{R} | \, y \geq 0\}$ besitzt f aber dieselben Minimalpunkte wie die stetig differenzierbare Funktion

$$f^2(z) = mz^\intercal z - 2z^\intercal \sum_{i=1}^m x^i + \sum_{i=1}^m \|x^i\|_2^2.$$

Kritische Punkte dieser Funktion sind genau die Lösungen der Gleichung

$$0 = \nabla(\, f^2(z)\,) = 2mz - 2\sum_{i=1}^m x^i,$$

also besitzt f^2 den eindeutigen kritischen Punkt

$$\bar{z} = \frac{1}{m}\sum_{i=1}^m x^i.$$

Tatsächlich ist $\bar{z}$ auch globaler Minimalpunkt von f^2 sowie von f, wie man durch folgende Argumentation sieht: nach Beispiel 1.2.31 existiert *zunächst ein globaler Minimalpunkt $\widetilde{z}$ von f und damit auch von f^2 auf $\mathbb{R}^n$. Aufgrund der Fermat'schen Regel (S. 2.4.2) muss $\widetilde{z}$ kritischer Punkt von f^2 sein. Der* einzige *kritische Punkt von f^2 ist aber das gerade berechnete $\bar{z}$, so dass nur $\widetilde{z} = \bar{z}$ gelten kann. Also ist $\bar{z}$ globaler Minimalpunkt von f auf $\mathbb{R}^n$.*

In Kapitel 2.5 werden wir außerdem nachweisen, dass f^2 eine auf $\mathbb{R}^n$ konvexe Funktion ist, was es ermöglichen wird, die globale Minimalität von $\bar{z}$ alternativ zu beweisen, ohne *zunächst die Lösbarkeit des zugrundeliegenden Optimierungsproblems zu zeigen.*

2.4.4 Beispiel (Beispiel 1.2.36 - Fortsetzung 3)
Die Zielfunktion $f(\lambda) = \lambda\bar{x} - \log(\lambda)$ mit $\bar{x} > 0$ des Problems P_{ML} aus Beispiel 1.2.36 erfüllt $f'(\lambda) = \bar{x} - \frac{1}{\lambda}$, besitzt als eindeutigen kritischen Punkt also $\bar{\lambda} = \frac{1}{\bar{x}}$.

Wie in Beispiel 2.4.3 lässt sich nun argumentieren, dass $\bar{\lambda} = \frac{1}{\bar{x}}$ globaler Minimalpunkt von P_{ML} und damit der gesuchte Maximum-Likelihood-Schätzer der Exponentialverteilung ist: nach Beispiel 1.2.41 existiert *zunächst ein globaler Minimalpunkt $\widetilde{\lambda}$ von P_{ML}. Da sich Optimierungsprobleme mit* offenen *zulässigen Mengen wie* unrestringierte *Optimierungsprobleme behandeln lassen, muss $\widetilde{\lambda}$ nach der Fermat'schen Regel ein kritischer Punkt von f sein.* Einziger *kritischer Punkt von f ist aber das gerade berechnete $\bar{\lambda}$, woraus die Behauptung folgt.*

In Kapitel 2.5 werden wir außerdem sehen, dass f auf der zulässigen Menge $(0, +\infty)$ von P_{ML} konvex ist, was es wieder ermöglichen wird, die globale Minimalität von $\bar{\lambda}$ alternativ zu beweisen, ohne *zunächst die Lösbarkeit von P_{ML} zu zeigen.*

Die in den vorausgegangenen Beispielen angedeutete alternative Beweismöglichkeit basiert darauf, dass der Zusammenhang zwischen kritischen Punkten und globalen Minimalpunkten erheblich übersichtlicher wird, falls $f \in C^1(\mathbb{R}^n, \mathbb{R})$ zusätzlich *konvex* ist.

2.4.5 Satz (Hinreichende Optimalitätsbedingung)
$f \in C^1(\mathbb{R}^n, \mathbb{R})$ sei konvex. Dann ist jeder kritische Punkt von f globaler Minimalpunkt von f.

Beweis. Der Punkt x sei kritisch für f, d.h. es gelte $\nabla f(x) = 0$. Mit Satz 2.2.2 folgt

$$\forall\, y \in \mathbb{R}^n : \quad f(y) \;\geq\; f(x) + \langle \underbrace{\nabla f(x)}_{=0}, y - x \rangle \;=\; f(x).$$

•

Abbildung 2.7 zeigt die in Satz 2.4.5 bewiesene Relation zwischen kritischen Punkten und globalen Minimalpunkten in einem Mengendiagramm.

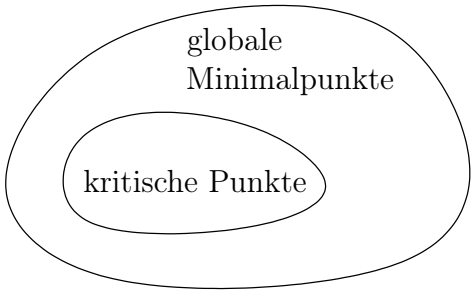


Abbildung 2.7: Hinreichende Optimalitätsbedingung

2.4.6 Korollar (Charakterisierung globaler Minimalpunkte)
$f \in C^1(\mathbb{R}^n, \mathbb{R})$ sei konvex. Dann sind die globalen Minimalpunkte genau die kritischen Punkte von f.

Beweis. Sätze 2.4.2 und 2.4.5. •

Zur Bestimmung globaler Minimalpunkte unrestringierter konvexer C^1-Probleme genügt es also nicht nur, lokale Minimalpunkte zu suchen (wie schon in S. 2.1.4 gesehen), sondern sogar nur kritische Punkte. Das globale Minimierungsproblem ist damit auf das Lösen der Gleichung $\nabla f(x) = 0$ zurückgeführt.

Insbesondere erhält man auch diese Aussage: falls f keinen kritischen Punkt besitzt, dann auch keinen globalen Minimalpunkt. Dazu muss allerdings *bewiesen* werden, dass f keinen kritischen Punkt besitzen kann (ein einfaches Beispiel hierfür ist $f(x) = e^x$).

2.5 C^2-Charakterisierung von Konvexität

Mehrdimensionale zweite Ableitung

Für eine nicht-leere offene Menge $U \subseteq \mathbb{R}^n$ und $f : U \to \mathbb{R}$ definiert man die zweite Ableitung als erste Ableitung des Gradienten (sofern sie existiert):

$$D^2 f(x) \;:=\; D\nabla f(x) = \begin{pmatrix} \partial_{x_1}\partial_{x_1} f(x) & \cdots & \partial_{x_n}\partial_{x_1} f(x) \\ \vdots & & \vdots \\ \partial_{x_1}\partial_{x_n} f(x) & \cdots & \partial_{x_n}\partial_{x_n} f(x) \end{pmatrix}.$$

Beispiel:

$$f(x) = x_1^2 + x_2 \;\Rightarrow\; D^2 f(x) = D\begin{pmatrix} 2x_1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

$D^2 f(\bar{x})$ heißt *Hessematrix* von f an $\bar{x}$ und ist stets eine (n,n)-Matrix. Falls alle Einträge von $D^2 f$ stetige Funktionen von x sind, nennt man f auf U *zweimal stetig differenzierbar*, kurz $f \in C^2(U, \mathbb{R})$. In diesem Fall ist $D^2 f(x)$ für $x \in U$ sogar symmetrisch (nach dem Satz von Schwarz). Die Forderung $f \in C^2(M, \mathbb{R})$ mit einer beliebigen Menge $M \subseteq \mathbb{R}^n$ bedeutet wieder, dass es eine offene Menge $U \supseteq M$ mit $f \in C^2(U, \mathbb{R})$ gibt. Für $n = 1$ gilt $D^2 f(\bar{x}) = f''(\bar{x})$.

Auch der nächste Satz wird in den Grundlagen der Analysis gezeigt.

2.5.1 Satz (Quadratische Approximation per Satz von Taylor im $\mathbb{R}^n$)
Für eine nicht-leere, offene und konvexe Menge $U \subseteq \mathbb{R}^n$ sei die Funktion $f : U \to \mathbb{R}$ zweimal differenzierbar an $x \in U$. Dann gilt für alle $y \in U$

$$f(y) \;=\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2}(y-x)^\intercal D^2 f(x)(y-x) + o(\|y-x\|^2).$$

Der Fehlerterm lässt sich dabei auch mit Hilfe des Lagrange-Restglieds *angeben:*

$$o(\|y-x\|^2) \;=\; \tfrac{1}{2}(y-x)^\intercal D^2 f(\xi)(y-x) - \tfrac{1}{2}(y-x)^\intercal D^2 f(x)(y-x),$$

wobei ξ ein nicht näher bekannter Punkt auf der Verbindungsstrecke zwischen x und y ist.

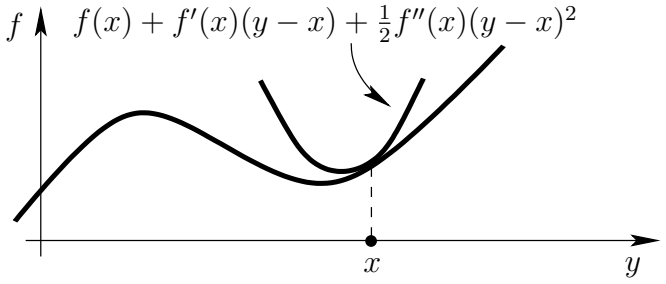


Abbildung 2.8: Quadratische Approximation von f um x für $n = 1$

Eine symmetrische (n, n)-Matrix A heißt *positiv semidefinit* (kurz: $A \succeq 0$), wenn

$$\forall\, x \in \mathbb{R}^n : \quad x^\intercal A x \;\geq\; 0$$

gilt, und *positiv definit* (kurz: $A \succ 0$), wenn die Ungleichung für alle $x \neq 0$ sogar strikt ist.

Positive (Semi-)Definitheit einer Matrix ist über diese Definitionen oft nur schwer überprüfbar. In der Linearen Algebra wird allerdings gezeigt, dass $A \succeq 0$ ($\succ 0$) genau dann gilt, wenn $\lambda \geq 0$ (> 0) für alle *Eigenwerte* λ von A gilt. Dies ist erheblich leichter zu überprüfen. Für $n = 1$ macht man sich außerdem leicht $A \succeq 0$ ($\succ 0$) $\Leftrightarrow$ $A \geq 0$ (> 0) klar.

Mit Hilfe von Satz 2.5.1 lässt sich ohne Konvexitätsannahmen folgende Optimalitätsbedingung zeigen (siehe [13]):

2.5.2 Satz (Notwendige Optimalitätsbedingung zweiter Ordnung)
Der Punkt $\bar{x}$ sei lokaler Minimalpunkt von $f : \mathbb{R}^n \to \mathbb{R}$, und f sei an $\bar{x}$ zweimal differenzierbar. Dann gilt $\nabla f(\bar{x}) = 0$ und $D^2 f(\bar{x}) \succeq 0$.

C^2-Charakterisierungen

2.5.3 Satz (C^2-Charakterisierung von Konvexität)
Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ sei die Funktion $f \in C^2(M, \mathbb{R})$ gegeben.

a) Falls

$$\forall\, x \in M: \quad D^2 f(x) \succeq 0$$

gilt, dann ist f auf M konvex.

b) Falls M außerdem offen ist, dann gilt auch die Umkehrung der Aussage in Teil a).

Beweis. Um Teil a) zu zeigen, wählen wir zunächst eine konvexe offene Obermenge U von M, auf der f zweimal stetig differenzierbar ist. Nach Satz 2.5.1 gilt dann insbesondere für alle $x, y \in M \subseteq U$ mit einem ξ auf der Strecke zwischen x und y

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2}(y-x)^\intercal D^2 f(\xi)(y-x) \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle \ , \end{aligned}$$

wobei wir $D^2 f(\xi) \succeq 0$ benutzt haben. Nach Satz 2.2.2 ist f damit konvex auf M.

Zum Beweis von Teil b) wähle ein $\bar{x} \in M$. Mit f ist auch die Funktion $F(x) := f(x) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle$ konvex auf M, in der $\bar{x}$ als fester Parameter zu interpretieren ist. Wegen $\nabla F(\bar{x}) = \nabla f(\bar{x}) - \nabla f(\bar{x}) = 0$ ist $\bar{x}$ kritischer Punkt von F. Da M offen ist, gelten für die Minimierung von F über M wie erwähnt analoge Aussagen wie für die *unrestringierte* Minimierung von F. Insbesondere ist $\bar{x}$ nach Korollar 2.4.6 ein globaler Minimalpunkt von F auf M. Satz 2.5.2 garantiert daher

$$0 \preceq D^2 F(\bar{x}) = D^2 f(\bar{x}).$$

•

Dass die Voraussetzung der Offenheit von M in Satz 2.5.3b) nicht nur beweistechnischer Natur ist, sieht man an der C^2-Funktion $f(x) = x_1^2 - x_2^2$, die nirgends eine positiv semidefinite Hessematrix besitzt, aber trotzdem auf der Menge $M = \mathbb{R} \times \{0\}$ konvex ist. M ist in diesem Beispiel natürlich nicht offen.

Die Voraussetzung der Offenheit von M in Satz 2.5.3b) lässt sich zur *Volldimensionalität* von M abschwächen, also im Wesentlichen zur Forderung,

dass M innere Punkte besitzt (s. [15]). Eine weitergehende Abschwächung ist nicht möglich.

2.5.4 Beispiel *Die Funktion $f(x) = (x-5)^2$ erfüllt $f''(x) = 2 \geq 0$ für alle $x \in \mathbb{R}$ und ist damit konvex auf $\mathbb{R}$.*

2.5.5 Beispiel *Die Funktion $f(x) = e^x$ erfüllt $f''(x) = e^x \geq 0$ für alle $x \in \mathbb{R}$ und ist damit konvex auf $\mathbb{R}$.*

2.5.6 Beispiel (Beispiel 1.1.5 - Fortsetzung 4)
Für den Gradienten der Funktion f^2 aus Beispiel 2.4.3 haben wir bereits die Darstellung

$$\nabla(\, f^2(z)\,) \;=\; 2mz - 2\sum_{i=1}^{m} x^i$$

hergeleitet, woraus

$$D^2(\, f^2(z)\,) \;=\; 2mE \quad \textit{(mit der } (n,n)\textit{-Einheitsmatrix } E\textit{)}$$

folgt. Damit besitzt $D^2 f^2(z)$ den n-fachen Eigenwert $2m \geq 0$. Hieraus folgt die Konvexität von f^2 auf $\mathbb{R}^n$.

Nach Korollar 2.4.6 ist jeder kritische Punkt von f^2 globaler Minimalpunkt von f^2 und damit auch von f, und der (einzige) kritische Punkt von f^2 ist das in Beispiel 2.4.3 berechnete arithmetische Mittel $\bar{z}$ der Punktewolke. Damit ist alternativ zur Argumentation in Beispiel 2.4.3 die globale Minimalität von $\bar{z}$ gezeigt, ohne zuvor die Lösbarkeit des Optimierungsproblems nachzuweisen.

2.5.7 Beispiel (Beispiel 1.2.36 - Fortsetzung 4)
Die Funktion $f(\lambda) = \lambda\bar{x} - \log(\lambda)$ mit $\bar{x} > 0$ aus Beispiel 2.4.4 erfüllt $f'(\lambda) = \bar{x} - \frac{1}{\lambda}$ und $f''(\lambda) = \frac{1}{\lambda^2} \geq 0$ für alle $\lambda \in (0, +\infty)$. Sie ist damit konvex auf $(0, +\infty)$. Die Konvexität von f korrespondiert zur Konkavität der Log-Likelihood-Funktion ℓ. Wir bemerken, dass die Likelihood-Funktion L selbst nicht *konkav ist.*

Der in Beispiel 2.4.4 berechnete kritische Punkt $\bar{\lambda} = \frac{1}{\bar{x}}$ von f ist nach Korollar 2.4.6 wieder globaler Minimalpunkt von P_{ML} und damit der gesuchte Maximum-Likelihood-Schätzer der Exponentialverteilung. Auch bei diesem Argument ist (alternativ zu dem in Beispiel 2.4.4) die Betrachtung der Lösbarkeit von P_{ML} unnötig.

2.5.8 Übung (Beispiel 1.1.6 - Fortsetzung 3)
Die Zielfunktion

$$f(z^1,\dots,z^k) \;=\; \sum_{i=1}^{m} \min_{\ell=1,\dots,k} \|z^\ell - x^i\| \;=\; \left\| \begin{pmatrix} \|z^{\ell(1)} - x^1\| \\ \vdots \\ \|z^{\ell(m)} - x^m\| \end{pmatrix} \right\|_1$$

mit $k \geq 2$ aus dem Problem P_1 der Clusteranalyse (Bsp. 1.1.6) ist nicht konvex auf $\mathbb{R}^{nk}$ (wenn die Indizes $\ell(i)$, $i = 1,\dots,m$, allerdings a priori bekannt sind, dann lässt sich mit einem anderen Argument als der Benutzung von S. 2.5.3 zeigen, dass f doch konvex ist).

2.5.9 Übung (Hinreichende Bedingung für strikte Konvexität)
Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ sei die Funktion $f \in C^2(M,\mathbb{R})$ gegeben, und es gelte

$$\forall\, x \in M: \quad D^2 f(x) \;\succ\; 0.$$

Dann ist f auf M strikt konvex.

Die Umkehrung der Aussage in Übung 2.5.9 ist falsch, wie das Beispiel der auf $M = \mathbb{R}$ strikt konvexen Funktion $f(x) = x^4$ mit $f''(0) = 0$ zeigt. Eine C^2-Charakterisierung von strikter Konvexität wird in [15] angegeben.

Im Folgenden bezeichne $\lambda_{\min}(A)$ den kleinsten Eigenwert einer symmetrischen Matrix A. Insbesondere gilt damit $D^2 f(x) \succeq 0$ ($\succ 0$) genau für $\lambda_{\min}(D^2 f(x)) \geq 0$ (> 0).

2.5.10 Satz (C^2-Charakterisierung von gleichmäßiger Konvexität)
Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ sei die Funktion $f \in C^2(M,\mathbb{R})$ gegeben.

a) Falls mit einer Konstanten $c > 0$

$$\forall\, x \in M: \quad \lambda_{\min}(D^2 f(x)) \;\geq\; c$$

gilt, dann ist f auf M gleichmäßig konvex.

b) Falls M außerdem offen ist, dann gilt auch die Umkehrung der Aussage in Teil a).

Beweis. Um Teil a) zu zeigen, konstruieren wir mit Hilfe der Konstanten c die Funktion $F(x) = f(x) - \frac{c}{2}\|x\|_2^2$ und zeigen deren Konvexität auf M. Per Definition ist f dann gleichmäßig konvex auf M.

Wegen $\|x\|_2^2 = x^\intercal x$ ist die Funktion F ebenfalls zweimal stetig differenzierbar auf M, so dass ihre Konvexität mit Hilfe von Satz 2.5.3a) nachgewiesen werden kann. In der Tat gilt für alle $x \in M$

$$D^2F(x) \;=\; D^2f(x) - cE,$$

wobei E die (n,n)-Einheitsmatrix bezeichnet. Jeder Eigenwert λ von $D^2f(x)$ erfüllt bekanntlich die Gleichung $\det(D^2f(x) - \lambda E) = 0$, so dass wir

$$\begin{aligned} 0 \;&=\; \det(D^2f(x) - \lambda E) \;=\; \det(\,(D^2f(x) - cE) - (\lambda - c)E\,) \\ &=\; \det(\,D^2F(x) - (\lambda - c)E\,) \end{aligned}$$

erhalten. Dies bedeutet, dass sich jeder Eigenwert von $D^2F(x)$ in der Form $\lambda - c$ mit einem Eigenwert λ von $D^2f(x)$ schreiben lässt. Insbesondere gilt $\lambda_{\min}(D^2F(x)) = \lambda_{\min}(D^2f(x)) - c$. Da der letzte Ausdruck nach Voraussetzung für alle $x \in M$ nicht-negativ ist, folgt die Konvexität von F auf M.

Beweis von Teil b): Übung. •

Die Voraussetzung der Offenheit von M in Satz 2.5.10b) lässt sich wieder zur *Volldimensionalität* von M abschwächen (s. [15]), aber nicht weiter.

2.5.11 Beispiel *Die Funktion $f(x) = (x-5)^2$ erfüllt $f''(x) = 2 > 0$ für alle $x \in \mathbb{R}$ und ist damit nicht nur strikt, sondern sogar gleichmäßig konvex auf $\mathbb{R}$.*

2.5.12 Beispiel *Die Funktion $f(x) = e^x$ erfüllt $f''(x) = e^x > 0$ für alle $x \in \mathbb{R}$ und ist damit strikt konvex auf $\mathbb{R}$. Allerdings gilt $\lim_{x\to-\infty} f''(x) = 0$, so dass sie nicht gleichmäßig konvex auf $\mathbb{R}$ ist.*

2.5.13 Beispiel (Beispiel 1.1.5 - Fortsetzung 5)
Für die Hessematrix des Quadrats der Funktion f aus Beispiel 1.1.5 haben wir in Beispiel 2.5.6 die Darstellung $D^2f^2(z) = 2mE$ und damit den n-fachen Eigenwert $2m$ für jedes $z \in \mathbb{R}^n$ hergeleitet. Insbesondere gilt dann $\lambda_{\min}(D^2f^2(z)) = 2m > 0$ für alle $z \in \mathbb{R}^n$, woraus die strikte und sogar gleichmäßige Konvexität von f^2 auf $\mathbb{R}^n$ folgt.

Auch ohne die bereits früher erfolgte Berechnung des globalen Minimalpunkts ist damit wegen Satz 2.3.3c) klar, dass f^2 (und damit auch f) genau einen globalen Minimalpunkt besitzt.

2.5.14 Beispiel (Beispiel 1.2.36 - Fortsetzung 5)
Die Funktion $f(\lambda) = \lambda\bar{x} - \log(\lambda)$ *mit* $\bar{x} > 0$ *aus Beispiel 1.2.36 erfüllt* $f''(\lambda) = \frac{1}{\lambda^2} > 0$ *für alle* $\lambda \in (0, +\infty)$*, aber* $\lim_{\lambda \to +\infty} f''(\lambda) = 0$*. Sie ist damit zwar strikt, aber nicht gleichmäßig konvex auf* $(0, +\infty)$*.*

Nach Beispiel 1.2.38 ist f *trotzdem koerziv. Dies zeigt, dass die Bedingung für Koerzivität aus Lemma 2.3.2 nur hinreichend, aber nicht notwendig ist.*

Konvexität und Monotonie der ersten Ableitung

Für $n = 1$, ein offenes Intervall $M \subseteq \mathbb{R}$ und eine Funktion $g \in C^1(M, \mathbb{R})$ ist aus der Analysis bekannt, dass g genau dann monoton wachsend auf M ist, wenn $g'(x) \geq 0$ für alle $x \in M$ gilt. Nach Satz 2.5.3 gilt für $n = 1$, ein offenes Intervall $M \subseteq \mathbb{R}$ und $f \in C^2(M, \mathbb{R})$ also, dass f genau dann konvex auf M ist, wenn f' auf M monoton wächst.

Im Folgenden werden wir sehen, dass eine solche Aussage auch ohne die C^2-Voraussetzung an f sowie allgemeiner für $n \geq 1$ gilt. Allerdings müssen wir zunächst definieren, was wir für $M \subseteq \mathbb{R}^n$ unter Monotonie der Funktion $\nabla f : M \to \mathbb{R}^n$ verstehen möchten. Dazu stellen wir fest, dass für $n = 1$ die Monotonie (im Sinne von monotonem Wachsen) von $g : M \to \mathbb{R}$ äquivalent zur Gültigkeit der Ungleichung $(g(y) - g(x))(y - x) \geq 0$ für alle $x, y \in M$ ist (Übung). Dies motiviert die folgende Definition.

2.5.15 Definition (Monotoner Operator)
Für eine nicht-leere konvexe Menge $M \subseteq \mathbb{R}^n$ *heißt* $g : M \to \mathbb{R}^n$ monoton *auf* M*, falls*

$$\forall\, x, y \in M : \quad \langle g(y) - g(x), y - x \rangle \;\geq\; 0$$

gilt.

2.5.16 Satz (Monotonie-Charakterisierung von Konvexität)
Auf einer konvexen Menge $M \subseteq \mathbb{R}^n$ *ist eine Funktion* $f \in C^1(M, \mathbb{R})$ *genau dann konvex, wenn* ∇f *auf* M *monoton ist.*

Beweis. Die Funktion $f \in C^1(M, \mathbb{R})$ sei konvex auf M. Nach Satz 2.2.2 gilt dann für alle $x, y \in M$

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle \quad \text{sowie} \quad f(x) \;\geq\; f(y) + \langle \nabla f(y), x - y \rangle.$$

Aus der Addition dieser beiden Ungleichungen folgt sofort die Monotonie von ∇f auf M.

Andererseits sei ∇f monoton auf M. Wir wählen beliebige $x, y \in M$, setzen $d := y - x$ sowie $x(t) := x + td$ für $t \in \mathbb{R}$ und betrachten die (auf einer offenen Obermenge des Intervalls $[0, 1]$ stetig differenzierbare) eindimensionale Einschränkung

$$\varphi_d : \; [0, 1] \to \mathbb{R}, \; t \mapsto f(x(t))$$

von f an x in Richtung d (vgl. auch [13]). Laut Kettenregel gilt $\varphi'_d(t) = \langle \nabla f(x(t)), d\rangle$ für jedes $t \in [0,1]$. Wegen $x(t) - x(0) = td$ erhalten wir damit aus der Monotonie von ∇f für jedes $t \in (0,1]$

$$\begin{aligned} \varphi'_d(t) - \varphi'_d(0) &= \langle \nabla f(x(t)) - \nabla f(x(0)), d\rangle \\ &= \tfrac{1}{t} \langle \nabla f(x(t)) - \nabla f(x(0)), x(t) - x(0)\rangle \ \geq \ 0. \end{aligned}$$

Nach dem Mittelwertsatz folgt hieraus mit einem $t \in (0,1)$

$$f(y) - f(x) \ = \ \varphi_d(1) - \varphi_d(0) \ = \ \varphi'_d(t) \ \geq \ \varphi'_d(0) \ = \ \langle \nabla f(x), y - x\rangle,$$

und Satz 2.2.2 liefert die Konvexität von f auf M. •

2.6 Dualität

Wir betrachten das restringierte Optimierungsproblem

$$P: \quad \min f(x) \quad \text{s.t.} \quad g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J$$

mit $f,\ g_i,\ h_j : \mathbb{R}^n \to \mathbb{R}$, $I = \{1, \ldots, p\}$, $J = \{1, \ldots, q\}$, $p, q \in \mathbb{N}_0$ und $q < n$ (wobei z.B. $p = 0$ dem Fall $I = \emptyset$ entspricht, also der Abwesenheit von Ungleichungsrestriktionen). Zunächst setzen wir weder Konvexität noch Differenzierbarkeit der beteiligten Funktionen voraus.

2.6.1 Definition (Lagrange-Funktion)
Die Funktion

$$L(x, \lambda, \mu) \ = \ f(x) + \sum_{i \in I} \lambda_i g_i(x) + \sum_{j \in J} \mu_j h_j(x)$$

(mit $\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_p \end{pmatrix}$ *und* $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_q \end{pmatrix}$*) heißt* Lagrangefunktion *von* P.

Es sei zunächst $p = 0$. Dann gilt $L(x, \mu) = f(x) + \sum_{j \in J} \mu_j h_j(x)$ und

$$\varphi(x) \ := \ \sup_{\mu \in \mathbb{R}^q} L(x, \mu) \ = \ \begin{cases} f(x), & \text{falls } \forall j \in J:\ h_j(x) = 0 \text{ (d.h. falls } x \in M) \\ +\infty, & \text{falls } x \notin M. \end{cases}$$

Beispiel:

$J = \{1\}$, $h_1(x) = x^2 - 1$, $f(x) = x$ führt zu der in Abbildung 2.9 skizzierten Funktion φ.

Statt P zu lösen kann man also formal auch die *unrestringierte* Funktion $\varphi(x)$ minimieren („formal“, weil φ nicht nach $\mathbb{R}$ abbildet, sondern nach $\overline{\mathbb{R}}$). Insbesondere gilt

$$v \ = \ \inf_{x \in M} f(x) \ = \ \inf_{x \in \mathbb{R}^n} \varphi(x) \ = \ \inf_{x \in \mathbb{R}^n} \sup_{\mu \in \mathbb{R}^q} L(x, \mu).$$

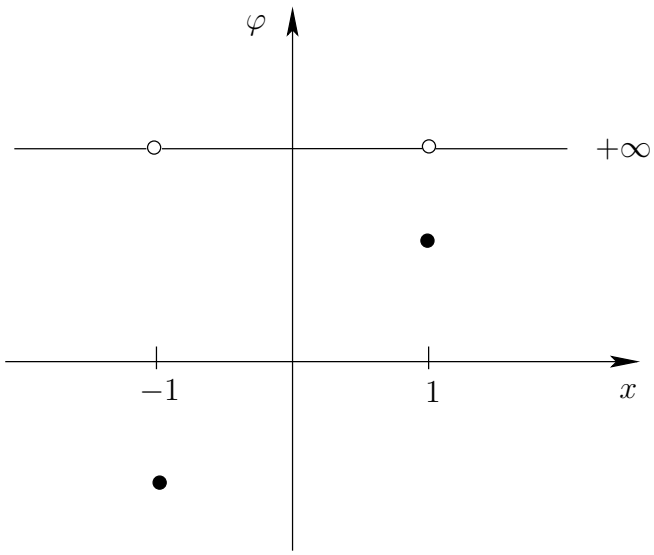


Abbildung 2.9: Beispiel für $\varphi(x) := \sup_{\mu\in\mathbb{R}^q} L(x,\mu)$

Nun sei $q = 0$. Dann gilt $L(x,\lambda) = f(x) + \sum_{i\in I} \lambda_i g_i(x)$,

$$\varphi(x) := \sup_{\lambda\geq 0} L(x,\lambda) = \begin{cases} f(x), & \text{falls } x \in M \\ +\infty, & \text{falls } x \notin M \end{cases}$$

und

$$v = \inf_{x\in M} f(x) = \inf_{x\in\mathbb{R}^n} \varphi(x) = \inf_{x\in\mathbb{R}^n} \sup_{\lambda\geq 0} L(x,\lambda).$$

Beispiel:

$I = \{1\}$, $g_1(x) = x^2 - 1$, $f(x) = x$ führt zu der in Abbildung 2.10 skizzierten Funktion φ.

Analog erhält man für beliebige $p, q \in \mathbb{N}_0$

$$\varphi(x) := \sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} L(x,\lambda,\mu) = \begin{cases} f(x), & \text{falls } x \in M \\ +\infty, & \text{falls } x \notin M. \end{cases}$$

und

$$v = \inf_{x\in M} f(x) = \inf_{x\in\mathbb{R}^n} \varphi(x) = \inf_{x\in\mathbb{R}^n} \sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} L(x,\lambda,\mu).$$

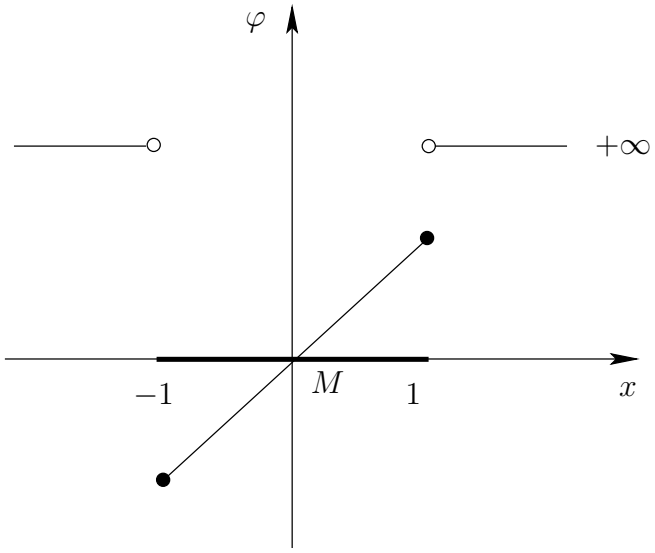


Abbildung 2.10: Beispiel für $\varphi(x) := \sup_{\lambda \geq 0} L(x, \lambda)$

Die *zentrale Frage* der Dualitätstheorie (und z.B. auch der Spieltheorie) lautet, ob man in obiger Formel „inf“ und „sup“ vertauschen darf, ob also auch die Formel

$$v = \inf_{x \in M} f(x) = \sup_{\lambda \geq 0,\ \mu \in \mathbb{R}^q} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu)$$

gilt.

Das folgende Beispiel zeigt, dass diese Vertauschung nicht immer möglich ist (vgl. Ü. 1.3.3d)).

2.6.2 Beispiel *Für $n = 1$, $p = 1$, $q = 0$, $f(x) = x^2$ und $g(x) = 1 - x^4 \leq 0$ berechnet man leicht $v = 1$. Andererseits gilt*

$$L(x, \lambda) = x^2 + \lambda(1 - x^4)$$

und damit

$$\inf_{x \in \mathbb{R}} L(x, \lambda) = \begin{cases} -\infty, & \lambda > 0 \\ 0, & \lambda = 0 \end{cases}$$

sowie

$$\sup_{\lambda \geq 0} \inf_{x \in \mathbb{R}} L(x, \lambda) = 0 < 1 = v.$$

Die Dualitätstheorie gibt Bedingungen an, unter denen solche Beispiele ausgeschlossen sind, unter denen also Gleichheit gilt. Dazu fassen wir den Ausdruck

$$\sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$$

als Maximalwert eines Optimierungsproblems auf:

$$D: \quad \max_{\lambda,\mu} \psi(\lambda,\mu) \quad \text{s.t.} \quad \lambda \geq 0$$

mit

$$\psi(\lambda,\mu) \ := \ \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu).$$

2.6.3 Definition (Lagrange-Dual)
Das Problem D heißt Lagrange-Dual *von P. Seinen Maximalwert bezeichnen wir mit*

$$v_D \ := \ \sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} \psi(\lambda,\mu)$$

und seine zulässige Menge mit

$$M_D \ := \ \{(\lambda,\mu)\in\mathbb{R}^p\times\mathbb{R}^q |\ \lambda\geq 0\}\,.$$

Analog bezeichnen wir P als primales *Problem mit Minimalwert $v_P := v$ und zulässiger Menge $M_P := M$.*

Die zentrale Frage der Dualitätstheorie lautet in dieser Notation, wann die Gleichheit

$$v_D \ = \ v_P$$

erfüllt ist.

Ohne weitere Voraussetzungen ist jedenfalls der folgende Satz richtig (vgl. Ü. 1.3.3c)):

2.6.4 Satz (Schwacher Dualitätssatz)
Es gilt $v_D \leq v_P$.

Beweis. Für alle $\bar{x}\in\mathbb{R}^n$, $\bar{\lambda}\geq 0$ und $\bar{\mu}\in\mathbb{R}^q$ gilt

$$\inf_{x\in\mathbb{R}^n} L(x,\bar{\lambda},\bar{\mu}) \ \leq \ L(\bar{x},\bar{\lambda},\bar{\mu}) \ \leq \ \sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} L(\bar{x},\lambda,\mu).$$

Die übliche Umformulierung dieser unendlich vielen Ungleichungen per Supremums- bzw. Infimumsbildung führt zu

$$\underbrace{\sup_{\bar{\lambda}\geq 0,\ \bar{\mu}\in\mathbb{R}^q} \inf_{x\in\mathbb{R}^n} L(x,\bar{\lambda},\bar{\mu})}_{v_D} \ \leq\ \underbrace{\inf_{\bar{x}\in\mathbb{R}^n} \sup_{\lambda\geq 0,\ \mu\in\mathbb{R}^q} L(\bar{x},\lambda,\mu)}_{v_P}.$$

•

Nach Satz 2.6.4 gilt

$$v_P - v_D \ \geq\ 0.$$

Der Wert $v_P - v_D$ heißt *Dualitätslücke*. In Beispiel 2.6.2 beträgt die Dualitätslücke $v_P - v_D = 1$.

In dieser Terminologie lautet die zentrale Frage der Dualitätstheorie, wann die Dualitätslücke verschwindet.

Um Dualität rechentechnisch nutzen zu können, muss das Dualproblem D zunächst handhabbar gemacht werden, denn es ist unklar, wie die Werte seiner Zielfunktion $\psi(\lambda,\mu) = \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ numerisch berechnet werden können.

Im Folgenden sei P ein konvexes Optimierungsproblem mit $f, g_i \in C^1(\mathbb{R}^n,\mathbb{R})$, $i \in I$, und h_j , $j \in J$, linear. Kurz gesagt sei P ab jetzt konvex und C^1. Dann ist die Lagrangefunktion

$$L(x,\lambda,\mu) \ =\ f(x) + \sum_{i\in I} \lambda_i g_i(x) + \sum_{j\in J} \mu_j h_j(x)$$

für beliebige fest vorgegebene $\lambda \geq 0$ und $\mu \in \mathbb{R}^q$ konvex und C^1 in der Variable x!

Folglich ist für alle $\lambda \geq 0,\ \mu \in \mathbb{R}^n$ der Wert $\psi(\lambda,\mu) = \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ der Minimalwert eines *unrestringierten konvexen* C^1-Problems. Falls dieses Infimum für alle $\lambda \geq 0,\ \mu \in \mathbb{R}^n$ als Minimalwert *angenommen* wird, lässt es sich mit Korollar 2.4.6 berechnen als

$$\psi(\lambda,\mu) \ =\ L(x^\star,\lambda,\mu),$$

wobei $x^\star$ einen globalen Minimalpunkt von $L(\cdot,\lambda,\mu)$ auf $\mathbb{R}^n$ bezeichnet, also eine Lösung $x^\star$ von

$$\nabla_x L(x,\lambda,\mu) \ =\ 0.$$

Das Dualproblem lässt sich dadurch folgendermaßen umformulieren:

$$D: \quad \max_{x,\lambda,\mu} L(x,\lambda,\mu) \quad \text{s.t.} \quad \nabla_x L(x,\lambda,\mu) = 0,\ \lambda \geq 0$$

(das sogenannte *Wolfe-Dual* von P), was im Gegensatz zum Lagrange-Dual ein numerisch gut handhabbares Problem ist.

Seine zulässige Menge wird nun zu

$$M_D \ = \ \{(x,\lambda,\mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q |\ \nabla_x L(x,\lambda,\mu) = 0,\ \lambda \geq 0\}\,.$$

Einen Punkt $x \in M_P = M$ nennen wir im Folgenden *primal zulässig*, und $(x,\lambda,\mu) \in M_D$ *dual zulässig*.

Schwache Dualität bedeutet jetzt

$$\sup_{(x,\lambda,\mu)\in M_D} L(x,\lambda,\mu) \ \leq \ \inf_{x\in M_P} f(x).$$

Aufgrund der formalen Definitionen für Suprema und Infima über leere Mengen ist diese Ungleichung bemerkenswerterweise sogar für $M_D = \emptyset$ und/oder $M_P = \emptyset$ sinnvoll.

2.6.5 Satz

a) $\bar{x}$ sei primal zulässig. Dann ist $f(\bar{x})$ eine Ober*schranke für den globalen Minimalwert von P:*

$$v_P \ \leq \ f(\bar{x}).$$

b) $(\bar{x},\bar{\lambda},\bar{\mu})$ sei dual zulässig. Dann ist $L(\bar{x},\bar{\lambda},\bar{\mu})$ eine Unter*schranke für v_P:*

$$v_P \ \geq \ L(\bar{x},\bar{\lambda},\bar{\mu}).$$

Beweis. Teil a) gilt natürlich auch ohne Dualitätstheorie:

$$f(\bar{x}) \overset{\bar{x}\ \in\ M_P}{\geq} \inf_{x\in M} f(x) \ = \ v_P\,.$$

Ferner ist aufgrund der schwachen Dualität

$$\begin{aligned} L(\bar{x},\bar{\lambda},\bar{\mu}) &\overset{(\bar{x},\bar{\lambda},\bar{\mu})\ \in\ M_D}{\leq} \sup_{(x,\lambda,\mu)\in M_D} L(x,\lambda,\mu) \\ &\leq \inf_{x\in M_P} f(x) \ = \ v_P\,, \end{aligned}$$

woraus Teil b) folgt. •

2.6.6 Beispiel (Abstand von einer Hyperebene)
Gesucht ist der Abstand von $z \in \mathbb{R}^n$ zu der Hyperebene $H = \{x \in \mathbb{R}^n | a^\intercal x = b\}$ mit $a \in \mathbb{R}^n \setminus \{0\}$ und $b \in \mathbb{R}$. Abbildung 2.11 illustriert das Problem für $n = 2$, $a = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $b = 2$, $z = 0$. Im Folgenden werden wir sehen, welche Aussagen

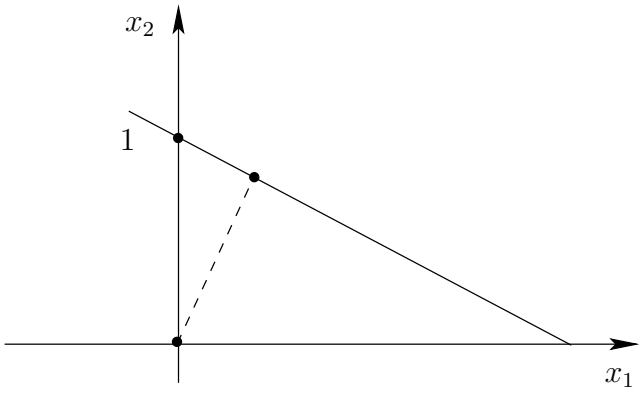


Abbildung 2.11: Abstand von einer Hyperebene im Fall $n = 2$

die Dualitätstheorie für das allgemeine Problem

$$P: \quad \min_{x \in \mathbb{R}^n} \|x - z\|_2 \quad s.t. \quad a^\intercal x = b$$

zulässt und dies an obigem speziellen Beispiel

$$\min_{x \in \mathbb{R}^2} \|x\|_2 \quad s.t. \quad x_1 + 2x_2 = 2$$

verdeutlichen. P ist durch Quadrieren der Zielfunktion nach Übung 1.3.5 äquivalent zu

$$Q: \quad \min (x - z)^\intercal (x - z) \quad s.t. \quad a^\intercal x = b.$$

In Q ist die Zielfunktion $f(x) := x^\intercal x - 2z^\intercal x + z^\intercal z$ offenbar konvex und C^1, und die Gleichungsrestriktion $h(x) = a^\intercal x - b$ ist linear. Also ist Q konvex und C^1 und damit für Dualitätsaussagen geeignet.

Die Lagrangefunktion von Q lautet

$$L(x, \mu) \;=\; x^\intercal x - 2z^\intercal x + z^\intercal z + \mu(a^\intercal x - b)$$

woraus

$$\nabla_x L(x, \mu) \;=\; 2x - 2z + \mu a$$

folgt. Damit kann man das Dualproblem zu Q aufstellen:

$$D: \quad \max_{x\in\mathbb{R}^n,\ \mu\in\mathbb{R}} x^\intercal x - 2z^\intercal x + z^\intercal z + \mu(a^\intercal x - b) \quad s.t. \quad 2(x-z)+\mu a = 0$$

$$\left(D: \quad \max_{x\in\mathbb{R}^2,\ \mu\in\mathbb{R}} x_1^2 + x_2^2 + \mu(x_1 + 2x_2 - 2) \quad s.t. \quad 2x + \mu\begin{pmatrix}1\\2\end{pmatrix} = 0\right).$$

Wegen $M_P = H$ bedeutet primale Zulässigkeit von x gerade $a^\intercal x = b$. Duale Zulässigkeit von (x,μ) liegt genau für $2(x-z)+\mu a = 0$ vor.

Oberschranke für v_P :

Da zum Beispiel $\bar{x} = b\,\dfrac{a}{a^\intercal a}$ ein primal zulässiger Punkt ist, liefert Satz 2.6.5a) die Abschätzung

$$\begin{aligned} v_Q &\le f\left(b\,\frac{a}{a^\intercal a}\right) = b^2\frac{a^\intercal a}{(a^\intercal a)^2} - 2z^\intercal\left(b\,\frac{a}{a^\intercal a}\right) + z^\intercal z \\ &= \frac{b^2}{a^\intercal a} - \frac{2b}{a^\intercal a}z^\intercal a + z^\intercal z. \end{aligned}$$

Es folgt

$$v_P = \sqrt{v_Q} \le \sqrt{\frac{b^2}{a^\intercal a} - \frac{2b}{a^\intercal a}z^\intercal a + z^\intercal z}$$

$$\left(v_Q \le \frac{4}{5}, \quad v_P \le \frac{2}{\sqrt{5}}\right).$$

(Im speziellen Beispiel ist eine andere Wahl eines primal zulässigen Punktes $\bar{x} = \begin{pmatrix}0\\1\end{pmatrix}$, woraus $v_Q \le f\begin{pmatrix}0\\1\end{pmatrix} = 1$ und $v_P \le 1$ folgt. Verglichen mit der eben berechneten ist dies allerdings eine schlechtere Oberschranke.)

Unterschranke für v_P :

In der Bedingung $2(x-z)+\mu a = 0$ für duale Zulässigkeit sind z und a vorgegeben, während Paare (x,μ) gesucht werden, die diese Bedingung erfüllen. Durch Isolieren von x erhält man

$$x = z - \frac{\mu}{2}a,$$

wodurch für jedes beliebig gewählte $\bar{\mu}\in\mathbb{R}$ das für duale Zulässigkeit dazu passende $\bar{x} = z - \frac{\bar{\mu}}{2}a$ angegeben wird. Es folgt:

$$\forall\ \bar{\mu}\in\mathbb{R}: \quad (\bar{x},\bar{\mu}) = \left(z - \frac{\bar{\mu}}{2}a,\ \bar{\mu}\right) \in M_D\,.$$

Nach Satz 2.6.5b) gilt also für alle $\bar{\mu}\in\mathbb{R}$

$$\begin{aligned} v_Q &\ge L(\bar{x},\bar{\mu}) \\ &= L\left(z - \frac{\bar{\mu}}{2}a, \bar{\mu}\right) \\ &= \left(z - \frac{\bar{\mu}}{2}a\right)^\intercal\left(z - \frac{\bar{\mu}}{2}a\right) - 2z^\intercal\left(z - \frac{\bar{\mu}}{2}a\right) + z^\intercal z + \bar{\mu}\left(a^\intercal\left(z - \frac{\bar{\mu}}{2}a\right) - b\right) \\ &= \bar{\mu}(z^\intercal a - b) - \frac{\bar{\mu}^2}{4}a^\intercal a \end{aligned}$$

$$\Bigg(v_Q \;\geq\; -2\bar{\mu} - \frac{5}{4}\bar{\mu}^2 \;, \quad \textit{z.B.} \quad \bar{\mu} = 0 \;\Rightarrow\; v_Q \geq 0 \quad \textit{(was ohnehin klar ist)},$$
$$\bar{\mu} = 1 \;\Rightarrow\; v_Q \geq -\frac{13}{4} \quad \textit{(was noch klarer ist)},$$
$$\bar{\mu} = -1 \;\Rightarrow\; v_Q \geq \frac{3}{4}\Bigg).$$

Wegen der für Q trivialerweise gültigen Abschätzung $v_Q \geq 0$ gilt für alle $\bar{\mu} \in \mathbb{R}$

$$v_Q \;\geq\; \max\left\{0,\, \bar{\mu}(z^\intercal a - b) - \frac{\bar{\mu}^2}{4} a^\intercal a\right\}$$

und damit

$$v_P \;=\; \sqrt{v_Q} \;\geq\; \sqrt{\max\left\{0,\, \bar{\mu}(z^\intercal a - b) - \frac{\bar{\mu}^2}{4} a^\intercal a\right\}}$$

$$\Bigg(v_P \;\geq\; \sqrt{\{\max\{0,\, -2\bar{\mu} - \frac{5}{4}\bar{\mu}^2\}}\;, \quad \textit{z.B.} \quad \bar{\mu} \;=\; -1 :$$
$$\underbrace{\frac{\sqrt{3}}{2}}_{\approx 0.86} \;\leq\; v_P \overset{\text{s.o.}}{\leq} \underbrace{\frac{2}{\sqrt{5}}}_{\approx 0.89}\Bigg).$$

Als best*mögliche Unterschranke für v_Q erhält man bei diesem Ansatz das Supremum der Schranken über alle Wahlen von $\bar{\mu} \in \mathbb{R}$,*

$$\sup_{\bar{\mu} \in \mathbb{R}} \bar{\mu}(z^\intercal a - b) - \frac{\bar{\mu}^2}{4} a^\intercal a,$$

was genau der Maximalwert des Dualproblems D ist. Zur Berechnung von v_D setzen wir

$$c(\bar{\mu}) \;:=\; \bar{\mu}(z^\intercal a - b) - \frac{\bar{\mu}^2}{4} a^\intercal a$$

und erhalten

$$c'(\bar{\mu}) \;=\; a^\intercal z - b - \frac{\bar{\mu}}{2} a^\intercal a,$$
$$c''(\bar{\mu}) \;=\; -\frac{a^\intercal a}{2} \;\overset{a \neq 0}{<}\; 0.$$

Folglich ist die Funktion c konkav. Nach Korollar 2.4.6 sind ihre globalen Maximalpunkte genau die Lösungen von

$$0 \;=\; c'(\bar{\mu}) \;=\; a^\intercal z - b - \frac{\bar{\mu}}{2} a^\intercal a$$

d.h. eindeutiger Maximalpunkt von c ist

$$\bar{\mu}^\star \;=\; 2\frac{a^\intercal z - b}{a^\intercal a}$$

mit Maximalwert

$$v_D \;=\; c(\bar{\mu}^\star) \;=\; 2\frac{(a^\intercal z - b)^2}{a^\intercal a} - \frac{(a^\intercal z - b)^2}{a^\intercal a} \;= \frac{(a^\intercal z - b)^2}{a^\intercal a}\,.$$

Als beste Unterschranke für v_P folgt daraus

$$v_P \;=\; \sqrt{v_Q} \;\geq\; \sqrt{v_D} \;=\; \frac{|a^\intercal z - b|}{\|a\|_2}$$

$$\left(v_P \;\geq\; \frac{|-2|}{\sqrt{5}} \;=\; \frac{2}{\sqrt{5}} \overset{\text{s.o.}}{\geq} \; v_P\right).$$

Im speziellen Beispiel folgt aus diesen Überlegungen $v_P = \frac{2}{\sqrt{5}}$, denn zufällig *hatte man oben auch die beste Oberschranke gefunden. Im allgemeinen Fall liefert die schwache Dualität hingegen für die Distanz nur die Abschätzung*

$$v_P \;\geq\; \frac{|a^\intercal z - b|}{\|a\|_2}\,.$$

*Wünschenswert wären nun noch eine Verbesserung dieser Abschätzung zu einer Gleichheit sowie eine Formel für den Minimal*punkt *von P.*

2.6.7 Lemma *P sei konvex und C^1, und es seien $\bar{x} \in \mathbb{R}^n$, $\bar{\lambda} \in \mathbb{R}^p$, $\bar{\mu} \in \mathbb{R}^q$ gegeben mit*

a) $\bar{x}$ primal zulässig,

b) $(\bar{x}, \bar{\lambda}, \bar{\mu})$ dual zulässig,

c) $f(\bar{x}) \;=\; L(\bar{x}, \bar{\lambda}, \bar{\mu})$.

Dann ist $\bar{x}$ globaler Minimalpunkt von P.

Beweis.

$$\begin{aligned} f(\bar{x}) \;&\overset{\text{a)}}{\geq}\; \inf_{x \in M_P} f(x) \;=\; v_P \\ &\geq\; v_D \;=\; \sup_{(x,\lambda,\mu) \in M_D} L(x, \lambda, \mu) \\ &\overset{\text{b)}}{\geq}\; L(\bar{x}, \bar{\lambda}, \bar{\mu}) \\ &\overset{\text{c)}}{=}\; f(\bar{x}). \end{aligned}$$

Demnach muss jede Ungleichheit in dieser Ungleichungskette mit Gleichheit erfüllt sein. Dies bedeutet insbesondere $f(\bar{x}) = v_P$ (und außerdem auch, dass die Dualitätslücke verschwindet). Damit ist $\bar{x}$ globaler Minimalpunkt von P. •

2.6.8 Beispiel (Beispiel 2.6.6 - Fortsetzung 1)
Wie oben gesehen ist $(\bar{x}, \bar{\mu})$ dual zulässig für $\bar{\mu} \in \mathbb{R}$ beliebig und $\bar{x} = z - \frac{\bar{\mu}}{2}a$. Außerdem bedeutet primale Zulässigkeit von $\bar{x}$, dass die Gleichung $a^\intercal \bar{x} = b$ erfüllt ist. Damit gleichzeitig die Bedingungen a) und b) aus Lemma 2.6.7 gelten, müssen $\bar{x}$ und $\bar{\mu}$ also das Gleichungssystem

$$\begin{aligned} \bar{x} &= z - \frac{\bar{\mu}}{2}\, a \\ a^\intercal \bar{x} &= b \end{aligned}$$

erfüllen. Die Lösung berechnet sich leicht zu

$$\begin{aligned} \mu^\star &= 2\, \frac{a^\intercal z - b}{a^\intercal a} \\ x^\star &= z - \frac{a^\intercal z - b}{a^\intercal a}\, a\,. \end{aligned}$$

Obwohl $(x^\star, \mu^\star)$ nun schon eindeutig bestimmt ist (z, a und b sind ja vorgegeben), muss in Lemma 2.6.7 auch noch Bedingung c) gelten, also

$$f(x^\star) \;=\; L(x^\star, \mu^\star).$$

Glücklicherweise erfüllen die bereits ermittelten $(x^\star, \mu^\star)$ diese Gleichung, denn

$$L(x^\star, \mu^\star) \;=\; f(x^\star) + \mu^\star \underbrace{(a^\intercal x^\star - b)}_{=\,0} \;=\; f(x^\star).$$

Insgesamt sind damit die Bedingungen a), b) und c) aus Lemma 2.6.7 erfüllt, so dass

$$x^\star \;=\; z - \frac{a^\intercal z - b}{a^\intercal a}\, a$$

globaler Minimalpunkt von Q mit Minimalwert

$$v_Q \;=\; \frac{(a^\intercal z - b)^2}{a^\intercal a}$$

ist. Wie gesehen, ist $x^\star$ dann auch globaler Minimalpunkt von P mit Minimalwert

$$v_P \;=\; \frac{|a^\intercal z - b|}{\|a\|_2}\,.$$

Dass wie im obigen Beispiel die Bedingung c) aus Lemma 2.6.7 automatisch von den Lösungen der Bedingungen a) und b) erfüllt wird, ist allerdings kein Glücksfall, sondern dieser Effekt tritt bei jedem Problem *ohne Ungleichungen* auf:

2.6.9 Korollar *P sei konvex und C^1, es gelte $I = \emptyset$, und es seien $\bar{x} \in \mathbb{R}^n$, $\bar{\mu} \in \mathbb{R}^q$ gegeben mit*

a) $\bar{x}$ primal zulässig,

b) $(\bar{x}, \bar{\mu})$ dual zulässig.

Dann ist $\bar{x}$ globaler Minimalpunkt von P.

Beweis. Es gilt

$$M_P \;=\; \{x \in \mathbb{R}^n |\; h_j(x) = 0,\; j \in J\}$$

und damit

$$L(\bar{x}, \bar{\mu}) \;=\; f(\bar{x}) + \sum_{j \in J} \bar{\mu}_j h_j(\bar{x}) \;\overset{\bar{x} \in M_P}{=}\; f(\bar{x}).$$

Insgesamt gelten also die Bedingungen a), b) und c) aus Lemma 2.6.7, woraus die Behauptung folgt. •

Da *Ungleichungen* auch strikt erfüllt sein können, kann man für sie nicht wie in Korollar 2.6.9 argumentieren. Bedingung c) aus Lemma 2.6.7 lässt sich aber trotzdem einfacher schreiben, wofür folgendes Konzept eingeführt wird (für das Konvexität zunächst *keine* Rolle spielt):

2.6.10 Definition (Karush-Kuhn-Tucker-Punkt)
Für ein C^1-Problem P heißt ein Punkt $\bar{x} \in \mathbb{R}^n$ Karush-Kuhn-Tucker-Punkt (KKT-Punkt) *mit Multiplikatoren $\bar{\lambda}$ und $\bar{\mu}$, falls folgendes System von Gleichungen und Ungleichungen erfüllt ist:*

$$\begin{aligned}
\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) &= 0 && (2.6.1)\\
\bar{\lambda}_i \, g_i(\bar{x}) &= 0,\; i \in I, && (2.6.2)\\
\bar{\lambda}_i &\geq 0,\; i \in I, && (2.6.3)\\
g_i(\bar{x}) &\leq 0,\; i \in I, && (2.6.4)\\
h_j(\bar{x}) &= 0,\; j \in J. && (2.6.5)
\end{aligned}$$

2.6.11 Lemma *Die Voraussetzungen von Lemma 2.6.7 sind für $(\bar{x}, \bar{\lambda}, \bar{\mu})$ genau dann erfüllt, wenn $\bar{x}$ KKT-Punkt von P mit Multiplikatoren $\bar{\lambda}$, $\bar{\mu}$ ist.*

Beweis. Es ist die Äquivalenz der Bedingungen a), b), c) aus Lemma 2.6.7 mit den Bedingungen (2.6.1) – (2.6.5) aus Definition 2.6.10 zu zeigen.

„⇒“:

Aus a) folgen (2.6.4) und (2.6.5), aus b) folgen (2.6.1) und (2.6.3), und aus c) folgt zunächst

$$\sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) + \sum_{j \in J} \bar{\mu}_j h_j(\bar{x}) \;=\; 0.$$

Da a) insbesondere $h_j(\bar{x}) = 0$, $j \in J$, bedeutet, reduziert sich diese Gleichung zu

$$\sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) \;=\; 0.$$

Ferner implizieren a) $g_i(\bar{x}) \leq 0$, $i \in I$, und b) $\bar{\lambda}_i \geq 0$, $i \in I$, so dass jeder Summand in obiger Summe nichtpositiv ist. Damit die Summe null ergibt, kann dann keiner der Summanden strikt negativ sein, so dass man

$$\bar{\lambda}_i \, g_i(\bar{x}) \;=\; 0, \; i \in I,$$

erhält, also (2.6.2).

„⇐“:

Aus (2.6.4) und (2.6.5) folgt a), aus (2.6.1) und (2.6.3) folgt b), und aus (2.6.2) und (2.6.5) folgt c). •

2.6.12 Satz (Hinreichende Optimalitätsbedingung für P konvex, C^1) *P sei konvex und C^1, und $\bar{x}$ sei KKT-Punkt (mit Multiplikatoren $\bar{\lambda}$, $\bar{\mu}$). Dann ist $\bar{x}$ globaler Minimalpunkt von P.*

Beweis. Lemma 2.6.11 und Lemma 2.6.7. •

Angemerkt sei, dass in Satz 2.6.12 die Forderung, $\bar{x}$ sei ein KKT-Punkt, per Lemma 2.6.11 und Lemma 2.6.7c) insbesondere die Forderung nach einer verschwindenden Dualitätslücke impliziert.

2.6.13 Beispiel (Beispiel 2.6.6 - Fortsetzung 2) *$\bar{x}$ ist KKT-Punkt von Q mit Multiplikator $\bar{\mu}$ genau dann, wenn die Gleichungen $2(\bar{x} - z) + \bar{\mu} a = 0$ und $a^\intercal \bar{x} - b = 0$ erfüllt sind. Wie oben folgt daraus $x^\star = z - \frac{a^\intercal z - b}{a^\intercal a} a$ und $\mu^\star = 2\frac{a^\intercal z - b}{a^\intercal a}$.*

Die Darstellung des Resultats aus Satz 2.6.12 als Mengendiagramm in Abbildung 2.12 führt auf ein analoges Bild wie in Abbildung 2.7.

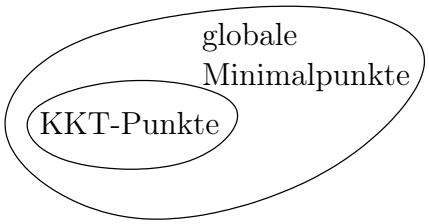


Abbildung 2.12: Hinreichende Optimalitätsbedingung

Dies lässt vermuten, dass auch das in Abbildung 2.6 illustrierte Resultat analog gilt, d.h. mit dem Begriff „KKT-Punkte“ anstellte von „kritische Punkte“. Das folgende Beispiel zeigt, dass dies leider *nicht* der Fall ist.

2.6.14 Beispiel *Betrachte das Problem*

$$P: \quad \min x \quad s.t. \quad x^2 \leq 0.$$

Dann ist P konvex und C^1 mit $M_P = S = \{0\}$.

Allerdings ist $\bar{x} = 0$ kein *KKT-Punkt von P, denn die Lagrangefunktion von P lautet*

$$L(x, \lambda) \;=\; x + \lambda x^2,$$

woraus

$$\nabla_x L(x, \lambda) \;=\; 1 + 2\lambda x$$

folgt. Die Gleichung des KKT-Systems

$$0 \;=\; \nabla_x L(\bar{x}, \lambda) \;=\; 1 + 2\lambda \cdot 0 \;=\; 1$$

ist somit für kein $\lambda \geq 0$ lösbar.

Dieses Beispiel zeigt, dass das Analogon zu Korollar 2.4.6, nämlich „jeder globale Minimalpunkt von P ist KKT-Punkt“ für restringierte Probleme *nicht* gilt.

In negativer Formulierung bedeutet dies auch: es ist möglich, dass ein restringiertes Problem zwar keinen KKT-Punkt besitzt, aber trotzdem lösbar ist. Anders ausgedrückt: selbst wenn man beweisen kann, dass ein Problem keine KKT-Punkte besitzt, bedeutet dies noch nicht, dass das Problem auch unlösbar ist.

2.6.15 Übung *Zeigen Sie, dass in Beispiel 2.6.14 starke Dualität im Sinne der Identität $v_P = v_D$ gilt. Warum lässt sich die Bedingung $f(\bar{x}) = L(\bar{x}, \bar{\lambda})$ aus Lemma 2.6.7c) hier trotzdem nicht erfüllen?*

Im Folgenden untersuchen wir, welche zusätzlichen Bedingungen garantieren, dass ein globaler Minimalpunkt KKT-Punkt ist.

Komplementarität

Wie im Beweis zu Lemma 2.6.11 gesehen, ist unter den Bedingungen a) und b) aus Lemma 2.6.7 (primale und duale Zulässigkeit) die Bedingung c) (Gleichheit von primalem und dualen Zielfunktionswert) gleichbedeutend mit

$$\bar{\lambda}_i \geq 0, \quad g_i(\bar{x}) \leq 0, \quad 0 = \bar{\lambda}_i \, g_i(\bar{x}), \quad i \in I.$$

Für jedes $i \in I$ heißt die Bedingung

$$\bar{\lambda}_i \geq 0, \quad g_i(\bar{x}) \leq 0, \quad \bar{\lambda}_i g_i(\bar{x}) = 0$$

Komplementaritätsbedingung.

Je nachdem, ob die Restriktion $g_i(x) \leq 0$ an $\bar{x}$ mit Gleichheit oder mit strikter Ungleichheit erfüllt ist, hat die Komplementaritätsbedingung unterschiedliche Konsequenzen:

1.Fall:

Für $g_i(\bar{x}) = 0$ ist $\bar{\lambda}_i g_i(\bar{x}) = 0$ für beliebige $\bar{\lambda}_i \geq 0$ erfüllt, d.h. die Komplementaritätsbedingung gilt automatisch.

2.Fall:

Für $g_i(\bar{x}) < 0$ impliziert die Bedingung $\bar{\lambda}_i g_i(\bar{x}) = 0$, dass $\bar{\lambda}_i = 0$ gilt.

Der Begriff *Komplementarität* bezieht sich darauf, dass mindestens eine der Zahlen $\bar{\lambda}_i$ und $g_i(\bar{x})$ verschwindet.

Die offensichtlich wichtige Unterscheidung, ob eine Ungleichung mit Gleichheit oder strikter Ungleichheit erfüllt ist, motiviert die folgende Definition.

2.6.16 Definition (Aktive-Index-Menge)
Zu $\bar{x} \in M$ heißt

$$I_0(\bar{x}) \;=\; \{i \in I |\; g_i(\bar{x}) = 0\}$$

Menge der aktiven Indizes *oder auch* Aktive-Index-Menge *von* $\bar{x}$.

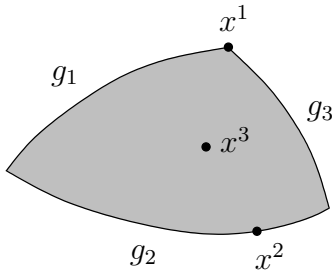


Abbildung 2.13: Aktive Indizes

Beispielsweise gilt in Abbildung 2.13

$$I_0(x^1) \;=\; \{1,3\}\,, \quad I_0(x^2) \;=\; \{2\}\,, \quad I_0(x^3) \;=\; \emptyset.$$

Geometrische Interpretation der KKT-Bedingungen

Da die Komplementaritätsbedingungen für alle $i \notin I_0(\bar{x})$ erzwingen, dass $\bar{\lambda}_i$ verschwindet, ist $\bar{x}$ KKT-Punkt mit Multiplikatoren $\bar{\lambda}, \bar{\mu}$ genau dann, wenn folgendes System erfüllt ist:

$$\begin{aligned}
\nabla f(\bar{x}) + \sum_{i \in I_0(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j \in J} \bar{\mu}_j \nabla h_j(\bar{x}) &= 0 \\
g_i(\bar{x}) &= 0,\ i \in I_0(\bar{x}), \\
g_i(\bar{x}) &< 0,\ i \notin I_0(\bar{x}), \\
h_j(\bar{x}) &= 0,\ j \in J, \\
\bar{\lambda}_i &\geq 0,\ i \in I_0(\bar{x}), \\
\bar{\lambda}_i &= 0,\ i \notin I_0(\bar{x}).
\end{aligned}$$

Um die geometrische Bedeutung dieser Bedingungen zu verstehen, setzen wir die Kenntnis der Tatsache voraus, dass Gradienten senkrecht auf Höhenlinien stehen und in die Richtung des steilsten Anstiegs zeigen (s. [13]).

Wir betrachten zunächst den Fall *ohne Ungleichungen*, d.h. $I = \emptyset$. Dann reduziert sich das KKT-System zu

$$\begin{aligned}
\nabla f(\bar{x}) + \sum_{j=1}^{q} \bar{\mu}_j \nabla h_j(\bar{x}) &= 0 \\
h_j(\bar{x}) &= 0,\ j \in J.
\end{aligned}$$

Für $n = 2$ und $q = 1$ zeigt Abbildung 2.14 Höhenlinien von $f(x) = x_1^2 + x_2^2$ und einer linearen Gleichungsrestriktion $h(x) = 0$.

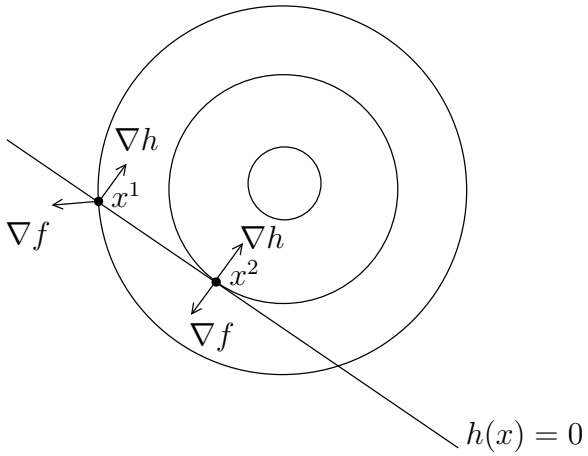


Abbildung 2.14: KKT-Punkte für $I = \emptyset$

Da die KKT-Bedingungen (neben der Zulässigkeit von $\bar{x}$) besagen, dass ∇f und ∇h an einem KKT-Punkt $\bar{x}$ linear abhängig sind, ist x^1 kein KKT-Punkt, während x^2 KKT-Punkt ist.

Bemerkenswert ist, dass zwar die Geometrie der zulässigen Menge M sich nicht ändert, wenn man h durch $-h$ ersetzt, aber $\nabla(-h) = -\nabla h$ in die zu ∇h entgegengesetzte Richtung zeigt. Obwohl sich also durch die Vorzeichenänderung von h nichts daran ändert, ob $\bar{x}$ zum Beispiel ein lokaler Minimalpunkt ist, ändern sich die KKT-Bedingungen. Dies wird allerdings dadurch kompensiert, dass die Multiplikatoren μ_j , $j \in J$, nicht vorzeichenbeschränkt sind. Die Ersetzung von h_j durch $-h_j$ kann man in den KKT-Bedingungen also einfach dadurch kompensieren, dass man $\bar{\mu}_j$ durch $-\bar{\mu}_j$ ersetzt.

Wir wenden uns nun dem Fall *ohne Gleichungen* zu, d.h. $J = \emptyset$. Dann lautet das KKT-System

$$\begin{aligned} \nabla f(\bar{x}) + \sum_{i \in I_0(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) &= 0 \\ g_i(\bar{x}) &= 0,\ i \in I_0(\bar{x}), \\ g_i(\bar{x}) &< 0,\ i \notin I_0(\bar{x}), \\ \bar{\lambda}_i &\geq 0,\ i \in I_0(\bar{x}). \end{aligned}$$

Dies bedeutet insbesondere, dass der Vektor $-\nabla f(\bar{x})$ in der Menge

$$\text{cone}\,\{\nabla g_i(\bar{x}),\ i \in I_0(\bar{x})\} \ := \ \left\{ \sum_{i \in I_0(\bar{x})} \lambda_i \nabla g_i(\bar{x}) \,\middle|\, \lambda \geq 0 \right\}$$

liegt, dem von den Vektoren $\nabla g_i(\bar{x})$, $i \in I_0(\bar{x})$, aufgespannten konvexen Kegel.

Für $n = 2$ und $p = 3$ zeigt Abbildung 2.15 eine durch konvexe Ungleichungen beschriebene zulässige Menge, Höhenlinien einer linearen Zielfunktion f (z.B. $f(x) = x_1 + x_2$) sowie einige von Gradienten aktiver Ungleichungen aufgespannte konvexe Kegel.

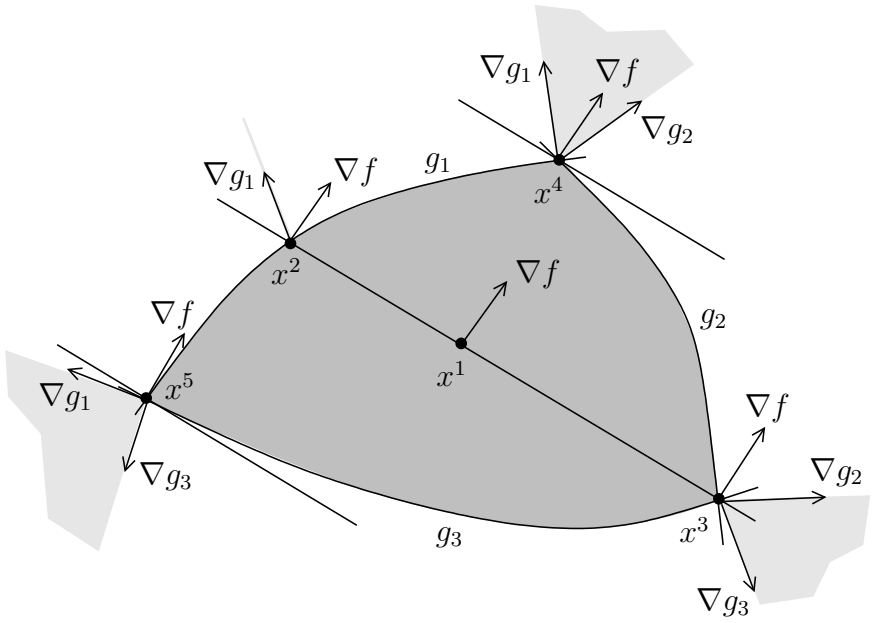


Abbildung 2.15: KKT-Punkte für $J = \emptyset$

Durch die obige geometrische Interpretation macht man sich leicht klar, dass von den Punkten x^1 bis x^5 in Abbildung 2.15 nur x^5 KKT-Punkt ist. Da P ein konvexes Optimierungsproblem ist, muss x^5 sogar globaler Minimalpunkt sein.

Ungünstig für dieses Konzept sind „Spitzen“ an der zulässigen Menge, wie in Abbildung 2.16 dargestellt (mit linearer Zielfunktion wie z.B. $f(x) = x_1 + x_2$).

Wie in Beispiel 2.6.14 ist $\bar{x}$ zwar globaler Minimalpunkt, aber kein KKT-Punkt. Durch diese geometrische Betrachtung lassen sich sogenannte *Con-*

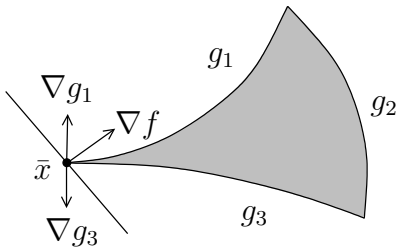


Abbildung 2.16: Spitze an der zulässigen Menge

straint Qualifications motivieren, unter denen globale Minimalpunkte KKT-Punkte sind.

Constraint Qualifications

2.6.17 Definition (Constraint Qualifications)

a) An $\bar{x} \in M$ gilt die Lineare-Unabhängigkeits-Bedingung (LUB)*, falls die Vektoren $\nabla g_i(\bar{x})$, $i \in I_0(\bar{x})$, $\nabla h_j(\bar{x})$, $j \in J$, linear unabhängig sind (d.h die Gleichung*

$$\sum_{i \in I_0(\bar{x})} \lambda_i \nabla g_i(\bar{x}) + \sum_{j \in J} \mu_j \nabla h_j(\bar{x}) \;=\; 0$$

hat nur die Lösung $\begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$).

b) An $\bar{x} \in M$ gilt die Mangasarian-Fromowitz-Bedingung (MFB)*, falls die Vektoren $\nabla h_j(\bar{x})$, $j \in J$, linear unabhängig sind und das System*

$$\langle \nabla g_i(\bar{x}), d \rangle < 0,\ i \in I_0(\bar{x}),\ \langle \nabla h_j(\bar{x}), d \rangle = 0,\ j \in J,$$

eine Lösung d besitzt (was laut [13] dazu äquivalent ist, dass das System

$$\begin{aligned} \sum_{i \in I_0(\bar{x})} \lambda_i \nabla g_i(\bar{x}) + \sum_{j \in J} \mu_j \nabla h_j(\bar{x}) &= 0, \\ \lambda &\geq 0 \end{aligned}$$

nur die Lösung $\begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ besitzt).

c) *M erfüllt die* Slaterbedingung (SB)*, falls die Vektoren $\nabla h_j(x)$, $j \in J$, für alle $x \in M$ linear unabhängig sind, und falls ein $x^\star$ existiert mit $g_i(x^\star) < 0$, $i \in I$, $h_j(x^\star) = 0$, $j \in J$.*

Bemerkungen:

- LUB und MFB sind *lokale* Bedingungen an *einem* $\bar{x} \in M$, SB ist eine *globale* Bedingung für ganz M.
- Für Constraint Qualifications spielt die Zielfunktion f keine Rolle.

2.6.18 Übung *LUB an $\bar{x} \in M$ impliziert MFB an $\bar{x} \in M$, aber die Umkehrung dieser Aussage gilt nicht.*

2.6.19 Lemma *Die h_j, $j \in J$, seien linear, d.h. $h_j(x) = a_j^\intercal x + b_j$, $j \in J$, und es seien $A = \begin{pmatrix} a_1^\intercal \\ \vdots \\ a_q^\intercal \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_q \end{pmatrix}$, also $h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_q(x) \end{pmatrix} = Ax + b$. Dann erfüllt M genau dann SB, wenn* $\operatorname{rang} A = q$ *gilt und wenn ein $x^\star$ existiert mit $g_i(x^\star) < 0$, $i \in I$, $h_j(x^\star) = 0$, $j \in J$.*

Beweis. Für alle $j \in J$ gilt $\nabla h_j(x) = a_j$ (unabhängig von x.) •

Die folgenden beiden Ergebnisse werden in [13] bewiesen.

2.6.20 Satz *Die g_i, $i \in I$, seien konvex und C^1, die h_j, $j \in J$, seien linear, und es gelte $M \neq \emptyset$. Dann sind die folgenden Aussagen äquivalent:*

a) *MFB gilt irgendwo in M.*

b) *MFB gilt überall in M.*

c) *M erfüllt SB.*

2.6.21 Satz (Satz von Karush-Kuhn-Tucker für nichtlineare Probleme) *P sei C^1, und $\bar{x} \in M$ sei ein lokaler Minimalpunkt von P, an dem MFB gilt. Dann ist $\bar{x}$ KKT-Punkt von P.*

Wegen Übung 2.6.18 kann man in Satz 2.6.21 statt MFB auch die leichter überprüfbare, aber stärkere LUB voraussetzen.

2.6.22 Satz (Satz von KKT für konvexe Probleme)
P sei konvex und C^1, M erfülle SB, und $\bar{x} \in M$ sei ein Minimalpunkt von P. Dann ist $\bar{x}$ KKT-Punkt von P.

Beweis. Satz 2.6.20 und Satz 2.6.21. •

Satz 2.6.22 impliziert insbesondere, dass unter SB in M im Falle der Lösbarkeit von P die Dualitätslücke verschwindet.

Zusammengefasst haben wir für konvexe Optimierungsprobleme P folgende Zusammenhänge gezeigt:

1. Fall: P unrestringiert

$$\bar{x} \text{ kritischer Punkt} \quad \begin{array}{c} \text{S. 2.4.5} \\ \Rightarrow \\ \Leftarrow \\ \text{S. 2.4.2} \end{array} \quad \bar{x} \text{ globaler Minimalpunkt}$$

2. Fall: P restringiert

$$\bar{x} \text{ KKT-Punkt} \quad \begin{array}{c} \text{S. 2.6.12} \\ \Rightarrow \\ \Leftarrow \\ \textbf{SB}, \text{ S. 2.6.22} \end{array} \quad \bar{x} \text{ globaler Minimalpunkt}$$

Während man im unrestringierten Fall also die Charakterisierung globaler Minimalpunkte als kritische Punkte erhält (Kor. 2.4.6), benötigt man im restringierten Fall für einen Teil der Charakterisierung die Slaterbedingung.

2.6.23 Korollar (Charakterisierung globaler Minimalpunkte unter SB)
P sei konvex und C^1, und M erfülle die SB. Dann sind die globalen Minimalpunkte von P genau die KKT-Punkte von P.

Beweis. Satz 2.6.12 und Satz 2.6.22. •

Da die SB in Optimierungsproblemen sehr häufig erfüllt ist, ist sie keine besonders einschränkende Voraussetzung für die Charakterisierungsaussage in Korollar 2.6.23.

Die Situation vereinfacht sich, wenn sowohl die Gleichungs-, als auch die Ungleichungsrestriktionen sämtlich linear sind. In diesem Fall ist im Satz von Karush-Kuhn-Tucker keine Constraint Qualification nötig (s. [13]).

2.6.24 Satz (Satz von KKT für lineare Nebenbedingungen)
f sei C^1, die g_i, $i \in I$, h_j, $j \in J$, seien linear, und $\bar{x} \in M$ sei ein lokaler Minimalpunkt von P. Dann ist $\bar{x}$ KKT-Punkt von P.

2.6.25 Korollar (Charakterisierung globaler Minimalpunkte bei lin. NB'n)
f sei konvex und C^1, und die g_i, $i \in I$, h_j, $j \in J$, seien linear. Dann sind die globalen Minimalpunkte von P genau die KKT-Punkte von P.

Beweis. Satz 2.6.12 und Satz 2.6.24. •

2.6.26 Korollar (Charakterisierung globaler Minimalpunkte bei LP's)
P sei ein lineares Optimierungsproblem. Dann sind die globalen Minimalpunkte von P genau die KKT-Punkte von P.

Beweis. P erfüllt alle Voraussetzungen von Korollar 2.6.25, denn die lineare Zielfunktion f ist konvex und C^1. •

Auf Korollar 2.6.26 basiert unter anderem der Simplex-Algorithmus der linearen Optimierung, denn sein Abbruchkriterium ist genau dann erfüllt, wenn es einen KKT-Punkt identifiziert hat.

2.6.27 Beispiel *Wir betrachten das Problem, die Distanz eines beliebigen Punktes $z \in \mathbb{R}^2$ zur Menge*

$$M = \{x \in \mathbb{R}^2 \,|\, x_1^2 + x_2^2 \leq 1,\ x_2 \geq 0\}$$

zu bestimmen, also den Optimalwert des Projektionsproblems

$$P: \quad \min \|x - z\|_2 \quad s.t. \quad \begin{array}{rcl} x_1^2 + x_2^2 & \leq & 1 \\ x_2 & \geq & 0. \end{array}$$

Abbildung 2.17 illustriert die geometrische Situation.

Durch Quadrieren der Zielfunktion erhalten wir das konvexe C^1-Problem

$$Q: \quad \min x^\intercal x - 2z^\intercal x + z^\intercal z \quad s.t. \quad \begin{array}{rcl} g_1(x) & = & x_1^2 + x_2^2 - 1 \ \leq \ 0 \\ g_2(x) & = & -x_2 \ \leq \ 0, \end{array}$$

dessen zulässige Menge M die SB erfüllt. Die Distanz von z zu M (also der Minimalwert von P) stimmt mit der Wurzel aus dem Minimalwert von Q überein, und die globalen Minimalpunkte von P sind mit denen von Q identisch. Nach Korollar 2.6.23 sind die globalen Minimalpunkte von Q außerdem genau die KKT-Punkte von Q. Um die Distanz von z zu M zu berechnen, reicht es also, die Zielfunktion von P an den KKT-Punkten von Q auszuwerten.

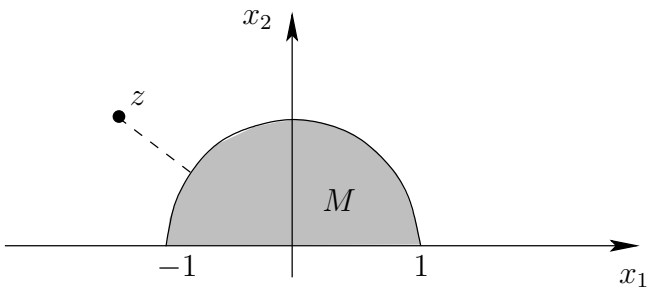


Abbildung 2.17: Ein Distanzproblem

Dazu geben wir zunächst die Gradienten der beteiligten Funktionen an:

$$\nabla f(x) = 2(x - z), \quad \nabla g_1(x) = 2x, \quad \nabla g_2(x) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Nun sei x ein KKT-Punkt von Q. In jedem Fall erfüllt x dann die beiden M definierenden Ungleichungen, und außerdem muss man zur Formulierung der KKT-Bedingungen entweder die Komplementaritätsbedingungen aufstellen oder aber eine Fallunterscheidung nach den möglichen aktiven Indexmengen durchführen. Da die Indexmenge $I = \{1, 2\}$ aus zwei Elementen besteht, besitzt sie vier verschiedene Teilmengen (allgemein bei p Elementen sind es 2^p Teilmengen). Wir betrachten jetzt diese Fallunterscheidungen.

1. Fall: $I_0(x) = \emptyset$

Das KKT-System lautet

$$\begin{aligned} 0 &= \nabla f(x) = 2(x - z), \\ 1 &> x^\intercal x, \\ 0 &< x_2 \,. \end{aligned}$$

Hieraus folgen die Gleichung $x = z$ sowie die Ungleichungen $z^\intercal z < 1$ und $z_2 > 0$. Damit ist auch klar, für welche z das KKT-System überhaupt lösbar ist, nämlich für

$$z \in M_< \;:=\; \left\{ x \in \mathbb{R}^n \middle|\; x_1^2 + x_2^2 < 1,\; x_2 > 0 \right\}.$$

Die Lösung und damit der globale Minimalpunkt von Q lautet dann natürlich

$$x^\star \;=\; z,$$

und der Minimalwert, also die Distanz von z zu M, beträgt offenbar null.

2.Fall: $I_0(x) = \{1\}$

Das KKT-System ist

$$\begin{aligned} 0 &= \nabla f(x) + \lambda_1 \nabla g_1(x) \;=\; 2(x - z) + \lambda_1 2x \\ 1 &= x^\intercal x \\ 0 &< x_2 \\ 0 &\leq \lambda_1 \,. \end{aligned}$$

Aus der ersten Gleichung folgt $(1+\lambda_1)x = z$. *Um die zweite Gleichung zu nutzen, bilden wir auf beiden Seiten das Skalarprodukt des Vektors mit sich selbst und erhalten*

$$(1+\lambda_1)^2 \underbrace{x^\intercal x}_{=1} \;=\; z^\intercal z \quad\Rightarrow\quad 1+\lambda_1 = \pm\|z\|_2 \quad\Rightarrow\quad \lambda_1 \;=\; \pm\|z\|_2 - 1.$$

Wegen $\lambda_1 \geq 0$ *kommt in dieser Bedingung nur die „+“-Alternative in Frage, und sie ist dann genau für* $\|z\|_2 \geq 1$ *lösbar. Um das zugehörige* x *zu bestimmen, setzen wir* $\lambda_1 = \|z\|_2 - 1$ *in die Gleichung* $(1+\lambda_1)x = z$ *ein und erhalten*

$$x^\star \;=\; \frac{z}{\|z\|_2}.$$

Aus der Bedingung $x_2^\star > 0$ *folgt nun noch* $z_2 > 0$.

Als Minimalwert von P *erhält man*

$$\|x^\star - z\|_2 \;=\; \left\| \left(\frac{1}{\|z\|_2} - 1 \right) z \right\|_2 \;=\; \left(1 - \frac{1}{\|z\|_2} \right) \|z\|_2 \;=\; \|z\|_2 - 1.$$

Zusammengefasst gilt: im Falle $\|z\|_2 \geq 1,\ z_2 > 0$ *ist* $x^\star = \frac{z}{\|z\|_2}$ *globaler Minimalpunkt mit* $\operatorname{dist}(z, M) = \|z\|_2 - 1$.

3.Fall: $I_0(x) = \{2\}$

Das KKT-System lautet

$$\begin{aligned} 0 &= \nabla f(x) + \lambda_2 \nabla g_2(x) \;=\; 2(x-z) + \lambda_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} \\ 1 &> x^\intercal x \\ 0 &= x_2 \\ 0 &\leq \lambda_2 \,. \end{aligned}$$

Aus den Gleichungen dieses Systems folgt $x_1 = z_1$, $x_2 = 0$ *sowie* $\lambda_2 = -2z_2$, *also insbesondere*

$$x^\star \;=\; \begin{pmatrix} z_1 \\ 0 \end{pmatrix}.$$

Der Minimalwert lautet damit $\|x^\star - z\| = \left\| \begin{pmatrix} 0 \\ -z_2 \end{pmatrix} \right\| = |z_2|$.

Die Ungleichungen geben Auskunft darüber, für welche z *das System lösbar ist: aus* $1 > (x^\star)^\intercal x^\star = z_1^2$ *folgt* $z_1 \in (-1, 1)$, *und aus* $0 \leq \lambda_2 = -2z_2$ *folgt* $z_2 \leq 0$.

Zusammengefasst erhalten wir für alle z *mit* $z_1 \in (-1, 1)$ *und* $z_2 \leq 0$ *den Minimalpunkt* $x^\star = \begin{pmatrix} z_1 \\ 0 \end{pmatrix}$ *mit* $\operatorname{dist}(z, M) = -z_2$.

4.Fall: $I_0(x) = \{1, 2\}$

Das KKT-System lautet

$$\begin{aligned} 0 &= \nabla f(x) + \lambda_1 \nabla g_1(x) + \lambda_2 \nabla g_2(x) \;=\; 2(x - z) + 2\lambda_1 x + \lambda_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} \\ 1 &= x^\intercal x \\ 0 &= x_2 \\ 0 &\leq \lambda_1 \\ 0 &\leq \lambda_2 \,. \end{aligned}$$

Aus der zweiten und dritten Gleichung folgen die beiden Lösungskandidaten

$$x^1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad x^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Für x^1 *ergibt die erste Gleichung*

$$z \;=\; (1 + \lambda_1) \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \frac{\lambda_2}{2} \begin{pmatrix} 0 \\ -1 \end{pmatrix} \;=\; -\begin{pmatrix} 1 + \lambda_1 \\ \frac{\lambda_2}{2} \end{pmatrix}$$

und damit $\lambda_1 = -z_1 - 1$ *und* $\lambda_2 = -2z_2$ *. Aus den Vorzeichenbedingungen and* λ_1 *und* λ_2 *folgt die Optimalität von* x^1 *für* $z_1 \leq -1$, $z_2 \leq 0$ *mit* $\operatorname{dist}(z, M) = \|x^1 - z\|_2 = \sqrt{(z_1 + 1)^2 + z_2^2}$ *.*

Analog überzeugt man sich von der Optimalität von x^2 *im Fall* $z_1 \geq 1$, $z_2 \leq 0$, *und zwar mit* $\operatorname{dist}(z, M) = \|x^2 - z\|_2 = \sqrt{(z_1 - 1)^2 + z_2^2}$ *.*

Als Zusammenfassung der Ergebnisse aller vier Fälle erhalten wir

$$\operatorname{dist}(z, M) \;=\; \begin{cases} 0, & \textit{falls} \;\; \|z\|_2 < 1, \;\; z_2 > 0 \\ \|z\|_2 - 1, & \textit{falls} \;\; \|z\|_2 \geq 1, \;\; z_2 > 0 \\ -z_2 \,, & \textit{falls} \;\; |z_1| < 1, \;\; z_2 \leq 0 \\ \sqrt{(z_1 + 1)^2 + z_2^2} \,, & \textit{falls} \;\; z_1 \leq -1, \;\; z_2 \leq 0 \\ \sqrt{(z_1 - 1)^2 + z_2^2} \,, & \textit{falls} \;\; z_1 \geq 1, \;\; z_2 \leq 0. \end{cases}$$

Abbildung 2.18 zeigt die Niveaulinien der Funktion $z \mapsto \operatorname{dist}(z, M)$.

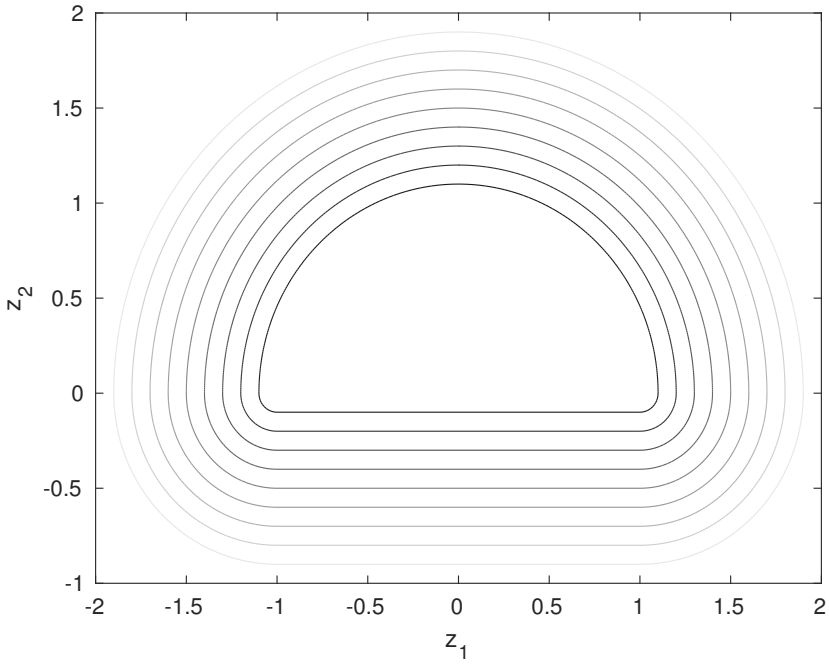


Abbildung 2.18: Niveaulinienbild von $\mathrm{dist}(z, M)$

2.7 Numerische Verfahren

Grundsätzlich lässt sich jedes Verfahren der nichtlinearen Optimierung auf konvexe Probleme anwenden, sofern die konvexen Probleme hinreichend glatt sind. Dabei gilt

- nach Satz 2.4.5 für jedes unrestringierte konvexe C^1-Problem P, dass jedes numerische Verfahren, das einen kritischen Punkt $x^\star$ erzeugt (d.h. $\nabla f(x^\star) = 0$), mit $x^\star$ auch einen globalen Minimalpunkt identifiziert (z.B. Gradientenverfahren, (Quasi-)Newtonverfahren, CG-Verfahren, Trust-Region-Verfahren, siehe [13]), und
- nach Satz 2.6.12 für jedes restringierte konvexe C^1-Problem P, dass jedes numerische Verfahren, das einen KKT-Punkt $x^\star$ erzeugt, mit $x^\star$ auch einen globalen Minimalpunkt identifiziert (z.B. Strafterm-, Barriere-, SQP-Verfahren, siehe [13]).

Konvexität verbessert dabei oft die Eigenschaften der Verfahren gegenüber dem nichtkonvexen Fall, etwa beim Newtonverfahren (s.u.) (*nicht* allerdings den Zigzagging-Effekt des Gradientenverfahrens).

Grundidee des Gradientenverfahrens

Wähle irgendeinen Startpunkt $x^0 \in \mathbb{R}^n$ sowie eine Toleranz $\varepsilon > 0$. Falls $\|\nabla f(x^0)\| \leq \varepsilon$, stopp ($x^0$ ist Approximation eines kritischen Punktes).

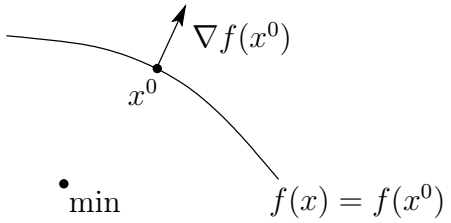


Abbildung 2.19: Gradient und Höhenlinie von f in x^0

Andernfalls nutzt man, dass $\nabla f(x^0)$ senkrecht auf der Niveaufläche $\{x \in \mathbb{R}^n |\ f(x) = f(x^0)\}$ steht und in Richtung des steilsten Anstiegs von f zeigt. Richtungen $d \in \mathbb{R}^n$ mit $\langle \nabla f(x^0), d\rangle > 0$ (spitzer Winkel) sind *Anstiegsrichtungen* (erster Ordnung) von f in x^0, Richtungen $d \in \mathbb{R}^n$ mit

$\langle \nabla f(x^0), d\rangle < 0$ (stumpfer Winkel) sind *Abstiegsrichtungen* (erster Ordnung). Zum Minimieren von f von x^0 aus kann man einen Schritt in die Richtung des steilsten Abstiegs, $d^0 = -\nabla f(x^0)$ unternehmen. Die Schrittweite $t^0 > 0$ wird zum Beispiel mit der Armijoregel bestimmt (s. [13]). Man setzt dann $x^1 = x^0 + t^0 d^0$ und prüft wieder $\|\nabla f(x^1)\| \leq \varepsilon$ usw.

Unter schwachen Voraussetzungen bricht das Gradientenverfahren nach endlich vielen Schritten ab. Dies kann wegen des Zigzagging-Effekts allerdings lange dauern, woran sich auch bei konvexen Funktionen nichts ändert.

Grundidee des Newtonverfahrens

Das Newtonverfahren ist zunächst ein Verfahren zur Nullstellensuche: zu lösen sei $g(x) = 0$ mit einer C^1-Funktion $g : \mathbb{R}^n \to \mathbb{R}^n$. Dazu wähle einen Startpunkt $x^0 \in \mathbb{R}^n$ und eine Toleranz $\varepsilon > 0$. Falls $\|g(x^0)\| \leq \varepsilon$, stopp ($x^0$ ist Approximation einer Nullstelle).

Ansonsten wird g um x^0 *linearisiert* und das linearisierte Problem gelöst. Für den Fall $n = 1$ ist das Vorgehen in Abbildung 2.20 illustriert. Linearisieren

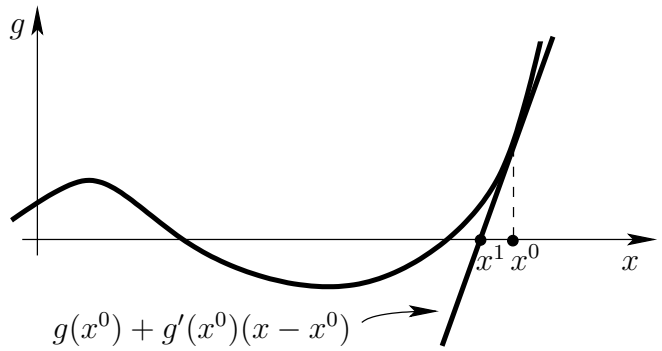


Abbildung 2.20: Newtonverfahren zur Nullstellensuche

von g bedeutet, dass die Tangente an g in x^0 berechnet wird, d.h g wird durch die Funktion $g(x^0) + g'(x^0)(x - x^0)$ approximiert. Lösen der Linearisierung bedeutet, dass eine Nullstelle der Tangente ermittelt wird: aus

$$0 \;=\; g(x^0) + g'(x^0)(x - x^0)$$

erhält man als neue Iterierte den Punkt

$$x^1 \;=\; x^0 - \frac{g(x^0)}{g'(x^0)}\,.$$

Dazu muss natürlich $g'(x^0) \neq 0$ gelten.

Für allgemeines n bedeutet die Nullstellensuche für die Linearisierung

$$0 \;=\; g(x^0) + \underbrace{Dg(x^0)}_{(n,n)\text{-Matrix}} (x - x^0),$$

woraus

$$x^1 \;=\; x^0 - (Dg(x^0))^{-1} g(x^0)$$

folgt. Dazu muss $Dg(x^0)$ invertierbar sein. Falls $\|g(x^1)\| \leq \varepsilon$, stopp usw.

Optimierungsprobleme löst man per Newtonverfahren, indem man einen kritischen Punkt sucht, d.h. das Nullstellenproblem $g(x) := \nabla f(x) = 0$ mit dem Newtonverfahren löst. In Iteration ν gilt dann

$$x^{\nu+1} \;=\; x^\nu - (D^2 f(x^\nu))^{-1} \nabla f(x^\nu).$$

Zur Suchrichtung $d^\nu = -(D^2 f(x^\nu))^{-1} \nabla f(x^\nu)$ kann man zusätzlich eine Schrittweite $t^\nu > 0$ bestimmen (z.B. per Armijoregel) und dann $x^{\nu+1} = x^\nu + t^\nu d^\nu$ setzen (man spricht dann vom *gedämpften* Newtonverfahren). Die Konvergenz dieses Verfahrens ist sehr schnell, falls x^0 nahe an einer Lösung liegt.

Nachteile des Newtonverfahrens sind im allgemeinen Fall:

- man weiß üblicherweise nicht, ob x^0 nahe genug an einem kritischen Punkt liegt,
- $D^2 f(x^\nu)$ ist nicht notwendigerweise invertierbar,
  (d.h. $D^2 f(x^\nu) d^\nu = -\nabla f(x^\nu)$ ist nicht notwendigerweise lösbar),
- d^ν ist nicht unbedingt eine Abstiegsrichtung, approximiert werden also beliebige kritische Punkte.

Der *letzte* Nachteil tritt bei konvexen Problemen nicht auf: für konvexes f gilt $D^2 f(x^\nu) \succeq 0$. Falls $D^2 f(x^\nu)$ invertierbar ist, muss demnach $D^2 f(x^\nu) \succ 0$ gelten. In der Linearen Algebra wird gezeigt, dass dann auch $(D^2 f(x^\nu))^{-1} \succ 0$ erfüllt ist. Für $\nabla f(x^\nu) \neq 0$ folgt

$$\langle \nabla f(x^\nu), d^\nu \rangle \;=\; -\nabla f(x^\nu)^\intercal (D^2 f(x^\nu))^{-1} \nabla f(x^\nu) \;<\; 0,$$

so dass d^ν Abstiegsrichtung erster Ordnung in x^ν ist. Damit kann man von der Konvergenz des Verfahrens gegen einen Minimalpunkt ausgehen. Ein alternatives Argument dafür ist, dass im konvexen Fall ohnehin alle kritischen Punkte globale Minimalpunkte sind, so dass die Konvergenz des Newtonverfahrens gegen einen kritischen Punkt die Konvergenz gegen einen globalen Minimalpunkt nach sich zieht.

Neben der Anwendung allgemeiner Verfahren der Nichtlinearen Optimierung auf konvexe Probleme existieren einige speziell auf Konvexität zugeschnittene Verfahren, die bis auf eine Ausnahme den allgemeinen Verfahren für kontinuierliche Optimierungsprobleme aber *nicht* überlegen sind. Für *nicht-glatte* konvexe Probleme gibt es eine Reihe von Verfahren (z.B. Subgradienten- und Bündelungsverfahren), auf die wir im Rahmen dieser Vorlesung nicht eingehen können (s. [1]).

Schnittebenenverfahren

Wir betrachten die Grundidee von Schnittebenenverfahren zunächst für unrestringierte Probleme.

Historisch lag der Grund zur Einführung von Schnittebenenverfahren darin, dass *lineare* Optimierungsprobleme seit Einführung des Simplex-Algorithmus gut lösbar waren, man also versuchen konnte, konvexe Probleme durch LPs zu approximieren.

Die Grundidee dazu besteht darin, dass der Graph einer konvexen C^1-Funktion f laut Satz 2.2.2 über jeder seiner Tangenten liegt (bzw. über jeder seiner Tangentialebenen, falls $n > 1$).

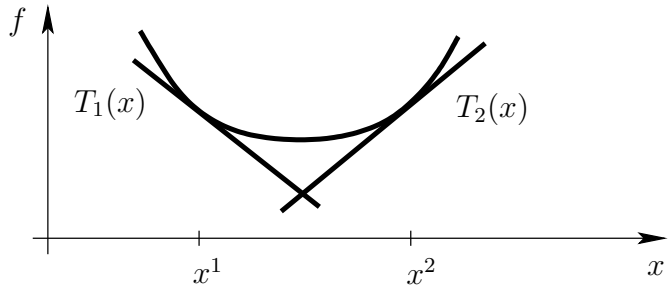


Abbildung 2.21: Idee des Schnittebenenverfahrens

Zu zwei gegebenen Punkten $x^1, x^2 \in \mathbb{R}^n$ wie in Abbildung 2.21 liegt der Graph von f sowohl über dem Graphen der Tangente $T_1(x) = f(x^1) + \langle \nabla f(x^1), x - x^1 \rangle$ als auch über dem von $T_2(x) = f(x^2) + \langle \nabla f(x^2), x - x^2 \rangle$. Für alle $x \in \mathbb{R}^n$ gilt also

$$f(x) \;\geq\; \max\{T_1(x),\, T_2(x)\}\,.$$

Analog kann man so für k Punkte $x^1, ..., x^k$ vorgehen, so dass f von unten stückweise linear approximiert wird. Statt f zu minimieren, löse jetzt das

approximierende Hilfsproblem

$$\min_x \max_{i=1,\dots,k} T_i(x).$$

Falls die Lösung x^{k+1} ein Abbruchkriterium (z.B. $\|\nabla f(x^{k+1})\| \leq \varepsilon$) erfüllt, stopp, ansonsten füge die Tangente

$$T_{k+1}(x) \;=\; f(x^{k+1}) + \langle \nabla f(x^{k+1}), x - x^{k+1} \rangle$$

zur Maximumsbildung hinzu und löse erneut.

Die Lösung des eigentlich nicht-glatten Hilfsproblems gelingt per Epigraph-Umformulierung (vgl. Ü. 1.3.7):

$$\begin{aligned}
\min_x \max_{i=1,\dots,k} T_i(x) \quad &\Leftrightarrow \quad \min_{x,\alpha} \alpha \quad \text{s.t.} \quad \max_{i=1,\dots,k} T_i(x) \leq \alpha \\
&\Leftrightarrow \quad \min_{x,\alpha} \alpha \quad \text{s.t.} \quad T_i(x) \leq \alpha,\ i = 1,\dots,k \\
&\Leftrightarrow \quad \min_{x,\alpha} \alpha \quad \text{s.t.} \quad \underbrace{f(x^i) + \langle \nabla f(x^i), x - x^i \rangle - \alpha}_{\text{linear in } (x,\alpha)} \leq 0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad i = 1,\dots,k.
\end{aligned}$$

Damit ist das approximierende Hilfsproblem zu einem äquivalenten LP umformuliert worden, das sich beispielsweise per Simplex-Algorithmus lösen lässt.

Im Hinblick auf den späteren Satz 3.5.2a) für nichtkonvexe Probleme halten wir an dieser Stelle fest, dass ein Optimalpunkt (x^{k+1}, α^{k+1}) des obigen LPs für den gesuchten Optimalwert $v = \min_{x \in \mathbb{R}^n} f(x)$ eine *Einschließung* liefert. Es gilt nämlich zum einen trivialerweise $v \leq f(x^{k+1})$, zum anderen aber auch $\alpha^{k+1} = \max_{i=1,\dots,k} T_i(x^{k+1}) \leq f(x)$ für alle $x \in \mathbb{R}^n$, also $\alpha^{k+1} \leq v$. Der Optimalwert v liegt also garantiert im Intervall $[\alpha^{k+1}, f(x^{k+1})]$.

Schnittebenenverfahren von Kelley (1960)

Das Schnittebenenverfahren von Kelley bedient sich ähnlicher Ideen, allerdings für restringierte Probleme der Form

$$\begin{array}{lllll} P: & \min c^\intercal x & \text{s.t.} & g_i(x) & \leq 0,\ i \in I\ (=\{1,...,p\}) \\ & & & Ax & \leq b \end{array}$$

mit konvexen und stetig differenzierbaren Funktionen g_i , $i \in I$.

Jedes konvexe Optimierungsproblem lässt sich in diese Form bringen, denn

- falls die Zielfunktion nichtlinear konvex ist, „verschiebe“ sie per Epigraph-Umformulierung in die Nebenbedingungen,
- falls Gleichungsnebenbedingungen vorliegen, sind diese linear und können daher als jeweils zwei lineare Ungleichungen geschrieben werden (Achtung: das macht man nur bei linearen Gleichungen, denn bei *nicht*-linearen Gleichungen zerstört diese Umformulierung wichtige Regularitätseigenschaften!).

Die Trennung zwischen linearen und nichtlinear konvexen Nebenbedingungen ist im Folgenden wichtig, da lineare Restriktionen nicht mehr linearisiert zu werden brauchen. Wir definieren daher

$$K \ = \ \{x \in \mathbb{R}^n |\ g_i(x) \leq 0,\ i \in I\}$$

und

$$L \ = \ \{x \in \mathbb{R}^n |\ Ax \leq b\}\,.$$

Offenbar gilt $M = K \cap L$. Im weiteren sei M nicht-leer und beschränkt (so dass P einen globalen Minimalpunkt besitzt).

Begleitendes Beispiel:

Zu lösen sei das Problem

$$\begin{array}{lllll} P: & \min x_2 & \text{s.t.} & x_1^2 + x_2^2 & \leq 1 \\ & & & x_2 & \geq e^{x_1} \\ & & & x_1 & \leq -\dfrac{1}{2} \\ & & & x_2 & \geq x_1\,. \end{array}$$

In diesem Problem gilt

$$\begin{aligned} c &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \\ g_1(x) &= x_1^2 + x_2^2 - 1, \\ g_2(x) &= e^{x_1} - x_2, \\ A &= \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}, \\ b &= \begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix}, \\ K &= \left\{ x \in \mathbb{R}^2 |\ x_1^2 + x_2^2 \leq 1,\ x_2 \geq e^{x_1} \right\}, \\ L &= \left\{ x \in \mathbb{R}^2 |\ x_1 \leq -\frac{1}{2},\ x_2 \geq x_1 \right\}. \end{aligned}$$

Die zulässige Menge von P ist in Abbildung 2.22 dargestellt.

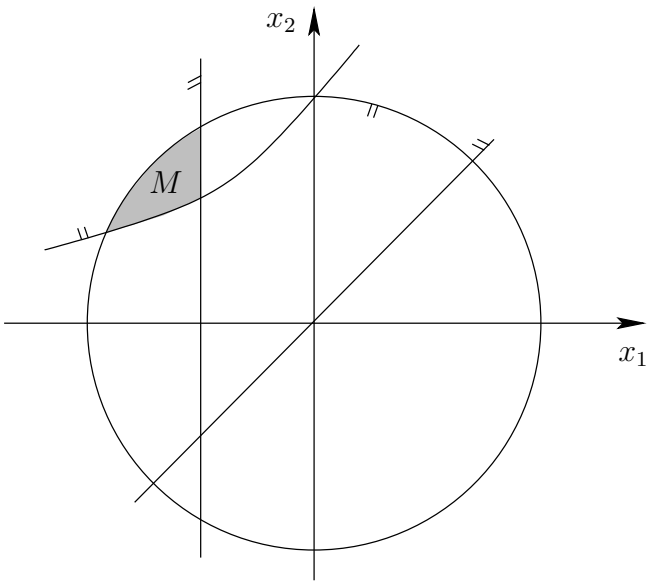


Abbildung 2.22: Zulässige Menge von P

Grundidee des Verfahrens ist es, M durch lineare Ungleichungen von außen zu approximieren und diese Approximation sukzessive zu verbessern. Da L ohnehin schon durch lineare Ungleichungen beschrieben ist, geschieht die Approximation nur für K.

Bestimme dazu zunächst ein konvexes Polyeder (d.h. eine durch Ungleichungen beschriebene Menge) B mit $K \subseteq B$ und setze $M^0 := B \cap L$. Dann ist M^0 ein konvexes Polyeder mit $M \subseteq M^0$. Wegen der Beschränktheit von M lässt sich B so wählen, dass M^0 ein konvexes Polytop wird, d.h. ein nicht-leeres und beschränktes konvexes Polyeder.

B darf eine sehr grobe Approximation von K sein, am einfachsten ist oft ein Quader (auch *Box* genannt, vgl. Kap. 3.3). Auch ohne geometrische Anschauung der Menge K lässt sich solch eine Box oft formal konstruieren. Im vorliegenden Fall kann man wie folgt vorgehen:

$$x \in K \quad \Rightarrow \quad 1 \;\geq\; x_1^2 + x_2^2 \;\geq\; x_1^2 \quad \Rightarrow \quad x_1 \in [-1,1]$$

und analog $x_2 \in [-1,1]$. Es folgt

$$K \subseteq B \;:=\; [-1,1]^2$$

und damit

$$M^0 \;=\; B \cap L \;=\; \left\{ x \in [-1,1]^2 \;\middle|\; x_1 \leq -\frac{1}{2},\; x_2 \geq x_1 \right\}.$$

Abbildung 2.23 zeigt die Menge M^0.

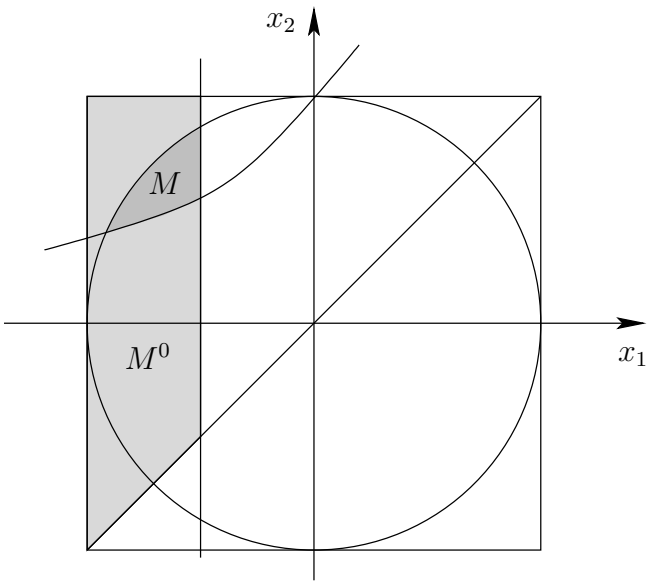


Abbildung 2.23: Startmenge M^0

Das Hilfsproblem

$$LP^0: \quad \min c^\intercal x \quad \text{s.t.} \quad x \in M^0$$

lässt sich zum Beispiel per Simplex-Algorithmus lösen. x^0 sei ein Minimalpunkt von LP^0. Wegen

$$c^\intercal x \ \geq \ c^\intercal x^0 \quad \forall\, x \in M^0$$

und $M \subseteq M^0$ folgt daraus

$$c^\intercal x \ \geq \ c^\intercal x^0 \quad \forall\, x \in M.$$

Falls also x^0 in M liegt, dann löst x^0 auch P.

Im Beispiel erhält man

$$x^0 \ = \ \begin{pmatrix} -1 \\ -1 \end{pmatrix} \notin M,$$

also ist x^0 nicht Lösung von P.

Im Fall $x^0 \notin M$ muss wegen $M^0 \subseteq L$ (mindestens) eine der Ungleichungen in K verletzt sein, d.h. es gilt $\max_{i \in I} g_i(x^0) > 0$. Wähle die (oder eine der) am meisten verletzten Ungleichungen, also ein $k \in I$ mit $g_k(x^0) = \max_{i \in I} g_i(x^0)$. Im Beispiel gilt $g_1(x^0) = 1$ und $g_2(x^0) = \frac{1}{e} + 1$, also $k = 2$.

Schnitt: Setze

$$M^1 \ = \ M^0 \cap \left\{ x \in \mathbb{R}^n \middle|\ g_k(x^0) + \langle \nabla g_k(x^0), x - x^0 \rangle \leq 0 \right\}.$$

Die neue Menge M^1 besitzt drei wichtige Eigenschaften:

1. $x^0 \notin M^1$, denn $g_k(x^0) + \langle \nabla g_k(x^0), x^0 - x^0 \rangle = g_k(x^0) > 0$, d.h. die alte Lösung x^0 wird „weggeschnitten“ und kann nicht noch einmal auftreten,

2. $M \subseteq M^1$, denn

$$\forall\, x \in M: \quad g_k(x^0) + \langle \nabla g_k(x^0), x - x^0 \rangle \overset{\text{S. 2.2.2}}{\leq} g_k(x) \ \leq \ 0,$$

3. M^1 ist wieder konvexes Polytop.

Im Beispiel erhält man die neue Ungleichung für M^1 wie folgt:

$$\begin{aligned} 0 \;\geq\; g_2(x^0) + \langle \nabla g_2(x^0), x - x^0 \rangle &= \frac{1}{e} + 1 + (\frac{1}{e}, -1) \begin{pmatrix} x_1 + 1 \\ x_2 + 1 \end{pmatrix} \\ &= \frac{1}{e} + 1 + \frac{x_1}{e} + \frac{1}{e} - x_2 - 1 \\ &= \frac{x_1}{e} - x_2 + \frac{2}{e}. \end{aligned}$$

In Abbildung 2.24 ist die Menge M^1 dargestellt.

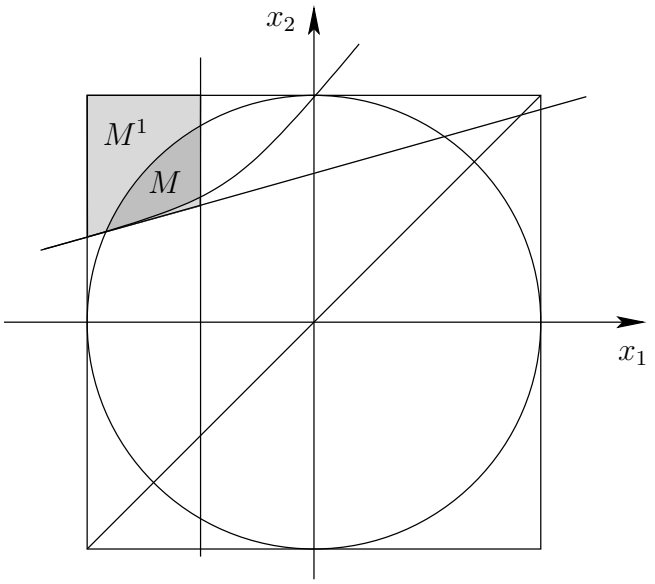


Abbildung 2.24: Menge M^1

Löse nun

$$LP^1: \quad \min c^\intercal x \quad \text{s.t.} \quad x \in M^1$$

usw. bis die Lösung x^ν eines Hilfsproblems LP^ν in M liegt. Wegen $M^\nu \supsetneq M$ für alle ν kann man allerdings nicht erwarten, dass irgendein x^ν tatsächlich in M liegt (alle Iterierten sind üblicherweise unzulässig, d.h. nicht in M), so dass man das Abbruchkriterium zu $\max_{i \in I} g_i(x^\nu) \leq \varepsilon$ mit einer Toleranz $\varepsilon > 0$ relaxiert und „fast zulässige“ Punkte akzeptiert (s. dazu auch Kap. 3.9). Das vollständige Verfahren ist in Algorithmus 2.1 angegeben.

Es ist klar, dass Algorithmus 2.1 eine Folge (x^ν) erzeugt, die sich M von außen nähert. Damit das Verfahren für eine beliebige Toleranz $\varepsilon > 0$ nach

Algorithmus 2.1: Schnittebenenverfahren von Kelley

Input : Stetig differenzierbares konvexes Minimierungsproblem P mit linearer Zielfunktion und nicht-leerer, beschränkter zulässiger Menge $M = K \cap L$.

Output : Approximation $\bar{x}$ eines globalen Minimalpunkts von P.

1 **begin**

2 Wähle ein konvexes Polyeder B mit $K \subseteq B$, so dass $M^0 := B \cap L$ konvexes Polytop ist.

3 Bestimme einen Minimalpunkt x^0 von

$$LP^0 : \quad \min c^\intercal x \quad \text{s.t.} \quad x \in M^0.$$

4 Wähle eine Toleranz $\varepsilon > 0$, und setze $\nu = 0$.

5 **while** $\max_{i \in I} g_i(x^\nu) > \varepsilon$ **do**

6 Wähle ein $k \in I$ mit $g_k(x^\nu) = \max_{i \in I} g_i(x^\nu)$.

7 Setze

$$M^{\nu+1} \;=\; M^\nu \cap \{x \in \mathbb{R}^n |\; g_k(x^\nu) + \langle \nabla g_k(x^\nu), x - x^\nu \rangle \leq 0\}\,.$$

8 Ersetze ν durch $\nu + 1$.

9 Bestimme einen Minimalpunkt x^ν von

$$LP^\nu : \quad \min c^\intercal x \quad \text{s.t.} \quad x \in M^\nu.$$

10 **end**

11 Setze $\bar{x} = x^\nu$.

12 **end**

endlich vielen Schritten abbricht, muss man noch ausschließen, dass die Folge (x^ν) einen positiven Abstand zu M behält.

2.7.1 Satz *Algorithmus 2.1 bricht nach endlich vielen Schritten ab.*

Beweis. Angenommen, der Algorithmus bricht nicht ab. Dann ist das Abbruchkriterium in Zeile 5 für kein $\nu \in \mathbb{N}$ erfüllt, d.h. das Verfahren erzeugt eine unendliche Folge (x^ν). Insbesondere wird in Zeile 6 unendlich oft ein $k_\nu \in I$ gewählt. Wegen $|I| < \infty$ tritt dabei mindestens ein Index $k \in I$ unendlich oft auf. Wir betrachten nun die Teilfolge der (x^ν), die zu diesen Schnittebenen gehören, und nennen sie wieder (x^ν).

Da alle x^ν in der kompakten Menge $B \cap L$ liegen, besitzt (x^ν) eine konvergente Teilfolge. OBdA sei bereits (x^ν) konvergent, und der Grenzwert sei $x^\star$.

Dann gilt $g_k(x^\nu) > \varepsilon$ für alle $\nu \in \mathbb{N}$ und daher nach dem Grenzübergang $g_k(x^\star) \geq \varepsilon$.

Wegen $x^{\nu+1} \in M^{\nu+1} \subseteq M^\nu$ (auch nach der Teilfolgenbildung) folgt

$$g_k(x^\nu) + \langle \nabla g_k(x^\nu), x^{\nu+1} - x^\nu \rangle \leq 0$$

und nach dem Grenzübergang $\nu \to \infty$

$$g_k(x^\star) + \underbrace{\langle \nabla g_k(x^\star), x^\star - x^\star \rangle}_{=0} \leq 0,$$

im Widerspruch zu $g_k(x^\star) \geq \varepsilon$. •

Verfahren von Frank-Wolfe (1956)

Das folgende Verfahren basiert auf einer grundlegend anderen Idee als Schnittebenenverfahren: es erzeugt eine Folge von zulässigen Iterierten (x^ν), bis x^ν eine *Optimalitätsbedingung* (ε-genau) erfüllt, während das Schnittebenenverfahren von Kelley aus der Optimalität im *Hilfs*problem die Optimalität im Originalproblem schließt, falls die Iterierte x^ν (ε-genau) zulässig ist.

Gegeben sei das Problem

$$P: \quad \min f(x) \quad \text{s.t.} \quad x \in M$$

mit nicht-leerer und konvexer zulässiger Menge M sowie konvexer Zielfunktion $f \in C^1(M, \mathbb{R})$.

2.7.2 Satz (Variationsformulierung konvexer Probleme)
Unter obigen Voraussetzungen an f und M gilt:

a) $\bar{x} \in \mathbb{R}^n$ ist genau dann globaler Minimalpunkt von P, wenn $\bar{x}$ globaler Minimalpunkt von

$$Q(\bar{x}): \quad \min_x \langle \nabla f(\bar{x}), x - \bar{x} \rangle \quad s.t. \quad x \in M$$

ist.

b) Es sei $\bar{x} \in M$. Dann besitzt $Q(\bar{x})$ den Optimalwert $v(\bar{x}) \leq 0$, und $\bar{x}$ ist globaler Minimalpunkt von P genau dann, wenn $v(\bar{x}) = 0$ gilt.

Beweis. Zum Beweis von Teil a) sei zunächst $\bar{x}$ ein globaler Minimalpunkt von $Q(\bar{x})$. Dann gilt insbesondere $\bar{x} \in M$, so dass $\bar{x}$ auch zulässig für P ist. Ferner hat man für alle $x \in M$

$$f(x) \overset{\text{S. 2.2.2}}{\geq} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \;\geq\; f(\bar{x}) + \langle \nabla f(\bar{x}), \bar{x} - \bar{x} \rangle \;=\; f(\bar{x}),$$

also ist $\bar{x}$ globaler Minimalpunkt von P.

Andererseits sei $\bar{x}$ ein globaler Minimalpunkt von P. Insbesondere ist $\bar{x}$ dann zulässig für $Q(\bar{x})$. Wir nehmen nun an, $\bar{x}$ sei kein globaler Minimalpunkt von $Q(\bar{x})$. Wegen $M \neq \emptyset$ existiert dann ein $y \in M$ mit

$$\langle \nabla f(\bar{x}), y - \bar{x} \rangle \;<\; \langle \nabla f(\bar{x}), \bar{x} - \bar{x} \rangle \;=\; 0,$$

Die Richtung $d := y - \bar{x}$ ist demnach Abstiegsrichtung (erster Ordnung) für f in $\bar{x}$, d.h. für hinreichend kleine $t > 0$ gilt $f(\bar{x} + td) < f(\bar{x})$. Außerdem verlässt man von $\bar{x}$ aus entlang d für kleine $t > 0$ nicht die zulässige Menge M, denn wegen $\bar{x}, y \in M$ und der Konvexität von M folgt für alle Schrittweiten $t \in [0, 1]$

$$\bar{x} + td \;=\; (1 - t)\bar{x} + ty \;\in\; M.$$

Für jedes hinreichend kleine $t > 0$ ist der Punkt $\bar{x} + td$ also zulässig für P und besitzt einen strikt kleineren Zielfunktionswert als $\bar{x}$. Dies widerspricht der Voraussetzung, $\bar{x}$ sei Minimalpunkt von P.

Die erste Behauptung von Teil b) folgt sofort aus $\bar{x} \in M$. Ferner folgt aus Teil a), dass jeder globale Minimalpunkt $\bar{x}$ von P auch globaler Minimalpunkt von $Q(\bar{x})$ ist, woraus $v(\bar{x}) = \langle \nabla f(\bar{x}), \bar{x} - \bar{x} \rangle = 0$ folgt. Falls andererseits $\bar{x} \in M$ kein globaler Minimalpunkt von P ist, dann ist $\bar{x}$ nach Teil a) auch kein globaler Minimalpunkt von $Q(\bar{x})$, es gibt wegen $M \neq \emptyset$ also ein $y \in M$ mit $\langle \nabla f(\bar{x}), y - \bar{x} \rangle < \langle \nabla f(\bar{x}), \bar{x} - \bar{x} \rangle = 0$. Wegen der Zulässigkeit von $\bar{x}$ für $Q(\bar{x})$ muss dann $v(\bar{x}) < 0$ gelten. •

Das Verfahren von Frank-Wolfe basiert auf der Lösung von Hilfsproblemen der Form $Q(\bar{x})$. Damit diese überhaupt lösbar sind, fordern wir im Folgenden, dass M nicht nur nicht-leer und konvex, sondern auch kompakt ist. Ferner wird das Verfahren nur dann numerisch interessant sein, wenn sich die Probleme $Q(\bar{x})$ schnell lösen lassen (also nicht z.B. durch Anwendung des Schnittebenenverfahrens von Kelley für jedes Hilfsproblem). Dies wird später auf weitere Voraussetzungen an M führen.

Der algorithmische Ansatz ist nun wie folgt. Wähle zunächst einen Startpunkt $x^0 \in M$ und bestimme den Minimalwert v^0 von $Q^0 := Q(x^0)$. Falls $v^0 = 0$, stopp, denn nach Satz 2.7.2b) ist x^0 globaler Minimalpunkt von P.

Da es eher unwahrscheinlich ist, dass man mit x^0 sofort einen globalen Minimalpunkt gefunden hat, ist die entscheidende Frage, wie man im Fall $v^0 < 0$ weiter vorgeht.

Dazu sei zu v^0 auch ein Minimalpunkt y^0 von Q^0 bestimmt. Wie im Beweis zu Satz 2.7.2a) ist die Suchrichtung $d^0 := y^0 - x^0$ dann eine zulässige Abstiegsrichtung für f in x^0, einen zulässigen Punkt mit kleinerem Zielfunktionswert findet man also auf jeden Fall in der Form $x^0 + td^0$ mit einem passenden (d.h. nicht zu großen) $t \in [0, 1]$.

Zur *Schrittweitensteuerung* (also zur Wahl der Schrittweite t) geht man folgendermaßen vor: wähle t^0 als Minimalpunkt von $f(x^0 + td^0)$ auf $[0, 1]$, d.h. löse das eindimensionale Problem

$$S(x^0, d^0): \quad \min_t \underbrace{f(x^0 + td^0)}_{\text{konvex in } t} \quad \text{s.t.} \quad 0 \le t \le 1.$$

Auch $S(x^0, d^0)$ muss zur numerischen Umsetzung schnell lösbar sein, oder man gibt sich mit einer Approximation an t^0 zufrieden.

Setze nun als neue Iterierte $x^1 := x^0 + t^0 d^0$ und löse Q^1 usw. Das Abbruchkriterium muss wieder relaxiert werden. Abbildung 2.25 zeigt die ersten zwei Iterationen des Verfahrens an einem Beispiel.

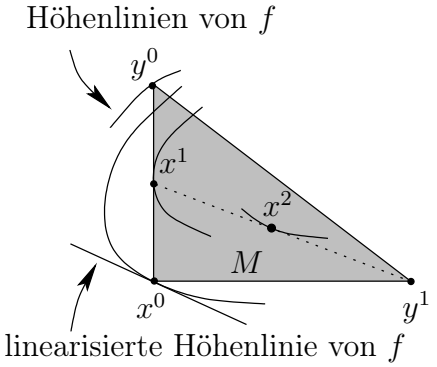


Abbildung 2.25: Beispiel für das Verfahren von Frank-Wolfe

Das vollständige Verfahren ist in Algorithmus 2.2 angegeben. Um es numerisch sinnvoll einsetzen zu können, müssen die Hilfsprobleme Q^ν und $S(x^\nu, d^\nu)$ schnell lösbar sein.

Algorithmus 2.2: Verfahren von Frank-Wolfe

Input : Konvexes Minimierungsproblem P mit stetig differenzierbarer Zielfunktion und nicht-leerer, kompakter zulässiger Menge M.

Output : Approximation $\bar{x}$ eines globalen Minimalpunkts von P (falls das Verfahren terminiert, vgl. S. 2.7.4).

begin
  Wähle einen Startpunkt $x^0 \in M$, eine Toleranz $\varepsilon > 0$, und setze $\nu = 0$.
  Bestimme einen Minimalpunkt y^0 und den Minimalwert v^0 von

$$Q^0: \quad \min_x \langle \nabla f(x^0), x - x^0 \rangle \quad \text{s.t.} \quad x \in M.$$

  while $v^\nu < -\varepsilon$ **do**
    Setze $d^\nu = y^\nu - x^\nu$.
    Wähle t^ν als Minimalpunkt von

$$S(x^\nu, d^\nu): \quad \min_t f(x^\nu + t d^\nu) \quad \text{s.t.} \quad 0 \le t \le 1.$$

    Setze $x^{\nu+1} = x^\nu + t^\nu d^\nu$.
    Ersetze ν durch $\nu + 1$.
    Bestimme einen Minimalpunkt y^ν und den Minimalwert v^ν von

$$Q^\nu: \quad \min_x \langle \nabla f(x^\nu), x - x^\nu \rangle \quad \text{s.t.} \quad x \in M.$$

  end
  Setze $\bar{x} = x^\nu$.
end

Für Q^ν sind dazu beispielsweise folgende Voraussetzungen geeignet:

- falls M ein konvexes Polytop ist (d.h. nur durch lineare Gleichungen und Ungleichungen beschrieben), dann ist Q^ν ein LP und damit etwa per Simplex-Algorithmus lösbar,

- falls M eine Kugel ist, dann lässt sich y^ν sogar in geschlossener Form berechnen.

Für $S(x^\nu, d^\nu)$ gilt:

- falls f konvex-quadratisch ist, dann ist t^ν in geschlossener Form berechenbar.

Daher wurde Algorithmus 2.2 im Jahr 1956 für konvex-quadratische Zielfunktionen mit linearen Nebenbedingungen formuliert.

2.7.3 Übung *Mit $A = A^\intercal \succ 0$ und $b \in \mathbb{R}^n$ sei $f(x) = \frac{1}{2}x^\intercal Ax + b^\intercal x$. Geben Sie einen geschlossenen Ausdruck für den eindeutigen Minimalpunkt t^ν von $S(x^\nu, d^\nu)$ an.*

2.7.4 Satz *∇f sei Lipschitz-stetig auf der nicht-leeren, konvexen und kompakten Menge M (vgl. Def. 3.9.1). Dann bricht Algorithmus 2.2 nach endlich vielen Schritten ab.*

Beweisskizze. Angenommen, der Algorithmus bricht nicht ab. Dann gilt $v^\nu < -\varepsilon$ für alle $\nu \in \mathbb{N}$, und es werden unendliche Folgen (x^ν), (y^ν) und (t^ν) erzeugt. Da die Folgen (x^ν) und (y^ν) in der kompakten Menge M enthalten sind, dürfen wir sie ohne Beschränkung der Allgemeinheit (d.h. nach eventuellem Übergang zu einer Teilfolge) als konvergent annehmen.

Durch die Schrittweitenwahl in Zeile 6 ist (t^ν) *effiziente* Schrittweitenfolge (s. [13]), d.h.

$$\exists\, c > 0 \quad \forall\, \nu \in \mathbb{N}: \quad f(x^\nu + t^\nu d^\nu) - f(x^\nu) \;\leq\; -c \left(\frac{\langle \nabla f(x^\nu), d^\nu \rangle}{\|d^\nu\|_2} \right)^2$$

(dies ist langwierig zu zeigen, und hier ist die Lipschitz-Stetigkeit von ∇f erforderlich). Es folgt

$$\underbrace{f(x^{\nu+1}) - f(x^\nu)}_{<0,\ \to 0} \;\leq\; -c \left(\frac{\langle \nabla f(x^\nu), d^\nu \rangle}{\|d^\nu\|_2} \right)^2 \leq 0$$

und daher per Sandwich-Theorem

$$\frac{\langle \nabla f(x^\nu), d^\nu \rangle}{\|d^\nu\|_2} \;\overset{\nu\to\infty}{\to}\; 0.$$

Der Zähler dieses Bruchs ist wegen

$$\langle \nabla f(x^\nu), d^\nu \rangle \;=\; \langle \nabla f(x^\nu), y^\nu - x^\nu \rangle \;=\; v^\nu \;<\; -\varepsilon$$

beschränkt, so dass der Nenner gegen unendlich gehen muss:

$$\|d^\nu\| \;\to\; \infty.$$

Dies ist jedoch ein Widerspruch, da x^ν und y^ν aus der beschränkten Menge M stammen und damit auch die Folge der $d^\nu = y^\nu - x^\nu$ beschränkt ist. •

Da im Beweis zu Satz 2.7.4 nur die Effizienz der Schrittweitenfolge (t^ν) benutzt wird und nicht die Tatsache, dass die t^ν exakte globale Minimalpunkte von $S(x^\nu, d^\nu)$ sind, gilt die Konvergenzaussage auch noch, wenn $S(x^\nu, t^\nu)$ nur inexakt (z.B. per Armijoregel) gelöst wird (s. [13]).

Primal-duale Innere-Punkte-Methoden

Gegeben sei das konvexe C^1-Problem

$$P: \quad \min\, f(x) \quad \text{s.t.} \quad g_i(x) \le 0,\; i \in I,$$

mit beschränkter zulässiger Menge M sowie

$$M_< \;:=\; \{x \in \mathbb{R}^n |\; g_i(x) < 0,\; i \in I\} \;\neq\; \emptyset$$

(d.h. M besitzt Slaterpunkte). Der Fall $J \neq \emptyset$ kann auch betrachtet werden, wird es hier der Übersichtlichkeit halber aber nicht.

Die primal-dualen Innere-Punkte-Methoden basieren auf einem rein primalen Verfahren, nämlich dem *Barriereverfahren* (siehe auch [13]). Seine Grundidee ist die Approximation von P durch unrestringierte Probleme, wobei am Rand $\mathrm{bd}M$ von M eine „Barriere“ errichtet wird, die die Zulässigkeit erzwingt.

Beispiel:

Für das Problem

$$\min -x \quad \text{s.t.} \quad x \le 0$$

ist $\beta(x) = -\log(-x)$ eine *Barrierefunktion*, d.h. $\beta : M_< \to \mathbb{R}$, und für alle $(x^\nu) \in M_<$ mit $\lim_\nu x^\nu = x^\star \in \mathrm{bd}M_<$ gilt $\lim_\nu \beta(x^\nu) = +\infty$. Wegen der Beschränktheit von M bedeutet dies gerade, dass β als koerziv auf $M_<$ vorausgesetzt wird. Abbildung 2.26 illustriert die Funktion β.

Die Strategie des Barriereverfahrens besteht darin, die Nähe zum Rand von M sukzessive weniger zu bestrafen, d.h. den Barriereparameter t gegen null streben zu lassen und dabei Lösungen von

$$\min_x\, B(t,x) \;:=\; -x - t\log(-x)$$

zu verfolgen (s. Abb. 2.27).

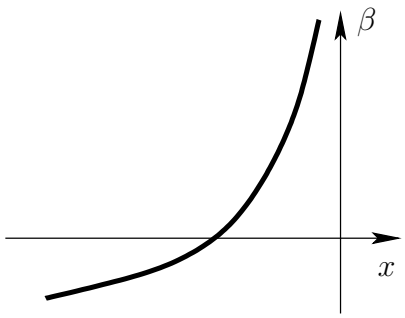


Abbildung 2.26: Logarithmische Barrierefunktion β

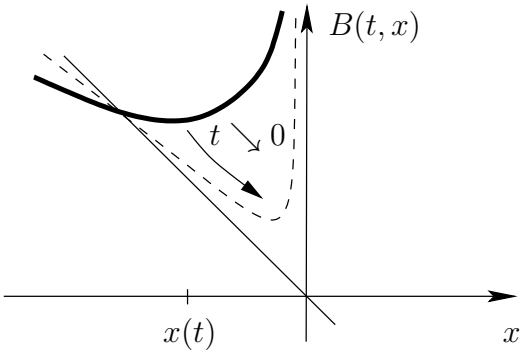


Abbildung 2.27: Probleme $B(t,x)$ mit $t \searrow 0$

Für allgemeine zulässige Mengen M von P ist

$$\beta(x) \;=\; -\sum_{i\in I} \log(-g_i(x))$$

Barrierefunktion, und die Hilfsprobleme

$$\min_x \; B(t,x) \;=\; f(x) - t\sum_{i\in I} \log(-g_i(x))$$

sind für $t \searrow 0$ zu lösen.

2.7.5 Satz *Für alle $t > 0$ ist $B(t,x)$ konvex in x und besitzt einen eindeutigen Minimalpunkt.*

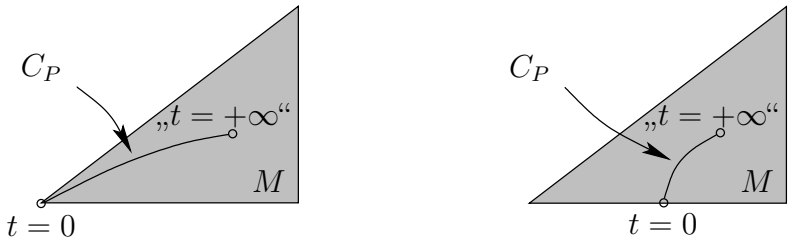
Beweis. Die Konvexität von $B(t,x)$ in x folgt aus der Konvexität und Monotonie von $-\log(-x)$. Die Beschränktheit von M impliziert Koerzivität von

$B(t,x)$ (in x) auf $M_<$, so dass die Existenz eines Minimalpunkts durch Korollar 1.2.40 garantiert ist. Seine Eindeutigkeit folgt mit Satz 2.3.3b) aus der strikten Konvexität der Funktion $-\log(-x)$, denn mit ihr ist auch $B(t,x)$ strikt konvex. •

Zu $t > 0$ bezeichne $x(t)$ den eindeutigen Minimalpunkt von $B(t,x)$. Die Menge

$$C_P \;=\; \{x(t)|\; t \in (0,+\infty)\}$$

heißt *primaler zentraler Pfad* von P. Abbildung 2.28 illustriert primale zentrale Pfade für zwei lineare Optimierungsprobleme.



$\min x_1$ s.t. $x \in M$ $\qquad$ $\min x_2$ s.t. $x \in M$

Abbildung 2.28: Primale zentrale Pfade

Unter schwachen Voraussetzungen ist $x^\star = \lim_{t\to 0} x(t)$ ein globaler Minimalpunkt von P (s. [13]). Das Problem bei der numerischen Umsetzung dieses Ansatzes besteht darin, dass die Funktion $B(t,x)$ für $t \searrow 0$ in der Nähe von $\mathrm{bd}M$ stark gekrümmt, „numerisch geknickt“ und damit schwer minimierbar ist. Dies spiegelt sich auch im Optimalitätskriterium wider: $x(t)$ ist die eindeutige Lösung von

$$\begin{aligned} 0 \;&=\; \nabla_x B(t,x) \;=\; \nabla_x \left(f(x) - t \sum_{i\in I} \log(-g_i(x)) \right) \\ &=\; \nabla f(x) - t \sum_{i\in I} \frac{1}{-g_i(x)} (-\nabla g_i(x)) \\ &=\; \nabla f(x) + \sum_{i\in I} \left(-\frac{t}{g_i(x)} \right) \nabla g_i(x). \end{aligned}$$

Für $x(t) \to x^\star \in \mathrm{bd}M$ gibt es mindestens ein $i \in I$ mit $g_i(x(t)) \to 0$, also ist $\lim_{t\to 0}(-\frac{t}{g_i(x(t))})$ „vom Typ $\frac{0}{0}$“ und die Optimalitätsbedingung damit für t nahe null „schlecht konditioniert“.

Als Ausweg macht man sich die Ähnlichkeit der obigen Optimalitätsbedingung zu den KKT-Bedingungen zu nutze und setzt für alle $t > 0$

$$\lambda_i(t) \;:=\; -\frac{t}{g_i(x(t))}, \quad i \in I.$$

Wegen $x(t) \in M_<$ gilt $g_i(x(t)) < 0$ für alle $i \in I$, und wegen $t > 0$ folgt daraus $\lambda_i(t) > 0$. Also ist $x(t)$ genau dann Minimalpunkt von $B(t, x)$, wenn ein $\lambda(t)$ existiert, so dass $(x(t), \lambda(t))$ das folgende System von Gleichungen und Ungleichungen löst:

$$\begin{aligned} \nabla f(x) + \sum_{i \in I} \lambda_i \nabla g_i(x) &= 0 \\ \lambda_i \, g_i(x) &= -t \\ \lambda_i &> 0 \\ g_i(x) &< 0, \quad i \in I. \end{aligned}$$

Dies ist gerade das KKT-System von P mit durch $-t$ gestörter Komplementaritätsbedingung! Die Lösung $(x(t), \lambda(t))$ ist für alle $t > 0$ eindeutig, $x(t)$ ist „primal innerer Punkt“, und $\lambda(t)$ ist „dual innerer Punkt“. Die Menge

$$C_{PD} \;=\; \{(x(t), \lambda(t)) |\; t \in (0, +\infty)\}$$

heißt *primal-dualer zentraler Pfad* von P.

Für $t \searrow 0$ geht das gestörte in das ungestörte KKT-System über, d.h. bei Konvergenz $\lim_{t \searrow 0}(x(t), \lambda(t)) = (x^\star, \lambda^\star)$ ist $x^\star$ KKT-Punkt von P mit Multiplikator $\lambda^\star$ und damit globaler Minimalpunkt. Entscheidender Vorteil gegenüber dem Barriereverfahren ist, dass nun für $t \searrow 0$ keine numerischen Probleme auftreten.

Geschickte Implementierungen von primal-dualen Innere-Punkte-Methoden lösen die Hilfsprobleme für große t nur grob, werden aber für $t \searrow 0$ immer exakter. Außerdem bestimmen die Verfahren selbst die Anpassung von t.

Dies kann sogar so geschehen, dass der Rechenaufwand zur Identifizierung einer ε-genauen Lösung selbst im worst case nur polynomial in der Problemdimension anwächst. Da dies insbesondere für lineare Optimierungsprobleme gilt, sind Innere-Punkte-Methoden in dieser Hinsicht dem Simplex-Algorithmus überlegen, der im worst case exponentiellen Rechenaufwand besitzt. Bei hochdimensionalen Problemen lässt sich diese Überlegenheit tatsächlich beobachten, so dass kommerzielle Softwarepakete solche Probleme nicht per Simplex-Algorithmus lösen, sondern mit primal-dualen Innere-Punkte-Methoden.

Von den in diesem Abschnitt behandelten speziell auf Konvexität zugeschnittenen Verfahren sind nur die primal-dualen Innere-Punkte-Methoden in der modernen Optimierung von praktischer Relevanz. Lösungsverfahren für allgemeine nichtlineare Optimierungsprobleme (s. [13]) sind heute so weit entwickelt, dass sie dem Schnittebenenverfahren von Kelley und dem Verfahren von Frank-Wolfe im allgemeinen auch für konvexe Probleme überlegen sind. Es ist dennoch sinnvoll, diese Verfahren zu kennen, denn ihre Grundgedanken tauchen in veränderter Form zum Beispiel in der (gemischt-)ganzzahligen nichtlinearen Optimierung (s. [11]) und bei der globalen Minimierung nichtkonvexer Funktionen wieder auf, dem Inhalt des nächsten Abschnitts.

Kapitel 3

Nichtkonvexe Optimierungsprobleme

3.1 Beispiele

In der Praxis sind Optimierungsprobleme häufig nicht konvex. Einfache Beispiele sind das Projektionsproblem (Bsp. 1.1.1) mit nichtkonvexer Menge M, das Problem der Cluster-Analyse (Bsp. 1.1.6) sowie die folgenden Probleme, die sogar mit Hilfe konvexer Funktionen gestellt sind.

3.1.1 Beispiel (Identifikation redundanter Ungleichungen)
Es sei $M = \{x \in \mathbb{R}^n | g_i(x) \leq 0,\ i \in I\} \neq \emptyset$ *mit konvexen Funktionen* g_i, $i \in I$. *Die Restriktion* $g_k(x) \leq 0$ *ist sicherlich dann redundant (d.h.* M *ändert sich nicht, wenn man die Bedingung* $g_k(x) \leq 0$ *weglässt), falls der Maximalwert* v *von*

$$R_k: \quad \max_x g_k(x) \quad s.t. \quad x \in M$$

negativ ist. R_k *ist kein konvexes Optimierungsproblem, sofern* g_k *nicht linear ist.*

In Abbildung 3.1 ist die Menge M *durch drei lineare Ungleichungen sowie*

$$g_4(x) \ = \ x_1^2 + x_2^2 - 1 \ \leq \ 0$$

beschrieben. Offenbar vergrößert man M *durch Weglassen von* $g_4(x) \leq 0$, *d.h.* g_4 *ist nicht redundant.*

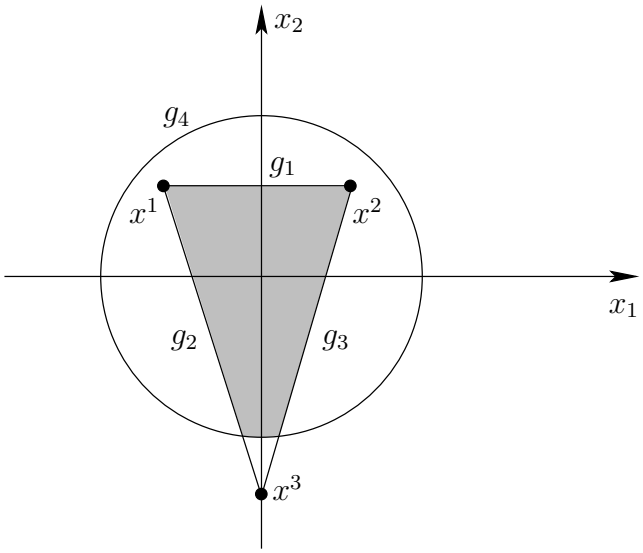


Abbildung 3.1: Identifikation von Redundanz als globales Problem

Die Punkte x^1 und x^2 sind lokale Maximalpunkte von R_4 mit $g_4(x^1) < 0$ und $g_4(x^2) < 0$. Falls man mit lokalen Suchverfahren nur x^1 und x^2 findet, aber nicht die Randpunkte x von M, an denen g_4 aktiv ist, so schließt man fälschlich $v < 0$, also Redundanz der Restriktion $g_4(x) \leq 0$.

3.1.2 Beispiel (Verifizierung von Konvexität)
Es seien $a, b \in \mathbb{R}$, $a < b$, $M = [a, b]$ und $f \in C^2(M, \mathbb{R})$ gegeben. Laut C^2-Charakterisierung ist f genau dann konvex auf der (volldimensionalen) Menge M, wenn $f''(x) \geq 0$ für alle $x \in M$ gilt. Zur Überprüfung des Kriteriums berechne den Minimalwert v von

$$K: \quad \min_x f''(x) \quad s.t. \quad x \in M$$

und teste auf $v \geq 0$. Das Optimierungsproblem K ist nicht konvex, denn selbst für konvexes f braucht f'' nicht konvex zu sein.

In Abbildung 3.2 sind x^1 und x^2 lokale Minimalpunkte von K mit $f''(x^1) > 0$ und $f''(x^2) < 0$. Falls man nur x^1 findet, schließt man fälschlich auf $v \geq 0$, also auf Konvexität von f.

Während im Projektionsproblem und in der Clusteranalyse eventuell auch gute lokale Lösungen akzeptabel sind, ist man in den Beispielen 3.1.1 und 3.1.2 auf die Bestimmung der globalen Lösung angewiesen. Allerdings benötigt man im ersten Fall auch ein Maß für „gute“ lokale Lösungen.

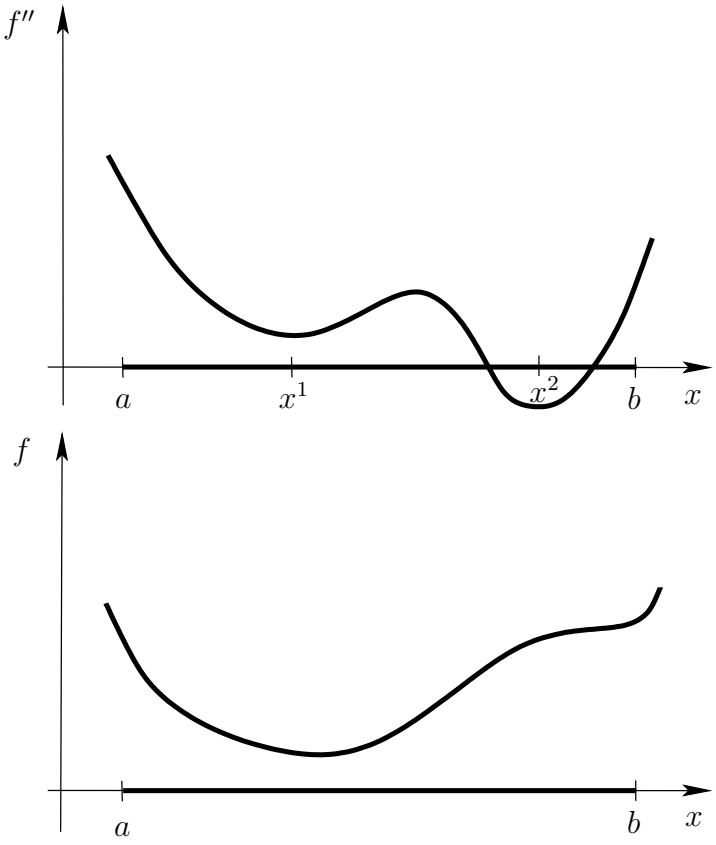


Abbildung 3.2: Verifizierung von Konvexität

Im vorliegenden Kapitel 3 werden wir sehen, dass beide Probleme sich in vielen Fällen mit gewissem Aufwand lösen lassen. Dazu werden wir im Folgenden stetig differenzierbare restringierte Optimierungsprobleme betrachten, also

$$P: \quad \min f(x) \quad \text{s.t.} \quad g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J,$$

mit stetig differenzierbaren Funktionen f, g_i, $i \in I$, h_j, $j \in J$, die nicht notwendigerweise konvex sind. Häufig werden wir außerdem die Lösbarkeit von P voraussetzen, was etwa per Satz von Weierstraß oder seinen Variationen aus Kapitel 1.2 zu überprüfen ist.

Zunächst sind wir durch Satz 2.6.21 in der Lage, den konzeptionellen Algorithmus 3.1 zur restringierten nichtlinearen globalen Minimierung anzugeben. Er benutzt folgende äquivalente Umformulierung der Aussage von Satz 2.6.21: An jedem lokalen Minimalpunkt $\bar{x}$ von P ist *entweder* MFB verletzt *oder* MFB ist erfüllt, und gleichzeitig ist $\bar{x}$ KKT-Punkt. Die Ver-

letzung von MFB tritt nur in Ausnahmefällen auf und wird daher auch als *degenerierter* Fall bezeichnet, was die Bezeichnung DEG in Algorithmus 3.1 erklärt.

Algorithmus 3.1: Konzeptioneller Algorithmus zur restringierten nichtlinearen globalen Minimierung

Input : Lösbares stetig differenzierbares restringiertes Optimierungsproblem P.
Output : Globaler Minimalpunkt $x^\star$ von P.

1 **begin**
2 Bestimme die Menge DEG der Punkte in M, an denen MFB verletzt ist.
3 Bestimme unter den Punkten in M, an denen MFB erfüllt ist, die Menge KKT aller KKT-Punkte.
4 Bestimme einen Minimalpunkt $x^\star$ von f in $DEG \cup KKT$.
5 **end**

Die Menge $DEG \cup KKT$ ist typischerweise nicht nur sehr viel kleiner als M, sondern sogar endlich. Dann ist in Zeile 4 die Minimierung von f auf $DEG \cup KKT$ durch den Vergleich endlich vieler Funktionswerte lösbar.

3.1.3 Beispiel *Abbildung 3.3 zeigt eine durch drei C^1-Ungleichungen beschriebene zulässige Menge M sowie Höhenlinien einer konkav-quadratischen Funktion f. Wegen der Spitze von M in x^3 sei dort MFB verletzt.*

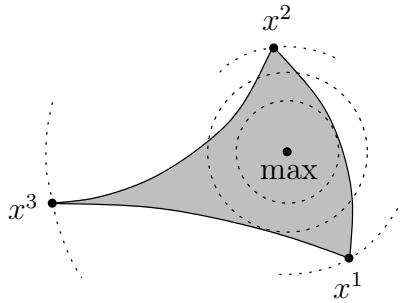


Abbildung 3.3: Kandidaten für globale Minimalpunkte

Zur Minimierung von f über M liefert Algorithmus 3.1 $KKT = \{x^1, x^2\}$ und $DEG = \{x^3\}$. Folglich liegt $x^\star = x^3$ in diesem Fall nicht in KKT, sondern in DEG.

Der Algorithmus 3.1 leidet unter folgendem grundlegenden Problem: sofern P nicht „sehr übersichtlich“ ist, kann man sich selbst bei endlichen Mengen DEG und KKT nicht sicher sein, ob man *alle* ihre Elemente bestimmt hat. Falls DEG und KKT aber nur unvollständig berechnet werden, besteht die Gefahr, dass die globalen Minimalpunkte übersehen werden.

Eine erheblich tragfähigere Lösungsstrategie besteht in der Ausnutzung der Tatsache, dass sich immerhin *glatte konvexe* Probleme mit aktuellen Optimierungsverfahren „leicht“ lösen lassen (s. Kap. 2). Daher geht man analog zu den Algorithmen 2.1 und 2.2 vor, in denen die zur Zeit der Publikation der Verfahren noch schwer lösbaren *konvexen* Probleme durch leicht lösbare *lineare* Probleme approximiert werden. Heute ist man hingegen in der Lage schwer lösbare *nichtkonvexe* Probleme durch leicht lösbare *konvexe* Probleme zu approximieren.

3.2 Konvexe Relaxierung

3.2.1 Definition (Konvex relaxierte Menge)
Es sei $\emptyset \neq M \subseteq \mathbb{R}^n$.

a) Jede konvexe Menge $\widehat{M} \subseteq \mathbb{R}^n$ *mit* $M \subseteq \widehat{M}$ *heißt* konvexe Relaxierung *von* M.

b) Der Durchschnitt aller konvexen Relaxierungen von M,

$$\widehat{\widehat{M}} := \bigcap \left\{ \widehat{M} \,\middle|\, \widehat{M} \supseteq M,\ \widehat{M} \text{ konvex} \right\}$$

heißt konvexe Hülle *von* M.

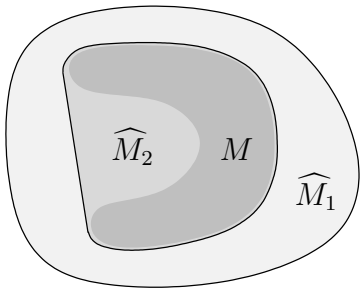


Abbildung 3.4: Konvexe Relaxierungen einer Menge

In Abbildung 3.4 sind $\widehat{M}_1$ und $\widehat{M}_2$ konvexe Relaxierungen von M, und es gilt $\widehat{M}_2 = \widehat{\widehat{M}}$.

Die konvexe Hülle ist die kleinstmögliche konvexe Relaxierung von M. Wegen der Konvexität von $\widehat{M} = \mathbb{R}^n$ und $M \subseteq \widehat{M}$ existiert sie für alle M. Außerdem ist sie eindeutig bestimmt.

3.2.2 Definition (Konvex relaxierte Funktion)
Es seien $\emptyset \neq X \subseteq \mathbb{R}^n$ konvex und $f : X \to \mathbb{R}$ gegeben.

a) Jede auf X konvexe Funktion $\widehat{f}$ mit

$$\forall\, x \in X : \ \widehat{f}(x) \ \leq \ f(x)$$

heißt konvexe Relaxierung *von f auf X.*

b) Eine konvexe Relaxierung $\widehat{\widehat{f}}$ von f auf X, die für alle anderen konvexen Relaxierungen $\widehat{f}$ von f auf X

$$\forall\, x \in X : \ \widehat{f}(x) \ \leq \ \widehat{\widehat{f}}(x)$$

erfüllt, heißt konvexe Hüllfunktion *von f auf X.*

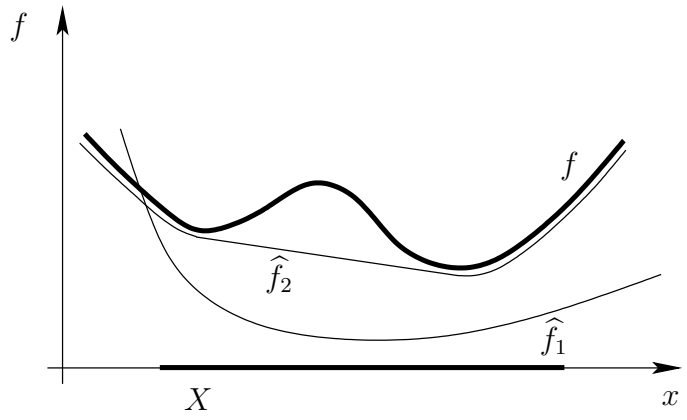


Abbildung 3.5: Konvexe Relaxierungen einer Funktion

In Abbildung 3.5 sind $\widehat{f}_1$ und $\widehat{f}_2$ konvexe Relaxierungen von f auf X, und es gilt $\widehat{f}_2 = \widehat{\widehat{f}}$.

Achtung: Im Gegensatz zu Mengen besitzt nicht jede Funktion f auf jeder Menge X eine konvexe Relaxierung. Zum Beispiel besitzt $f(x) = -x^2$ *keine* konvexe Relaxierung auf $X = \mathbb{R}$.

3.2.3 Satz *Es seien $\emptyset \neq X \subseteq \mathbb{R}^n$ konvex (z.B. $X = \mathbb{R}^n$), $g_i : X \to \mathbb{R}$, $i \in I$, beliebig und $M = \{x \in X |\ g_i(x) \leq 0,\ i \in I\}$.*

a) Falls für jedes $i \in I$ die Funktion $\widehat{g}_i$ eine konvexe Relaxierung von g_i auf X ist, dann ist die Menge

$$\widehat{M} \ = \ \{x \in X |\ \widehat{g}_i(x) \leq 0,\ i \in I\}$$

eine konvexe Relaxierung von M.

b) Selbst wenn für jedes $i \in I$ die Funktion $\widehat{\widehat{g}}_i$ die konvexe Hüllfunktion von g_i auf X ist, kann trotzdem

$$\widehat{\widehat{M}} \ \neq \ \left\{x \in X \,\middle|\, \widehat{\widehat{g}}_i(x) \leq 0,\ i \in I\right\}$$

gelten.

Beweis.

a) Aus der Konvexität der Funktionen $\widehat{g}_i$, $i \in I$, sowie der Menge X folgt die Konvexität von $\widehat{M}$. Nun sei $x \in M$. Dann gilt für alle $i \in I$

$$\widehat{g}_i(x) \ \leq \ g_i(x) \ \leq \ 0,$$

also $x \ \in \ \widehat{M}$, so dass $\widehat{M}$ auch eine Relaxierung von M ist.

b) Beispielsweise für $n = |I| = 1$, $X = [0, 2]$ und $g(x) = 1 - x^2$ gilt $M = [1, 2]$, $\widehat{\widehat{g}}(x) = 1 - 2x$ und $\{x \in X |\ \widehat{\widehat{g}}(x) \leq 0\} = [\frac{1}{2}, 2]$, aber $\widehat{\widehat{M}} = M = [1, 2]$.

•

Satz 3.2.3a) gibt die übliche Methode an, konvexe Relaxierungen von Mengen zu konstruieren. Satz 3.2.3b) zeigt, dass sich diese Methode allerdings nicht auf die Konstruktion von konvexen Hüllen übertragen lässt.

3.2.4 Definition (Konvex relaxiertes Optimierungsproblem)
Für $\emptyset \neq X \subseteq \mathbb{R}^n$ konvex seien $f : X \to \mathbb{R}$ und $M \subseteq X$ beliebig sowie das Optimierungsproblem

$$P: \quad \min f(x) \quad s.t. \quad x \in M$$

gegeben.

a) $\widehat{f}$ sei eine konvexe Relaxierung von f auf X, und $\widehat{M} \subseteq X$ sei eine konvexe Relaxierung von M. Dann heißt

$$\widehat{P}: \quad \min \widehat{f}(x) \quad s.t. \quad x \in \widehat{M}$$

konvexe Relaxierung *von P (auf X).*

b) $\widehat{\widehat{f}}$ sei die konvexe Hüllfunktion von f auf X, und $\widehat{\widehat{M}}$ sei die konvexe Hülle von M. Dann heißt

$$\widehat{\widehat{P}}: \quad \min \widehat{\widehat{f}}(x) \quad s.t. \quad x \in \widehat{\widehat{M}}$$

konvexes Hüllproblem *von P (auf X).*

Beachten Sie, dass in Definition 3.2.4a) zwar $\widehat{M} \subseteq X$ gefordert werden muss, damit $\widehat{f}$ auf $\widehat{M}$ definiert ist, dass aber in Definition 3.2.4b) die Bedingung $\widehat{\widehat{M}} \subseteq X$ automatisch gilt, weil X selbst eine konvexe Relaxierung von M darstellt.

3.2.5 Satz *Mit den Voraussetzungen aus Definition 3.2.4 seien alle im Folgenden auftretenden Optimierungsprobleme P, $\widehat{P}$ und $\widehat{\widehat{P}}$ lösbar mit globalen Minimalwerten v, $\widehat{v}$ bzw. $\widehat{\widehat{v}}$.*

a) Für den Minimalwert $\widehat{v}$ jeder konvexen Relaxierung $\widehat{P}$ von P gilt $\widehat{v} \leq v$.

b) Für den Minimalwert $\widehat{\widehat{v}}$ des konvexen Hüllproblems $\widehat{\widehat{P}}$ von P und den Minimalwert $\widehat{v}$ jeder konvexen Relaxierung $\widehat{P}$ gilt $\widehat{v} \leq \widehat{\widehat{v}} \leq v$ (d.h $\widehat{\widehat{v}}$ ist die beste per konvexer Relaxierung erzielbare Unterschranke von v).

c) Es gilt nicht notwendigerweise $\widehat{\widehat{v}} = v$.

d) Falls M konvex ist, gilt $\widehat{\widehat{v}} = v$.

e) Falls f linear ist, gilt $\widehat{\widehat{v}} = v$.

Beweis. a) Es gilt

$$\widehat{v} \;=\; \min_{x\in\widehat{M}} \widehat{f}(x) \;\overset{M\subseteq\widehat{M}}{\leq}\; \min_{x\in M} \widehat{f}(x) \;\overset{M\subseteq X}{\leq}\; \min_{x\in M} f(x) \;=\; v.$$

b) Übung.

c) Beispielsweise für $n = 1$, $X = [-2, 2]$, $M = [-2, -1] \cup [1, 2]$ und $f(x) = x^2$ gilt $v = 1$, aber wegen $\widehat{\widehat{M}} = X$ und $\widehat{\widehat{f}}(x) = x^2$ erhalten wir $\widehat{\widehat{v}} = 0 < 1 = v$.

d) Wegen Teil b) ist nur $v \leq \widehat{\widehat{v}}$ zu zeigen. Dazu stellen wir fest, dass die konstante Funktion $\widehat{f}(x) = v$ eine konvexe Relaxierung von f auf der konvexen Menge M ist. Wegen $M \subseteq X$ gilt für die konvexe Hüllfunktion $\widehat{\widehat{f}}$ von f auf X also

$$\forall\, x \in M :\; v \;=\; \widehat{f}(x) \;\leq\; \widehat{\widehat{f}}(x)$$

und damit $v \leq \min_{x\in M} \widehat{\widehat{f}}(x)$. Die Konvexität von M impliziert schließlich $\widehat{\widehat{M}} = M$, also

$$v \;\leq\; \min_{x\in M} \widehat{\widehat{f}}(x) \;=\; \min_{x\in\widehat{\widehat{M}}} \widehat{\widehat{f}}(x) \;=\; \widehat{\widehat{v}}.$$

e) Mit passendem $c \in \mathbb{R}^n$ und $d \in \mathbb{R}$ sei $f(x) = c^\intercal x + d$. Wegen Teil b) ist wieder nur $v \leq \widehat{\widehat{v}}$ zu zeigen. In der Tat gilt nach Definition des Minimalwerts v

$$\forall\; x \in M :\quad v \;\leq\; f(x) \;=\; c^\intercal x + d,$$

so dass die Menge $\widehat{M} := \{x \in \mathbb{R}^n \,|\, v \leq c^\intercal x + d\}$ eine konvexe Relaxierung von M bildet (hier geht nur die *Konkavität* von f ein). Dass die konvexe Hülle $\widehat{\widehat{M}}$ von M insbesondere in $\widehat{M}$ enthalten ist, kann man explizit ausschreiben zu

$$\forall\; x \in \widehat{\widehat{M}} :\quad v \;\leq\; c^\intercal x + d \;=\; f(x) \;=\; \widehat{\widehat{f}}(x),$$

wobei wir für die letzte Gleichung die *Konvexität* von f benutzt haben, und woraus sofort $v \leq \widehat{\widehat{v}}$ folgt. •

Der Grund für die nicht garantierte Gleichheit von $\widehat{\widehat{v}}$ und v in Satz 3.2.5c) liegt darin, dass $\widehat{\widehat{f}}$ auf $\widehat{\widehat{M}} \setminus M$ bessere Werte annehmen kann als f auf M.

Zumindest dieser Effekt lässt sich durch eine vorgeschaltete Epigraphumformulierung von P umgehen:

3.2.6 Übung *Unter den Voraussetzungen von Satz 3.2.5 stimmt v mit dem Optimalwert des konvexen Problems*

$$\min_{x,\alpha} \alpha \quad s.t. \quad (x,\alpha) \in \widehat{\widehat{\operatorname{epi}}}\,(f,M)$$

überein.

Erinnert man sich aus Satz 3.2.3b) allerdings daran, dass im Allgemeinen noch nicht einmal eine funktionale Darstellung von $\widehat{\widehat{M}}$ durch die konvexen Hüllfunktionen $\widehat{\widehat{g}}_i$, $i \in I$, möglich ist, und damit auch keine funktionale Darstellung des Epigraphen aus Übung 3.2.6, besteht im Folgenden kein Anlass, tatsächlich konvexe Hüllprobleme zu bestimmen. Stattdessen werden wir nur noch mit konvexen Relaxierungen arbeiten.

Für die Relation $\widehat{v} \leq v$ in Satz 3.2.5a) spielt die *Konvexität* der Relaxierung $\widehat{P}$ von P übrigens keine Rolle, aber unter Konvexität ist $\widehat{v}$ mit den Methoden aus Kapitel 2 numerisch *berechenbar*.

Zu klären ist im nächsten Schritt, ob und wie man konvexe Relaxierungen von Funktionen auf Mengen numerisch konstruieren kann.

3.3 Intervallarithmetik

Motivation und Anwendungen

Grundidee der Intervallarithmetik ist die Bestimmung schnell berechenbarer Ober- und Unterschranken für Funktionswerte.

Am Beispiel der Berechnung von e^π wird der Grundgedanke deutlich: die ungenaue Berechnung der Form

$$\pi \approx 3.141 \;\Rightarrow\; e^\pi \;\approx\; e^{3.141} \;\approx\; 23.14$$

birgt gewisse Unsicherheiten über das Ergebnis. Es ist beispielsweise nicht klar, wie weit 23.14 tatsächlich vom gewünschten Wert e^π entfernt ist.

Eine bessere Aussage erhält man wie folgt (vgl. Abb. 3.6):

$$\begin{array}{rcl} \pi & \in & [3.141, 3.142] \\ e^{3.141} & > & 23.127, \quad e^{3.142} < 23.15 \quad (\text{„Rundung nach außen“}) \\ \Rightarrow\; e^\pi & \in & [23.127, 23.15]\,. \end{array}$$

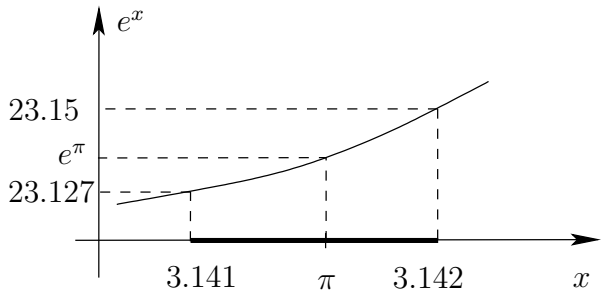


Abbildung 3.6: Berechnung von e^π

Allgemeiner ist für $a \leq b$ und eine stetige Funktion $f : [a, b] \to \mathbb{R}$ die Bildmenge $\text{bild}(f, [a, b]) := \{f(x) |\; x \in [a, b]\}$ immer ein nicht-leeres und kompaktes Intervall (aufgrund des Zwischenwertsatzes und des Satzes von Weierstraß). Die ebenfalls gebräuchliche Bezeichnung $f([a, b])$ für $\text{bild}(f, [a, b])$ vermeiden wir hier, da sie später im Rahmen der Intervallarithmetik zu Verwirrung führen würde.

Gesucht ist ein möglichst kleines Intervall $[c, d]$ mit $\text{bild}(f, [a, b]) \subseteq [c, d]$, d.h. (vgl. Abb. 3.7)

$$\forall\, x \in [a, b]: \quad c \;\leq\; f(x) \;\leq d.$$

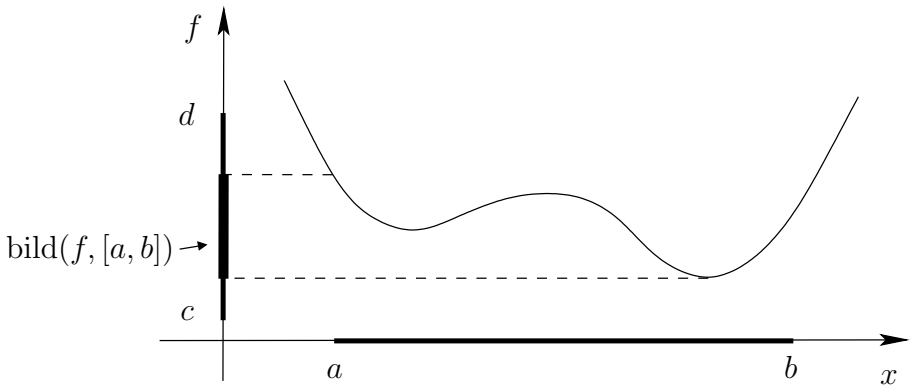


Abbildung 3.7: Einschließung der Bildmenge $\text{bild}(f, [a, b])$

Offenbar gilt $c \leq \min\limits_{x\in[a,b]} f(x)$ und $d \geq \max\limits_{x\in[a,b]} f(x)$ d.h. die Bestimmung von c und d beinhaltet zwei globale Optimierungsprobleme. Mit Methoden der Intervallarithmetik kann man die Lösung dieser Probleme *umgehen*, erhält allerdings oft nur grobe Schranken c und d.

Bei der numerischen Berechnung von c und d ist häufig eine Rundung erforderlich. Zur Sicherheit benutzt man dann nicht die übliche Rundungsregel, sondern „Runden nach außen“, d.h. c wird nach unten, d nach oben gerundet. In diesem Fall spricht man von „sicherer Intervallarithmetik“. Unter Nutzung der üblichen Rundung spricht man von „unsicherer Intervallarithmetik“.

Neben der Behandlung von Rundungsfehlern sind andere Anwendungen der Intervallarithmetik zum Beispiel:

- die Toleranzanalyse in technischen und physikalischen Systemen,
- die verlässliche Lösung von Gleichungen,
- der Computerbeweis der Keplerschen Vermutung (1611/1998/2003/2014).

Intervallgrundrechenarten

Die Intervallarithmetik basiert darauf, zunächst die Grundrechenarten von Zahlen auf Intervalle zu übertragen. Dazu führen wir folgende Notation ein:

$$\mathbb{R} \;=\; \{[a,b] \,|\, a,b \in \mathbb{R},\; a \leq b\}$$

bezeichnet die Menge aller nicht-leeren und kompakten Intervalle in $\mathbb{R}$. Elemente von $\mathbb{R}$ werden mit Großbuchstaben und die Intervallgrenzen mit den entsprechenden Kleinbuchstaben wie folgt bezeichnet:

$$X \in \mathbb{R} \;\Rightarrow\; X = [\underline{x}, \overline{x}]\,.$$

Für $\underline{x}, \overline{x} \in \mathbb{R}^n$ ist

$$X \;=\; [\underline{x}, \overline{x}] \;=\; \{x \in \mathbb{R}^n |\; \underline{x} \leq x \leq \overline{x}\}$$

ein nicht-leeres und kompaktes n-dimensionales Intervall, kurz *Box* genannt. $\mathbb{R}^n$ bezeichnet die Menge aller Boxen. Alternativ kann man eine Box auch per kartesischem Produkt als $X = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_n, \overline{x}_n]$ schreiben.

Wir leiten im Folgenden „natürliche“ Definitionen der Grundrechenarten auf $\mathbb{R}$ her. Die Systematik dafür besteht darin, die Bildmengen der entsprechenden Operationen für „Intervall-Inputs“ auszurechnen. Formal aufgeschrieben bestimmt man also für eine (stetige) Operation $f : \mathbb{R}^2 \to \mathbb{R},\ (x,y) \mapsto f(x,y)$, (wie $f(x,y) = x + y$) eine entsprechende Operation $F : \mathbb{R}^2 \to \mathbb{R}$ durch $(X,Y) \mapsto \text{bild}(f, X \times Y)$, wobei die Menge $\text{bild}(f, X \times Y)$ explizit ausgerechnet wird. Da f stetig ist und $\mathbb{R}^2$ nur aus nicht-leeren und kompakten Mengen besteht, garantieren der Zwischenwertsatz sowie der Satz von Weierstraß wieder, dass F tatsächlich nach $\mathbb{R}$ abbildet, als Output also stets ein nicht-leeres und kompaktes Intervall liefert.

Addition

Für $X, Y \in \mathbb{R}$ wird die (Minkowski-)Summe $X + Y$ als Menge aller auftretenden Summen von Elementen aus X und Y definiert:

$$X + Y \;:=\; \{x + y |\; x \in X,\; y \in Y\}\,.$$

In die obige Systematik ordnet sich dies durch die Wahl $f(x,y) = x + y$ und $F(X,Y) = \text{bild}(f, X \times Y) = \{x + y |\, (x,y) \in X \times Y\}$ ein.

Die Darstellung von $X + Y$ muss nun noch explizit gemacht werden: mit $X = [\underline{x}, \overline{x}]$ und $Y = [\underline{y}, \overline{y}]$ gilt

$$X + Y \;=\; \left[\underline{x} + \underline{y}, \overline{x} + \overline{y}\right],$$

denn für alle $x \in X$ und $y \in Y$ ergibt die Addition der beiden Ungleichungen $\underline{x} \leq x \leq \overline{x}$ und $\underline{y} \leq y \leq \overline{y}$ die neue Ungleichung

$$\underline{x} + \underline{y} \ \leq \ x + y \ \leq \overline{x} + \overline{y},$$

woraus $X + Y \subseteq [\underline{x} + \underline{y}, \overline{x} + \overline{y}])$ folgt.

Da außerdem die Punkte $\underline{x} + \underline{y}$ und $\overline{x} + \overline{y}$ in $X + Y$ liegen und $X + Y$ als Intervall eine konvexe Menge ist, folgt die Behauptung. Insgesamt erhält man also die Definition

$$[\underline{x}, \overline{x}] \ + \ \left[\underline{y}, \overline{y}\right] \ = \ \left[\underline{x} + \underline{y}, \overline{x} + \overline{y}\right].$$

Man macht sich leicht klar, dass obige Formel ebenfalls für $X, Y \in \mathbb{IR}^n$ gilt, womit auch die Addition von Boxen jeder Dimension definiert ist.

Addition mit einem Vektor

Für $x \in \mathbb{R}^n$ und $Y = [\underline{y}, \overline{y}] \in \mathbb{IR}^n$ ist es naheliegend, die Addition

$$x \ + \ \left[\underline{y}, \overline{y}\right] \ = \ \left[x + \underline{y}, x + \overline{y}\right]$$

zu definieren. In der Tat liefert obige Systematik dieses Ergebnis. Ein alternatives Vorgehen wäre es, den Vektor x mit der einpunktigen Box $[x, x]$ zu *identifizieren* und dann die obige Rechenregel für zwei Intervalle anzuwenden. Aus Gründen der Übersichtlichkeit werden wir im Folgenden aber auf diese Identifizierung durchgängig *verzichten*.

Die Definitionen der restlichen arithmetischen Operationen werden auf dieselbe systematische Weise hergeleitet.

Subtraktion

Für $X, Y \in \mathbb{IR}^n$ setzen wir

$$-X \ := \ \{-x |\ x \in X\} \ = \ [-\overline{x}, -\underline{x}]$$

und

$$X - Y \ := \ X + (-Y) \ = \ \left[\underline{x} - \overline{y}, \overline{x} - \underline{y}\right].$$

Multiplikation

Für $X, Y \in \mathbb{IR}$ setzen wir

$$X \cdot Y \;:=\; \{x\,y |\; x \in X,\; y \in Y\} \;=\; \square\left\{\underline{x}\,\underline{y},\, \underline{x}\,\overline{y},\, \overline{x}\,\underline{y},\, \overline{x}\,\overline{y}\right\},$$

wobei wir für eine beschränkte Menge reeller Zahlen $A \subseteq \mathbb{R}$ die in der Literatur zur Intervallarithmetik gebräuchliche *Intervallhülle*

$$\square A \;:=\; [\inf A,\, \sup A]$$

von A einführen. Beachten Sie allerdings, dass für die in unseren Anwendungen nur benötigten *endlichen* Mengen $A \subseteq \mathbb{R}$ die Identität $\square A = \mathrm{conv}(A)$ gilt.

Mangels einer Multiplikationsvorschrift für Vektoren $x, y \in \mathbb{R}^n$, die ein Ergebnis in $\mathbb{R}^n$ liefert, definieren wir entsprechend auch keine Multiplikation von Boxen $X, Y \in \mathbb{IR}^n$. Selbstverständlich ließe sich eine Intervall-Entsprechung des Skalarproduktes definieren (Übung), die man aber nicht als arithmetische Operation bezeichnen würde.

Multiplikation mit einem Skalar

Für $x \in \mathbb{R}$ mit $x \geq 0$ und $Y = [\underline{y}, \overline{y}] \in \mathbb{IR}^n$ definieren wir

$$x \cdot \left[\underline{y}, \overline{y}\right] \;:=\; \left[x\underline{y}, x\overline{y}\right]$$

und für $x < 0$

$$x \cdot \left[\underline{y}, \overline{y}\right] \;:=\; (-x) \cdot (-Y) \;=\; \left[x\overline{y}, x\underline{y}\right].$$

Division

Für $X \in \mathbb{IR}$ mit $0 \notin X$ setzen wir

$$\frac{1}{X} \;:=\; \left\{\frac{1}{x}\middle|\; x \in X\right\} \;=\; \left[\frac{1}{\overline{x}}, \frac{1}{\underline{x}}\right]$$

und für $X, Y \in \mathbb{IR}$ mit $0 \notin Y$

$$\frac{X}{Y} \;:=\; X \cdot \left(\frac{1}{Y}\right) \;=\; \square\left\{\frac{\underline{x}}{\overline{y}}, \frac{\underline{x}}{\underline{y}}, \frac{\overline{x}}{\overline{y}}, \frac{\overline{x}}{\underline{y}}\right\}.$$

Nachdem nun die Grundrechenarten für Intervalle definiert sind, halten wir zunächst fest, dass sich *nicht alle Rechenregeln aus* $\mathbb{R}$ *auf Intervalle übertragen*. Beispielsweise gilt für $X \in \mathbb{IR}$ im Allgemeinen weder $X - X = [0, 0]$

noch $X/X = [1,1]$, sondern ist nur für $\underline{x} = \overline{x}$ wahr. Auch liefert $X \cdot X$ nicht für jedes $X \in \mathbb{IR}$ das Bild von X unter der Funktion $f(x) = x^2$. Dies liegt am später erklärten *Abhängigkeitseffekt*, der immer dann auftritt, wenn *dieselbe* Intervallvariable *mehrfach* in Rechenvorschriften auftaucht.

Mit Hilfe der Grundrechenarten lassen sich *rationale intervallwertige Funktionen* definieren, d.h. Funktionen $F : \mathbb{IR}^n \to \mathbb{IR}$, in deren Funktionsvorschrift nur die Intervall-Grundrechenarten auftreten.

3.3.1 Beispiel *Für* $F : \mathbb{IR}^2 \to \mathbb{IR}$, $F(X_1, X_2) = ([1,2]X_1 + [0,1])X_2$ *und* $X_1 = X_2 = [0,1]$ *sei* $F(X_1, X_2)$ *zu berechnen. Dies geschieht schrittweise wie folgt:*

$$\begin{aligned} V_1 &= [1,2]X_1 = [1,2][0,1] = \square\{0,1,2\} = [0,2] \\ V_2 &= V_1 + [0,1] = [0,2] + [0,1] = [0,3] \\ F(X_1,X_2) &= V_2 X_2 = [0,3][0,1] = [0,3]. \end{aligned}$$

Zusätzlich zu den Grundrechenarten lassen sich auch elementare (d.h. in der Softwareumgebung vordefinierte) Funktionen f in natürlicher Weise zu intervallwertigen Funktionen erweitern, nach unserer obigen Systematik nämlich wieder durch die explizite Berechnung von $F(X) := \mathrm{bild}(f, X)$. Beispielsweise liefert die Monotonie der Funktionen exp und log sofort

$$\mathrm{EXP}(X) := [\exp(\underline{x}), \exp(\overline{x})]$$

für $X \in \mathbb{IR}$ und

$$\mathrm{LOG}(X) := [\log(\underline{x}), \log(\overline{x})]$$

für $X \in \mathbb{IR}$ mit $\underline{x} > 0$. Wegen der mangelnden Monotonie der Funktion sin ist die Berechnung von $\mathrm{SIN}(X)$ für $X \in \mathbb{IR}$ als $\mathrm{bild}(\sin, X)$ weniger einfach, durch passende Fallunterscheidungen aber trotzdem durchführbar (Übung).

Natürliche Intervallerweiterung

Für die folgende Definition sei daran erinnert, dass für $f : \mathbb{R}^k \to \mathbb{R}$ und $g : \mathbb{R}^n \to \mathbb{R}^k$ die Funktion $f \circ g : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto (f \circ g)(x) := f(g(x))$ *Komposition* von f und g genannt wird.

3.3.2 Definition (Faktorisierbare Funktion)
Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ *heißt* faktorisierbar*, wenn die Funktionsvorschrift von* f *sich in endlich viele Elementaroperationen, bestehend aus den Grundrechenarten, elementaren Funktionen und Kompositionen, zerlegen lässt.*

3.3.3 Beispiel *Die Funktion* $f(x) = \dfrac{\sin(e^{x_1+3x_2})+5}{\|x\|_2+1}$ *ist faktorisierbar, Lösungen von Differentialgleichungen sind es meistens* nicht*, und* $f(x) = \displaystyle\sum_{k=0}^{\infty} \frac{x^k}{(2k)!}$ *ist es ebenfalls* nicht*.*

3.3.4 Definition (Intervallerweiterung)

a) Für $f : \mathbb{R}^n \to \mathbb{R}$ *heißt* $F : \mathbb{IR}^n \to \mathbb{IR}$ Intervallerweiterung von f*, falls* $\forall x \in \mathbb{R}^n : F([x,x]) = [f(x), f(x)]$ *gilt.*

b) Für eine faktorisierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ *heißt* $F : \mathbb{IR}^n \to \mathbb{IR}$, $F(X_1, \ldots, X_n) := f(X_1, \ldots, X_n)$ natürliche Intervallerweiterung von f*.*

In Teil b) dieser Definition ist mit dem Ausdruck $f(X_1, \ldots, X_n)$ diejenige Funktionsvorschrift gemeint, die entsteht, wenn man in der Funktionsvorschrift von f jedes Auftreten einer Variable x_i durch X_i ersetzt und dann alle auftretenden Elementaroperationen intervallwertig interpretiert. Entscheidend für den Aufbau der Intervallarithmetik ist, dass damit *nicht* die Menge $\text{bild}(f, X_1 \times \ldots \times X_n)$ gemeint ist. Wie das nachfolgende Beispiel 3.3.6 zeigen wird, stimmen $f(X_1, \ldots, X_n)$ und $\text{bild}(f, X_1 \times \ldots \times X_n)$ im Allgemeinen auch nicht überein. Dass für $n = 1$ insbesondere $f(X)$ und $\text{bild}(f, X)$ nicht übereinzustimmen brauchen, ist der Grund für unsere Notation des Bildes von X unter f.

3.3.5 Beispiel *Die oben definierten Funktionen* EXP *und* LOG *sind Intervallerweiterungen von* exp *bzw.* log*. Dasselbe gilt für die Intervallgrundrechenarten, etwa ist* $F([\underline{x}, \overline{x}], [\underline{y}, \overline{y}]) = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$ *eine Intervallerweiterung von* $f(x, y) = x + y$*. Die Funktion* $F : \mathbb{IR}^2 \to \mathbb{IR}$ *aus Beispiel 3.3.1 kann hingegen* nicht *die Intervallerweiterung einer Funktion* $f : \mathbb{R}^2 \to \mathbb{R}$ *sein.*

3.3.6 Beispiel *Der Verlauf der Funktion* $f(x) = x - x^2$ *ist in Abbildung 3.8 dargestellt. Man überlegt sich leicht, dass diese Funktion* $\text{bild}(f, [0,1]) = \left[0, \frac{1}{4}\right]$ *erfüllt.*

Da f offenbar faktorisierbar ist, besitzt f die natürliche Intervallerweiterung $F(X) = X - X \cdot X$*, wobei wir den Ausdruck* x^2 *als* $x \cdot x$ *aufgefasst haben. Einsetzen von* $X = [0,1]$ *liefert*

$$F([0,1]) \;=\; [0,1] - [0,1][0,1] \;=\; [0,1] - [0,1] \;=\; [-1,1],$$

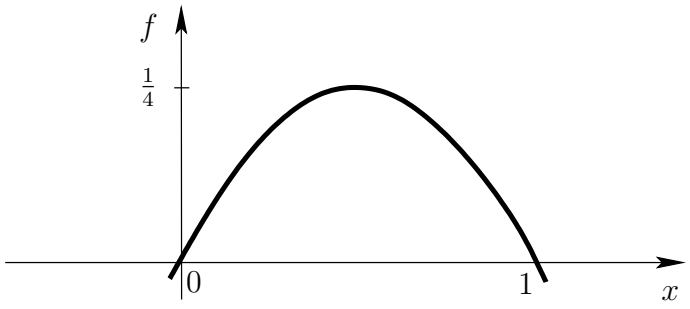


Abbildung 3.8: Schranken des Bildbereichs von f auf $[0,1]$

so dass $F([0,1])$ nicht *mit* $\mathrm{bild}(f,[0,1])$ *übereinstimmt.*

Es kommt sogar noch schlimmer: die Funktion f *lässt sich zwar äquivalent durch die Funktionsvorschrift* $\widetilde{f}(x) = x(1-x)$ *darstellen, trotz* $f = \widetilde{f}$ *stimmt aber die natürliche Intervallerweiterung* $\widetilde{F}(X) = X(1-X)$ *von* $\widetilde{f}$ *nicht mit der natürlichen Intervallerweiterung* F *von* f *überein, denn für* $X = [0,1]$ *gilt*

$$\widetilde{F}([0,1]) \;=\; [0,1](1-[0,1]) \;=\; [0,1][0,1] \;=\; [0,1] \;\neq\; [-1,1] \;=\; F([0,1]).$$

Im Folgenden werden wir gelegentlich nur die explizite Unter- oder Obergrenze eines Intervalls $F(X)$ benötigen und schreiben dafür

$$F(X) \;=\; \left[\underline{F}(X), \overline{F}(X)\right].$$

Etwa gilt in Beispiel 3.3.6 $\underline{F}([0,1]) = -1$, $\underline{\widetilde{F}}([0,1]) = 0$, und wir haben $\overline{\mathrm{EXP}}([\underline{x},\overline{x}]) = \exp(\overline{x})$ usw.

Abhängigkeitseffekt

Beispiel 3.3.6 zeigt, dass die Ergebnisse der Intervallarithmetik nicht nur von den Eigenschaften einer Funktion f abhängen, sondern *auch* von ihrer expliziten Darstellung. Der Grund dafür liegt im sogenannten *Abhängigkeitseffekt*, mit dem man immer dann rechnen muss, wenn eine Intervallvariable mehr als einmal in einer Funktionsvorschrift auftritt. Er begründet sich damit, dass eine mehrfach auftretende Variable genauso behandelt wird wie mehrere unabhängige Variablen.

Beispielsweise wird $X(1-X)$ genauso aufgefasst wie $X(1-Y)$ mit der Zusatzbedingung $X = Y$. Die letztere Abhängigkeit von X und Y wird allerdings nicht berücksichtigt. Dies führt dazu, dass bei der Betrachtung der beiden Faktoren x und $1-x$ Abhängigkeiten wie $x = 1 \Rightarrow 1-x = 0$ und $x = \frac{1}{2} \Rightarrow 1-x = \frac{1}{2}$ unter den Tisch fallen und bei der Berechnung von $X(1-X)$ insbesondere auch Produkte wie $1 \cdot 1$ auftreten, die bei der Auswertung von $x(1-x)$ gar nicht vorkommen.

Analog werden $X - X$ wie $X - Y$ mit $X = Y$ und $\frac{X}{X}$ wie $\frac{X}{Y}$ mit $X = Y$ behandelt, weshalb nicht $X - X = [0,0]$ oder $X/X = [1,1]$ gilt. Auch das Produkt $X \cdot X$ liefert nicht immer das Bild von X unter $f(x) = x^2$ (z.B. für $X = [-1,1]$), obwohl wir das Produkt von X und Y als exaktes Bild des Produktoperators definiert haben. Letzteres stimmt aber wieder nur für zwei unabhängige Inputs X und Y, während $X \cdot X$ wie $X \cdot Y$ mit $X = Y$ behandelt wird.

Als Faustregel gilt: je öfter eine Variable auftritt, desto größer wird das per Intervallarithmetik berechnete Intervall. Beispiel 3.3.6 zeigt etwa, dass $X - X \cdot X$ (X tritt dreimal auf) für $X = [0,1]$ ein erheblich größeres Intervall liefert als $X(1-X)$ (X tritt zweimal auf).

Häufig auftretende Funktionen, die unter dem Abhängigkeitseffekt leiden, lassen sich per Intervallarithmetik besser behandeln, indem man sie ebenfalls als elementare Funktionen auffasst, beispielsweise alle sogenannten Monome x^k, $k \in \mathbb{N}$. Etwa besitzt $f(x) = x \cdot x$ die Intervallerweiterung $F(X) = X \cdot X$ mit $F([-1,1]) = [-1,1]$, während es die Auffassung von $\mathrm{sqr}(x) = x^2$ als elementare Funktion erlaubt, die Intervallerweiterung

$$\mathrm{SQR}([\underline{x},\overline{x}]) \;=\; \begin{cases} [\min\{\underline{x}^2,\overline{x}^2\}, \max\{\underline{x}^2,\overline{x}^2\}]\,, & \text{falls } \; 0 \notin [\underline{x},\overline{x}] \\ [0, \max\{\underline{x}^2,\overline{x}^2\}]\,, & \text{falls } \; 0 \in [\underline{x},\overline{x}] \end{cases}$$

mit $\mathrm{SQR}([-1,1]) = [0,1]$ anzugeben. Auch Funktionen wie $\mathrm{norm}(x) = \|x\|_2$ werden oft als elementar aufgefasst.

Einschließungseigenschaft

Trotz dieser eher „unangenehmen“ Eigenschaften der natürlichen Intervallerweiterung erfüllt sie aber den von uns verfolgten Hauptzweck, als Ergebnis $F(X)$ eine *Obermenge* von $\mathrm{bild}(f,X)$ zu liefern, und damit eine garantierte Unterschranke an den Minimalwert sowie eine garantierte Oberschranke an

den Maximalwert von f auf X. Etwa in Beispiel 3.3.6 gilt

$$\left[0, \tfrac{1}{4}\right] \;=\; \text{bild}(f, [0,1]) \;\subseteq\; \begin{cases} F([0,1]) \;=\; [-1,1] \\ \widetilde{F}([0,1]) \;=\; [0,1]. \end{cases}$$

Dass dies tatsächlich immer gilt, werden wir im Folgenden herleiten. Dazu müssen wir zunächst noch die Intervallerweiterung der Komposition zweier Funktionen etwas genauer angeben.

Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ heißt faktorisierbar, wenn jede Komponente f_j, $j = 1, \ldots, m$, als Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$ faktorisierbar ist, und ihre natürliche Intervallerweiterung wird dann komponentenweise definiert, also als

$$F(X_1, \ldots, X_n) \;:=\; \begin{pmatrix} F_1(X_1, \ldots, X_n) \\ \vdots \\ F_m(X_1, \ldots, X_n) \end{pmatrix}$$

mit den natürlichen Intervallerweiterungen F_j der Funktionen f_j, $j = 1, \ldots, m$.

Falls $f : \mathbb{R}^k \to \mathbb{R}$ und $g : \mathbb{R}^n \to \mathbb{R}^k$ faktorisierbare Funktionen mit natürlichen Intervallerweiterungen $F : \mathbb{IR}^k \to \mathbb{IR}$ und $G : \mathbb{IR}^n \to \mathbb{IR}^k$ sind, dann definiert man naheliegenderweise

$$(F \circ G)(X) \;:=\; F(G(X))$$

als natürliche Intervallerweiterung von $f \circ g$. Dies ist lediglich die Formalisierung der offensichtlichen Tatsache, dass man als Intervallerweiterung von $f(g(x)$ den Ausdruck $F(G(X)$ benutzt.

3.3.7 Übung *Zeigen Sie, dass für jede faktorisierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ihre in Definition 3.3.4b) eingeführte natürliche Intervallerweiterung tatsächlich nicht nur so* heißt*, sondern auch eine Intervallerweiterung von f im Sinne von Definition 3.3.4a)* ist*.*

3.3.8 Definition (Monotone Intervallerweiterung)
Eine intervallwertige Funktion $F : \mathbb{IR}^n \to \mathbb{IR}$ heißt monoton *(oder* inklusionsisoton*), falls*

$$\forall\, X, Y \in \mathbb{IR}^n \;\; \textit{mit} \;\; X \subseteq Y : \quad F(X) \;\subseteq\; F(Y)$$

gilt.

3.3.9 Satz *Die Intervall-Grundrechenarten, die intervallwertigen elementaren Funktionen sowie die Komposition von Funktionen sind monoton.*

Beweis. Für jede Intervall-Grundrechenart oder eine intervallwertige elementare Funktion f gilt mit dazu passenden Mengen X per Definition $f(X) = \text{bild}(f, X)$. Wegen $\text{bild}(f, X) \subseteq \text{bild}(f, Y)$ für $X \subseteq Y$ folgt daraus ihre Monotonie.

Für Funktionen f und g, die selbst keine Kompositionen enthalten, ist damit auch die Monotonie von $F \circ G$ klar. Da eine Komposition außerdem nicht als eine „innerste“ Operation in einer Funktionsvorschrift auftreten kann, folgt durch ein rekursives Argument die Monotonie jeder Komposition $F \circ G$. •

Satz 3.3.9 impliziert das folgende Ergebnis.

3.3.10 Korollar *Für jede faktorisierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist ihre natürliche Intervallerweiterung $F : \mathbb{IR}^n \to \mathbb{IR}$ monoton.*

3.3.11 Satz *Für jede faktorisierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$, ihre natürliche Intervallerweiterung $F : \mathbb{IR}^n \to \mathbb{IR}$ und alle $X \in \mathbb{IR}^n$ gilt*

$$\text{bild}(f, X) \ \subseteq \ F(X).$$

Beweis. Für alle $x \in X$ gilt

$$[f(x), f(x)] \ = \ F([x, x]) \overset{F \text{ monoton}}{\subseteq} F(X)$$

und damit $f(x) \in F(X)$ sowie $\text{bild}(f, X) \subseteq F(X)$. •

Satz 3.3.11 bedeutet gerade, dass die natürliche Intervallerweiterung gültige Schranken für $\text{bild}(f, X)$ liefert. Wie gesehen können sie wegen des Abhängigkeitseffekts aber sehr grob sein.

Taylor-Modelle

Verbesserte Schranken kann man unter anderem mit sogenannten *Taylor-Modellen* erzielen, deren Idee im Folgenden kurz erklärt wird. Für stetig differenzierbares $f : M \to \mathbb{R}$ mit konvexer Menge $M \subseteq \mathbb{R}^n$ gilt nach dem Mittelwertsatz

$$\forall\, x, \widetilde{x} \in M : \quad f(x) \ = \ f(\widetilde{x}) + \langle \nabla f(y), x - \widetilde{x} \rangle$$

mit einem y auf der Verbindungsstrecke von x und $\widetilde{x}$. Über y ist nicht viel bekannt, aber immerhin muss y selbst auch in M liegen, denn es liegt in der konvexen Hülle der Elemente x und $\widetilde{x}$ von M, und M ist konvex. Da mit der Abkürzung $g := \nabla f$

$$\forall\, x, \widetilde{x} \in M: \quad f(x) \;=\; f(\widetilde{x}) + \sum_{i=1}^{n} g_i(y)(x_i - \widetilde{x}_i)$$

gilt, folgt

$$\forall\, x, \widetilde{x} \in M: \quad f(x) \;\in\; f(\widetilde{x}) + \sum_{i=1}^{n} \mathrm{bild}(g_i, M)(x_i - \widetilde{x}_i).$$

Wählt man insbesondere $M \in \mathbb{IR}^n$ und ist außerdem $g = \nabla f$ faktorisierbar mit natürlicher Intervallerweiterung G, so braucht man die Mengen $\mathrm{bild}(g_i, M)$, $i = 1, \ldots, n$, nicht explizit zu berechnen, sondern kann wegen $\mathrm{bild}(g_i, M) \subseteq G_i(M)$

$$\forall\, x, \widetilde{x} \in M: \quad f(x) \;\in\; f(\widetilde{x}) + \sum_{i=1}^{n} G_i(M)(x_i - \widetilde{x}_i)$$

schließen. Damit gilt für alle $X \in \mathbb{IR}^n$ mit $X \subseteq M$ sowie jedes $\widetilde{x} \in M$

$$\mathrm{bild}(f, X) \;\subseteq\; f(\widetilde{x}) + \sum_{i=1}^{n} G_i(M)([\underline{x}_i, \overline{x}_i] - \widetilde{x}_i).$$

Dann ist $F(X) := f(\widetilde{x}) + \sum_{i=1}^{n} G_i(M)([\underline{x}_i, \overline{x}_i] - \widetilde{x}_i)$ für $X \subseteq M$ zwar keine Intervallerweiterung von f, liefert aber trotzdem Schranken für $\mathrm{bild}(f, X)$, und insbesondere folgt

$$\mathrm{bild}(f, M) \;\subseteq\; F(M) \;=\; f(\widetilde{x}) + \sum_{i=1}^{n} G_i(M)([\underline{m}_i, \overline{m}_i] - \widetilde{x}_i).$$

Wähle als $\widetilde{x}$ etwa den Mittelpunkt von M, also $\widetilde{x} = \frac{m+\overline{m}}{2}$.

3.3.12 Beispiel *Für $f(x) = x(1-x)$ und $M = [0,1]$ gilt $g(x) = f'(x) = 1 - 2x$ und damit $G([0,1]) = 1 - 2[0,1] = [-1,1]$. Mit $\widetilde{x} = \frac{1}{2}$ folgt*

$$\begin{aligned}
\mathrm{bild}(f, [0,1]) \;&\subseteq\; F([0,1]) \;:=\; f\left(\tfrac{1}{2}\right) + G([0,1])\left([0,1] - \tfrac{1}{2}\right) \\
&=\; \tfrac{1}{4} + [-1,1]\left([0,1] - \tfrac{1}{2}\right) \\
&=\; \tfrac{1}{4} + [-1,1]\left[-\tfrac{1}{2}, \tfrac{1}{2}\right] \\
&=\; \tfrac{1}{4} + \left[-\tfrac{1}{2}, \tfrac{1}{2}\right] \\
&=\; \left[-\tfrac{1}{4}, \tfrac{3}{4}\right].
\end{aligned}$$

Vergleichen Sie dies mit den entsprechenden Ergebnissen aus Beispiel 3.3.6.

Diese Idee eines Taylor-Modells lässt sich analog auf Taylorentwicklungen höherer Ordnung verallgemeinern.

Weitere Bezeichnungen

Zum späteren Gebrauch führen wir abschließend für eine Box $X = [\underline{x}, \overline{x}] \in \mathbb{IR}^n$ noch folgende Bezeichnungen ein:

- $m(X) = \dfrac{\underline{x} + \overline{x}}{2}$ ist der *Boxmittelpunkt*, und
- $w(X) = \|\overline{x} - \underline{x}\|_2$ ist die *Boxweite*.

Beachten Sie, dass $w(X)$ die Länge der Boxdiagonale bezeichnet und damit ein Maß für die Boxgröße ist. Es sei darauf hingewiesen, dass in der Literatur zur Intervallarithmetik auch andere Wahlen der Norm zur Definition von $w(X)$ üblich sind. Dabei entspricht zum Beispiel die ℓ_∞-Norm gerade der größten Kantenlänge von X, und die ℓ_1-Norm liefert die Summe der Kantenlängen entlang der n Koordinatenrichtungen, von denen man ein (dimensionsabhängiges) Vielfaches als „Umfang“ von X, also also Summe aller Kantenlängen, auffassen kann. Im Falle $n = 1$ gilt für jede der drei angesprochenen Normen natürlich $w(X) = \overline{x} - \underline{x}$.

3.3.13 Übung *Für eine Menge $M \subseteq \mathbb{R}^n$ bezeichnet $\sup_{x,y \in M} \|x - y\|_2$ den* Durchmesser *von M. Zeigen Sie, dass der Durchmesser einer Box X mit ihrer Boxweite übereinstimmt.*

Aus numerischer Sicht, insbesondere bei der Behandlung von Rundungsfehlern bei der Auswertung von f an einem $x \in \mathbb{R}^n$ durch Intervallarithmetik, wäre eine entscheidende nächste Frage, ob und wie schnell die Einschließungsintervalle $F(X)$ für $\mathrm{bild}(f, X)$ klein werden, wenn man kleiner werdende Einschließungsboxen X für x benutzt, also für $w(X) \to 0$. Da sich diese Frage für unsere Anwendung in der globalen Optimierung aber nicht stellen wird, verweisen wir für weiterführende Darstellungen der Intervallarithmetik auf [9].

3.4 Konvexe Relaxierung per αBB-Methode

Mit Hilfe der Intervallarithmetik lassen sich in verschiedener Weise konvexe Relaxierungen von Funktionen konstruieren. Wir konzentrieren uns hier auf die sogenannte *αBB-Methode* von Adjiman/Androulakis/Floudas (1998, [3]).

Dazu seien $X \in \mathbb{IR}^n$ und $f \in C^2(X, \mathbb{R})$. Gesucht ist eine konvexe Relaxierung von f auf X, also eine konvexe Funktion $\widehat{f} : X \to \mathbb{R}$ mit $\widehat{f}(x) \leq f(x)$ für alle $x \in X$ (vgl. Def. 3.2.2a)).

Grundidee der αBB-Methode ist es, zunächst eine möglichst einfache gleichmäßig konvexe Funktion ψ zu konstruieren, die auf X nicht-positiv ist. Addiere dann ein genügend großes Vielfaches $\alpha\psi$ auf f, d.h. setze

$$\widehat{f}_\alpha(x) \;=\; f(x) + \alpha\psi(x)$$

mit hinreichend großem $\alpha \geq 0$, um Nichtkonvexitäten in f zu kompensieren.

Mit der Darstellung $X = [\underline{x}, \overline{x}]$ ist eine einfache gleichmäßig konvexe und auf X nicht-positive Funktion durch

$$\psi(x) \;=\; \frac{1}{2}(\underline{x} - x)^\intercal(\overline{x} - x)$$

gegeben, denn wegen

$$\forall\, x \in X: \quad \forall\, i = 1, \ldots, n: \quad \underline{x}_i \;\leq\; x_i \;\leq\; \overline{x}_i$$

gilt für alle $x \in X$

$$\psi(x) \;=\; \frac{1}{2}\sum_{i=1}^{n} \underbrace{(\underline{x}_i - x_i)}_{\leq 0}\underbrace{(\overline{x}_i - x_i)}_{\geq 0} \;\leq\; 0,$$

und außerdem erfüllt die Hessematrix für alle $x \in \mathbb{R}^n$

$$D^2\psi(x) \;=\; \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \;=\; E.$$

Damit erhalten wir $\lambda_{\min}(D^2\psi(x)) = 1$, so dass ψ nach Satz 2.5.10a) gleichmäßig konvex auf $\mathbb{R}^n$ ist.

Wir halten ferner fest, dass ψ in jedem Eckpunkt von X verschwindet, denn

$$y \;\text{ Eckpunkt von }\; X \;\Leftrightarrow\; \forall\, i: \quad y_i \in \{\underline{x}_i, \overline{x}_i\} \;\Rightarrow\; \psi(y) = 0.$$

3.4.1 Übung *Der eindeutige Minimalpunkt der gleichmäßig konvexen Funktion*

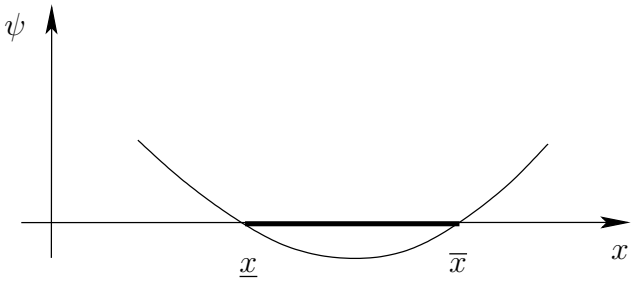
$$\psi(x) \;=\; \frac{1}{2}(\underline{x}-x)^{\mathsf{T}}(\overline{x}-x),$$

ist der Mittelpunkt $m(X)=\frac{1}{2}(\underline{x}+\overline{x})$ *der Box* X*, und der Minimalwert lautet*

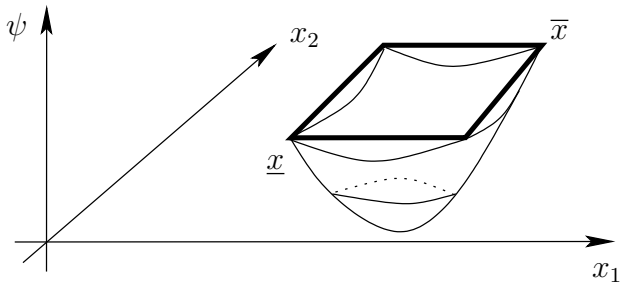
$$\min_{x\in\mathbb{R}^n}\psi(x) \;=\; \min_{x\in X}\psi(x) \;=\; \psi\left(m(X)\right) \;=\; -\frac{1}{8}\,w(X)^2, \tag{3.4.1}$$

wobei $w(X)=\|\overline{x}-\underline{x}\|_2$ *die Boxweite bezeichnet.*

Abbildung 3.9 zeigt die Gestalt von ψ für $n=1$ und $n=2$.



$n=1$



$n=2$

Abbildung 3.9: Die Funktion ψ

Man setzt also

$$\widehat{f}_\alpha(x) \;=\; f(x)+\alpha\psi(x) \;=\; f(x)+\frac{\alpha}{2}(\underline{x}-x)^{\mathsf{T}}(\overline{x}-x)$$

mit einem noch zu bestimmenden Parameter α.

3.4.2 Lemma

a) Für alle $x \in X$ und $\alpha \geq 0$ gilt $\widehat{f}_\alpha(x) \leq f(x)$.

b) Für jeden Eckpunkt y von X und jedes $\alpha \in \mathbb{R}$ gilt $\widehat{f}_\alpha(y) = f(y)$.

c) Für alle $\alpha \geq 0$ beträgt die maximale Abweichung zwischen f und $\widehat{f}_\alpha$ auf X

$$\max_{x \in X} \left(f(x) - \widehat{f}_\alpha(x)\right) \;=\; \frac{\alpha}{8}\, w(X)^2 \,.$$

Beweis.

a)

$$\forall\, x \in X,\ \alpha \geq 0 : \quad f(x) - \widehat{f}_\alpha(x) \;=\; -\underbrace{\alpha}_{\geq 0}\underbrace{\psi(x)}_{\leq 0} \;\geq\; 0.$$

b) Für jeden Eckpunkt y von X und alle $\alpha \in \mathbb{R}$ gilt

$$f(y) - \widehat{f}_\alpha(y) \;=\; -\alpha \underbrace{\psi(y)}_{=0} \;=\; 0.$$

c) Für alle $\alpha \geq 0$ gilt

$$\max_{x \in X} \left(f(x) - \widehat{f}_\alpha(x)\right) \;=\; -\alpha \min_{x \in X} \psi(x) \;\overset{(3.4.1)}{=}\; \frac{\alpha}{8}\, w(X)^2 \,.$$

•

Lemma 3.4.2c) besagt, dass die maximale Abweichung zwischen f und $\widehat{f}_\alpha$ auf X nur von α und von der Boxweite $w(X)$ abhängt (und z.B. überhaupt nicht von f oder von anderen geometrischen Eigenschaften von X).

Nach Lemma 3.4.2a) ist $\widehat{f}_\alpha$ für alle $\alpha \geq 0$ eine Relaxierung von f auf X. Wir wählen α nun außerdem so groß, dass die Nichtkonvexitäten in f vom (gleichmäßig) konvexen Term $\alpha\psi$ kompensiert werden. Wegen $f \in C^2$ und $\psi \in C^2$ ist auch $\widehat{f}_\alpha = f + \alpha\psi$ eine C^2-Funktion. Laut C^2-Charakterisierung von Konvexität (S. 2.5.3) ist $\widehat{f}_\alpha$ genau dann konvex auf der (volldimensionalen) Box X, wenn $D^2\widehat{f}_\alpha(x) \succeq 0$ für alle $x \in X$ gilt, wenn also für alle $x \in X$ sämtliche Eigenwerte von $D^2\widehat{f}_\alpha(x)$ nicht-negativ sind.

Aufgrund von

$$D^2\widehat{f}_\alpha(x) \;=\; D^2 f(x) + \alpha D^2\psi(x) \;=\; D^2 f(x) + \alpha E$$

kann man nun versuchen, α so zu wählen, dass alle Eigenwerte von $D^2 f(x) + \alpha E$ nicht-negativ sind. Das ist tatsächlich möglich, denn die Eigenwerte von $D^2 f(x) + \alpha E$ und $D^2 f(x)$ hängen sehr einfach zusammen (wie wir auch bereits in Beweis von Satz 2.5.10b) gesehen hatten):

$$\begin{aligned} \widehat{\lambda} \text{ EW von } D^2\widehat{f}_\alpha(x) &\Leftrightarrow \det(D^2\widehat{f}_\alpha(x) - \widehat{\lambda}E) = 0 \\ &\Leftrightarrow \det(D^2 f(x) + \alpha E - \widehat{\lambda}E) = 0 \\ &\Leftrightarrow \det(D^2 f(x) - (\widehat{\lambda} - \alpha)E) = 0 \\ &\Leftrightarrow \widehat{\lambda} - \alpha \text{ EW von } D^2 f(x), \end{aligned}$$

d.h. die Eigenwerte von $D^2\widehat{f}_\alpha(x)$ erhält man, indem man die Eigenwerte von $D^2 f(x)$ um α nach rechts verschiebt (vgl. Abb. 3.10).

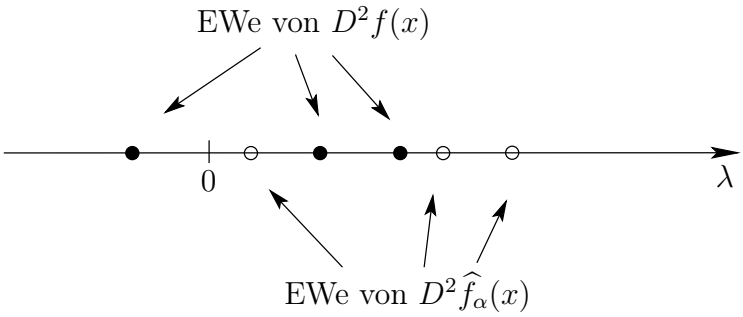


Abbildung 3.10: Verschiebung von Eigenwerten

Wenn insbesondere $\lambda_{\min}$ den *kleinsten* Eigenwert von $D^2 f(x)$ bezeichnet, dann ist $\lambda_{\min} + \alpha$ der kleinste Eigenwert von $D^2\widehat{f}_\alpha(x)$, und wir erhalten

$$D^2\widehat{f}_\alpha(x) \succeq 0 \;\Leftrightarrow\; \lambda_{\min} + \alpha \geq 0.$$

Da allerdings auch $\lambda_{\min}$ von x abhängig ist ($\lambda_{\min}(x)$ ist der kleinste Eigenwert von $D^2 f(x)$), muss man genauer sagen: $\widehat{f}_\alpha(x)$ ist genau dann konvex auf X, wenn

$$\forall\, x \in X: \quad \lambda_{\min}(x) + \alpha \;\geq\; 0$$

gilt, oder äquivalent

$$\min_{x \in X} \lambda_{\min}(x) + \alpha \;\geq\; 0$$

und damit

$$\alpha \;\geq\; -\min_{x \in X} \lambda_{\min}(x).$$

Da für die Relaxierungseigenschaft $\widehat{f}_\alpha \leq f$ außerdem $\alpha \geq 0$ nötig ist, folgt insgesamt:

3.4.3 Satz *Für alle* $\alpha \geq \max\{0, -\min_{x\in X} \lambda_{\min}(x)\}$ *ist* $\widehat{f}_\alpha = f + \alpha\psi$ *eine konvexe Relaxierung von* f *auf* X*.*

Ein grundlegendes Problem dafür, globale Optimierungsprobleme mit Hilfe konvexer Relaxierungen per αBB-Methode zu lösen, liegt nun darin, dass man für die passende Wahl von α scheinbar wiederum ein globales Optimierungsproblem lösen muss, nämlich

$$\min \lambda_{\min}(x) \quad \text{s.t.} \quad x \in X.$$

Abhilfe schafft hierfür die Intervallarithmetik, mit deren Hilfe sich eine garantierte Unterschranke β von $\lambda_{\min}(x)$ auf X bestimmen lässt, so dass man die Lösung des globalen Optimierungsproblems umgeht. Nach der Berechnung von β setzt man

$$\alpha := \max\{0, -\beta\},$$

denn damit folgt $\alpha \geq 0$ und $\alpha \geq -\beta \geq -\min_{x\in X} \lambda_{\min}(x)$. Nach Satz 3.4.3 ist $\widehat{f}_\alpha$ also eine konvexe Relaxierung von f auf X.

Das so gewählte α wird im Allgemeinen größer sein als das per globaler Optimierung von $\lambda_{\min}(x)$ über X erzielbare, und damit muss man nach Lemma 3.4.2c) einen größeren Maximalabstand zwischen f und $\widehat{f}_\alpha$ auf X in Kauf nehmen. Dafür ist diese Wahl von α numerisch leicht umsetzbar.

Wir berechnen β zunächst für zwei einfache Spezialfälle.

$\underline{n = 1:}$

Wegen $D^2 f(x) = f''(x)$ ist $\lambda(x) = f''(x)$ einziger Eigenwert von $D^2 f(x)$ und damit $\lambda_{\min}(x) = f''(x)$. Falls f'' faktorisierbar ist, bilde die natürliche Intervallerweiterung F'' von f'' und wähle $\beta := \underline{F''}(X)$, also als Untergrenze von $F''(X)$. Dann gilt $\beta \leq \min_{x\in X} f''(x) = \min_{x\in X} \lambda_{\min}(x)$.

3.4.4 Beispiel *Die Funktion* $f(x) = x^3 - x$ *ist auf* $X = [-1, 1]$ *nicht konvex (vgl. Abb. 3.11). Eine konvexe Relaxierung von* f *auf* X *berechnet man per* αBB*-Methode wie folgt: setze* $\widehat{f}_\alpha(x) = f(x) + \alpha\psi(x)$ *mit*

$$\psi(x) = \frac{1}{2}(x+1)(x-1) = \frac{x^2-1}{2}$$

und

$$\alpha \;=\; \max\{0, -\beta\},$$

wobei

$$\beta \;\leq\; \min_{x\in[-1,1]} f''(x) \;=\; \min_{x\in[-1,1]} 6x$$

zu bestimmen ist. Es gilt $F''(X) = 6X$ *und damit* $F''([-1,1]) = [-6,6]$. *Es folgt* $\beta = \underline{F}''([-1,1]) = -6$ *und* $\alpha = 6$. *Damit ist*

$$\widehat{f}_6(x) \;=\; f(x) + 6\psi(x) \;=\; x^3 - x + 3x^2 - 3$$

eine konvexe Relaxierung von f *auf* $[-1,1]$ *(vgl. Abb. 3.11).*

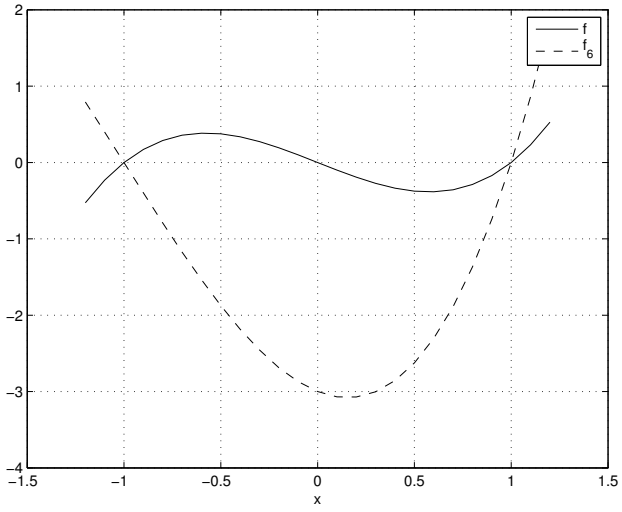


Abbildung 3.11: Eine konvexe Relaxierung

$n > 1$, f separabel:

Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *separabel*, wenn sie als Summe von n Funktionen geschrieben werden kann, die jeweils nur von einem der n Argumente abhängen:

$$f(x) \;=\; \sum_{i=1}^{n} f_i(x_i).$$

Zum Beispiel ist $f^1(x) = x_1^2 + x_2^3 + e^{x_3}$ separabel, aber $f^2(x) = x_1^2 x_2 + e^{x_3}$ nicht.

Für separable Funktionen gilt offenbar

$$\nabla f(x) \;=\; \begin{pmatrix} f_1'(x_1) \\ \vdots \\ f_n'(x_n) \end{pmatrix}$$

und

$$D^2 f(x) \;=\; \begin{pmatrix} f_1''(x_1) & & 0 \\ & \ddots & \\ 0 & & f_n''(x_n) \end{pmatrix}$$

$$\text{(z.B. } D^2 f^1(x) \;=\; \begin{pmatrix} 2 & 0 & 0 \\ 0 & 6x_2 & 0 \\ 0 & 0 & e^{x_3} \end{pmatrix} \text{).}$$

Daher ist für jede separable C^2-Funktion die Hessematrix $D^2 f(x)$ eine Diagonalmatrix. Da bei Diagonalmatrizen die Eigenwerte mit den Diagonalelementen identisch sind, folgt in diesem Fall für den kleinsten Eigenwert

$$\lambda_{\min}(x) \;=\; \min_{i=1,\ldots,n} f_i''(x_i)$$

und damit nach Übung 1.3.3a) sowie Übung 1.3.6c)

$$\begin{aligned} \min_{x\in X} \lambda_{\min}(x) &= \min_{x\in X} \min_{i=1,\ldots,n} f_i''(x_i) \\ &= \min_{i=1,\ldots,n} \min_{x\in X} f_i''(x_i) \\ &= \min_{i=1,\ldots,n} \min_{x_i\in[\underline{x}_i,\overline{x}_i]} f_i''(x_i). \end{aligned}$$

Falls alle f_i'' faktorisierbar sind, bilde die natürlichen Intervallerweiterungen F_i'', wähle $\beta_i := \underline{F_i''}([\underline{x}_i, \overline{x}_i])$, $i = 1, ..., n$, und setze $\beta := \min_{i=1,\ldots,n} \beta_i$. Dann gilt $\beta \le \min_{x\in X} \lambda_{\min}(x)$.

Allgemeiner Fall:

Im allgemeinen Fall lautet die Hessematrix von f

$$D^2 f(x) \;=\; \begin{pmatrix} \partial_{x_1}\partial_{x_1} f(x) & \cdots & \partial_{x_n}\partial_{x_1} f(x) \\ \vdots & & \vdots \\ \partial_{x_1}\partial_{x_n} f(x) & \cdots & \partial_{x_n}\partial_{x_n} f(x) \end{pmatrix} \;=:\; A \quad (\text{eigentlich } \; A(x)).$$

In diesem Fall liegen leider keine geschlossenen Formeln für die Eigenwerte mehr vor, sondern die Eigenwerte sind Lösungen der Gleichung

$$\det(A - \lambda E) \;=\; 0$$

(für eine Motivation dieser Gleichung vgl. den Anhang von [10]). Das ist zur Anwendung der Intervallarithmetik ungünstig, denn sie gibt Schranken für explizite Funktionen an, nicht für implizite Lösungen von Gleichungen. Man behilft sich stattdessen mit geschlossenen Formeln für *Schranken* an die Eigenwerte, an die per Intervallarithmetik nochmals Schranken berechnet werden.

3.4.5 Definition (Gerschgorin-Kreisscheiben)
Für die (n,n)-Matrix A setze

$$r_i \;:=\; \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1,\ldots,n.$$

Dann heißt

$$\{\lambda \in \mathbb{C}|\ |\lambda - a_{ii}| \leq r_i\}$$

Gerschgorin-Kreisscheibe *von A.*

Der folgende Satz wird in der Numerischen Linearen Algebra bewiesen.

3.4.6 Satz (Satz von Gerschgorin)
Es sei A eine (n,n)-Matrix mit Einträgen aus $\mathbb{C}$. Dann liegen alle Eigenwerte von A in der Menge

$$\bigcup_{i=1}^{n} \{\lambda \in \mathbb{C}|\ |\lambda - a_{ii}| \leq r_i\}\,.$$

3.4.7 Beispiel *Für*

$$A \;=\; \begin{pmatrix} 1 & 2 & 1 \\ 3 & 0 & 0 \\ -1 & 1 & -2 \end{pmatrix}$$

gilt $r_1 = 3$, $r_2 = 3$ und $r_3 = 2$. Die Gerschgorin-Kreisscheiben lauten also $\{\lambda \in \mathbb{C}|\,|\lambda - 1| \leq 3\}$, $\{\lambda \in \mathbb{C}|\,|\lambda| \leq 3\}$ und $\{\lambda \in \mathbb{C}|\,|\lambda + 2| \leq 2\}$ (vgl. Abb. 3.12). Tatsächlich berechnen die Eigenwerte von A sich zu $\lambda_1 = 3$ und $\lambda_{2/3} = -2 \pm i$.

Da für $f \in C^2$ die Hessematrix $A = D^2 f(x)$ symmetrisch ist, liefert ein Ergebnis der linearen Algebra, dass alle Eigenwerte von A *reell* sind. Die Betrachtung von Gerschgorin-Kreisscheiben in der komplexen Zahlenebene ist für unsere Zwecke also gar nicht nötig, sondern anstelle der Kreisscheiben kann man ihre Schnitte mit der reellen Achse, die *Gerschgorin-Intervalle* $[a_{ii} - r_i, a_{ii} + r_i]$ benutzen.

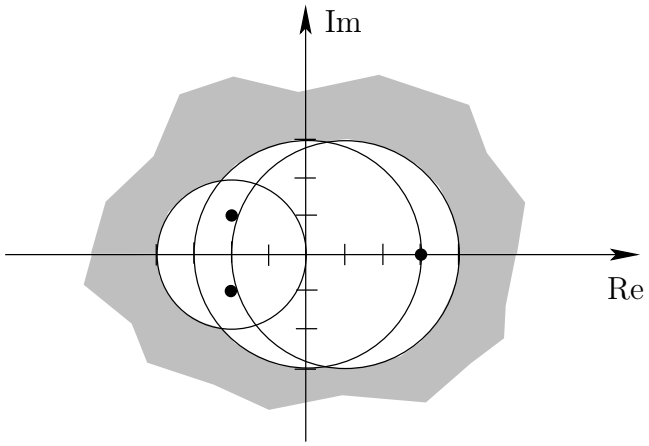


Abbildung 3.12: Gerschgorin-Kreisscheiben

3.4.8 Korollar *Es sei A eine symmetrische (n,n)-Matrix mit Einträgen aus $\mathbb{R}$. Dann liegen alle Eigenwerte von A in der Menge*

$$\bigcup_{i=1}^{n} [a_{ii} - r_i \;\; a_{ii} + r_i].$$

3.4.9 Beispiel *Für die symmetrische Matrix*

$$A \;=\; \begin{pmatrix} 1 & 2 & -1 \\ 2 & 0 & 0 \\ -1 & 0 & -2 \end{pmatrix}$$

gilt $r_1 = 3$, $r_2 = 2$ und $r_3 = 1$. Die Gerschgorin-Intervalle lauten also $[-2, 4]$, $[-2, 2]$ und $[-3, -1]$ (vgl. Abb. 3.13). Die tatsächlichen Eigenwerte sind $\lambda_1 \approx -2.5$, $\lambda_2 \approx -1.2$ und $\lambda_3 \approx 2.7$.

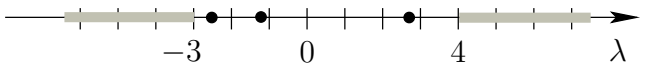


Abbildung 3.13: Gerschgorin-Intervalle

Eine Unterschranke für alle Eigenwerte und damit auch für $\lambda_{\min}$ ist die kleinste der Intervallgrenzen:

$$\lambda_{\min} \;\geq\; \min_{i=1,\ldots,n} (a_{ii} - r_i).$$

Berücksichtigt man wieder die x-Abhängigkeit, also $A(x) = D^2 f(x)$, so folgt

$$\forall\, x \in X: \quad \lambda_{\min}(x) \;\geq\; \min_{i=1,\ldots,n} (a_{ii}(x) - r_i(x))$$

mit

$$r_i(x) \;=\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}(x)|.$$

Mit Übung 1.3.3a) erhalten wir daraus

$$\begin{aligned} \min_{x\in X} \lambda_{\min}(x) &\geq \min_{x\in X} \min_{i=1,\ldots,n} (a_{ii}(x) - r_i(x)) \\ &= \min_{i=1,\ldots,n} \min_{x\in X} (a_{ii}(x) - r_i(x)). \end{aligned}$$

Falls alle Einträge a_{ij} von $A(x) = D^2 f(x)$ faktorisierbar sind, bilde deren natürliche Intervallerweiterungen $A_{ij}(X)$ sowie die Intervallerweiterungen $R_i(X)$ von $r_i(x)$, wähle für jedes $i = 1, ..., n$

$$\beta_i \;:=\; \underline{A_{ii}(X) - R_i(X)}$$

und setze $\beta := \min_{i=1,\ldots,n} \beta_i$. Dann gilt $\beta \leq \min_{x\in X} \lambda_{\min}(x)$.

Als natürliche Intervallerweiterung der in den Funktionen r_i auftretenden Betragsfunktion $\mathrm{abs}(x) := |x|$ auf $\mathbb{R}$ benutzen wir dabei

$$\mathrm{ABS}([\underline{x}, \overline{x}]) \;=\; \begin{cases} [\min\{|\underline{x}|, |\overline{x}|\}, \max\{|\underline{x}|, |\overline{x}|\}], & \text{falls } \; 0 \notin [\underline{x}, \overline{x}] \\ [0, \max\{|\underline{x}|, |\overline{x}|\}], & \text{falls } \; 0 \in [\underline{x}, \overline{x}], \end{cases}$$

also

$$A_{ii}(X) - R_i(X) \;=\; A_{ii}(X) - \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{ABS}(A_{ij}(X)).$$

Damit gilt expliziter für jedes $i = 1, \ldots, n$

$$\begin{aligned} \beta_i \;=\; \underline{A_{ii}(X) - R_i(X)} \;&=\; \underline{A_{ii}(X)} - \overline{R_i(X)} \\ &=\; \underline{A_{ii}(X)} - \sum_{\substack{j=1 \\ j\neq i}}^{n} \overline{\mathrm{ABS}(A_{ij}(X))} \\ &=\; \underline{A}_{ii}(X) - \sum_{\substack{j=1 \\ j\neq i}}^{n} \max\left\{|\underline{A}_{ij}(X)|, |\overline{A}_{ij}(X)|\right\}. \end{aligned}$$

3.5 Gleichmäßig verfeinerte Gitter

Wir wenden uns nun der Frage zu, wie man konvexe Relaxierungen für die globale Optimierung verwenden kann.

3.5.1 Beispiel *Betrachte nochmals die Funktion $f(x) = x^3 - x$ auf $X = [-1, 1]$ aus Beispiel 3.4.4. v bezeichne den globalen Minimalwert von*

$$P: \quad \min f(x) \quad s.t. \quad x \in X.$$

Wir sind in der Lage, gültige Schranken *an v anzugeben.*

Oberschranke:

$$\forall\, x \in X: \quad v \;\leq\; f(x),$$

zum Beispiel gilt $v \leq f(0) = 0$.

Unterschranke:

Es gilt $v \geq \widehat{v}$, wobei $\widehat{v}$ Minimalwert von

$$\widehat{P}: \quad \min \widehat{f}(x) \quad s.t. \quad x \in X$$

mit einer konvexen Relaxierung $\widehat{f}$ von f auf X ist (vgl. S. 3.2.5a)).

In Beispiel 3.4.4 haben wir gesehen, dass $\widehat{f}(x) = x^3+3x^2-x-3$ eine konvexe Relaxierung von f auf X ist. Die kritischen Punkte von $\widehat{f}$ berechnen sich wie folgt:

$$0 \;\stackrel{!}{=}\; \widehat{f}'(x) \;=\; 3x^2 + 6x - 1 \;\Leftrightarrow\; x^2 + 2x - \frac{1}{3} \;=\; 0$$

und damit

$$x_{1/2} \;=\; -1 \pm \sqrt{1 + \frac{1}{3}} \;=\; -1 \pm \frac{2}{\sqrt{3}}.$$

Einziger kritischer Punkt im Inneren von X ist also $\widehat{x} = \frac{2}{\sqrt{3}} - 1$ mit $\widehat{f}(\widehat{x}) > -3.08$.

Damit haben wir einen KKT-Punkt des konvexen Problems $\widehat{P}$ gefunden und brauchen die Randpunkte von X nicht mehr zu untersuchen. Der Punkt $\widehat{x} = \frac{2}{\sqrt{3}} - 1$ ist also globaler Minimalpunkt von $\widehat{P}$ mit $\widehat{v} > -3.08$. Insgesamt folgt $v \in [-3.08, 0]$.

Statt die Oberschranke an v mit *irgend*einem Punkt $x \in X$ (in Bsp. 3.5.1: $x = 0$) zu bestimmen, gibt es bessere Alternativen:

- Bestimme mit einem Verfahren der nichtlinearen Optimierung (s. [13]) einen lokalen Minimalpunkt x^{lok} von f auf X. Dann gilt $v \leq v^{\text{lok}} := f(x^{\text{lok}})$, und man kann die Hoffnung hegen, eine gute Oberschranke von v zu erzielen.

  Im Beispiel 3.5.1 liefert dieser Ansatz folgende Verbesserung: Aus $0 = f'(x) = 3x^2 - 1$ folgt nach den üblichen Berechnungen $x^{\text{lok}} = \frac{1}{\sqrt{3}}$ mit $v^{\text{lok}} = (\frac{1}{\sqrt{3}})^3 - \frac{1}{\sqrt{3}} < -0.38$, also $v \in [-3.08, -0.38]$.

- Setze $\widehat{x}$ in f ein, in der Hoffnung, dass ein globaler Minimalpunkt der Relaxierung auch für die Originalfunktion einen niedrigen Wert liefert:

  $$v \;\leq\; f(\widehat{x}).$$

  In Beispiel 3.5.1 führt dies zu $(\frac{2}{\sqrt{3}} - 1)^3 - (\frac{2}{\sqrt{3}} - 1) < -0.15$, also $v \in [-3.08, -0.15]$.

Während man bei der ersten der obigen Alternativen mit gewissem Aufwand (Anwendung eines NLO-Verfahrens) im Allgemeinen eine bessere Oberschranke für v findet, besteht der Vorteil der zweiten Alternative darin, dass er zum einen mit wenig Aufwand zu realisieren ist (nämlich durch eine Funktionsauswertung), und dass man bei αBB-Relaxierungen außerdem etwas über den Abstand der Schranken weiß:

$$v \;\in\; [\widehat{v},\, f(\widehat{x})] \;=\; \left[\widehat{f}_\alpha(\widehat{x}),\, f(\widehat{x})\right]$$

mit

$$w([\widehat{v},\, f(\widehat{x})]) \;=\; f(\widehat{x}) - \widehat{f}_\alpha(\widehat{x}) \;\leq\; \max_{x \in X} \left(f(x) - \widehat{f}_\alpha(x)\right) \overset{\text{L. 3.4.2c)}}{=} \frac{\alpha}{8}\, w(X)^2\,.$$

Damit ist folgendes Ergebnis gezeigt:

3.5.2 Satz *Es seien $X \in \mathbb{IR}^n$, $f \in C^2(X, \mathbb{R})$, $D^2 f$ faktorisierbar, v der globale Minimalwert von*

$$P: \quad \min f(x) \quad s.t. \quad x \in X,$$

$\widehat{f}_\alpha$ eine per αBB-Methode konstruierte konvexe Relaxierung von f auf X, $\widehat{x}$ ein globaler Minimalpunkt von $\widehat{f}_\alpha$ auf X sowie $\widehat{v}$ der Minimalwert $\widehat{f}_\alpha(\widehat{x})$. Dann gilt

$$v \in [\widehat{v}, f(\widehat{x})] \quad mit \quad w([\widehat{v}, f(\widehat{x})]) \leq \frac{\alpha}{8} w(X)^2 .$$

Satz 3.5.2 impliziert um so bessere Schranken für v, je kleiner die Box X ist, und zwar *unabhängig* davon, wie grob α gewählt ist (s.u.). Für $w(X) \to 0$ gilt dann sogar $\widehat{v} \nearrow v$. Wegen der Konvexität von X und Satz 3.2.5d) steht dies *nicht* im Widerspruch zu Satz 3.2.5c).

Da X fest vorgegeben ist, kann man zwar die Größe von X nicht ändern, aber man kann X in kleinere Boxen unterteilen (vgl. Abb. 3.14):

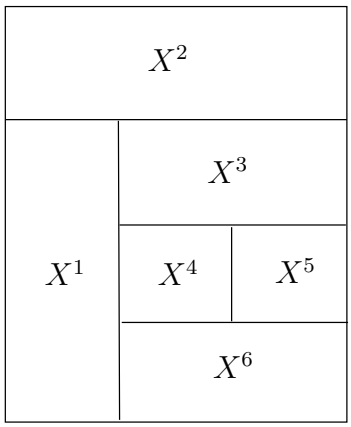


Abbildung 3.14: Parkettierung einer Box

3.5.3 Definition (Parkettierung)
Für $k \in \mathbb{N}$ bilden $X^1, \ldots, X^k \in \mathbb{IR}^n$ eine Parkettierung *von X, falls*

a) $\bigcup_{\ell=1}^{k} X^\ell = X$ *und*

b) $\forall j \neq \ell: \quad \mathrm{int}(X^j) \cap \mathrm{int}(X^\ell) = \emptyset$

gilt.

Zu jeder Parkettierung von X gibt es (mindestens) eine Teilbox $X^{\ell_{\text{glob}}}$ die einen globalen Minimalpunkt x^{glob} von f auf X enthält. Man findet den globalen Minimalwert also, indem man alle Minimalwerte von f auf X^ℓ, $\ell = 1, ..., k$, miteinander vergleicht:

$$v \;=\; \min_{x\in X} f(x) \;=\; \min_{x\,\in\,\bigcup_{\ell=1}^{k} X^\ell} f(x) \;=\; \min_{\ell=1,\ldots,k}\; \min_{x\in X^\ell} f(x)\,,$$

wobei wir Übung 1.3.4 benutzt haben.

Mit $v^\ell := \min_{x\in X^\ell} f(x)$ heißt dies $v = \min_{\ell=1,\ldots,k} v^\ell$, was in Abbildung 3.15 für $n = 1$ und $k = 4$ illustriert ist.

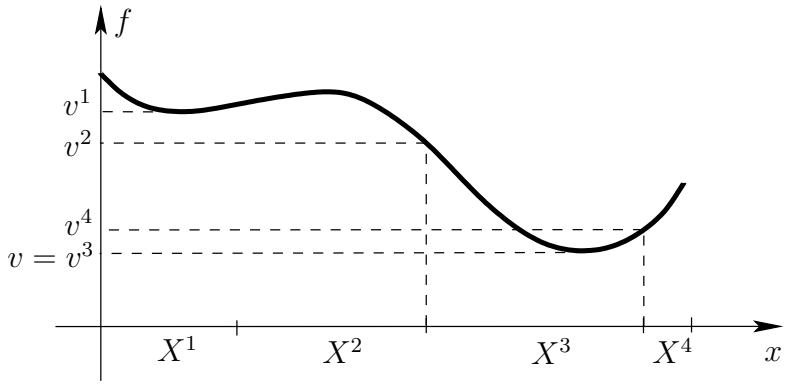


Abbildung 3.15: Minimalwerte auf Teilboxen

Grundidee des Folgenden ist es, nun auf *jeder* Teilbox X^ℓ Unterschranken $\widehat{v}^\ell$ an v^ℓ per αBB-Methode zu berechnen. Aus $v = \min_{\ell=1,\ldots,k} v^\ell$ und $v^\ell \geq \widehat{v}^\ell$, $\ell = 1, ..., k$, folgt dann

$$v \;\geq\; \min_{\ell=1,\ldots,k} \widehat{v}^\ell.$$

Da die X^ℓ, $\ell = 1, ..., k$, kleiner als X sind, sollte diese Unterschranke an v besser (d.h. größer) als ein $\widehat{v}$ sein, das wie in Satz 3.5.2 für die gesamte Box X berechnet wird.

Zur Berechnung der $\widehat{v}^\ell$ sei

$$\forall\, \ell = 1, ..., k: \quad \widehat{f}^{\,\ell}_{\alpha_\ell}(x) \;:=\; f(x) + \frac{\alpha_\ell}{2}(\underline{x}^\ell - x)^{\intercal}(\overline{x}^\ell - x)$$

mit

$$\alpha_\ell \;\geq\; \max\left\{0,\; -\min_{x\in X^\ell} \lambda_{\min}(x)\right\}$$

eine konvexe Relaxierung von f auf X^ℓ.

Falls $\alpha \geq \max\{0, -\min_{x\in X} \lambda_{\min}(x)\}$ gilt, dann darf man $\alpha_\ell := \alpha$ für alle $\ell = 1, ..., k$ setzen, denn

$$\begin{aligned} X^\ell \subseteq X \;&\Rightarrow\; \min_{x\in X^\ell} \lambda_{\min}(x) \;\geq\; \min_{x\in X} \lambda_{\min}(x) \\ &\Rightarrow\; \max\left\{0, -\min_{x\in X^\ell} \lambda_{\min}(x)\right\} \;\leq\; \max\left\{0, -\min_{x\in X} \lambda_{\min}(x)\right\} \;\leq\; \alpha. \end{aligned}$$

Im Folgenden benutzen wir der Einfachheit halber diese Wahl ($\alpha_\ell := \alpha$ für alle ℓ), obwohl die Bestimmung von neuen α_ℓ für alle X^ℓ zu besseren Schranken führen könnte. Es seien also

$$\widehat{f}^\ell_\alpha(x) \;=\; f(x) + \frac{\alpha}{2}(\underline{x}^\ell - x)^\intercal(\overline{x}^\ell - x),$$

$\widehat{x}^\ell$ ein globaler Minimalpunkt von $\widehat{f}^\ell_\alpha$ auf X^ℓ und $\widehat{v}^\ell$ der zugehörige Minimalwert. Dann gilt nach Satz 3.5.2 für alle $\ell = 1, ..., k$

$$v^\ell \;\in\; [\widehat{v}^\ell, f(\widehat{x}^\ell)] \quad \text{mit} \quad w([\widehat{v}^\ell, f(\widehat{x}^\ell)]) \;\leq\; \frac{\alpha}{8}\, w(X^\ell)^2\,.$$

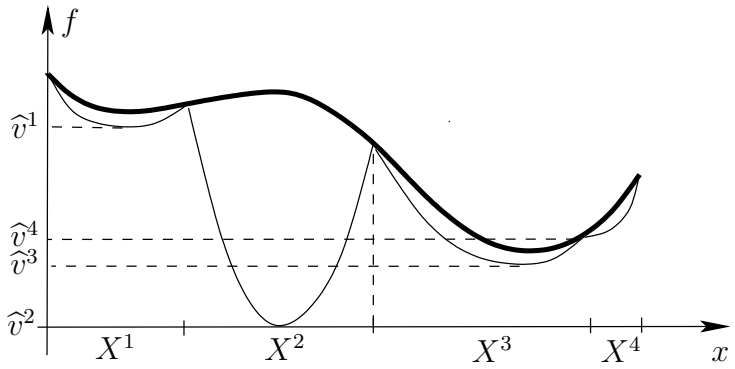


Abbildung 3.16: Relaxierung auf Teilboxen

Wie oben bereits gesehen, bildet der Ausdruck $\min_{\ell=1,...,k} \widehat{v}^\ell$ eine Unterschranke an v. Um ihn genauer untersuchen zu können, konzentrieren wir uns auf einen Index $\ell_\star$, an dem das Minimum realisiert wird. Es seien also $\widehat{v}^{\ell_\star} = \min_{\ell=1,...,k} \widehat{v}^\ell$ und $\widehat{x}^{\ell_\star}$ ein globaler Minimalpunkt von $\widehat{f}^{\ell_\star}_\alpha$ in $X^{\ell_\star}$. Dann folgt $v \geq \widehat{v}^{\ell_\star}$ und wegen $\widehat{x}^{\ell_\star} \in X$ auch $v \leq f(\widehat{x}^{\ell_\star})$ (Abb. 3.16 illustriert, dass aber nicht notwendigerweise $v = v^{\ell_\star}$ gilt, dass $X^{\ell_\star}$ also nicht einen globalen Minimalpunkt von f auf X zu enthalten braucht!). Insgesamt erhalten wir

$$v \;\in\; [\widehat{v}^{\ell_\star}, f(\widehat{x}^{\ell_\star})] \quad \text{mit} \quad w([\widehat{v}^{\ell_\star}, f(\widehat{x}^{\ell_\star})]) \;\leq\; \frac{\alpha}{8}\, w(X^{\ell_\star})^2\,.$$

Daher ist man in der Lage, die Genauigkeit der Berechnung von v (d.h. die Länge des Einschließungsintervalls für v) über die Boxgrößen zu steuern: mit einer vom Benutzer vorgegebenen Toleranz $\varepsilon > 0$ gilt $w([\widehat{v}^{\ell\star}, f(\widehat{x}^{\ell\star})]) \leq \varepsilon$ sicher dann, wenn man im Fall $\alpha > 0$ für die Boxweite von $X^{\ell\star}$ die Abschätzung

$$w(X^{\ell\star}) \;\leq\; \sqrt{\frac{8\varepsilon}{\alpha}}$$

garantieren kann. Im Trivialfall $\alpha = 0$ spielt die Boxweite keine Rolle, und jede Toleranz $\varepsilon > 0$ wird eingehalten.

Eine kleine Länge des Einschließungsintervalls für v impliziert ihrerseits das folgende Ergebnis.

3.5.4 Lemma *Der Wert $\widehat{v}^{\ell\star}$ und der Punkt $\widehat{x}^{\ell\star}$ seien berechnet wie oben, und mit einem $\varepsilon > 0$ gelte $w([\widehat{v}^{\ell\star}, f(\widehat{x}^{\ell\star})]) \leq \varepsilon$. Dann erfüllt der Punkt $\widetilde{x} := \widehat{x}^{\ell\star}$ die Bedingungen*

$$\widetilde{x} \;\in\; X \quad \textit{und} \quad v \;\leq\; f(\widetilde{x}) \;\leq\; v + \varepsilon,$$

ist also ein „fast optimaler“ zulässiger Punkt von P.

Beweis. Wegen $\widetilde{x} = \widehat{x}^{\ell\star} \in X^{\ell\star} \subseteq X$ gelten die Bedingungen $\widetilde{x} \in X$ und $v \leq f(\widetilde{x})$. Aus

$$\varepsilon \;\geq\; w([\widehat{v}^{\ell\star}, f(\widehat{x}^{\ell\star})]) \;=\; f(\widetilde{x}) - \widehat{v}^{\ell\star} \;\geq\; f(\widetilde{x}) - v$$

folgt außerdem $f(\widetilde{x}) \leq v + \varepsilon$. •

Zur Generierung einer hinreichend kleinen Box $X^{\ell\star}$ verfolgen wir zunächst den Ansatz, X *gleichmäßig* in Teilboxen $X^1, \ldots, X^k$ zu parkettieren, etwa durch rekursive Unterteilung der Boxen an ihren Mittelpunkten $m(X^\ell) = \frac{\underline{x}^\ell + \overline{x}^\ell}{2}$, $\ell = 1, ..., k$.

3.5.5 Beispiel *Für $n = 2$ betrachte die Box X in Abbildung 3.17. Nach einem gleichmäßigen Unterteilungsschritt am Boxmittelpunkt gilt für die vier entstandenen Teilboxen*

$$\forall\ \ell = 1, ..., 4: \quad w(X^\ell) \;=\; \frac{w(X)}{2}$$

und nach einem weiteren Verfeinerungsschritt

$$\forall\ \ell = 1, ..., 16: \quad w(X^\ell) \;=\; \frac{w(X)}{4}\,.$$

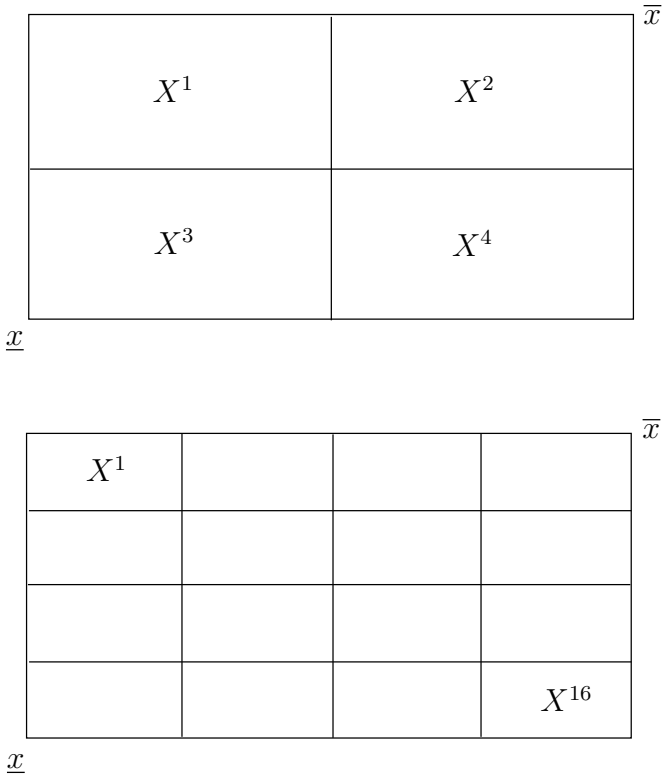


Abbildung 3.17: Gleichmäßige Unterteilung an Boxmittelpunkten

Nach r Verfeinerungsschritten haben alle Boxen die Weite $\frac{w(X)}{2^r}$.

Insbesondere folgt für die Box $X^{\ell_\star}$

$$f(\widehat{x}^{\ell_\star}) - \widehat{v}^{\ell_\star} \;\leq\; \frac{\alpha}{8}\, w(X^{\ell_\star})^2 \;=\; \frac{\alpha}{8}\, \frac{w(X)^2}{4^r}\,.$$

Um eine Toleranz $\varepsilon > 0$ in der Berechnung von v zu erreichen, reicht es (im Falle $\alpha > 0$) also, genügend oft zu verfeinern, d.h. r hinreichend groß zu wählen:

$$\begin{aligned} \frac{\alpha}{8}\, \frac{w(X)^2}{4^r} \;\overset{!}{\leq}\; \varepsilon \;&\Leftrightarrow\; \left(\frac{1}{4}\right)^r \;\leq\; \frac{8\varepsilon}{\alpha w(X)^2} \\ &\Leftrightarrow\; r \;\geq\; \log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right) / \log\left(\frac{1}{4}\right). \end{aligned}$$

Das minimale $r \in \mathbb{N}$ mit dieser Eigenschaft lautet

$$r \;=\; \left\lceil \log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right) / \log\left(\frac{1}{4}\right) \right\rceil,$$

wobei $\lceil x \rceil$ die obere Gauß-Klammer *von x, also das kleinste $z \in \mathbb{N}$ mit $z \geq x$ bezeichnet.*

Man macht sich leicht klar, dass die Berechnung von r in Beispiel 3.5.5 unabhängig von der Dimension n stets dasselbe Ergebnis liefert. Damit können wir Algorithmus 3.2 zur globalen Minimierung von f auf X formulieren.

Algorithmus 3.2: Globale Minimierung einer box-restringierten Funktion per gleichmäßiger Gitterverfeinerung

---

Input : $X = [\underline{x}, \overline{x}] \in \mathbb{IR}^n$, $f \in C^2(X, \mathbb{R})$ mit faktorisierbarer Hessematrix $D^2 f$, Toleranz $\varepsilon > 0$.

Output : $\widetilde{x} \in X$ mit $v \leq f(\widetilde{x}) \leq v + \varepsilon$.

1 **begin**

2 Berechne ein $\alpha \geq \max\{0, -\min_{x \in X} \lambda_{\min}(x)\}$.

3 Setze

$$r \;=\; \begin{cases} \left\lceil \log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right) / \log\left(\frac{1}{4}\right) \right\rceil, & \text{falls } \alpha > 0, \\ 0, & \text{falls } \alpha = 0, \end{cases}$$

und unterteile X r-mal gleichmäßig in Teilboxen $X^1, \ldots, X^k$.

4 Berechne für jedes $\ell = 1, ..., k$ den Minimalwert $\widehat{v}^\ell$ von

$$\widehat{f}^\ell_\alpha(x) \;=\; f(x) + \frac{\alpha}{2}(\underline{x}^\ell - x)^\intercal(\overline{x}^\ell - x)$$

auf X^ℓ (z.B. mit einem Verfahren der nichtlinearen Optimierung).

5 Wähle ein $\ell_\star$ mit $\widehat{v}^{\ell_\star} = \min_{\ell=1,\ldots,k} \widehat{v}^\ell$.

6 Falls in Zeile 4 noch nicht geschehen, berechne einen Minimalpunkt $\widehat{x}^{\ell_\star}$ von $\widehat{f}^{\ell_\star}_\alpha$ auf $X^{\ell_\star}$.

7 Setze $\widetilde{x} := \widehat{x}^{\ell_\star}$.

8 **end**

Algorithmus 3.2 ist für die Praxis meist nicht empfehlenswert, da die Anzahl k der Boxen beim Verfeinern in Zeile 3 exponentiell wächst: es gilt $k = (2^n)^r$. Für kleine Toleranzen ε müsste man daher in Zeile 4 eine üblicherweise nicht handhabbar große Anzahl von Teilproblemen lösen.

3.5.6 Beispiel *Für $n = 3$ betrachte die „Einheitsbox“*

$$X = \left[-\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right]$$

mit Weite $w(X) = 2\sqrt{3}$. Für die nicht ungewöhnliche Parameterwahlen $\alpha = 1$ und $\varepsilon = 10^{-3}$ folgt dann

$$r = \left\lceil \log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right) / \log\left(\frac{1}{4}\right) \right\rceil = 6.$$

Damit ist in Algorithmus 3.2 eine Verfeinerung in $(2^3)^6 = 262.144$ Boxen nötig!

3.6 Branch-and-Bound für box-restringierte Probleme

In diesem Abschnitt geben wir ein *praxistaugliches* Verfahren an, das für das box-restringierte Problem

$$P: \quad \min f(x) \quad \text{s.t.} \quad x \in X$$

mit $X \in \mathbb{IR}^n$, $f \in C^2(X, \mathbb{R})$, $D^2 f$ faktorisierbar, und für jede Toleranz $\varepsilon > 0$ in „handhabbar endlich vielen“ Schritten einen Punkt $\widetilde{x} \in X$ mit $v \leq f(\widetilde{x}) \leq v + \varepsilon$ erzeugt, also eine Approximation eines globalen Minimalpunkts von P.

Es basiert auf den Überlegungen in Kapitel 3.5, jedoch ohne die Parkettierung von X durch gleichmäßige Verfeinerung zu konstruieren. Hauptmotivation für das Folgende ist, die Anzahl der Teilboxen stattdessen so klein wie möglich zu halten. Wesentliche Strategien dazu sind:

- in jedem Schritt wird nur *eine* Teilbox verfeinert und diese wird nur *halbiert* („branching“),
- die Berechnung von Schranken auf den Teilboxen („bounding“) ermöglicht die „vielversprechendste“ Wahl der zu unterteilenden Teilbox, sowie den Ausschluss von Boxen, in denen garantiert keine besseren als die bereits bekannten zulässigen Punkte liegen („Ausloten“, „pruning“, „fathoming“). Im Gegensatz zur *gleichmäßigen* Gitterverfeinerung aus Kapitel 3.5 führt dies zu einer *adaptiven* Gitterverfeinerung.

Halbierung von Boxen:

Die Box $X^\ell = [\underline{x}^\ell, \overline{x}^\ell]$ mit $\ell \in \{1, ..., k\}$ sei in zwei gleich große Teilboxen zu zerlegen. Dazu definieren wir zunächst die eindimensionalen Intervalle $X_i^\ell = [\underline{x}_i^\ell, \overline{x}_i^\ell]$, $i = 1, \ldots, n$, so dass

$$X^\ell \;=\; X_1^\ell \times \ldots \times X_n^\ell \tag{3.6.2}$$

gilt. Nun wähle ein $i \in \{1, ..., n\}$ und definiere $X^{\ell,1}$ als diejenige Box, die aus (3.6.2) hervorgeht, wenn dort X_i^ℓ durch das Intervall $[\underline{x}_i^\ell, m(X_i^\ell)]$ ersetzt wird, sowie $X^{\ell,2}$ als diejenige Box, die aus (3.6.2) hervorgeht, wenn dort X_i^ℓ

durch das Intervall $[m(X_i^\ell), \overline{x}_i^\ell]$ ersetzt wird. Ausgeschrieben bedeutet dies

$$X^{\ell,1} = \left[\begin{pmatrix} \underline{x}_1^\ell \\ \vdots \\ \underline{x}_i^\ell \\ \vdots \\ \underline{x}_n^\ell \end{pmatrix}, \begin{pmatrix} \overline{x}_1^\ell \\ \vdots \\ \frac{\underline{x}_i^\ell + \overline{x}_i^\ell}{2} \\ \vdots \\ \overline{x}_n^\ell \end{pmatrix} \right], \qquad X^{\ell,2} = \left[\begin{pmatrix} \underline{x}_1^\ell \\ \vdots \\ \frac{\underline{x}_i^\ell + \overline{x}_i^\ell}{2} \\ \vdots \\ \underline{x}_n^\ell \end{pmatrix}, \begin{pmatrix} \overline{x}_1^\ell \\ \vdots \\ \overline{x}_i^\ell \\ \vdots \\ \overline{x}_n^\ell \end{pmatrix} \right].$$

Abbildung 3.18 zeigt die Halbierung von $X^\ell \in \mathbb{R}^2$ mit den Wahlen $i = 1$ und $i = 2$.

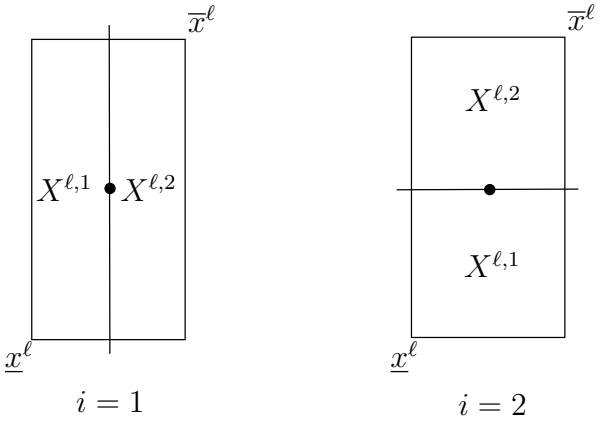


Abbildung 3.18: Halbierung von Boxen

Je nach Auswahlregel für $i \in \{1, ..., n\}$ können die neuen Boxen „langgezogen“ (Halbierung der kürzesten Kante) oder „eher würfelförmig“ (Halbierung der längsten Kante) werden. Beispielsweise durch sukzessive Halbierung entlang der kürzesten Kante kann man allerdings nicht dafür sorgen, dass die Weiten der entstehenden Boxen immer geringer werden. Im Folgenden halbieren wir Boxen daher immer entlang der (bzw. einer) längsten Kante, wählen also ein $i \in \{1, ..., n\}$ mit

$$\overline{x}_i^\ell - \underline{x}_i^\ell = \max_{j=1,...,n} \left(\overline{x}_j^\ell - \underline{x}_j^\ell \right) \quad (= \|\overline{x}^\ell - \underline{x}^\ell\|_\infty).$$

Schranken:

Die Hauptidee zur Verringerung der Anzahl betrachteter Teilboxen basiert auf der Kenntnis irgendeines zulässigen Punkts $\widetilde{x} \in X$, etwa des Boxmittelpunkts $m(X)$. Sein Zielfunktionswert $\widetilde{v} = f(\widetilde{x})$ bildet zunächst natürlich eine

Oberschranke für v. Der Wert $\widetilde{v}$ kann aber auch dazu benutzt werden, gewisse Teilboxen X^ℓ von der weiteren Betrachtung auszuschließen. Gilt nämlich $\widetilde{v} \leq v^\ell$, so folgt daraus

$$f(\widetilde{x}) \;=\; \widetilde{v} \;\leq\; v^\ell \;=\; \min_{x \in X^\ell} f(x)$$

und damit $f(\widetilde{x}) \leq f(x)$ für alle $x \in X^\ell$. Also enthält die Teilbox X^ℓ keinen Punkt x mit besserem Zielfunktionswert als $\widetilde{x}$ und braucht nicht weiter betrachtet zu werden.

Anstelle der Ungleichung $\widetilde{v} \leq v^\ell$ mit dem schwer berechenbaren Wert v^ℓ lässt sich als hinreichendes Kriterium zum Ausloten der Box X^ℓ auch die Ungleichung $\widetilde{v} \leq \widehat{v}^\ell$ nutzen, denn wegen $\widehat{v}^\ell \leq v^\ell$ impliziert sie $\widetilde{v} \leq v^\ell$.

Für die Funktion f in Abbildung 3.19 lässt sich daher die Box X^{ℓ_1} ausloten. Die Box X^{ℓ_2} enthält zwar ebenfalls keine Punkte x mit besserem Zielfunktionswert als $\widetilde{x}$, allerdings lässt sich dies anhand der (dafür zu groben) Unterschranke $\widehat{v}^{\ell_2}$ nicht entscheiden.

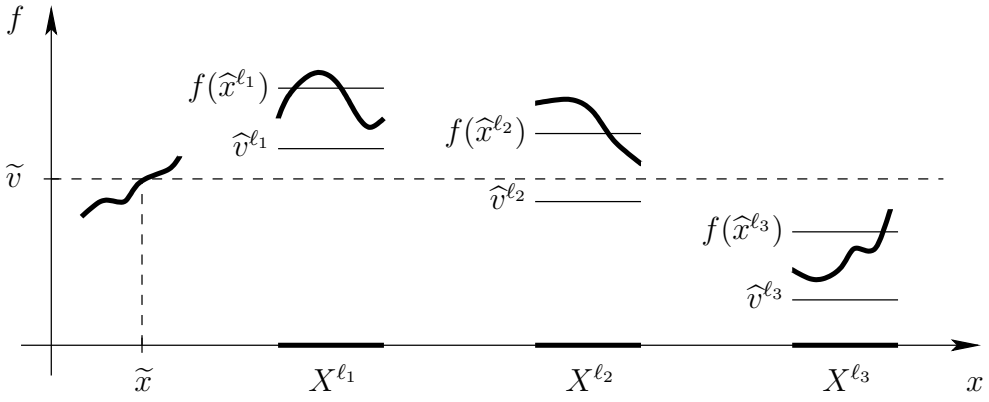


Abbildung 3.19: Schranken im Branch-and-Bound-Verfahren

Offensichtlich lassen sich um so mehr Teilboxen ausloten, je kleiner die Schranke $\widetilde{v}$ ist. Daher passt man sie im Laufe des Verfahrens jedesmal an, wenn ein zulässiger Punkt $x' \in X$ mit besserem Zielfunktionswert als $\widetilde{x}$, also $f(x') < f(\widetilde{x}) = \widetilde{v}$, generiert wird. Den vorher besten bekannten zulässigen Punkt $\widetilde{x}$ ersetzt man dann durch den neuen Punkt x' sowie $\widetilde{v}$ durch $f(x')$.

Etwa in der Situation von Abbildung 3.19 kann man $x' := \widehat{x}^{\ell_3}$ setzen. Nach der entsprechenden Anpassung von $\widetilde{v}$ auf $f(\widehat{x}^{\ell_3})$ wird es möglich, auch die Box X^{ℓ_2} auszuloten.

Weiterer Effekt dieser Anpassung des Werts von $\widetilde{v}$ ist, dass er eine immer genauere Oberschranke für v bildet. Um auch eine Unterschranke für v zu erhalten, berechnen wir zur gegebenen Parkettierung $X^1, \ldots, X^k$ wieder das Minimum aller Minimalwerte $\widehat{v}^\ell$ von $\widehat{f}_\alpha^\ell$ auf X^ℓ, d.h. wir bestimmen ein $\ell_\star$ mit $\widehat{v}^{\ell_\star} = \min_{\ell=1,\ldots,k} \widehat{v}^\ell$. Dann gilt $v \in [\widehat{v}^{\ell_\star}, \widetilde{v}]$. Falls die Länge $w([\widehat{v}^{\ell_\star}, \widetilde{v}]) = \widetilde{v} - \widehat{v}^{\ell_\star}$ dieses Einschließungsintervalls für v unter einer vorgegebenen Toleranz $\varepsilon > 0$ liegt, terminiert das Verfahren. Ansonsten versucht man, die Schranken zu verbessern.

Um die Unterschranke $\widehat{v}^{\ell_\star}$ an v zu verbessern, ist es naheliegend, die Box $X^{\ell_\star}$ zu halbieren, die den „schlechten“ Minimalwert $\widehat{v}^{\ell_\star}$ zu verantworten hat. Die Lösung der Relaxierungen der beiden neuen Teilprobleme liefert unter anderem auch neue zulässige Punkte, die ihrerseits die Oberschranke $\widetilde{v}$ verbessern können. Wir werden später sehen, dass die Länge des Einschließungsintervalls $[\widehat{v}^{\ell_\star}, \widetilde{v}]$ nach endlich vielen Unterteilungsschritten tatsächlich unter jede vorgegebene Toleranz $\varepsilon > 0$ fällt.

Algorithmus 3.3 realisiert diese Überlegungen durch die Pflege einer Liste von Boxen aus der aktuellen Parkettierung, in denen noch ein besserer als der bislang beste bekannte zulässige Punkt $\widetilde{x}$ liegen kann. Um die nötigen Vergleiche durchführen zu können, werden die Teilboxen X^ℓ gemeinsam mit der zugehörigen Schranke $\widehat{v}^\ell$ als Paar $(X^\ell, \widehat{v}^\ell)$ in der Liste abgespeichert.

Dass die Liste in Algorithmus 3.3 tatsächlich komplett geleert wird, wie es in den Zeilen 18 und 21 abgefragt wird, kann tatsächlich auftreten. In diesem Fall sind nämlich alle Teilboxen der letzten Parkettierung von X ausgelotet worden, und dies geschieht genau für $\widehat{v}' \geq \widetilde{v}$ für alle Teilboxen X'. Daraus folgt

$$\min_{x \in X'} f(x) \;=\; v' \;\geq\; \widehat{v}' \;\geq\; \widetilde{v} \;=\; f(\widetilde{x})$$

für jede Teilbox X' und damit $\min_{x \in X} f(x) \geq f(\widetilde{x})$. Also gilt Liste $= \emptyset$ genau dann, wenn man mit $\widetilde{x}$ einen globalen Minimalpunkt genau identifiziert hat. Dieser Fall kann beispielsweise dann natürlicherweise auftreten, wenn P einen globalen Minimalpunkt in einer Ecke von X besitzt und diese Ecke auch als Optimalpunkt einer Relaxierung auf einer Teilbox identifiziert wird.

Im Gegensatz zu Algorithmus 3.2 ist zunächst nicht klar, ob in Algorithmus 3.3 das Abbruchkriterium in Zeile 21 stets nach endlich vielen Schritten erfüllt wird. Dies beantwortet der folgende Satz, der noch etwas Vorbereitung benötigt.

Algorithmus 3.3: Globale Minimierung einer box-restringierten Funktion per Branch and Bound - αBB, 1998

Input : $X = [\underline{x}, \overline{x}] \in \mathbb{IR}^n$, $f \in C^2(X, \mathbb{R})$ mit faktorisierbarer Hessematrix $D^2 f$, Toleranz $\varepsilon > 0$.

Output : $\widetilde{x} \in X$ mit $v \leq f(\widetilde{x}) \leq v + \varepsilon$.

begin
  Berechne ein $\alpha \geq \max\{0, -\min_{x \in X} \lambda_{\min}(x)\}$.
  Setze $\widetilde{x} = m(X)$ und $\widetilde{v} = f(\widetilde{x})$.
  Setze $X^\star = X$, $\widehat{v}^\star = -\infty$ und Liste $= (X^\star, \widehat{v}^\star)$.
  repeat
    Halbiere $X^\star$ entlang einer längsten Kante und nenne die neuen Boxen X^1, X^2, d.h. wähle ein $i \in \{1, \ldots, n\}$ mit

$$\overline{x}_i^\star - \underline{x}_i^\star = \max_{j=\{1,\ldots,n\}} \left(\overline{x}_j^\star - \underline{x}_j^\star\right)$$

    und setze

$$X^1 \;=\; X_1^\star \times \ldots \times [\underline{x}_i^\star, m(X_i^\star)] \times \ldots \times X_n^\star$$

    sowie

$$X^2 \;=\; X_1^\star \times \ldots \times [m(X_i^\star), \overline{x}_i^\star] \times \ldots \times X_n^\star \,.$$

    Streiche $(X^\star, \widehat{v}^\star)$ aus Liste.
    for $\ell = 1, 2$ **do**
      Berechne einen Minimalpunkt $\widehat{x}^\ell$ und den Minimalwert $\widehat{v}^\ell$ von $\widehat{f}_\alpha^\ell(x) = f(x) + \frac{\alpha}{2}\left(\underline{x}^\ell - x\right)^\intercal \left(\overline{x}^\ell - x\right)$ auf X^ℓ (z.B. mit einem Verfahren der nichtlinearen Optimierung).
      if $\widehat{v}^\ell < \widetilde{v}$ **then**
        Füge das Paar $\left(X^\ell, \widehat{v}^\ell\right)$ zu Liste hinzu.
        if $f(\widehat{x}^\ell) < \widetilde{v}$ **then**
          Setze $\widetilde{x} = \widehat{x}^\ell$ und $\widetilde{v} = f(\widehat{x}^\ell)$.
          Streiche alle Paare $(X', \widehat{v}')$ mit $\widehat{v}' \geq \widetilde{v}$ aus Liste.
        end
      end
    end
    if Liste $\neq \emptyset$ **then**
      Wähle ein $(X^\star, \widehat{v}^\star)$ mit minimalem $\widehat{v}^\star$ aus Liste.
    end
  until Liste $= \emptyset$ **or** $\widetilde{v} - \widehat{v}^\star \leq \varepsilon$;
end

Der bei der Halbierung der Boxen entstehende Baum ist binär, d.h. jeder Knoten hat höchstens zwei Kinder und genau einen Elternknoten (außer der Wurzel). Je tiefer man in den Baum absteigt, desto kleiner werden offenbar die Teilboxen. Für den folgenden Satz benötigen wir eine Präzisierung dieser Aussage. Dazu nummerieren wir die durch sukzessive Verzweigungen entstehenden Stufen des Baumes, beginnend mit der Wurzel als Stufe 0.

3.6.1 Lemma *Jede Box aus Stufe N des Verzweigungsbaums besitzt höchstens die Weite* $\left(1-\frac{3}{4n}\right)^{\frac{N}{2}} \cdot w(X)$.

Beweis. Es seien Y eine Box und Z eine der beiden aus Y hervorgehenden Boxen laut Halbierungsregel in Zeile 6. Dann gilt für das Verhältnis der quadrierten Diagonalenlängen

$$\begin{aligned}\frac{\|\overline{z}-\underline{z}\|_2^2}{\|\overline{y}-\underline{y}\|_2^2} &= \frac{\sum_{j=1}^n(\overline{z}_j-\underline{z}_j)^2}{\|\overline{y}-\underline{y}\|_2^2} = \frac{\sum_{j=1}^n(\overline{y}_j-\underline{y}_j)^2-(\overline{y}_i-\underline{y}_i)^2+\left(\frac{\overline{y}_i-\underline{y}_i}{2}\right)^2}{\|\overline{y}-\underline{y}\|_2^2} \\ &= 1-\frac{3}{4}\frac{(\overline{y}_i-\underline{y}_i)^2}{\|\overline{y}-\underline{y}\|_2^2} \le 1-\frac{3}{4n},\end{aligned}$$

wobei die Ungleichung wegen der Maximaleigenschaft von $\overline{y}_i-\underline{y}_i$ aus

$$\|\overline{y}-\underline{y}\|_2^2 = \sum_{j=1}^n(\overline{y}_j-\underline{y}_j)^2 \le n(\overline{y}_i-\underline{y}_i)^2$$

folgt. Beim Wechsel in ein tiefere Stufe verkürzen sich die Diagonalenlängen also mindestens um den Faktor $\sqrt{1-\frac{3}{4n}}$. Daraus folgt die Behauptung. •

Beispielsweise liegt der Verkürzungsfaktor für $n=2$ unter $\sqrt{\frac{5}{8}} \approx 0.79$ und für $n=3$ unter $\frac{\sqrt{3}}{2} \approx 0.89$. Offensichtlich gilt $\sqrt{1-\frac{3}{4n}} \to 1$ für $n \to \infty$. Dies quantifiziert den verlangsamenden Effekt hoher Dimensionen auf die Konvergenz des Verfahrens.

3.6.2 Satz *Algorithmus 3.3 bricht nach endlich vielen Schritten ab.*

Beweis. Wir zeigen, dass nach endlich vielen Iterationen das Abbruchkriterium in Zeile 21 zutrifft. Falls nach endlich vielen Iterationen Liste $= \emptyset$ gilt, ist dies sicherlich der Fall. Wir dürfen im Folgenden also Liste $\neq \emptyset$ für alle betrachteten Iterationen annehmen.

Für jede Iteration wird demnach in Zeile 19 eine Box $X^\star$ mit zugehöriger Schranke $\widehat{v}^\star$ aus der Liste gewählt. $X^\star$ muss in dieser oder in einer früheren Iteration in Zeile 11 zur Liste hinzugefügt worden sein, woraufhin in den Zeilen 12 und 13 der Wert $f(\widehat{x}^\star)$ in die Bestimmung von $\widetilde{v}$ eingegangen ist. In der aktuellen Iteration gilt also $\widetilde{v} \leq f(\widehat{x}^\star)$ und damit in Zeile 21

$$\widetilde{v} - \widehat{v}^\star \;\leq\; f(\widehat{x}^\star) - \widehat{v}^\star \;\leq\; \frac{\alpha}{8}\, w(X^\star)^2.$$

Damit $\widetilde{v} - \widehat{v}^\star$ nach endlich vielen Schritten wie gefordert unter ε liegt, genügt es also zu zeigen, dass die Weite der in Zeile 19 gewählten Box $X^\star$ nach hinreichend vielen Iterationen genügend klein ist. Falls $X^\star$ in Stufe N des Verzweigungsbaums liegt, gilt nach Lemma 3.6.1

$$w(X^\star) \;\leq\; \left(1 - \frac{3}{4n}\right)^{\frac{N}{2}} w(X)$$

und damit

$$\widetilde{v} - \widehat{v}^\star \;\leq\; \frac{\alpha}{8}\left(1 - \frac{3}{4n}\right)^{N} w(X)^2.$$

Die Toleranz $\widetilde{v} - \widehat{v}^\star \leq \varepsilon$ wird also durch $\frac{\alpha}{8}(1 - \frac{3}{4n})^N w(X)^2 \leq \varepsilon$ garantiert, woraus

$$N \;\geq\; \frac{\log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right)}{\log\left(1 - \frac{3}{4n}\right)}$$

folgt, was zum ersten Mal in der Stufe

$$N \;=\; \left\lceil \frac{\log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right)}{\log\left(1 - \frac{3}{4n}\right)} \right\rceil$$

eintritt.

Es bleibt also die Frage, ob man im Laufe der Iteration garantiert nach endlich vielen Schritten die Stufe N des Verzweigungsbaums erreicht. Im besten Fall gelangt man bereits nach N Iterationen in diese Stufe, wenn nämlich in jeder Iteration eine tiefere Stufe erreicht wird.

Im schlimmsten Fall konstruiert das Verfahren zunächst alle Teilboxen auf Stufe $N-1$, bevor es in Stufe N wechselt. Dazu sind

$$1 + 2 + \ldots + 2^{N-1} \;=\; \frac{2^N - 1}{2 - 1} \;=\; 2^N - 1$$

Iterationen erforderlich. Spätestens in Iteration 2^N erreicht das Verfahren dann aber Stufe N des Baumes. •

Aus diesem Beweis folgt sofort:

3.6.3 Korollar *Mit* $N = \left\lceil \frac{\log\left(\frac{8\varepsilon}{\alpha w(X)^2}\right)}{\log\left(1-\frac{3}{4n}\right)} \right\rceil$ *benötigt Algorithmus 3.3 im besten Fall höchstens* N *Schritte, im schlechtesten Fall höchstens* 2^N *Schritte.*

Die Aussage in Korollar 3.6.3 über den besten Fall ist nicht besonders hilfreich, da man a priori nicht weiß, ob ein guter oder ein schlechter Fall vorliegt. In der Praxis beobachtet man allerdings häufig Laufzeiten der Größenordnung N, und weniger häufig Laufzeiten der Größenordnung 2^N.

3.6.4 Beispiel *Für die in Beispiel 3.5.6 benutzten Daten* $n = 3$, $w(X) = 2\sqrt{3}$, $\alpha = 1$ *und* $\varepsilon = 10^{-3}$ *folgt* $N = 26$ *und* $2^N = 67.108.864$.

Bemerkungen zu Algorithmus 3.3:

- Wenn vor Anwendung des Verfahrens bereits ein Punkt $\widetilde{x} \in X$ mit niedrigerem Zielfunktionswert als $m(X)$ bekannt ist, etwa als beobachtetes Ausgangsszenario oder als Ergebnis einer Heuristik, ersetzt man Zeile 3 durch $\widetilde{v} = f(\widetilde{x})$. Durch Ausloten kann die Liste dann üblicherweise erheblich kürzer gehalten werden, und auch der Wert $\widetilde{v} - \widehat{v}^\star$ ist von vornherein kleiner. Beides kann zu einer schnelleren Terminierung des Verfahrens führen.

- Eine Neuberechnung der α-Werte auf jeder Teilbox kann zu deutlich besseren Schranken und damit zu erheblicher Senkung der Iterationszahl führen. Es besteht allerdings ein Trade-off zum Aufwand der α-Berechnungen, d.h. die CPU-Zeit kann sich dabei eventuell verlängern.

- Auch bei vorzeitigem Abbruch des Verfahrens erhält man brauchbare Informationen, nämlich *gültige* Ober- und Unterschranken an v sowie eine grobe Lokalisierung der globalen Minimalpunkte durch Ausloten. Gegebenenfalls kann außerdem ein bekannter Startpunkt zumindest durch einen neuen „besten bekannten Punkt“ $\widetilde{x}$ ersetzt werden.

- Die Relaxierung von f per αBB-Technik kann durch andere Techniken ersetzt werden, sofern Abschätzungen für den maximalen Fehler per Boxgröße bekannt sind (s. Kap. 3.9). Für einige nichtkonvexe Funktionen sind sogar explizite Formeln für die Hüllfunktionen bekannt (s. [3]).

3.7 Branch-and-Bound für konvex restringierte Probleme

Es sei

$$P: \quad \min f(x) \quad \text{s.t.} \quad x \in M$$

mit zulässiger Menge

$$M \;=\; \{x \in X |\; g_i(x) \leq 0,\; i \in I,\; h_j(x) = 0,\; j \in J\}$$

ein nichtkonvexes C^2-Problem mit $X \in \mathbb{R}^n$, $|I| < \infty$, $|J| < n$. In diesem Abschnitt setzen wir die Funktionen g_i, $i \in I$, als konvex voraus, und die Funktionen h_j, $j \in J$, als linear, so dass M eine konvexe Menge ist und die Nichtkonvexität von P einzig durch die Nichtkonvexität (bzw. die nicht bekannte Konvexität) von f bedingt ist.

Fast alle Resultate und Bemerkungen aus dem rein box-restringierten Fall übertragen sich auf konvex restringierte Probleme. Ein erster wesentlicher Unterschied besteht allerdings darin, dass in der konvexen Relaxierung $\widehat{P}^\ell$ von P auf X^ℓ die Funktion $\widehat{f}_\alpha^\ell$ nicht auf X^ℓ, sondern auf der Menge

$$M^\ell \;=\; M \cap X^\ell \;=\; \{x \in X^\ell |\; g_i(x) \leq 0,\; i \in I,\; h_j(x) = 0,\; j \in J\}$$

zu minimieren ist (vgl. Def. 3.2.4a)). Da die Menge M^ℓ wieder konvex ist, braucht sie nicht zu einer Menge $\widehat{M}^\ell$ relaxiert zu werden. Jeder Optimalpunkt von

$$\widehat{P}^\ell: \qquad \min \widehat{f}_\alpha^\ell(x) \quad \text{s.t.} \quad x \in M^\ell$$

liegt daher auch in M und kann damit zur Verbesserung der Oberschranke $\widetilde{v}$ benutzt werden.

Zweiter wesentlicher Unterschied ist, dass Mengen M^ℓ leer sein können, wie Abbildung 3.20 zeigt.

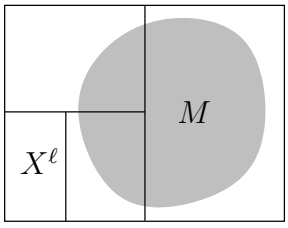


Abbildung 3.20: Leere zulässige Menge M^ℓ

Ein relaxiertes Teilproblem $\widehat{P}^\ell$ kann also wegen Inkonsistenz unlösbar sein, und die zugehörige Teilbox X^ℓ wird dann natürlich ausgelotet. Da $\widehat{P}^\ell$ ein konvexes Optimierungsproblem ist, kann man bei der Meldung eines NLP-Lösers für $\widehat{P}^\ell$, er finde keine zulässigen Punkte, darauf vertrauen, dass tatsächlich $M^\ell = \emptyset$ gilt, und damit wie üblich $\widehat{v}^\ell = +\infty$ setzen.

Auch die Initialisierung von $\widetilde{v}$ durch die Auswertung von f an $m(X)$ kann daran scheitern, dass $m(X)$ nicht in M liegt. Falls kein zulässiger Startpunkt bekannt ist, setzt man daher $\widetilde{v} = +\infty$ und lässt Algorithmus 3.4 zulässige Punkte und damit endliche Werte für $\widetilde{v}$ selbst generieren. Falls M nicht leer ist, geschieht dies für mindestens eine der ersten beiden Teilboxen von X.

Falls M andererseits leer ist, wird den ersten beiden Teilboxen von X jeweils das Infimum $\widehat{v} = +\infty$ zugeordnet. Außerdem muss $\widetilde{v}$ ebenfalls zu $+\infty$ initialisiert worden sein, so dass Algorithmus 3.4 mit der Meldung der Unlösbarkeit terminiert.

Ansonsten völlig analog zu Satz 3.6.2 beweist man:

3.7.1 Satz *Algorithmus 3.4 bricht nach endlich vielen Schritten ab.*

Auch Korollar 3.6.3 und die obigen Bemerkungen zu Algorithmus 3.3 gelten entsprechend. Die Voraussetzung zweimal stetig differenzierbarer Funktionen g_i, $i \in I$, ist hier nicht wesentlich und kann zu einer Glattheitsvoraussetzung abgeschwächt werden, unter der die Optimierungsprobleme in den Zeilen 9 und 12 algorithmisch behandelt werden können.

Algorithmus 3.4: Globale Minimierung einer konvex restringierten Funktion per αBB

Input : $X = [\underline{x}, \overline{x}] \in \mathbb{R}^n$, $f \in C^2(X, \mathbb{R})$ mit faktorisierbarer Hessematrix $D^2 f$, konvexe $g_i \in C^2(X, \mathbb{R})$, $i \in I$, lineare h_j, $j \in J$, Toleranz $\varepsilon > 0$.

Output : $\widetilde{x} \in M$ mit $v \le f(\widetilde{x}) \le v + \varepsilon$ oder Meldung der Unlösbarkeit.

begin
  Berechne ein $\alpha \ge \max\{0, -\min_{x \in X} \lambda_{\min}(x)\}$.
  if $\widetilde{x} \in M$ bekannt **then** setze $\widetilde{v} = f(\widetilde{x})$ **else** setze $\widetilde{v} = +\infty$.
  Setze $X^\star = X$, $\widehat{v}^\star = -\infty$ und Liste $= (X^\star, \widehat{v}^\star)$.
  repeat
    Halbiere $X^\star$ entlang einer längsten Kante und nenne die neuen Boxen X^1, X^2.
    Streiche $(X^\star, \widehat{v}^\star)$ aus Liste.
    for $\ell = 1, 2$ **do**
      Berechne das Infimum $\widehat{v}^\ell$ von $\widehat{f}_\alpha^\ell(x)$ auf $M^\ell = M \cap X^\ell$.
      if $\widehat{v}^\ell < \widetilde{v}$ **then**
        Füge das Paar $\left(X^\ell, \widehat{v}^\ell\right)$ zu Liste hinzu.
        Berechne einen Minimalpunkt $\widehat{x}^\ell$ von $\widehat{f}_\alpha^\ell(x)$ auf M^ℓ.
        if $f(\widehat{x}^\ell) < \widetilde{v}$ **then**
          Setze $\widetilde{x} = \widehat{x}^\ell$ und $\widetilde{v} = f(\widehat{x}^\ell)$.
          Streiche alle Paare $(X', \widehat{v}')$ mit $\widehat{v}' \ge \widetilde{v}$ aus Liste.
        end
      end
    end
    if Liste $\neq \emptyset$ **then**
      Wähle ein $(X^\star, \widehat{v}^\star)$ mit minimalem $\widehat{v}^\star$ aus Liste.
    end
  until Liste $= \emptyset$ **or** $\widetilde{v} - \widehat{v}^\star \le \varepsilon$;
end
case $\widetilde{v} < +\infty$
  $\widetilde{x}$ ist Approximation eines Minimalpunkts von P.
case $\widetilde{v} = +\infty$
  P ist unlösbar wegen Inkonsistenz.

3.8 Branch-and-Bound für nichtkonvexe Probleme

Es sei

$$P: \quad \min f(x) \quad \text{s.t.} \quad x \in M$$

mit zulässiger Menge

$$M = \{x \in X |\ g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J\}$$

ein nichtkonvexes C^2-Problem mit $X \in \mathbb{R}^n$, $|I| < \infty$, $|J| < n$. Im einfachsten Fall rührt die Nichtkonvexität daher, dass nur *eine* der beteiligten Funktionen die Konvexitätsannahmen verletzt:

- f ist nicht konvex (bzw. man weiß nicht, ob f konvex ist) oder
- ein g_i ist nicht konvex oder
- ein h_j ist nicht linear.

Üblicherweise kommen mehrere solcher Verletzungen zusammen.

Für den Branch-and-Bound-Ansatz benötigt man wieder eine konvexe Relaxierung $\widehat{P}$ von P, wobei natürlich nur diejenigen Funktionen modifiziert werden, bei denen Konvexität bzw. Linearität nicht ohnehin klar ist (vgl. die Mengen K und L in Alg. 2.1):

- falls f nicht konvex ist, berechne $\alpha \geq \max\{0, -\min_{x \in X} \lambda_f(x)\}$, wobei $\lambda_f(x)$ kleinster Eigenwert von $D^2 f(x)$ ist.
- falls g_i nicht konvex ist, berechne $\beta_i \geq \max\{0, -\min_{x \in X} \lambda_{g_i}(x)\}, i \in I$,
- falls h_j nicht linear ist, spalte die Gleichung $h_j(x) = 0$ in zwei Ungleichungen $h_j(x) \leq 0$ und $-h_j(x) \leq 0$ auf und relaxiere diese gegebenenfalls, d.h berechne $\gamma_j^+ \geq \max\left\{0, -\min_{x \in X} \lambda_{h_j}(x)\right\}$, falls h_j nicht konvex ist und $\gamma_j^- \geq \max\left\{0, -\min_{x \in X} \lambda_{-h_j}(x)\right\}$, falls $-h_j$ nicht konvex ist.

Beispiele für die Behandlung von Gleichungen:

Betrachte $h(x) = x_2 - \sin(x_1)$ auf $X = [0, 2\pi] \times [-1, 1]$. Dann ist weder h noch $-h$ konvex, und beide Funktionen sind konvex zu relaxieren.

Für $h(x) = 1 - x_1^2 - x_2^2$ ist zwar h nicht konvex, aber $-h$ ist konvex. Daher braucht nur h konvex relaxiert zu werden.

Mit den benötigten konvexen Relaxierungen der Funktionen konstruiert man nun die konvexe Relaxierung $\widehat{P}$ von P auf X mit zulässiger Menge $\widehat{M}$. Ersetzt man die Box X durch eine Teilbox $X^\ell = [\underline{x}^\ell, \overline{x}^\ell]$, so entstehen entsprechend die Restriktionen der zulässigen Menge $\widehat{M}^\ell$ des Teilproblems $\widehat{P}^\ell$.

3.8.1 Beispiel *Wir betrachten das Problem*

$$P: \quad \min\ f(x) \qquad s.t. \qquad \begin{array}{l} g_1(x) \le 0,\ g_2(x) \le 0,\ x \in X \\ h_1(x) = 0,\ h_2(x) = 0,\ h_3(x) = 0 \end{array}$$

mit f, g_1 konvex, g_2 nichtkonvex, h_1 linear, h_2 nichtlinear konvex und h_3 weder konvex noch konkav. Zu berechnen sind dann β_2, γ_2^-, γ_3^+ und γ_3^-, und man erhält

$$\begin{array}{lll} \widehat{P}: \quad \min\ f(x) & s.t. \quad x \in X, & g_1(x) \le 0 \\ & & g_2(x) + \dfrac{\beta_2}{2}(\underline{x} - x)^\intercal(\overline{x} - x) \le 0 \\ & & h_1(x) = 0 \\ & & h_2(x) \le 0 \\ & & -h_2(x) + \dfrac{\gamma_2^-}{2}(\underline{x} - x)^\intercal(\overline{x} - x) \le 0 \\ & & h_3(x) + \dfrac{\gamma_3^+}{2}(\underline{x} - x)^\intercal(\overline{x} - x) \le 0 \\ & & -h_3(x) + \dfrac{\gamma_3^-}{2}(\underline{x} - x)^\intercal(\overline{x} - x) \le 0. \end{array}$$

Für das Branch-and-Bound-Verfahren treten bei einer nichtkonvexen zulässigen Menge von P zwei neue Effekte auf:

Inkonsistenz gegebenenfalls nicht sofort erkennbar

Wie im konvex restringierten Fall können wegen Inkonsistenz unlösbare relaxierte Teilprobleme $\widehat{P}^\ell$ auftreten, wie Abbildung 3.21 für eine nichtkonvexe Menge M illustriert.

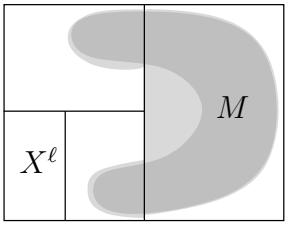


Abbildung 3.21: Leere zulässige Menge $\widehat{M}^\ell$

Da $\widehat{M}^\ell$ eine konvexe Menge ist, lässt sich ihre Inkonsistenz numerisch wieder leicht ermitteln. Nach Satz 3.2.3a) gilt $\widehat{M}^\ell \supseteq M^\ell = M \cap X^\ell$. Aus $\widehat{M}^\ell = \emptyset$ folgt also $M^\ell = \emptyset$ (vgl. Abb. 3.21), und die Teilbox X^ℓ kann dann ausgelotet werden.

Umgekehrt impliziert $M^\ell = \emptyset$ *nicht* immer, dass auch $\widehat{M}^\ell = \emptyset$ gilt, wie Abbildung 3.22 illustriert.

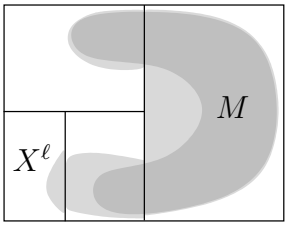


Abbildung 3.22: $M^\ell = \emptyset$ impliziert nicht $\widehat{M}^\ell = \emptyset$

Die Box X^ℓ wird dann trotz $M^\ell = \emptyset$ nicht ausgelotet und verbleibt in der Liste. Durch weitere Unterteilungen von X^ℓ lässt sich die Inkonsistenz aller entsprechenden Teilmengen von M^ℓ aber nach endlich vielen Schritten identifizieren, wie das folgende Resultat zeigt.

3.8.2 Satz *Zu einer gegebenen Teilbox X^ℓ sei $M^\ell = M \cap X^\ell = \emptyset$. Dann gilt nach endlich vielen Unterteilungen $\widehat{M}^k = \emptyset$ für jede Teilbox X^k von X^ℓ.*

Beweis. Wir führen einen Widerspruchsbeweis und nehmen dazu an, dass eine unendliche Folge von Teilboxen $X^k \subseteq X^\ell$ mit $\widehat{M}^k \neq \emptyset$ existiert. Dann gilt $\lim_k w(X^k) = 0$, und man kann (gegebenenfalls nach Wahl einer Teilfolge) ohne Beschränkung der Allgemeinheit annehmen, dass $X_{k+1} \subseteq X_k$, $k \in \mathbb{N}$, gilt. Die Boxmittelpunkte $m(X^k)$ besitzen damit einen Limes $x^\star \in X^\ell$. Offenbar gilt auch $x^\star \in X^k$ für alle $k \in \mathbb{N}$.

Wegen $M^\ell = \emptyset$ muss an $x^\star$ mindestens eine Restriktion verletzt sein, etwa gelte $g_i(x^\star) > 0$ mit einem $i \in I$. Die Stetigkeit von g_i impliziert, dass für alle hinreichend großen $k \in \mathbb{N}$ mit $x^\star \in X^k$ auch alle anderen Elemente von X^k diese Ungleichung erfüllen. Nach dem Satz von Weierstraß gilt also $\min_{x \in X^k} g_i(x) = c > 0$. Für hinreichend große $k \in \mathbb{N}$ erfüllt jedes $x \in X^k$ daher nach Lemma 3.4.2c)

$$(\widehat{g}_i)^k_{\beta_i}(x) \;\geq\; g_i(x) - \frac{\beta_i}{8}\, w(X^k)^2 \;\geq\; c - \frac{\beta_i}{8}\, w(X^k)^2 \;\geq\; \frac{c}{2} \;>\; 0.$$

Daraus folgt $\widehat{M}^k = \emptyset$, im Widerspruch zur Annahme. •

Mögliche Unzulässigkeit von Minimalpunkten der Relaxierungen

Ein *wesentliches* Problem bei nichtkonvex restringierten Problemen besteht darin, dass ein Optimalpunkt $\widehat{x}^\ell$ von $\widehat{P}^\ell$ ist nicht notwendigerweise zulässig für P zu sein braucht, denn etwa Randpunkte von $\widehat{M}^\ell$ (an denen sich Minimalpunkte gerne aufhalten) liegen nicht notwendigerweise auch in M^ℓ (vgl. Abb. 3.21).

Dies hat zwei Konsequenzen. Einerseits kann man (wie bei jedem Äußere-Approximations-Verfahren) nur erwarten, dass die vom Algorithmus erzeugten Punkte $\widetilde{x}$ *asymptotisch* zulässig sind, dass die nach endlich vielen Schritten erzeugte Approximation eines optimalen Punktes also nicht notwendigerweise in M liegt. Dies *könnte* man (wie bei Äußere-Approximations-Verfahren üblich) dadurch abfangen, dass man durch eine weitere Toleranz $\varepsilon_M > 0$ und eine Straftermfunktion für M eine „maximal erlaubte Unzulässigkeit“ von $\widetilde{x}$ definiert, etwa durch

$$\varrho(\widetilde{x}) \;:=\; \sum_{i \in I} g_i^+(\widetilde{x}) + \sum_{j \in J} |h_j(\widetilde{x})| \;\leq\; \varepsilon_M\,,$$

wobei $g_i^+(\widetilde{x}) = \max\{0, g_i(\widetilde{x})\}$ den „positiven Anteil“ von $g_i(\widetilde{x})$ bezeichnet. Mit einem ähnlichen Argument wie im Beweis von Satz 3.8.2 lässt sich zeigen, dass dieses Kriterium für jedes gegebene $\varepsilon_M > 0$ nach endlich vielen Boxunterteilungen erfüllt ist.

Die zweite Konsequenz ist allerdings weitreichender und lässt sich mit dieser Konstruktion nicht behandeln: Für $\widehat{x}^\ell \notin M$ ist der Wert $f(\widehat{x}^\ell)$ nicht notwendigerweise eine Oberschranke für v und kann daher nicht zum Update von $\widetilde{v}$ benutzt werden (s. Abb. 3.23).

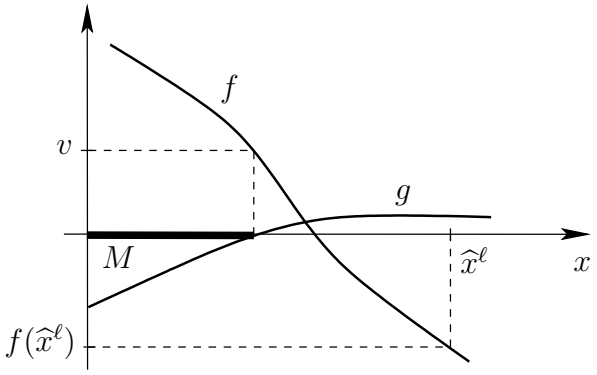


Abbildung 3.23: Unzulässiger Punkt liefert keine Oberschranke für v

Als Ausweg überprüft man natürlich zunächst, ob nicht doch $\widehat{x}^\ell \in M$ gilt (durch Einsetzen in die Funktionen g_i, $i \in I$, h_j, $j \in J$). Falls ja, nutzt man $f(\widehat{x}^\ell)$ zum Update der Oberschranke $\widetilde{v}$ von v.

Falls nein, sind verschiedene *Heuristiken* zum weiteren Vorgehen verbreitet:

a) Man kann darauf hoffen, dass der Algorithmus irgendwann selbst zulässige Punkte bestimmt. Dies ist leider nicht garantiert.

b) Man kann versuchen, einen lokalen Minimalpunkt x^ℓ von f auf $M^\ell = M \cap X^\ell$ per NLP-Löser zu bestimmen und den Wert $f(x^\ell)$ zum Update von $\widetilde{v}$ zu benutzen. Neben dem Aufwand dieses Ansatzes bestehen die Probleme, dass ein Punkt $x^\ell \in M^\ell$ bei nichtkonvexen Restriktionen nicht notwendigerweise gefunden wird, selbst wenn er existiert, und dass unklar ist, ob mit den so bestimmten Oberschranken das Abbruchkriterium $\widetilde{v} - \widehat{v}^\star \leq \varepsilon$ nach endlich vielen Schritten erfüllt wird.

c) Man lässt im obigen Sinne „fast zulässige“ Punkte für den Update von $\widetilde{v}$ zu, also $\widehat{x}^\ell$ mit $\varrho(\widehat{x}^\ell) < \varepsilon_M$. Abbildung 3.24 zeigt, dass auch hierbei beliebig falsche Werte von $\widetilde{v}$ entstehen können. Auf einen Ansatz, den entstehenden Fehler mit Lipschitz-Konstanten abzuschätzen, werden wir in Kapitel 3.9 eingehen. Auch dies wird aber nicht zum Ziel führen.

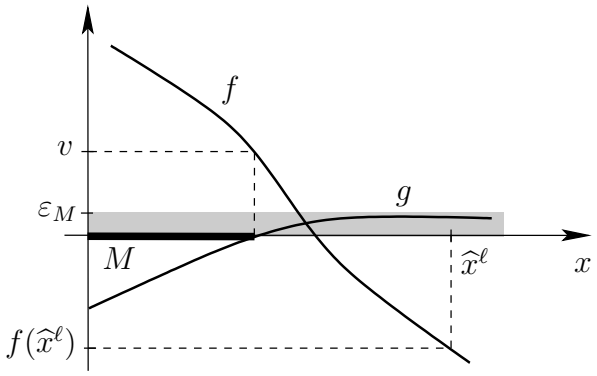


Abbildung 3.24: ε_M-zulässiger Punkt liefert keine Oberschranke für v

Ein deterministischer Ansatz zur Bestimmung eines zulässigen Punkts im Fall $\widehat{x}^\ell \notin M$ wird in [8] vorgestellt, kann im Rahmen dieser Vorlesung aber nicht diskutiert werden. Benutzt man diese Technik zur Ausgestaltung von Zeile 14 in Algorithmus 3.5, so gelten analoge Bemerkungen wie zu Algorithmus 3.3, und es lassen sich wieder der Abbruch von Algorithmus 3.5 nach endlich vielen Schritten sowie Komplexitätsresultate zeigen.

3.8.3 Übung *Nach Satz 3.2.5c) kann für den Optimalwert des konvexen Hüllproblems eines Optimierungsproblems mit nichtkonvexer zulässiger Menge $\widehat{\widehat{v}} < v$ gelten. In einem solchen Fall behalten die Optimalwerte $\widehat{v}$ aller konvex relaxierten Probleme wegen $\widehat{v} \leq \widehat{\widehat{v}} < v$ einen positiven Abstand von v. Warum steht dies nicht im Widerspruch dazu, dass sich für die iterativ verfeinerten Relaxierungen aus Algorithmus 3.5 trotzdem ein Konvergenzresultat zeigen lässt?*

Algorithmus 3.5: Globale Minimierung eines nichtkonvexen Problems per αBB

Input : $X = [\underline{x}, \overline{x}] \in \mathbb{R}^n$, f, g_i, $h_j \in C^2(X, \mathbb{R})$, $i \in I$, $j \in J$, mit
- f konvex oder $D^2 f$ komp.weise fakt.bar,
- g_i konvex oder $D^2 g_i$ komp.weise fakt.bar, $i \in I$,
- h_j linear oder
  - h_j konvex oder $D^2 h_j$ komp.weise fakt.bar, $j \in J$,
  - $-h_j$ konvex oder $-D^2 h_j$ komp.weise fakt.bar, $j \in J$,

Toleranz $\varepsilon > 0$.

Output : $\widetilde{x} \in M$ mit $v \le f(\widetilde{x}) \le v + \varepsilon$ oder Meldung der Unlösbarkeit.

begin
  Berechne die erforderlichen α, β_i, $i \in I$, $\gamma_j^{\pm}$, $j \in J$.
  if $\widetilde{x} \in M$ bekannt **then** setze $\widetilde{v} = f(\widetilde{x})$ **else** setze $\widetilde{v} = +\infty$.
  Setze $X^\star = X$, $\widehat{v}^\star = -\infty$ und Liste $= (X^\star, \widehat{v}^\star)$.
  repeat
    Halbiere $X^\star$ entlang einer längsten Kante und nenne die neuen Boxen X^1, X^2.
    Streiche $(X^\star, \widehat{v}^\star)$ aus Liste.
    for $\ell = 1, 2$ **do**
      Berechne das Infimum $\widehat{v}^\ell$ von $\widehat{f}_\alpha^\ell(x)$ auf $\widehat{M}^\ell$.
      if $\widehat{v}^\ell < \widetilde{v}$ **then**
        Füge das Paar $\left(X^\ell, \widehat{v}^\ell\right)$ zu Liste hinzu.
        Berechne einen Minimalpunkt $\widehat{x}^\ell$ von $\widehat{f}_\alpha^\ell(x)$ auf $\widehat{M}^\ell$.
        if $\widehat{x}^\ell \notin M$ **then**
          Versuche, $\widehat{x}^\ell$ durch einen zulässigen Punkt zu ersetzen.
        end
        if $\widehat{x}^\ell \in M$ **and** $f(\widehat{x}^\ell) < \widetilde{v}$ **then**
          Setze $\widetilde{x} = \widehat{x}^\ell$ und $\widetilde{v} = f(\widehat{x}^\ell)$.
          Streiche alle Paare $(X', \widehat{v}')$ mit $\widehat{v}' \ge \widetilde{v}$ aus Liste.
        end
      end
    end
    if Liste $\neq \emptyset$ **then**
      Wähle ein $(X^\star, \widehat{v}^\star)$ mit minimalem $\widehat{v}^\star$ aus Liste.
    end
  until Liste $= \emptyset$ **or** $\widetilde{v} - \widehat{v}^\star \le \varepsilon$;
end
case $\widetilde{v} < +\infty$
  $\widetilde{x}$ ist Approximation eines Minimalpunkts von P.
case $\widetilde{v} = +\infty$
  P ist unlösbar wegen Inkonsistenz.

3.9 Lipschitz-Optimierung

3.9.1 Definition (Lipschitz-Stetigkeit)
Für $M \subseteq \mathbb{R}^n$ heißt $f : M \to \mathbb{R}$ Lipschitz-stetig*, falls eine Konstante $L > 0$ mit*

$$\forall\, x, y \in M : \quad |f(x) - f(y)| \;\leq\; L \cdot \|x - y\|_2$$

existiert. L heißt dann Lipschitz-Konstante *für f auf M.*

Für $x = y$ ist die Lipschitz-Bedingung uninteressant. Für $x \neq y$ besagt sie, dass die Sekante durch die Punkte $(x, f(x))$ und $(y, f(y))$ an den Graphen von f eine „betraglich beschränkte Steigung“ besitzt. Das Auftreten des Betrags erklärt sich dadurch, dass für $n > 1$ nicht alle Argumente x und y durch „$\leq$“ vergleichbar sind und daher nicht klar ist, ob man die Sekantensteigung von x aus in Richtung y oder in entgegengesetzter Richtung messen soll. Im betraglichen Ausdruck $|f(x) - f(y)| \,/\, \|x - y\|_2$ spielt dies jedoch keine Rolle. Für eine Lipschitz-stetige Funktion ist dieser Ausdruck also für jede Wahl von $x, y \in M$ durch die gleiche Konstante $L > 0$ beschränkt. Die „Variation“ von f ist in diesem Sinne also beschränkt, und man spricht manchmal auch von „Dehnungsbeschränktheit“.

Beispiele:

- Die Funktion $f(x) = \sqrt[3]{x}$ ist auf $M = [-1, 1]$ nicht Lipschitz-stetig. Abbildung 3.25 illustriert, warum dies geometrisch klar ist: man kann mit den Wahlen $x = 0$ und $y^\nu = \frac{1}{\nu}$ für $\nu \to \infty$ beliebig steile Sekanten an den Graphen von f erzeugen.

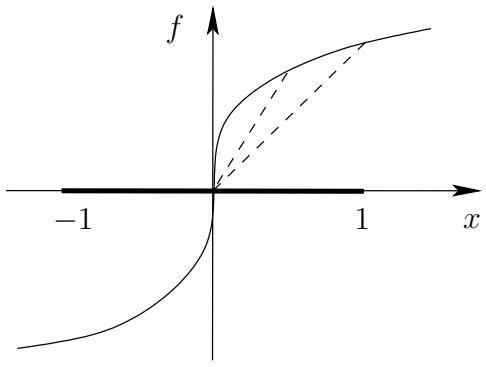


Abbildung 3.25: Nicht Lipschitz-stetige Funktion

- Die Funktion $f(x) = x^2$ ist auf $M = \mathbb{R}$ nicht Lipschitz-stetig. Auch dies ist geometrisch klar, formal sieht man es wie folgt: für alle $x, y \in \mathbb{R}$ gilt

$$|f(x) - f(y)| \;=\; |x^2 - y^2| \;=\; |x + y| \cdot |x - y|.$$

  Da der Ausdruck $|x + y|$ durch passende Wahlen von $x, y \in \mathbb{R}$ beliebig groß wird, findet man keine Lipschitz-Konstante L.

  Allgemeiner sieht man mit diesem Argument, dass $f(x) = x^2$ auf jeder beschränkten Menge M Lipschitz-stetig ist, und auf jeder unbeschränkten Menge M nicht.

Bemerkungen:

- Aus Lipschitz-Stetigkeit von f auf M folgt Stetigkeit von f auf M.
- Nach dem Satz von Rademacher sind Lipschitz-stetige Funktionen in einem gewissen Sinne „fast überall“ differenzierbar.
- Die Funktion $f(x) = |x|$ ist auf $\mathbb{R}$ nicht differenzierbar, aber konvex und Lipschitz-stetig.
- Die Funktion $f(x) = -|x|$ ist auf $\mathbb{R}$ weder differenzierbar noch konvex, aber Lipschitz-stetig.
- Für eine Lipschitz-Konstante L von f auf M ist auch jedes $L' > L$ Lipschitz-Konstante von f auf M. Im Allgemeinen ist man daran interessiert, möglichst kleine oder sogar die kleinste Lipschitz-Konstante zu identifizieren, da dies den meisten Informationsgehalt über das Verhalten von f auf M liefert.
- Lipschitz-Stetigkeit lässt sich auch bezüglich beliebiger anderer Normen $\|\cdot\|$ anstelle von $\|\cdot\|_2$ in Definition 3.9.1 betrachten.

3.9.2 Lemma *Es seien $M \subseteq \mathbb{R}^n$ nicht-leer, kompakt und konvex sowie $f \in C^1(M, \mathbb{R})$. Dann ist f auf M Lipschitz-stetig, und jedes $L > 0$ mit*

$$L \;\geq\; \max_{x \in M} \|\nabla f(x)\|_2$$

ist Lipschitz-Konstante von f auf M.

Beweis. Nach dem Mittelwertsatz gibt es für alle $x, y \in M$ ein z auf der Verbindungsstrecke zwischen x und y mit

$$f(x) \;=\; f(y) + \langle \nabla f(z), x - y \rangle,$$

wobei die Konvexität von M auch $z \in M$ impliziert. Mit der Cauchy-Schwarz-Ungleichung folgt daraus

$$\begin{aligned} |f(x) - f(y)| &= |\langle \nabla f(z), x - y \rangle| \\ &\leq \|\nabla f(z)\|_2 \cdot \|x - y\|_2 \\ &\leq \left(\sup_{z \in M} \|\nabla f(z)\|_2 \right) \cdot \|x - y\|_2 \,. \end{aligned}$$

Wegen $f \in C^1(M, \mathbb{R})$ sind ∇f und damit auch $\|\nabla f\|_2$ auf M stetig. Da M als nicht-leer und kompakt vorausgesetzt ist, liefert der Satz von Weierstraß, dass das Supremum angenommen wird. Die Unabhängigkeit der Zahl

$$\max_{z \in M} \|\nabla f(z)\|_2$$

von x und y liefert nun die Behauptung. •

3.9.3 Übung *Da man nur an möglichst kleinen Lipschitz-Konstanten interessiert ist, wäre es in Lemma 3.9.2 naheliegend, direkt* $L := \max_{x \in M} \|\nabla f(x)\|_2$ *als Lipschitz-Konstante zu definieren. In welchem Trivialfall würde dann aber ein formales Problem entstehen?*

3.9.4 Übung *Zeigen Sie anhand der Menge* $M = \{0\} \times \mathbb{R}$ *und der Funktion* $f(x) = x_1$*, dass es möglich sein kann, die Größe der in Lemma 3.9.2 angegebenen Lipschitz-Konstanten noch zu unterbieten.*

3.9.5 Übung *Zeigen Sie unter den Voraussetzungen von Lemma 3.9.2, dass* $L = \max_{x \in M} \|\nabla f(x)\|_1$ *eine Lipschitz-Konstante für* f *auf* M *bezüglich* $\|\cdot\|_\infty$ *ist.*

Aus Lemma 3.9.2 erhält man sofort folgendes Resultat:

3.9.6 Satz *Es seien* $M \in \mathbb{R}^n$*,* $f \in C^1(M, \mathbb{R})$*, und* $g := \nabla f$ *sei faktorisierbar mit natürlicher Intervallerweiterung* G*. Bezeichnet zudem* NORM *die natürliche Intervallerweiterung der Funktion* $\mathrm{norm}(x) := \|x\|_2$*, dann ist* $\overline{\mathrm{NORM}}(G(M))$ *eine Lipschitz-Konstante von* f *auf* M*.*

3.9.7 Beispiel *Auf $M = [0, 2]$ sei eine Lipschitz-Konstante für $f(x) = x^2$ zu bestimmen. Grafisch überzeugt man sich leicht davon, dass keine Sekante eine höhere Steigung als 4 (nämlich die Tangentensteigung bei $x = 2$) aufweisen kann, und dass diese Steigung beliebig gut durch Sekantensteigungen approximiert werden kann. Daher ist $L = 4$ die bestmögliche Lipschitz-Konstante.*

Nach Satz 3.9.6 berechnet man eine Lipschitz-Konstante wie folgt: es gilt $g(x) = \nabla f(x) = f'(x) = 2x$ und damit $G(X) = 2X$. Daraus folgt $\mathrm{NORM}(G(X)) = \mathrm{ABS}(2X)$ sowie

$$\mathrm{NORM}(G(M)) \;=\; \mathrm{ABS}(2[0,2]) \;=\; \mathrm{ABS}([0,4]) \;=\; [0,4].$$

Wir erhalten aus Satz 3.9.6 also die Lipschitz-Konstante $\overline{\mathrm{NORM}}(G(M)) = 4$, was mit der grafisch ermittelten besten Lipschitz-Konstante übereinstimmt. Wegen des Abhängigkeitseffekts kann man im Allgemeinen natürlich nicht erwarten, dass Satz 3.9.6 immer eine beste *Lipschitz-Konstante liefert.*

In einem nächsten Schritt lassen sich mit Hilfe von Lipschitz-Konstanten gültige Ober- und Unterschranken für Funktionen auf Mengen ermitteln:

3.9.8 Lemma *Für $X \subseteq \mathbb{R}^n$ und $f : X \to \mathbb{R}$ sei L eine Lipschitz-Konstante, und $y \in X$ sei gegeben. Dann gilt*

$$\forall\, x \in X: \quad f(x) \;\in\; [f(y) - L\|x - y\|_2,\; f(y) + L\|x - y\|_2]\,.$$

Beweis. Für alle $x \in X$ gilt

$$f(x) - f(y) \;\le\; |f(x) - f(y)| \;\le\; L\|x - y\|_2$$

und damit $f(x) \le f(y) + L\|x - y\|_2$. Analog folgt aus

$$f(y) - f(x) \;\le\; |f(x) - f(y)| \;\le\; L\|x - y\|_2$$

die Ungleichung $f(x) \ge f(y) - L\|x - y\|_2$ und damit insgesamt die Behauptung.

•

3.9.9 Beispiel *Nach Beispiel 3.9.7 ist $L = 4$ Lipschitz-Konstante von $f(x) = x^2$ auf $X = [0, 2]$. Mit $y = 1$ folgt aus Lemma 3.9.8*

$$\forall\, x \in [0,2]: \quad f(x) \;\in\; [1 - 4 \cdot |x - 1|,\; 1 + 4 \cdot |x - 1|]\,.$$

Lemma 3.9.8 lässt sich auch folgendermaßen interpretieren: der Graph von f, also die Menge $\{(x, f(x))|\, x \in X\}$ ist für jedes beliebige $y \in X$ in der Menge $\{(x, \alpha) \in X \times \mathbb{R}|\, |\alpha - f(y)| \leq L\|x - y\|_2\}$ enthalten. Für $n = 1$ ist dies ein (Teil eines) Doppelkegels mit Scheitelpunkt $(y, f(y))$, für $n > 1$ allerdings nicht (dann ist nur das Komplement dieser Menge ein Doppelkegel).

Direkte Anwendung auf Algorithmus 3.5

Für das Branch-and-Bound-Verfahren 3.5 bietet die Kenntnis einer Lipschitz-Konstante von f die Möglichkeit für einen Update der Oberschranke $\widetilde{v}$ von v im Fall der Unzulässigkeit von $\widehat{x}^\ell$ in Zeile 14. Dazu wird der „falsche“ Wert $f(\widehat{x}^\ell)$ mit Hilfe der Lipschitz-Konstante von f zu einer garantierten, aber möglicherweise sehr groben Oberschranke korrigiert. Die Bestimmung eines zugehörigen zulässigen Punktes ist dabei aber nicht ohne weiteres möglich, und auch die Konvergenz des Verfahrens kann mit diesen Oberschranken nicht garantiert werden.

L sei dazu eine Lipschitz-Konstante von f auf X, und

$$M \;=\; \{x \in X|\; g_i(x) \leq 0,\; i \in I,\; h_j(x) = 0,\; j \in J\}$$

sei nicht leer. Aus Lemma 3.9.8 mit $y = \widehat{x}^\ell$ folgt dann

$$\forall\; x \in M: \quad v \;\leq\; f(x) \;\leq\; f(\widehat{x}^\ell) + L\|x - \widehat{x}^\ell\|_2,$$

also ist für jede beliebige Wahl von $x \in M$ die Zahl $f(\widehat{x}^\ell) + L\|x - \widehat{x}^\ell\|_2$ eine Oberschranke für v, die zur Verbesserung von $\widetilde{v}$ herangezogen werden kann. Die *beste* so ermittelbare Oberschranke ist

$$f(\widehat{x}^\ell) + L \min_{x \in M} \|x - \widehat{x}^\ell\|_2 \;=\; f(\widehat{x}^\ell) + L\,\mathrm{dist}(\widehat{x}^\ell, M),$$

aber zu ihrer Bestimmung wäre die Berechnung der Distanz $\mathrm{dist}(\widehat{x}^\ell, M)$ von $\widehat{x}^\ell$ zu M erforderlich, d.h. die Lösung eines weiteren globalen Optimierungsproblems.

Wegen $M \neq \emptyset$ gibt es jedenfalls irgendwo in X einen Punkt $x \in M$. Schlimmstenfalls besitzt dieser maximalen Abstand von $\widehat{x}^\ell$ in X, also ist eine gültige, aber vermutlich grobe Oberschranke für v auch

$$f(\widehat{x}^\ell) + L \max_{x \in X} \|x - \widehat{x}^\ell\|_2.$$

Die Maximierung von $\|x - \widehat{x}^\ell\|_2$ über X löst man beispielsweise durch Kenntnis der Tatsache, dass ein Maximalpunkt sicher in einer Ecke von X zu finden

ist (Eckensatz der konvexen Maximierung), was auf 2^n Kandidaten führt. Wegen Übungen 1.3.5 und 1.3.2 erhalten wir sogar die explizite Darstellung

$$\begin{aligned}\max_{x\in X} \|x-\widehat{x}^\ell\|_2 &= \sqrt{\max_{x\in X} \sum_{i=1}^n (x_i-\widehat{x}_i^\ell)^2} \\ &= \sqrt{\sum_{i=1}^n \max_{x_i\in[\underline{x}_i\,,\overline{x}_i]} (x_i-\widehat{x}_i^\ell)^2} \\ &= \sqrt{\sum_{i=1}^n \big(\max\{\widehat{x}_i^\ell-\underline{x}_i\,,\ \overline{x}_i-\widehat{x}_i^\ell\}\big)^2}.\end{aligned}$$

Diese Möglichkeit zum Update der Oberschranke $\widetilde{v}$ lässt sich leider *nicht* auf beliebige Teilboxen X^ℓ übertragen, da man dafür zunächst

$$M^\ell \ = \ \{x\in X^\ell |\ g_i(x)\le 0,\ i\in I,\ h_j(x)=0,\ j\in J\} \ \neq\ \emptyset$$

überprüfen müsste.

Auch die Kenntnis einer „Oberschranke der Unzulässigkeit“ von $\widehat{x}^\ell$ im Sinne von

$$\varrho(\widehat{x}^\ell) \ = \ \sum_{i\in I} g_i^+(\widehat{x}^\ell) + \sum_{j\in J} |h_j(\widehat{x}^\ell)| \ \le\ \varepsilon_M$$

hilft nicht in einfacher Weise weiter, was wir uns kurz für den Fall $I=\{1\}$ und $J=\emptyset$ überlegen, also für $\varrho(\widehat{x}^\ell)=g^+(\widehat{x}^\ell)$. Zu vermuten wäre zunächst, dass die Relaxierung $M_{\varepsilon_M}=\{x\in\mathbb{R}^n|\, g(x)\le\varepsilon_M\}$ der Menge $M=\{x\in\mathbb{R}^n|\, g(x)\le 0\}$ sich nur wenig von M unterscheidet. Zum Beispiel sollte es möglich sein, für jeden Punkt $\widehat{x}^\ell\in M_{\varepsilon_M}$ die Distanz $\operatorname{dist}(\widehat{x}^\ell,M)$ etwa in der Form

$$\operatorname{dist}(\widehat{x}^\ell,M) \ \le\ \gamma\,\varepsilon_M \tag{3.9.3}$$

mit einer von $\widehat{x}^\ell$ unabhängigen Konstante $\gamma>0$ abzuschätzen. Dann erhielte man eine Oberschranke $\widetilde{v}$ nach obigen Überlegungen durch

$$\forall\, x\in M:\quad v\ \le\ f(x)\ \le\ f(\widehat{x}^\ell)+L\operatorname{dist}(\widehat{x}^\ell,M)\ \le\ f(\widehat{x}^\ell)+L\,\gamma\,\varepsilon_M\,.$$

Die Abschätzung in (3.9.3) hängt leider mit dem „Gegenteil“ einer Lipschitz-Abschätzung zusammen, nämlich mit der Abschätzung des Abstands der Argumente durch ein Vielfaches des Abstands der Funktionswerte von g^+. Etwas genauer sieht man dies wie folgt: falls ein $\gamma>0$ existiert, so dass es zu jedem $\widehat{x}^\ell\in M_{\varepsilon_M}$ ein $x\in M$ mit

$$\|x-\widehat{x}^\ell\|_2 \ \le\ \gamma\,|g^+(x)-g^+(\widehat{x}^\ell)|$$

gibt, dann folgt wegen der Zulässigkeit der $x \in M$ und der Nichtnegativität von $g^+(\widehat{x}^\ell)$

$$\|x - \widehat{x}^\ell\|_2 \;\leq\; \gamma\, g^+(\widehat{x}^\ell) \;=\; \gamma\, \varrho(\widehat{x}^\ell) \;\leq\; \gamma\, \varepsilon_M$$

und damit

$$\operatorname{dist}(\widehat{x}^\ell, M) \;=\; \inf_{x \in M} \|x - \widehat{x}^\ell\|_2 \;\leq\; \gamma\, \varepsilon_M\,.$$

Für lineare und konvexe Funktionen sind Abschätzungen wie in (3.9.3) tatsächlich möglich und als *Hoffman-Lemma* oder *Fehlerschranken* bekannt, wobei die Konstante γ als *Hoffman-Konstante* bezeichnet wird [12]. Für den uns interessierenden nichtkonvexen Fall zeigt allerdings wieder Abbildung 3.24, dass die Angabe von Fehlerschranken auf grundsätzliche Probleme stößt.

Eine Variation von Algorithmus 3.5

Unabhängig von αBB führt die Kenntnis von Lipschitz-Konstanten allerdings auf eine weitere Möglichkeit, Relaxierungen von Funktionen zu konstruieren, die dann aber nicht notwendigerweise konvex sind. Im Folgenden diskutieren wir nur den Fall $n = 1$, in dem eine C^1-Funktion $f : \mathbb{R} \to \mathbb{R}$ auf $X = [\underline{x}, \overline{x}] \in \mathbb{IR}$ zu minimieren ist. Dabei wird die Funktion

$$\psi(x) \;=\; |x - m(X)| - \frac{w(X)}{2} \;=\; \left| x - \frac{\underline{x} + \overline{x}}{2} \right| - \frac{\overline{x} - \underline{x}}{2}$$

die Rolle von $\psi(x) = \frac{1}{2}(\underline{x} - x)(\overline{x} - x)$ aus dem αBB-Ansatz übernehmen. Sie ist in Abbildung 3.26 skizziert.

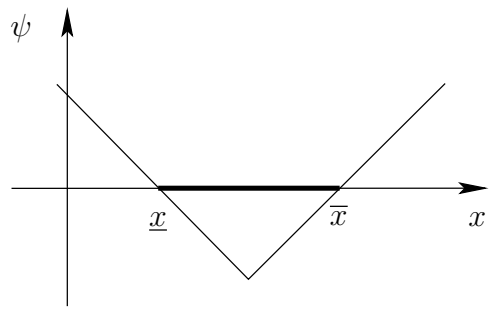


Abbildung 3.26: Die Funktion ψ in der Lipschitz-Optimierung

3.9.10 Satz *Es seien $X \in \mathbb{IR}$, $f \in C^1(X, \mathbb{R})$ und L eine Lipschitz-Konstante von f auf X mit $L \geq \max_{x \in X} |f'(x)|$. Dann gelten mit*

$$\widehat{f}_L(x) \;:=\; f(x) + L \cdot \psi(x)$$

die folgenden Aussagen:

a) $\forall\, x \in X: \quad \widehat{f}_L(x) \;\leq\; f(x)$,

b) $\widehat{f}_L(\underline{x}) = f(\underline{x})$ *und* $\widehat{f}_L(\overline{x}) = f(\overline{x})$,

c) $\max\limits_{x \in X} (f(x) - \widehat{f}_L(x)) \;=\; \dfrac{L}{2}\, w(X)$,

d) $\widehat{f}_L$ *ist monoton fallend auf* $[\underline{x},\, m(X))$ *und monoton wachsend auf* $(m(X), \overline{x}]$,

e) $m(X)$ *ist globaler Minimalpunkt von* $\widehat{f}_L$ *auf* X *mit Minimalwert* $f(m(X)) - \frac{L}{2}\, w(X)$.

Beweis.

a) Für alle $x \in [\underline{x},\, m(X)]$ gilt

$$\psi(x) \;=\; m(X) - x - \frac{w(X)}{2} \;=\; \underline{x} - x \;\leq\; 0,$$

und für alle $x \in [m(X), \overline{x}]$

$$\psi(x) \;=\; x - m(X) - \frac{w(X)}{2} \;=\; x - \overline{x} \;\leq\; 0.$$

Mit $L \geq 0$ folgt die Behauptung.

b) Die Behauptung ergibt sich sofort aus $\psi(\underline{x}) = \underline{x} - \underline{x} = 0$ und $\psi(\overline{x}) = \overline{x} - \overline{x} = 0$.

c) Für alle $x \in X$ gilt

$$f(x) - \widehat{f}_L(x) \;=\; -L\psi(x),$$

also

$$\max_{x \in X} (f(x) - \widehat{f}_L(x)) \;=\; -L \min_{x \in X} \psi(x).$$

Da ψ auf $[\underline{x},\, m(X))$ Steigung -1 und auf $(m(X), \overline{x}]$ Steigung $+1$ besitzt, ist $m(X)$ Minimalpunkt von ψ mit Wert $\psi(m(X)) = -\frac{\overline{x}-\underline{x}}{2}$. Daraus folgt

$$\max_{x \in X} (f(x) - \widehat{f}_L(x)) \;=\; -L \left(-\frac{\overline{x} - \underline{x}}{2} \right) \;=\; \frac{L}{2}\, w(X).$$

d) Auf $[\underline{x},\, m(X))$ gilt

$$\widehat{f}_L'(x) \;=\; f'(x) - L \;\leq\; |f'(x)| - L \;\leq\; \max_{x\in X} |f'(x)| - L \;\leq\; 0,$$

und auf $(m(X), \overline{x}]$

$$\widehat{f}_L'(x) \;=\; f'(x) + L \;\geq\; -|f'(x)| + L \;\geq\; -\max_{x\in X} |f'(x)| + L \;\geq\; 0.$$

e) Aus d) folgt sofort, dass $m(X)$ globaler Minimalpunkt von $\widehat{f}_L$ auf X ist.

•

Mit Hilfe von Satz 3.9.10 kann man völlig analog zu Algorithmus 3.3 ein Branch-and-Bound-Verfahren angeben, das nach endlich vielen Schritten abbricht.

Vorteile:

- Statt $f \in C^2$ ist nur $f \in C^1$ erforderlich,
- in Zeile 9 muss kein Verfahren der nichtlinearen Optimierung bemüht werden, um $\widehat{x}^\ell$ und $\widehat{v}^\ell$, $\ell = 1, 2$, zu berechnen, sondern Satz 3.9.10e) gibt Formeln dafür an.

Nachteile:

- Die Struktur von f wird schlecht ausgenutzt, denn beispielsweise sind Minimalpunkte $\widehat{x}^\ell$ der Relaxierungen am Rand von Boxen ausgeschlossen,
- die Schranken $\widehat{v}^\ell$ sind üblicherweise viel schlechter als bei αBB-Relaxierungen,
- die Anzahl der nötigen Iterationen ist im Vergleich zu αBB üblicherweise viel höher, die CPU-Zeit pro Iteration allerdings auch erheblich kürzer,
- die Verallgemeinerung auf den Fall $n > 1$ ist nicht klar.

Als verfeinerte Möglichkeiten, die Steigungsinformation der zugrundeliegenden Funktion auszunutzen, existieren etwa *zentrische Formen*, *Neumaier-Unterschätzer* und ihre Kombination, die sogenannten *Kites*. Details hierzu finden sich beispielsweise in [9].

Literaturverzeichnis

[1] W. Alt, *Numerische Verfahren der konvexen, nichtglatten Optimierung,* Teubner, 2004.

[2] M.S. Bazaraa, H.D. Sherali, C.M. Shetty, *Nonlinear Programming,* Wiley, 1993.

[3] C.A. Floudas, *Deterministic Global Optimization,* Kluwer, 2000.

[4] O. Güler, *Foundations of Optimization,* Springer, 2010.

[5] J.-B. Hiriart-Urruty, C. Lemaréchal, *Fundamentals of Convex Analysis,* Springer, 2001.

[6] R. Horst, H. Tuy, *Global Optimization,* Springer, 1996.

[7] H.Th. Jongen, K. Meer, E. Triesch, *Optimization Theory,* Kluwer, 2004.

[8] P. Kirst, O. Stein, P. Steuermann, *Deterministic upper bounds for spatial branch-and-bound methods in global minimization with nonconvex constraints,* TOP, Vol. 23 (2015), 591-616.

[9] A. Neumaier, *Interval Methods for Systems of Equations,* Cambridge University Press, 1990.

[10] S. Nickel, O. Stein, K.-H. Waldmann, *Operations Research,* 2. Auflage, Springer-Gabler, 2014.

[11] O. Stein, *Gemischt-ganzzahlige Optimierung I und II,* Skript, Karlsruher Institut für Technologie (KIT), 2016.

[12] O. Stein, *Konvexe Analysis,* Skript, Karlsruher Institut für Technologie (KIT), 2017.

[13] O. Stein, *Nichtlineare Optimierung I und II,* Skript, Karlsruher Institut für Technologie (KIT), 2016.

[14] O. Stein, *Parametrische Optimierung,* Skript, Karlsruher Institut für Technologie (KIT), 2016.

[15] O. Stein, *Twice differentiable characterizations of convexity notions for functions on full dimensional convex sets,* Schedae Informaticae, Vol. 21 (2012), 55-63.

Index