TECHNISCHE UNIVERSITÄT BERLIN

Fakultät II – Institut für Mathematik

Nichtlineare Optimierung

Vorlesung im Sommersemester 2015

Fredi Tröltzsch

Das Skript wurde gemeinsam von D. Hömberg und F. Tröltzsch erarbeitet.
Grundlage der Vorlesung ist das Buch

W. Alt, Nichtlineare Optimierung, 2. Auflage, VIEWEG 2011

Inhaltsverzeichnis

1 Optimierungsaufgaben

1.1 Empfohlene Literatur

1. Alt, W., *Nichtlineare Optimierung*. Vieweg, Braunschweig/Wiesbaden 2002.
2. Geiger, C., Kanzow, C., *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer-Verlag, Berlin 2002.
3. Gill, P.E., Murray, W., und M.H. Wright, *Practical Optimization*. Academic Press, London 1981.
4. Kelley, C.T., *Iterative Methods for Optimization.* SIAM, Philadelphia 1999.
5. Nocedal, J. und Wright, S.J., *Numerical Optimization.* Springer, New York 1997.
6. Spelluci, P., *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser, Basel 1993.
7. Luenberger, D.G., *Optimization by Vector Space Methods.* Wiley, 1969.
8. Luenberger, D.G., *Linear and Nonlinear Programming.* Addison Wesley, London 1984.
9. Großmann, C. und Terno, J, *Numerik der Optimierung.* Teubner-Verlag, Stuttgart 1993.
10. Moré, J.J. and Wright, S.J., *Optimization Software Guide.* SIAM, Philadelphia 1993.
11. Ulbrich M. und Ulbrich, S., *Nichtlineare Optimierung.* Birkhäuser-Verlag 2012.

1.2 Übersicht

In diesem Kurs behandeln wir verschiedene Klassen von Optimierungsaufgaben, die sich durch die Art der gegebenen Nebenbedingungen sowie weitere Eigenschaften wie Konvexität oder Nichtkonvexität der gegebenen Funktionen unterscheiden. Wir beginnen mit einfachsten Extremwertaufgaben ohne Nebenbedingungen und schließen mit der Optimierung nichtlinearer Funktionen bei nichtlinearen Gleichungen und Ungleichungen als Nebenbedingung ab.

Alle im Kurs behandelten Optimierungsaufgaben können in der folgenden allgemeinen Gestalt dargestellt werden:

$$\text{(P)} \qquad \boxed{\min_{x \in \mathscr{F}} f(x)}$$

Hierin ist $f : D \to \mathbb{R}$ eine gegebene Funktion, die **Zielfunktion** mit einer gegebenen offenen Menge $D \subset \mathbb{R}^n$ als Definitionsbereich. Die Menge $\mathscr{F} \subset D$ ist der sogenannte **zulässige Bereich.** Ihre Elemente heißen **zulässige Punkte**. Im Fall $\mathscr{F} = D$ heißt (P) **unrestringierte** oder **freie Optimierungsaufgabe**. Das mutet für eine echte Teilmenge $D \subset \mathbb{R}^n$ etwas eigenartig an, lässt sich aber leicht einsehen, denn Lösungen des Problems können nicht am Rand von D liegen.

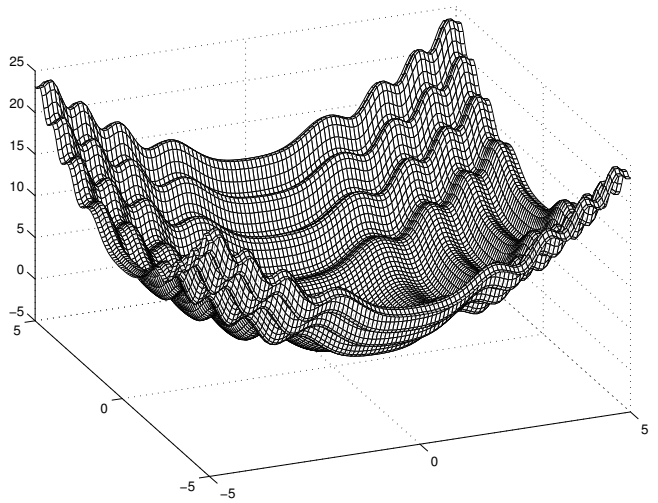
Einige einfache Beispiele unrestringierter Aufgaben zum warm werden:

Beispiel 1.2.1

- $f(x) = x^2, f : \mathbb{R} \to \mathbb{R}$ *hat genau ein Min bei* $\tilde{x} = 0$*;* $\mathscr{F} = D = \mathbb{R}$.
- $f(x) = x, f : \mathbb{R} \to \mathbb{R}$ *ist nicht nach unten beschränkt; die Minimumaufgabe ist unlösbar;* $\mathscr{F} = D = \mathbb{R}$.

Selbst einfache Aufgaben sind also unlösbar, wenn sie falsch gestellt sind.

Beispiel 1.2.2
$f(x_1,x_2) = \frac{1}{2}(x_1^2+x_2^2) - \cos(x_1^2) - \cos(x_2^2), f : \mathbb{R}^2 \to \mathbb{R}$ *hat ein (strenges)* **globales** *Minimum und mehrere (strenge)* **lokale** *Minima und Maxima.*



Bei solchen Aufgaben wird ein numerisches Verfahren in der Regel gegen irgendeines der vielen lokalen Minima oder Maxima konvergieren.

Ist $\mathscr{F}$ durch Nebenbedingungen gegeben, so heißt (P) **Optimierungsproblem mit Nebenbed.** oder **restringiertes Optimierungsproblem**. In der Regel ist $\mathscr{F}$ durch Gleichungen und Ungleichungen definiert.

Wir unterscheiden im Wesentlichen folgende Typen von restringierten Aufgaben:

- **Aufgaben mit Gleichungsrestriktionen**

  $$\min f(x) \text{ bei } h(x) = 0,$$

  mit $f : \mathbb{R}^n \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}^m$.

  Probleme dieses Typs haben Sie bereits im Grundkurs Analysis II behandelt. Hauptresultat war dabei die *Regel der Lagrange-Multiplikatoren.* Eine wichtige Unterklasse bilden Aufgaben mit **linearen Nebenbedingungen** des Typs

  $$\min f(x) \text{ bei } Ax = b,$$

  mit einer (m,n)-Matrix A und $b \in \mathbb{R}^m$.

- **Aufgaben mit Gleichungs- und Ungleichungsrestriktionen** In gewissem Sinne fängt die "richtige" Optimierung erst mit der Vorgabe von Ungleichungsrestriktionen an. Die allgemeinste Klasse von Aufgaben, die wir dabei behandeln, hat die Form

$$\min f(x) \text{ bei } h(x) = 0,\ g(x) \leq 0.$$

  Dabei ist $g : \mathbb{R}^n \to \mathbb{R}^p$ eine weitere, i.a. nichtlineare Funktion. Die Ungleichungsrelation $\leq$ ist dabei komponentenweise zu verstehen.

  Besonders schön gestaltet sich die Theorie im *Fall linearer Restriktionen* – hier sind h und g affin-lineare Abbildungen – und quadratischer Funktion f der Form

$$f(x) = a^\top x + x^\top C x$$

  mit positiv semidefiniter Matrix C. Dann spricht man von **quadratischer Optimierung**. Quadratische Optimierungsaufgaben sind deshalb von besonderem Interesse, weil sie erstens oft auftreten und zweitens die Grundlage zur Lösung allgemeiner nichtlinearer Optimierungsaufgaben sind.

Beispiel 1.2.3

$$\begin{aligned} &\min_{(x\in\mathbb{R})} x^3 \\ &\textit{bei} \quad x \geq 1 \end{aligned}$$

Es handelt sich um eine Aufgabe mit einer linearen Ungleichungsrestriktion, $D = \mathbb{R}$, $\mathscr{F} = [1,\infty)$, *die Lösung ist* $\tilde{x} = 1$.

Wir sehen bereits an diesem einfachen Beispiel, dass die von Extremwertaufgaben bekannte notwendige Optimalitätsbedingung $f'(x) = 0$ hier nicht greift.

Für alle Klassen von Aufgaben werden wir jeweils folgende Fragestellungen untersuchen:

- Existenz und Eindeutigkeit von Lösungen
- Notwendige Optimalitätsbedingungen
- Hinreichende Optimalitätsbedingungen
- Numerische Verfahren zur Lösung von Optimierungsaufgaben.

1.3 Einige Grundbegriffe

Bezeichnungen: Für $x \in \mathbb{R}^n$ heißt

$\|x\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$	euklidische Norm
$B(x,r) = \{y \in \mathbb{R}^n \mid \|y-x\| < r\}$	offene Kugel
$\overline{B}(x,r) = cl\, B(x,r)$	abgeschlossene Kugel
x_i	i-te Komponente von x
x^k	k-tes Glied einer Folge $\left(x^k\right)_{k=1}^\infty$.

Definition 1.3.1 *Ein Punkt $\tilde{x} \in \mathscr{F}$ heißt*

- **lokales Minimum** *von f in $\mathscr{F}$ oder* **lokale Lösung** *von (P), wenn $\exists r > 0$, so dass*

$$f(x) \geq f(\tilde{x}) \qquad \forall x \in \mathscr{F} \cap B(\tilde{x}, r)$$

- *analog* **strenges lokales Min.**, *wenn entsprechend*

$$f(x) > f(\tilde{x}) \qquad \forall x \in \mathscr{F} \cap B(\tilde{x}, r), \quad x \neq \tilde{x}$$

- *analog* **globales Minimum** *bzw.* **globale Lösung**, *wenn*

$$f(x) \geq f(\tilde{x}) \qquad \forall x \in \mathscr{F}$$

- **strenges globales Min.** *bzw.* **strenge globale** *Lösung, wenn*

$$f(x) > f(\tilde{x}) \qquad \forall x \in \mathscr{F}, \quad x \neq \tilde{x}$$

  gilt.

Bei nichtlinearen Optimierungsaufgaben können viele lokale oder globale Minima auftreten, wie etwa bei $f(x) = \sin x, f(x) = x \sin\left(\frac{1}{x}\right)$.

Bemerkung: Wir entwickeln unsere Theorie für den Fall der *Minimierung* von f. Den Fall der Suche nach Maxima $\tilde{x}$,

$$f(x) \leq f(\tilde{x}) \qquad \forall x \in \mathscr{F}$$

führen wir wegen der Äquivalenz zu $-f(x) \geq -f(\tilde{x})$ auf die Minimierung von $\tilde{f} := -f$ zurück.

Das folgende Problem ist eine der am häufigsten gelösten Optimierungsaufgaben:

Beispiel 1.3.1 (Lineare Regression) *Gesucht ist eine lineare Funktion*

$$\eta(\xi) = x_1 \xi + x_2$$

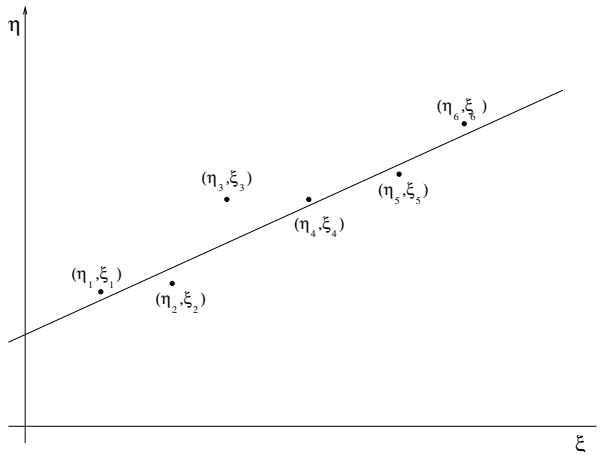
mit unbekannten Koeffizienten x_1, x_2, die am besten zu gegebenen Wertepaaren $(\xi_i, \eta_i), i = 1, \ldots, m$ (z.B. Messwerten) passt. Wir setzen

$$\eta(\xi) = g(x_1, x_2, \xi) = x_1 \xi + x_2$$

und wollen x_1, x_2 so wählen, dass die Zielfunktion

$$\begin{aligned} f(x) = f(x_1, x_2) &= \sum_{i=1}^{m} (\eta_i - g(x_1, x_2, \xi_i))^2 \\ &= \sum_{i=1}^{m} (\eta_i - x_1 \xi_i - x_2)^2 \end{aligned}$$

minimiert wird. Dabei ist f ein Polynom zweiten Grades in x_1, x_2, also eine quadratische Zielfunktion.



1.4 Existenz von Lösungen

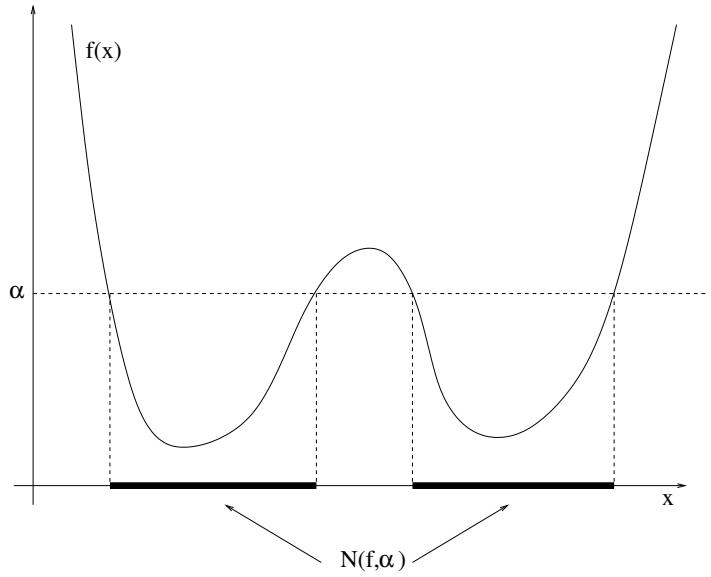
Grundlage für die meisten Existenzbeweise ist der bekannte

Satz 1.4.1 (Weierstraß) *Ist $f : \mathbb{R}^n \supset D \to \mathbb{R}$ stetig und $K \subset D$ nichtleer und kompakt, dann nimmt f auf K sein Infimum (bzw. sein Supremum) an, d. h., es existiert ein globales Minimum (bzw. Maximum) von f auf K.*

Definition 1.4.1 *Es sei $f : D \to \mathbb{R}, D \subset \mathbb{R}^n, \alpha \in \mathbb{R}$. Die Mengen*

$$N(f,\alpha) = \{x \in D \quad | \quad f(x) \leq \alpha\}$$

heißen **Niveaumengen** *von f.*



Satz 1.4.2 *$D \subset \mathbb{R}^n, f : D \to \mathbb{R}$ stetig und $\mathscr{F} \subset D$ abgeschlossen. Für mindestens ein $w \in \mathscr{F}$ sei die Niveaumenge*

$$N(f, f(w)) = \{x \in D \,|\, f(x) \leq f(w)\}$$

kompakt. Dann existiert (mindestens) ein globales Minimum von f auf $\mathscr{F}$.

Beweis: Es sei $\alpha = \inf_{x\in\mathscr{F}} f(x)$. Offenbar gilt $\alpha \le f(w)$. Die Menge $\mathscr{F} \cap N(f, f(w))$ ist kompakt, und nur in dieser Menge können Elemente von $\mathscr{F}$ liegen, deren Funktionswerte kleiner oder gleich $f(w)$ sind. Somit

$$\alpha = \inf_{x\in\mathscr{F}\cap N(f,f(w))} f(x) = f(\tilde{x}),$$

wobei $\tilde{x} \in \mathscr{F}$ wegen des Satzes von Weierstraß existiert. □

Typische Anwendungen dieses Prinzips sind die folgenden zwei Aussagen:

Folgerung 1.4.1 *Es seien $D, \mathscr{F}$ wie in Satz 1.4.1, $f : \mathbb{R}^n \to \mathbb{R}$ stetig. Zusätzlich habe f die Eigenschaft*

$$\lim_{\|x\|\to\infty} f(x) = \infty.$$

Dann besitzt die Aufgabe

$$\min_{x\in\mathscr{F}} f(x)$$

mindestens eine globale Lösung.

Beweis: Wegen $\lim_{\|x\|\to\infty} f(x) = \infty$ sind alle Niveaumengen $N(f,\alpha)$ kompakt (Übungsaufgabe). Der Rest ist Folgerung aus dem letzten Satz. □

Definition 1.4.2 *Eine (n,n)-Matrix H heißt* **positiv semidefinit**, *wenn $x^\top Hx \ge 0$ für alle $x \in \mathbb{R}^n$ gilt, sowie* **positiv definit**, *wenn*

$$x^\top Hx > 0 \quad \forall x \in \mathbb{R}^n, x \ne 0$$

Man zeigt mit einem Kompaktheitsschluss, dass positive Definitheit äquivalent ist zur Existenz eines $\alpha > 0$, so dass

$$x^\top Hx \ge \alpha \|x\|^2 \quad \forall x \in \mathbb{R}^n$$

(Übungsaufgabe). Offenbar gilt dann $x^\top Hx \to \infty, \|x\| \to \infty$. Damit liegen wir im Bereich von Folgerung 1.4.1

Beispiel 1.4.1 (Unrestringierte quadratische Optimierungsaufgabe)

Wir betrachten

$$(QU) \qquad \boxed{\min_{x\in\mathbb{R}^n} f(x) = \tfrac{1}{2}x^\top Hx + b^\top x}$$

mit gegebenem $b \in \mathbb{R}^n$ und positiv definiter (n,n)-Matrix H. Sie zeigen leicht, dass $\lim_{\|x\|\to\infty} f(x) = \infty$ gilt (Übungsaufgabe). Wegen Folgerung 1.4.1 hat (QU) damit mindestens eine globale Lösung.

Beispiel 1.4.2 (Lineare Regression aus Bsp 1.3.1)

Ausmultiplizieren der Zielfunktion ergibt

$$\begin{aligned} f(x) &= \sum_{i=1}^{m} (\eta_i - (x_1\xi_i + x_2))^2 \\ &= \underbrace{\sum_{i=1}^{m} \eta_i^2}_{c} \quad \underbrace{-2\sum_{i=1}^{m} \eta_i(\xi_i x_1 + x_2)}_{b^\top x} \quad + \underbrace{\sum_{i=1}^{m} (x_1\xi_i + x_2)^2}_{\frac{1}{2}x^\top Hx} \\ &= \tfrac{1}{2}x^\top Hx + b^\top x + c \end{aligned}$$

$$\text{mit} \quad H = 2\begin{pmatrix} \sum_1^m \xi_i^2 & \sum_1^m \xi_i \\ \sum_1^m \xi_i & m \end{pmatrix} \quad b = -2\begin{pmatrix} \sum_1^m \xi_i\eta_i \\ \sum_1^m \eta_i \end{pmatrix}.$$

Sind mindestens zwei der ξ_i verschieden, so ist H positiv definit (Übungsaufgabe).

Damit ist die Aufgabe der linearen Regression in diesem Fall lösbar. Sind alle ξ_i gleich, so ist sie auch nicht sinnvoll gestellt!

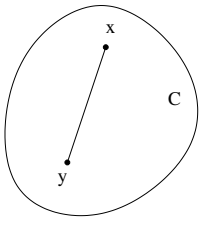
1.5 Konvexe Optimierungsaufgaben

Unter allen Optimierungsaufgaben haben konvexe die schönsten Eigenschaften! Es gibt dazu auch eine gut ausgebaute **konvexe Analysis** (z. B. siehe Webster, R., Convexity. Oxford University Press 1994, oder Rockafellar, R.T., Convex Analysis. Princeton University Press 1970).

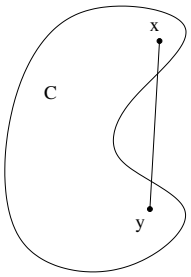
Definition 1.5.1 *Eine Menge $C \subset \mathbb{R}^n$ heißt konvex, falls für je 2 beliebige $x, y \in C$ auch die Strecke*

$$[x,y] = \{z = (1-t)x + ty \,|\, 0 \le t \le 1\}$$

in C enthalten ist: $x, y \in C \Rightarrow [x,y] \subset C$.



konvexe Menge



nicht konvexe Menge

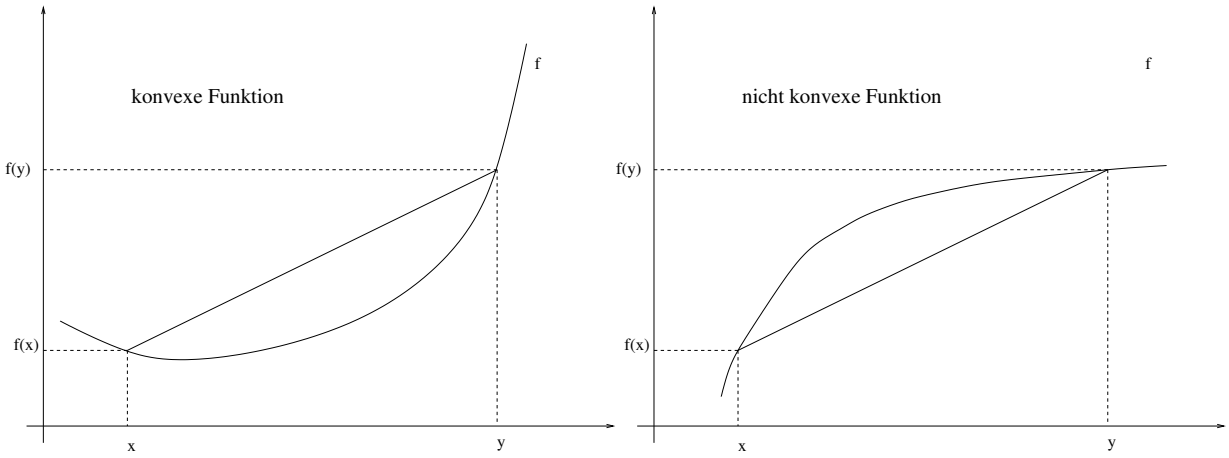
Definition 1.5.2 *Sei $C \subset \mathbb{R}^n$ konvex und nichtleer, $C \subset D$. Eine Funktion $f : D \to \mathbb{R}$ heißt* **konvex auf** *C, wenn*

$$f((1-t)x + ty) \le (1-t)f(x) + tf(y) \quad \forall x, y \in C,\ \forall t \in [0,1].$$

Gilt die verschärfte Beziehung

$$f((1-t)x+ty) < (1-t)f(x)+tf(y) \quad \forall x,y \in C, x \neq y,\ \forall t \in]0,1[\,,$$

so heißt f **strikt** *oder* **streng** *konvex auf C.*



Beispiel 1.5.1 *$f(x) = x$ ist konvex, $f(x) = x^2$ streng konvex.*

Definition 1.5.3 *Nun betrachten wir $f : D \to \mathbb{R}, D \subset \mathbb{R}^n$ offen, nichtleer; $\mathscr{F} \subset D$ sei konvex. Ist f konvex auf $\mathscr{F}$, so heißt das Problem*

$$\min_{x \in \mathscr{F}} f(x) \tag{P}$$

konvexe Optimierungsaufgabe.

Schon der nächste Satz zeigt, welche schöne Eigenschaften konvexe Probleme haben.

Satz 1.5.1 *Die obige Aufgabe (P) sei eine konvexe Optimierungsaufgabe. Dann ist jedes lokale Minimum von (P) auch ein globales. Die Menge aller Lösungen von (P) ist konvex.*

Beweis:

(i) Es sei $\tilde{x}$ lokale Lösung, d. h., mit einem $r > 0$ gilt

$$f(x) \geq f(\tilde{x}) \quad \forall x \in \mathscr{F} \cap B(\tilde{x}, r)\,. \tag{$*$}$$

Zu zeigen ist

$$f(y) \geq f(\tilde{x}) \quad \forall y \in \mathscr{F}\,.$$

Wir wählen ein $y \in \mathscr{F}$ und betrachten $\tilde{x}+t(y-\tilde{x})$ für kleine $t > 0$.

Es gilt $\tilde{x}+t(y-\tilde{x}) \in \mathscr{F} \quad \forall t \in [0,1]$: Denn

$$\tilde{x}+t(y-\tilde{x}) = (1-t)\tilde{x}+ty \in \mathscr{F},$$

da $\mathscr{F}$ konvex ist. Es gilt auch $\tilde{x}+t(y-\tilde{x}) \in B(\tilde{x},r)$ für alle hinreichend kleinen t, d. h. $t \in [0,t_0], t_0 > 0$. Deshalb wegen $(*)$

$$f(\tilde{x}) \underset{\substack{\uparrow \\ (*)}}{\leq} f((1-t)\tilde{x}+ty) \leq (1-t)f(\tilde{x})+tf(y).$$

Durch Umstellen ergibt sich $f(\tilde{x}) \leq f(y)$.

(ii) Nun seien x und $\tilde{x}$ Lösungen, d.h. $f(\tilde{x}) = f(x) = \tilde{\alpha} = \min f$. Dann

$$f((1-t)\tilde{x}+tx) \leq (1-t)f(\tilde{x})+tf(x) = \tilde{\alpha}$$

$\Rightarrow$ auch $(1-t)\tilde{x}+tx$ ist Lösung. □

Satz 1.5.2 *Sei $D \subset R^n, \mathscr{F} \subset D$ konvex, $\mathscr{F} \neq \emptyset, f : D \to \mathbb{R}$ streng konvex. Hat (P) eine Lösung $\tilde{x}$, dann ist $\tilde{x}$ eindeutig bestimmt und ein strenges Minimum von f in $\mathscr{F}$.*

Beweis: Es seien x, y zwei Minima von f, also nach dem letzten Satz globale Minima. Damit gilt

$$f(x) = f(y) = \alpha = \min_{x \in \mathscr{F}} f(x).$$

Angenommen, es gilt $x \neq y$. Dann liefert $z = \frac{1}{2}(x+y)$ einen kleineren Wert als α, denn

$$f(z) = f\left(\frac{1}{2}x+\frac{1}{2}y\right) \underset{\substack{\uparrow \\ \text{strenge Konvexität}}}{<} \frac{1}{2}f(x)+\frac{1}{2}f(y) = \frac{1}{2}\alpha+\frac{1}{2}\alpha = \alpha.$$

Außerdem gilt $z \in \mathscr{F}$ und insgesamt widerspricht das der Optimalität von x, y. Damit $x = y$, strenges Minimum. □

Beispiel 1.5.2 $f(x) = \frac{1}{2}x^\top Hx + b^\top x$

Ist H positiv definit, so ist f streng konvex (Übungsaufg.).

Folgerung: Sind zwei der Werte ξ_i verschieden, so ist das Zielfunktional bei der Aufgabe der linearen Regression streng konvex und daher die Lösung eindeutig bestimmt.

1.6 Weitere wichtige Beispiele von Optimierungsaufgaben

Beispiel 1.6.1 (Nichtlineare Regression) *Bei der linearen Regression war eine affin-lineare Funktion $\eta(\xi) = x_1\xi + x_2$ zu bestimmen. Allgemeiner kann η eine nichtlineare Funktion von ξ sein, gegeben durch einen nichtlinearen Ansatz*

$$\eta(\xi) = g(x_1, x_2, \xi)$$

oder allgemeiner

$$\eta(\xi) = g(x_1, \ldots, x_n, \xi)$$

mit einem unbekannten Vektor $x \in \mathbb{R}^n$, z. B.

$$g = x_1 e^{\xi x_2} + x_3.$$

$\Rightarrow$ *Minimierung von*

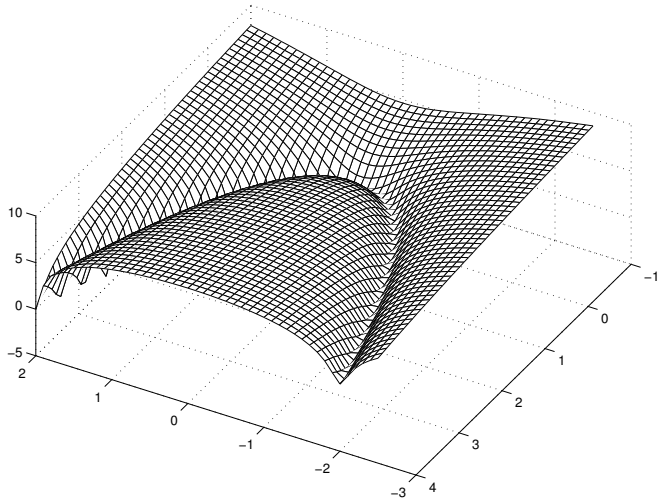
$$\boxed{f(x) = \sum_{i=1}^{m} (\eta_i - g(x, \xi_i))^2}$$

$$\textit{Typ:} \qquad f(x) = \sum_{i=1}^{m} (f_i(x))^2 \qquad f_i = \eta_i - g(\cdot, \xi_i)\,.$$

Nun betrachten wir noch einige berühmte pathologische Testfunktionen, an denen gern Algorithmen getestet werden.

Beispiel 1.6.2 (Rosenbrock-Funktion) *(„Banana shaped valley“)*

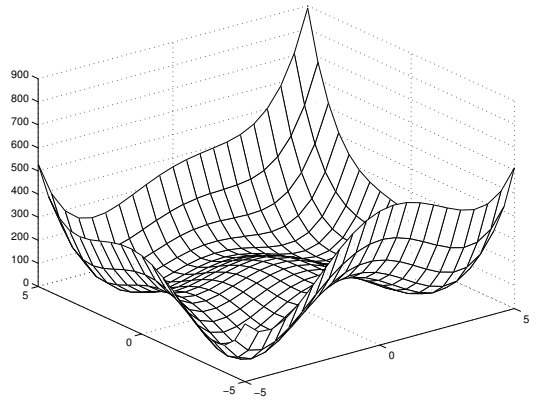
$$\boxed{f(x_1,x_2) = \underbrace{100(x_2 - x_1^2)^2}_{\substack{\textit{definiert}\\ \textit{das Tal}\\ \textit{(Parabel)}}} + \underbrace{(1-x_1)^2}_{\textit{kippt leicht an}}}$$



Beispiel 1.6.3 (Himmelblau)

$$\boxed{f(x_1,x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2}$$

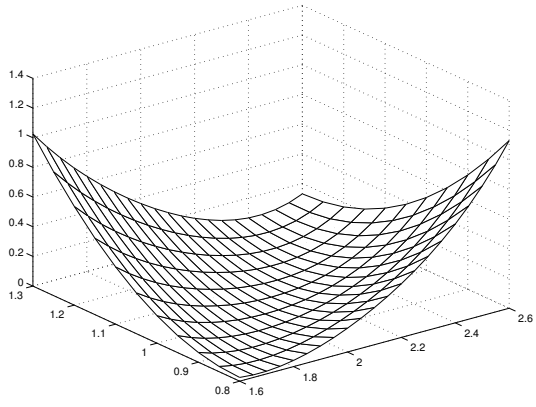
4 lokale Minimalstellen, die zugleich globale Minimalstellen mit Funktionswert 0 *sind; 4 Sattelpunkte und ein lokales Maximum bei* $(-0.270845, -0.923039)^\top$.



Beispiel 1.6.4 (Bazaraa-Shetty)

$$f(x_1,x_2) = (x_1-2)^4 + (x_1-2x_2)^2$$

Globales Min. bei $(2,1)$. *Die Hesse-Matrix ist an dieser Stelle singulär, was bei manchen Algorithmen zu Problemen führen kann.*



Beispiel 1.6.5

$$f(x_1,\cdots,x_5) = 2x_1^2 + 2x_2^2 + x_3^2 + x_4^2 + \tfrac{1}{2}x_5^2 - 4(x_1+x_2) - 2(x_3+x_4) - x_5 + 6.5$$

Globales Min. bei $\tilde{x} = (1,1,1,1,1)^\top, f(\tilde{x}) = 0$ *(Übungsaufgabe).*

Beispiel 1.6.6 (Dixon)

$$f(x_1,\cdots,x_{10}) = (1-x_1)^2 + (1-x_{10})^2 + \sum_{i=1}^{9}(x_i^2 - x_{i+1})^2$$

Globales Minimum bei $\tilde{x} = (1,\cdots,1)^\top$.

1.7 Numerische Algorithmen für Optimierungsaufgaben

Wir werden die in der Vorlesung zu untersuchenden Optimierungsverfahren numerisch durch iterative Verfahren lösen, die teilweise nach endlich vielen Schritten eine Lösung ermitteln oder deren Lösungsfolge einem Grenzwert zustrebt:

$$\lim_{k\to\infty} x^k = \tilde{x}.$$

Dabei werden wir Optimierungsaufgaben verschiedener Struktur untersuchen (z. B. linear-quadratische Aufgaben, nichtlineare Funktionale mit linearen Restriktionen, allgemeine nichtlineare Probleme, nicht jedoch lineare oder diskrete Optimierungsaufgaben.)

Es gibt zahlreiche kommerzielle Codes zur Lösung von Optimierungsaufgaben.

1.8 Optimierungs-Software

1.8.1 Programmbibliotheken

Empfehlenswert und bei uns verfügbar:

- NAG-Library (Numerical Algorithms Group) Fortran Codes
- minpack (ist public domain software)

1.8.2 Interaktive Programmsysteme

- MATLAB (MATrix LABoratory) kommerziell
- Scilab (SCIentific, LABoratory) kostenlos von INRIA, Paris www.inria.fr

Übersichten numerischer Codes:

- Entscheidungshilfen im Internet: Hans D. Mittelmann, http://plato.la.asu.edu/guide.html
- Software-Guide: Moré and Wright, [9]
- Sammlung von Codes im Internet: NEOS (NEOS Server for Optimization)

2 Ableitungsfreie Verfahren

Oft ist die Berechnung der Ableitung von f so aufwendig oder – bei nicht differenzierbarem f – unmöglich, dass man Verfahren entwickelt hat, die ohne Ableitungen auskommen.

Stellen Sie sich etwa folgende, in der Praxis sehr häufig auftretende Situation vor: Der Vektor x wird in ein System mehrerer partieller Differentialgleichungen eingesetzt, etwa die Navier-Stokes-Gleichungen zur Berechnung einer Strömung und die Wärmeleitgleichung zur Berechnung der Temperatur in der Strömung. Das alles dreidimensional bei turbulenter Strömung. Das entstandene Geschwindigkeitsfeld wird in ein Integral eingesetzt, das den Funktionswert $f(x)$ liefert.

Dann ist es in der Regel nur mit enormem Aufwand möglich, die Ableitung $f'(x)$ einigermaßen genau zu berechnen. Man muss ohne sie auskommen.

Wir behandeln hier kurz zwei ableitungsfreie Verfahren, um die unrestringierte Aufgabe

$$\min_{x\in\mathbb{R}^n} f(x) \tag{PU}$$

numerisch zu lösen.

2.1 Simplexverfahren von Nelder und Mead

2.1.1 Grundkonstruktionen

Bemerkung: Das Verfahren hat nichts mit der Simplexmethode der linearen Optimierung zu tun! Der Name "Simplexmethode" leitet sich von der Konstruktion von Simplexen ab:

Definition 2.1.1 *$x^0,\dots,x^n \in \mathbb{R}^n$ seien affin unabhängig, d. h. $x^i - x^0$, $i = 1,\dots,n$ sind linear unabhängig. Die konvexe Hülle der Punkte $x^0,\dots,x^n$*

$$S = \left\{ \sum_{i=0}^{n} \lambda_i x^i \mid \lambda_i \geq 0, i = 0,\dots,n, \sum_{i=0}^{n} \lambda_i = 1 \right\}$$

heißt (n-dimensionales) **Simplex** *mit den Ecken $x^0,\dots,x^n$.*

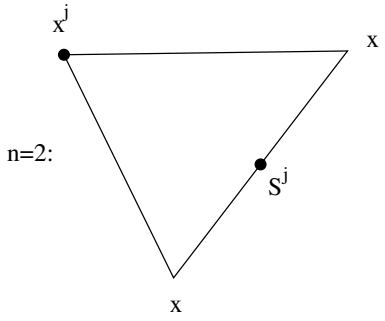
- Beim Start des Verfahrens wird ein Simplex vorgegeben.
- Man ermittelt eine Ecke mit dem größten Funktionswert,

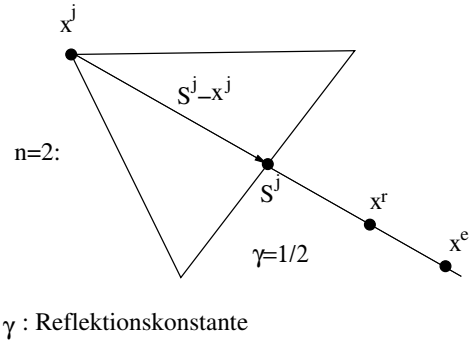
$$f(x^m) = \max\left\{f(x^0),\dots,f(x^n)\right\}$$

- Danach wird ein neuer Punkt ermittelt, der einen i.a. kleineren Funktionswert ergibt und x^m ersetzt.

Dazu werden folgende Konstruktionen benutzt:

Def. $\boxed{s^j = \frac{1}{n} \sum_{\substack{i=0 \\ i \neq j}}^{n} x^i}$ Schwerpunkt der (anderen) Ecken bzgl. x^j





γ : Reflektionskonstante

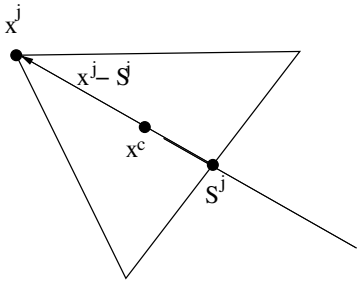
Konstruktionsprinzipien:

- **Reflektion** von x^j an s^j

$$x^r = s^j + \gamma(s^j - x^j),\ 0 < \gamma \leq 1$$

- Dieses eben konstruierte x^r kann weiter nach außen bewegt werden:
  Expansion von x^r in Richtung $s^j - x^j$ (d. h. in Richtung $x^r - s^j$)

$$x^e = s^j + \beta(x^r - s^j),\ \beta > 1 \quad \text{Expansionskonstante.}$$



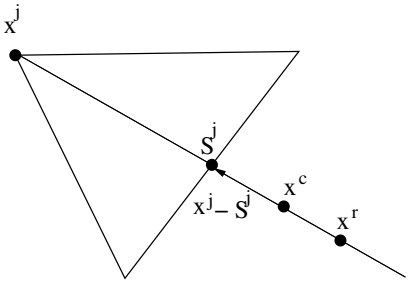
- **Kontraktion** (3 Typen)

  (i) **Partielle Kontraktion innen**

$$x^c = s^j + \alpha(x^j - s^j) \qquad 0 < \alpha < 1 \quad \text{Kontraktionskonstante}$$

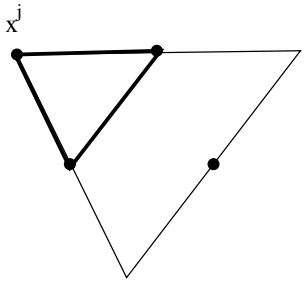
  (ii) **Partielle Kontraktion außen**

$$x^c = s^j + \alpha(x^r - s^j)$$



(iii) **Totale Kontraktion**
Ersetze alle x^i außer x^j durch

$$\hat{x}^i = x^i + \frac{1}{2}(x^j - x^i) = \frac{1}{2}(x^i + x^j)$$



2.1.2 Ablauf des Verfahrens

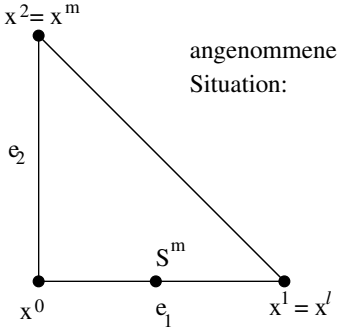
Die einfachste Variante läuft so ab:

Vorab werden gewählt	$\alpha \in (0,1)$	Kontraktionskonstante
	$\beta > 1$	Expansionskonstante
	$\gamma \in (0,1]$	Reflektionskonstante

Folgende Schritte laufen ab:

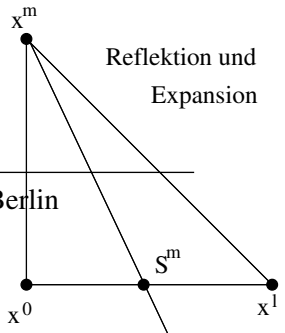
1. **Wahl eines Startpunkts** $x^0 \in R^n$, Festlegung der anderen n Ecken des **Startsimplexes** durch
$$x^j = x^0 + e^j,\ j = 1,\ldots,n,$$
wobei e^j den j-ten Standardeinheitsvektor bezeichnet.

2. Bestimme (die) Ecken mit **maximalem und minimalem Funktionswert:** x^m, x^l mit
$$\begin{aligned} f(x^m) &= \max\left\{f(x^0),\ldots,f(x^n)\right\} \\ f(x^l) &= \min\left\{f(x^0),\ldots,f(x^n)\right\} \end{aligned}$$
und bestimme den **Schwerpunkt der Ecken** bezügl. x^m
$s^m = \frac{1}{n}\sum_{\substack{i=0 \\ i\neq m}}^{n} x^i$



3. **Reflektion** von x^m am Schwerpunkt s^m
$$x^r = s^m + \gamma(s^m - x^m)$$

4. **Aufbau des neuen Simplexes**
Dazu eine Fallunterscheidung



(i)

$$\boxed{f(x^r) < f(x^l)}$$

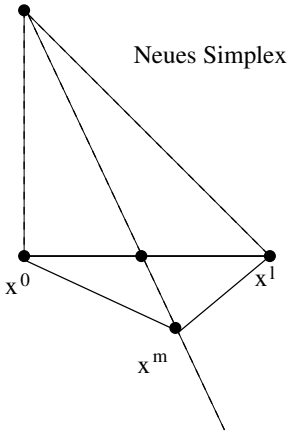
Dann war die Richtung gut, und man probiert noch etwas mehr: Expansion von x^r

$$x^e = s^m + \beta(x^r - s^m)$$

Man ersetzt x^m durch den besseren der beiden Punkte:

$$\tilde{x}^m = \begin{cases} x^e, & f(x^e) < f(x^r) \\ x^r, & f(x^r) \leq f(x^e) \end{cases}$$

$$\underline{x^m := \tilde{x}^m}$$



(ii)

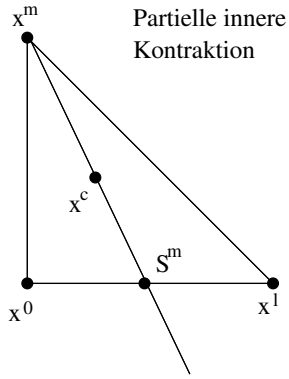
$$\boxed{f(x^l) \leq f(x^r) \leq \max\left\{f(x^j),\, j \neq m\right\}}$$

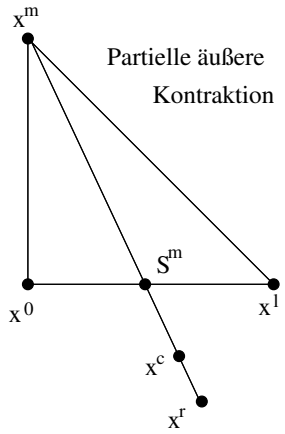
Nichts gewonnen, nichts verloren – ersetze x^m durch x^r

$$\underline{x^m := x^r}$$

(iii)

$$\boxed{f(x^r) > \max\left\{f(x^j),\, j \neq m\right\}}$$





- Wenn $f(x^r) \geq f(x^m)$: **Partielle innere Kontraktion**

$$x^c = s^m + \alpha(x^m - s^m)$$

- Wenn $f(x^r) < f(x^m)$: **Partielle äußere Kontraktion**

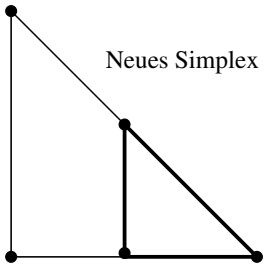
$$x^c = s^m + \alpha(x^r - s^m)$$

- Wenn $f(x^c) < f(x^m)$, dann ersetze x^m durch x^c

$$x^m := x^c$$

- Wenn $f(x^c) \geq f(x^m)$, dann führe eine **totale Kontraktion** bezüglich x^l aus:

$$x^i := \frac{1}{2}\left(x^i + x^l\right),\ i \neq l$$



5. Gehe mit dem neu ermittelten Simplex (Ecken $\{x^0,\ldots,x^n\}$) **zu Schritt 2.**
   Das Verfahren erzeugt Eckenfolgen $\{x^{(k,0)},\ldots,x^{(k,n)}\}_{k=1}^{\infty}$ und stellt sicher, dass

$$f\left(x^{(k+1,l)}\right) \leq f\left(x^{(k,l)}\right)$$

   gilt. In gewissem Sinne kann man $x^{(k,l)}$ als den aktuellen Iterationspunkt bezeichnen. Allgemeine Konvergenzsätze gibt es nicht.

   Empirische Untersuchungen zeigen $0.4 \leq \alpha \leq 0.6,\ 2 \leq \beta \leq 3, \gamma = 1$ sind zu empfehlen.

Verfügbare Codes:	EO4CCF	(NAG)
	fmins	(MATLAB 5)
	fminsearch	(MATLAB 6)

Das Verhalten des Verfahrens wird am Beispiel der Rosenbrock-Funktion deutlich:

Startpunkt: $(-1.9,2)^\top$	EO4CCF:	stoppt nach 186 Funktionsauswertungen bei $(1.000011, 1.000023)^\top$
	fmins:	$(1.00002, 1.00003)^\top$ nach 210 Auswertungen

2.2 Mutations-Selektions-Verfahren

- Zufällige "Mutation" der aktuellen Iterierten
- Auswahl der "brauchbaren" Iterierten

Diese Verfahren gehören zur Klasse von Methoden der stochastischen Suche.

Verfahrensgrundprinzip:

1. Wähle Startpunkt $x^0 \in \mathbb{R}^n$
   $k := 0$

2. Berechne neuen Punkt $v^{(k)}$ durch zufällige Änderung von x^k (Zufallszahlen), z. B.

$$\boxed{v_i^{(k)} = x_i^k + \delta_k \left(r_i^{(k)} - 0.5 \right)} \quad i = 1, \cdots, n$$

$r_i^{(k)}$: Zufallszahlen aus $[0,1]$
δ_k : Schrittweiten

3.

$$x^{k+1} = \begin{cases} v^{(k)} & \text{falls} \quad f\left(v^{(k)}\right) < f\left(x^k\right) \\ x^k & \text{sonst} \end{cases}$$

Numerisches Resultat: Für Beispiel 1.6.2 (Rosenbrock) werden für eine gewisse Implementierung 2776 S

2.3 Anwendung: Nichtlineare Regression

Siehe Beispiel 1.6.1; Gegeben sind 10 Messwertpaare:

ξ_i	1	2	3	4	5	6	7	8	9	10
η_i	1	1.1	1.2	1.35	1.55	1.75	2.5	3	3.7	4.5

Ansatz: $\eta(\xi) = g(x,\xi) = x_1 e^{\xi x_2}$

$$\text{(P)} \quad \boxed{\min f(x) = f(x_1, x_2) = \sum_{i=1}^{10} \left(\eta_i - x_1 e^{\xi_i x_2} \right)^2}$$

Anwendung der MATLAB-Implementierung fmins des Nelder-Mead-Verfahrens ergibt

$$\tilde{x} = \begin{pmatrix} 0.632067 \\ 0.195061 \end{pmatrix} \quad f(\tilde{x}) = 0.19041.$$

Weitere Beispiele werden in den Übungen diskutiert.

3 Probleme ohne Restriktionen – Theorie

3.1 Optimalitätsbedingungen

3.1.1 Bedingungen erster Ordnung

Im gesamten Kapitel 3 wird vorausgesetzt:

$D \subset \mathbb{R}^n$ offen, nichtleer

$f : D \to \mathbb{R}$ mit gewissen Differenzierbarkeitsannahmen

Wir betrachten die unrestringierte Aufgabe

$$\text{(PU)} \qquad \boxed{\min_{x \in D} f(x)}$$

Satz 3.1.1 *(Fermat) f besitze in $\tilde{x} \in D$ ein lokales Minimum und sei an der Stelle $\tilde{x}$ differenzierbar. Dann gilt die* **notwendige Bedingung 1. Ordnung**

$$\nabla f(\tilde{x}) = 0. \tag{3.1}$$

Die Aussage ist Grundwissen aus der Analysis.

Bemerkung: Bei uns sind Vektoren stets Spaltenvektoren. Deshalb ist $f'(x)$ ein Zeilenvektor und $\nabla f(x)$ ein Spaltenvektor. Es gilt

$$\nabla f(x) = f'(x)^\top.$$

Beispiel 3.1.1

$$f(x) = \tfrac{1}{2} x^\top H x + b^\top x \qquad \begin{array}{l} H \in \mathbb{R}^{(n,n)}, \quad \underline{\textit{symmetrisch}} \\ b \in \mathbb{R}^n. \end{array}$$

Hier gilt

$$\nabla f(x) = Hx + b\,.$$

Eine Lösung der Aufgabe

$$\min_{x \in \mathbb{R}^n} f(x)$$

muss also die Gleichung $Hx = -b$ erfüllen. Ist H außerdem positiv definit, so hat das 2 Effekte. Erstens existiert eine Lösung (Bsp 1.4.1). Außerdem ist unser Gleichungssystem eindeutig lösbar. Damit ist

$$\tilde{x} = -H^{-1} b$$

die eindeutig bestimmte Lösung.

Anwendung: Lineare Regression (Fortsetzg. Bsp 1.4.2)

Wir hatten

$$H = 2\begin{pmatrix} \sum_1^m \xi_i^2 & \sum_1^m \xi_i \\ \sum_1^m \xi_i & m \end{pmatrix}, \quad b = -2\begin{pmatrix} \sum_1^m \xi_i \eta_i \\ \sum_1^m \eta_i \end{pmatrix}$$

erhalten. Ist H positiv definit, dann ergibt sich für die Lösung der Regressionsaufgabe das Gleichungssystem

$$\begin{array}{rcl} \left(\sum_1^m \xi_i^2\right) x_1 + \left(\sum_1^m \xi_i\right) x_2 = \sum_1^m \xi_i \eta_i \\ \sum_1^m \xi_i x_1 \quad + \quad m \quad x_2 = \sum_1^m \eta_i \,. \end{array}$$

Bestimmen Sie die Lösung!

Definition 3.1.1 *Ist f in $\tilde{x} \in D$ differenzierbar und gilt $\nabla f(\tilde{x}) = 0$, so heißt $\tilde{x}$* **stationärer Punkt** *von f.*

Bemerkung: Optimierungsalgorithmen berechnen in der Regel stationäre Punkte. Diese müssen keineswegs lokale oder globale Minima (Maxima) ergeben, denken Sie etwa an $f(x) = x^3$ bei $x = 0$.

Beispiel 3.1.2 *Rosenbrock-Funktion: Hat genau einen stationären Punkt bei $\tilde{x} = \binom{1}{1}$*

$$\nabla f(x) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}.$$

Ist die Zielfunktion f differenzierbar, so heißt (PU) **glatte** oder **differenzierbare** Optimierungsaufgabe. Bei vielen Anwendungen ist f nicht überall differenzierbar, ein typisches Beispiel ist

$$f(x) = \|x\| \quad \text{bei} \quad x = 0.$$

Mit **nichtglatter Optimierung** werden wir uns kaum befassen. Allerdings geben wir folgendes nützliches Resultat an.

Definition 3.1.2 *f heißt in $x \in D$ in Richtung $h \in \mathbb{R}^n$* **richtungsdifferenzierbar**, *wenn die* **Richtungsableitung**

$$f'(x,h) := \lim_{t \downarrow 0} \quad \frac{f(x+th) - f(x)}{t}$$

existiert. Gilt dies für alle Richtungen h, so heißt f **richtungsdifferenzierbar** *an der Stelle x.*

Laut dieser Definition hängt die Richtungsableitung nicht nur von der Richtung h sondern auch von deren Betrag ab. Deshalb wird in der Analysis noch $\|h\|$ gefordert. In der Optimierung ist das insbesondere aus numerischer Sicht nicht sinnvoll. Deshalb unterscheidet sich diese Definition von der aus der Analysis.

Satz 3.1.2 *Ist $\tilde{x}$ ein lokales Minimum von (PU) und ist f an der Stelle $\tilde{x} \in D$ richtungsdifferenzierbar, dann gilt die* **Variationsungleichung**

$$\boxed{f'(\tilde{x},h) \geq 0 \quad \forall h \in \mathbb{R}^n} \tag{3.2}$$

Beweis: D ist offen, damit $\exists r > 0$:

$$f(x) \geq f(\tilde{x}) \quad \forall x \in B(\tilde{x},r).$$

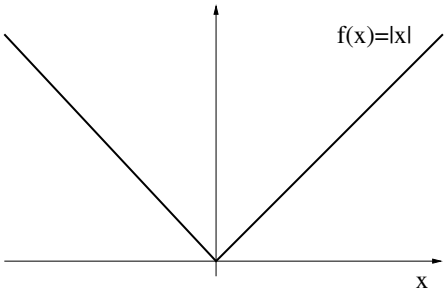
Sei $h \in \mathbb{R}^n$ beliebig, aber fest. Dann gilt $\tilde{x}+th \in B(\tilde{x},r)$ für betragsmäßig kleine t, somit

$$\begin{array}{rlll} & f(\tilde{x}+th)-f(\tilde{x}) & \geq 0 & \\ \Rightarrow & \dfrac{f(\tilde{x}+th)-f(\tilde{x})}{t} & \geq 0 \quad \Rightarrow & f'(\tilde{x},h) \geq 0 \end{array}$$

□

(3.2) ist intuitiv einleuchtend. In $\tilde{x}$ liegt ein lok. Minimum vor, also kann keine Richtung existieren, in der es abwärts geht!

Beispiel 3.1.3 $f(x) = |x|$ *hat lokales Min. bei* $\tilde{x} = 0$.



f ist bei $\tilde{x} = 0$ nicht differenzierbar, aber die Richtungsableitung existiert:

$$\begin{array}{rll} & \dfrac{f(th)-f(0)}{t} & = \dfrac{|th|}{t} = |h|, \quad t > 0 \\ \Rightarrow & f'(0,h) & = |h| \geq 0 \quad \forall h \subset \mathbb{R}. \end{array}$$

3.1.2 Notwendige Bedingungen zweiter Ordnung

Satz 3.1.3 *f sei in einer Umgebung von $\tilde{x} \in D$ zweimal stetig differenzierbar. Ist $\tilde{x}$ lokales Minimum von (PU), so muss neben der notwendigen Bedingung erster Ordnung auch*

$$\boxed{h^\top f''(\tilde{x})h \geq 0 \quad \forall h \in \mathbb{R}^n} \tag{3.3}$$

erfüllt sein, d. h., $f''(\tilde{x})$ muss **positiv semidefinit** *sein.*

Beweis: Bekannt aus der Analysis; Beweisskizze: Wir setzen für beliebiges aber festes h,

$$F(t) = f(\tilde{x}+th).$$

F hat lokales Minimum bei $t = 0$ und ist vom Typ C^2. Taylorentwicklung:

$$F(t) = F(0) + \underbrace{F'(0)t}_{=0} + \tfrac{1}{2}F''(\vartheta t)t^2,\ 0 < \vartheta < 1$$
$$\Rightarrow \quad 0 \le \frac{F(t)-F(0)}{t^2} = \frac{1}{2}F''(\vartheta t).$$

Nun $t \downarrow 0$; Stetigkeit von $F'' \Rightarrow F''(0) = h^\top f''(\tilde{x})h \ge 0$. □

Beispiel 3.1.4 $f(x) = \frac{1}{2}x^\top Hx + b^\top x$;

$$f''(x) = H.$$

Soll (PU) für dieses f eine Lösung haben, dann muss H positiv semidefinit sein.

Beispiel 3.1.5 *Rosenbrock-Funktion*

$$\begin{array}{lcl} fx_1 & = & -400x_1(x_2 - x_1^2) - 2(1-x_1) \\ fx_2 & = & 200(x_2 - x_1^2) \\ fx_1x_1 & = & -400(x_2 - x_1^2) + 800x_1^2 + 2 \\ fx_1x_2 & = & fx_2x_1 = -400x_1 \\ fx_2x_2 & = & 200 \end{array}$$

$$\Rightarrow \quad f''(1,1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

ist nach dem Satz von Sylvester positiv definit.

3.1.3 Hinreichende Bedingungen zweiter Ordnung

Aus der Gültigkeit von notwendigen Bedingungen erster und zweiter Ordnung kann man bekanntlich nicht auf lokale Optimalität schließen (Bsp: $f(x) = x^3$ bei $x = 0$). Dazu zieht man nach Möglichkeit hinreichende Bedingungen 2. Ordnung zu Rate.

Im Weiteren sagen wir "f ist in U aus der Klasse C^2", kurz "aus C^2", wenn f zweimal stetig differenzierbar in U ist.

Satz 3.1.4 *f sei aus C^2 in einer Umgebung von $\tilde{x} \in D$. Die notwendige Bedingung $\nabla f(\tilde{x}) = 0$ sowie*

$$h^\top f''(z)h \ge 0 \quad \forall h \in \mathbb{R}^n \tag{3.4}$$

sei erfüllt für alle $z \in B(\tilde{x}, r)$ mit einem $r > 0$. Dann ist $\tilde{x}$ lokales Minimum von (PU).

Beweis: Sei $x \in B(\tilde{x}, r)$ beliebig. Dann

$$\begin{array}{rcl} f(x) - f(\tilde{x}) & = & \underbrace{f'(\tilde{x})(x-\tilde{x})}_{0} + \tfrac{1}{2}\underbrace{(x-\tilde{x})}_{h} f''\underbrace{(\tilde{x} + \vartheta(x-\tilde{x}))}_{z}(x-\tilde{x}),\ \vartheta \in (0,1) \\ & \ge & 0 \qquad \text{wegen (3.4).} \end{array}$$

Offenbar gilt $z \in B(\tilde{x}, r)$. Deshalb ist $\tilde{x}$ ist lokales Min.

□

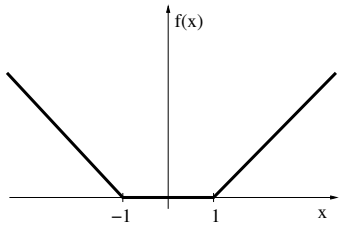
Beispiel 3.1.6 (Lineare Regression) *Hier hängt $f''(x) = H$ nicht von x ab. Ist H nur positiv semidefinit und erfüllt $\tilde{x}$ die notwendige Bedingung $H\tilde{x} + b = 0$, dann ist $\tilde{x}$ lokales Min. Sind zwei der Messpunkte ξ_i verschieden, dann ist H positiv definit, und wir haben Existenz und Eindeutigkeit.*

Beispiel 3.1.7 *Auf folgende Funktionen passt der Satz:*

(i) $f(x) \equiv 0 \quad \forall x \in \mathbb{R}^n$
Jedes $x \in \mathbb{R}^n$ ist lokales Minimum

(ii) $f(x) = \max\{0, |x| - 1\}$

Alle $x \in [-1, 1]$ sind lokale Minima, aber
$x = -1, x = 1$ passen nicht in die Theorie ...



Solche etwas pathologischen Fälle schließt man durch etwas schärfere Bedingungen aus, die strenge lokale Minima implizieren:

Satz 3.1.5 *f sei aus C^2 in einer Umgebung von $\tilde{x} \in D$, es gelte $\nabla f(\tilde{x}) = 0$, und $f''(\tilde{x})$ sei positiv definit, d. h.*

$$h^\top f''(\tilde{x}) h > 0 \quad \forall h \in \mathbb{R}^n, h \neq 0. \tag{3.5}$$

Dann existieren $r > 0$ und $\alpha > 0$, so dass die **quadratische Wachstumsbedingung**

$$\boxed{f(x) \geq f(\tilde{x}) + \alpha \|x - \tilde{x}\|^2 \quad \forall x \in B(\tilde{x}, r)}$$

erfüllt ist. Damit ist $\tilde{x}$ strenges lokales Minimum von (PU).

Beweis: Wie in der Vorlesung Analysis! Skizze:

- Aus dem bereits erwähnten Kompaktheitsschluss folgt die Äquivalenz von (3.5) mit

  $$h^\top f''(\tilde{x}) h \geq \tilde{\alpha} \|h\|^2 \quad \forall h \in \mathbb{R}^n,$$

  für ein $\tilde{\alpha} > 0$.

- Dann folgt

$$\begin{aligned} f(x) &= f(\tilde{x}) + \underbrace{\nabla f(\tilde{x})^\top (x - \tilde{x})}_{=0} + \tfrac{1}{2}(x - \tilde{x})^\top f''(\tilde{x} + \vartheta(x - \tilde{x}))(x - \tilde{x}), \quad \vartheta \in (0,1) \\ &= f(\tilde{x}) + \underbrace{\frac{1}{2}(x - \tilde{x})^\top f''(\tilde{x})(x - \tilde{x})}_{\geq \frac{1}{2}\tilde{\alpha}\|x - \tilde{x}\|^2} + \underbrace{\frac{1}{2}(x - \tilde{x})^\top [f''(\tilde{x} + \vartheta(x - \tilde{x})) - f''(\tilde{x})](x - \tilde{x})}_{\geq -\frac{\tilde{\alpha}}{4}\|x - \tilde{x}\|^2, \text{ wenn } \|x - \tilde{x}\| \text{ hinreichend klein}} \\ &\geq f(\tilde{x}) + \tfrac{\tilde{\alpha}}{4}\|x - \tilde{x}\|^2 \qquad (f \in C^2!) \end{aligned}$$

$\alpha := \frac{\tilde{\alpha}}{4}$ □

Offenbar gilt $h^\top f''(z)h \geq 0 \; \forall h, \; \forall z \in B(\tilde{x}, r)$, also impliziert (3.5) die Bedingung (3.4).

Beispiel 3.1.8 *Lineare Regression mit positiv definitem H.*

Beispiel 3.1.9 *Rosenbrock-Funktion bei* $\tilde{x} = [1,1]^\top$.

$$f''(1,1) = \begin{pmatrix} 802 & -400 \\ 400 & 200 \end{pmatrix}$$

ist positiv definit, also ist $\tilde{x}$ *strenges lokales Minimum.*

Beispiel 3.1.10 $\quad f(x) = x^{2p}, \quad p \in \mathbb{N}, \quad x \in \mathbb{R}.$

$\tilde{x} = 0$ *ist lokales Minimum, aber die Regel "*$f'(\tilde{x}) = 0 \wedge f''(\tilde{x}) > 0 \Rightarrow$ *lokales Minimum" funktioniert nur bei* $p = 1$*:*

$$\begin{array}{rcll} f'(x) & = & 2px^{2p-1} & \Rightarrow \quad f'(0) = 0 \\ f''(x) & = & 2p(2p-1)x^{2p-2} & \Rightarrow \quad f''(0) > 0 \quad \textit{falls} \quad p = 1 \\ & & & \quad\;\; f''(0) = 0 \quad \textit{falls} \quad p > 1. \end{array}$$

Satz 3.1.5 ist nur für $p = 1$ *anwendbar, Satz 3.1.4 stets.*

3.2 Konvexe Optimierungsaufgaben

Wir untersuchen die konvexe Aufgabe

$$\text{(P)} \qquad \min_{x \in \mathscr{F}} f(x)$$

mit konvexem $f : D \to \mathbb{R}$ und nichtleerer konvexer Menge $\mathscr{F} \subset D$.

Jede lokale Lösung von (P) ist damit eine globale.

Charakterisierung der Konvexität von f durch Ableitungen:

Satz 3.2.1 f *sei differenzierbar in* $\tilde{x} \in D$*. Dann ist* f *genau dann konvex auf* $\mathscr{F}$*, wenn*

$$\boxed{f(y) \geq f(x) + f'(x)(y-x) \quad \forall x, y \in \mathscr{F}} \tag{3.6}$$

Beweis: $\Rightarrow$: Ist f konvex, dann haben wir mit $t \in [0,1]$

$$t\,f(y) + (1-t)\,f(x) \geq f(t\,y + (1-t)\,x) = f(x + t\,(y-x)),$$

also nach einer Taylorentwicklung

$$t(f(y) - f(x)) + f(x) \geq f(x) + f'(x)t(y-x) + o(t).$$

Nach Division durch t und Grenzübergang $t \downarrow 0$ ergibt sich

$$f(y) - f(x) \geq f'(x)(y-x).$$

$\Leftarrow$: Es sei $z = tx + (1-t)y \in \mathscr{F}$ und die im Satz gegebene Beziehung erfüllt. Dann schreiben wir die folgenden 2 Ungleichungen auf:

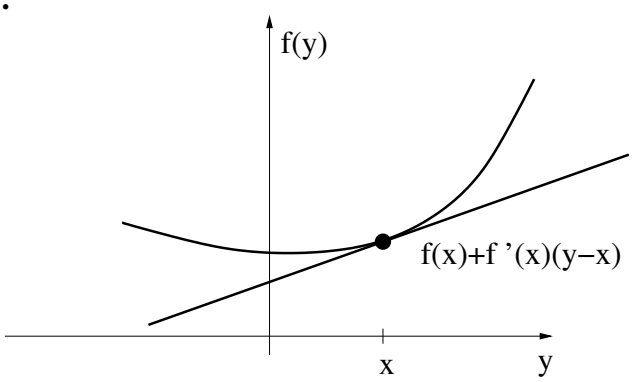
$$\begin{aligned} f'(z)(y-z) + f(z) &\leq f(y) \qquad |\cdot(1-t) \\ f'(z)(x-z) + f(z) &\leq f(x) \qquad |\cdot t \end{aligned}$$

Nach Multiplikation wie angegeben und Addition folgt

$$f'(z)\underbrace{((1-t)y + tx - z)}_{=0} + f(z) \leq (1-t)f(y) + tf(x).$$

Nach Konstruktion von z ist das offenbar zur Definition der Konvexität äquivalent. □

Illustration für $n = 1$:



Satz 3.2.2 *(i) f sei differenzierbar in D (aber nicht notwendig konvex) und $\mathscr{F}$ konvex. Ist $\tilde{x} \in \mathscr{F}$ Lösung der Optimierungsaufgabe (P), dann muss die* **Variationsungleichung**

$$\boxed{f'(\tilde{x})(x-\tilde{x}) \geq 0 \quad \forall x \in \mathscr{F}} \tag{3.7}$$

erfüllt sein.

(ii) Ist (P) eine konvexe Optimierungsaufgabe und erfüllt $\tilde{x} \in \mathscr{F}$ die Variationsungleichung (3.7)*, dann ist $\tilde{x}$ (globale) Lösung von (P).*

Beweis:

(i) Ist $\tilde{x}$ lokales Minimum, dann wissen wir

$$f(\tilde{x} + t(x-\tilde{x})) - f(\tilde{x}) \geq 0$$

für alle hinreichend kleinen $t > 0$, denn $\tilde{x} + t(x-\tilde{x})$ ist dann zulässig. Die Ungleichung bleibt nach Division durch t erhalten und nun folgt wie im Beweis von Satz 3.1.2 mit $h = x - \tilde{x}$ nach Grenzübergang $t \to 0$ die Ungleichung $f'(\tilde{x})(x-\tilde{x}) \geq 0$.

(ii) Gilt (3.7) und ist $y \in \mathscr{F}$ gegeben, so folgt wegen Konvexität und (3.6)

$$f(y) - f(\tilde{x}) \geq f'(\tilde{x})(y-\tilde{x}) \geq 0.$$

Damit ist $\tilde{x}$ globale Lösung. □

Strenge Konvexität kann man auch so charakterisieren:

Satz 3.2.3 *Sei D offen, $\mathscr{F} \subset D$ nichtleer und konvex, $f : D \to \mathbb{R}$ differenzierbar auf D. Dann ist f genau dann streng konvex auf $\mathscr{F}$, wenn*

$$f(y) > f(x) + f'(x)(y-x) \quad \forall x, y \in \mathscr{F}, x \neq y. \tag{3.8}$$

Beweis:

(i) f sei streng konvex. Dann ist f insbesondere konvex und

$$f(y) \geq f(x) + f'(x)(y-x). \tag{3.9}$$

Nehmen wir an, (3.8) gilt nicht. Dann gibt es ein Paar $(x,y), x \neq y$, so dass in (3.9) Gleichheit gilt. Wegen strenger Konvexität von f,

$$\begin{aligned} f\left(\frac{1}{2}x+\frac{1}{2}y\right) &< \frac{1}{2}\underbrace{f(y)}+\frac{1}{2}f(x) \\ &\qquad = \text{rechte Seite von (3.9) wegen Gleichheit} \\ &= \frac{1}{2}f(x)+\frac{1}{2}f'(x)(y-x)+\frac{1}{2}f(x) \\ &= f(x)+f'(x)\left(\frac{1}{2}y+\frac{1}{2}x-x\right) = f(x)+\underbrace{f'(x)(z-x)}_{\leq f(z)-f(x)} \text{ mit } z=(x+y)/2 \\ &\leq f(x)+f(z)-f(x) \quad \text{wegen Satz 3.2.1} \\ &= f\left(\frac{1}{2}x+\frac{1}{2}y\right), \end{aligned}$$

ein Widerspruch. Dabei haben wir in der Mitte die Ungleichung $f(z)-f(x) \geq f'(x)(z-x)$ angewendet.

(ii) Die andere Richtung zeigt man analog zum Beweis von Satz 3.2.1. □

Geometrisch ist die Aussage sehr einleuchtend.

Satz 3.2.4 *Sei f differenzierbar auf D und streng konvex auf $\mathscr{F}$ sowie $\tilde{x} \in \mathscr{F}$. Dann gilt: $\tilde{x}$ ist genau dann strenges lokales Minimum von (P) und damit auch strenges globales Minimum wenn die Variationsungleichung*

$$f'(\tilde{x})(x-\tilde{x}) \geq 0 \quad \forall x \in \mathscr{F}$$

erfüllt ist.

Beweis:

(i) $\Rightarrow$: Natürlich muss die Variationsungl. erfüllt sein.

(ii) $\Leftarrow$: aus der Variationsungl. und der strengen Konvexität folgt

$$f(x) \underset{\substack{\uparrow \\ (3.8)}}{\geq} f(\tilde{x}) + \underbrace{f'(\tilde{x})(x-\tilde{x})}_{\substack{\geq 0 \\ \text{Var.-ungl.}}} \geq f(\tilde{x}) \qquad \forall x \in \mathscr{F},\ x \neq \tilde{x}$$

□

Man kann Konvexität auch über zweite Ableitungen charakterisieren, als Verallgemeinerung der aus dem eindimensionalen Fall bekannten Beziehung

$$\text{“}f''(x) > 0 \Rightarrow f \text{ konvex”}$$

Satz 3.2.5 *Sei $D \subset \mathbb{R}^n$ offen, $\mathscr{F} \subset D$ konvex und nichtleer; $f : D \to \mathbb{R}$ aus C^2. Dann gilt*

(i) Ist $f''(x)$ positiv semidefinit für alle $x \in \mathscr{F}$, so ist f konvex auf $\mathscr{F}$. Ist $\mathscr{F}$ offen, so gilt auch die Umkehrung.

(ii) Ist $f''(x)$ positiv definit für alle $x \in \mathscr{F}$, so ist f streng konvex auf $\mathscr{F}$.

Beweis: Es seien $x, y \in \mathscr{F}$. Dann gilt mit einem $\vartheta \in (0,1)$:

$$f(y) - f(x) - f'(x)(y-x) = \frac{1}{2}(y-x)^\top f''(x+\vartheta(y-x))(y-x). \tag{3.10}$$

(i) Wegen pos. Semidefinitheit gilt dann, dass die rechte Seite von (3.10) nichtnegativ ist. Das heißt aber Konvexität nach Satz 3.2.1.

Umgekehrt sei $\mathscr{F}$ offen, $x \in \mathscr{F}$ beliebig, f konvex. Wir zeigen die positive Semidefinitheit von $f''(x)$. Dazu sei $d \in \mathbb{R}^n$ beliebig, $t \in \mathbb{R}, |t|$ hinreichend klein. Dann haben wir $x + td \in \mathscr{F}$ und schließlich gilt wegen der Charakterisierung der Konvexität durch die erste Ableitung

$$\begin{aligned} f(x+td) &\geq f(x) + f'(x)(td) \\ f(x) &\geq f(x+td) + f'(x+td)(-td). \end{aligned}$$

Wir haben zuerst eine Taylorentwicklung am Punkt x in Richtung $h = td$ vorgenommen, danach am Punkt $x+td$ in Richtung $h = -td$. Addition $\Rightarrow$

$$[f'(x+td) - f'(x)](td) \geq 0$$

$$\begin{aligned} \Rightarrow \quad d^\top f''(x)d &= \lim_{t \to 0} \frac{1}{t} \left[f'(x+td) - f'(x) \right] d \\ &= \lim_{t \to 0} \underbrace{\frac{1}{t^2}}_{\geq 0} \underbrace{[f'(x+td) - f'(x)](td)}_{\geq 0} \geq 0. \end{aligned}$$

(ii) Ist $f''(x)$ positiv definit, so steht in (3.10) für $y \neq x$ sofort eine (streng) positive rechte Seite. Nach Satz 3.2.3 ist damit f streng konvex.

□

Eine weitere Verschärfung des Begriffs der Konvexität ist:

Definition 3.2.1 *Eine Funktion f heißt* **gleichmäßig konvex** *auf einer konvexen Menge $F \subset D \subset \mathbb{R}^n$, wenn mit einem $\alpha > 0$ gilt*

$$(1-t)f(x)+tf(y) \geq f((1-t)x+ty)+t(1-t)\alpha\|x-y\|^2 \quad \forall x,y \in \mathscr{F}, t \in [0,1].$$

Damit kann man zeigen:

- Gleichmäßige Konvexität ist bei differenzierbarem f äquivalent zu

$$f(y)-f(x) \geq f'(x)(y-x)+\alpha\|x-y\|^2 \quad \forall x,y \in \mathscr{F}.$$

- Für $f \in C^2$ folgt gleichmäßige Konvexität aus der **gleichmäßigen positiven Definitheit**, d.h.

$$h^\top f''(x)h \geq \beta\|h\|^2 \quad \forall h \in \mathbb{R}^n,$$

  wobei $\beta > 0$ nicht von $x \in \mathscr{F}$ abhängt.

  Ist $\mathscr{F}$ offen, dann folgt umgekehrt aus der gleichmäßigen Konvexität von f auf $\mathscr{F}$, dass f'' gleichmäßig positiv definit ist.

4 Probleme ohne Restriktionen – Verfahren

4.1 Grundlagen

Wir untersuchen im gesamten Kapitel numerische Verfahren zur Lösung der unrestringierten Aufgabe

$$\text{(PU)} \qquad \boxed{\min_{x \in R^n} f(x)}$$

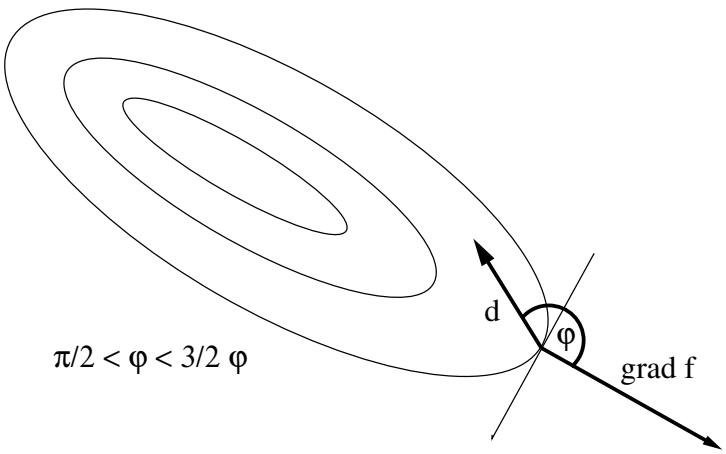
Da wir wissen, dass an der Stelle einer Lösung $\tilde{x}$ die Gleichung

$$\nabla f(\tilde{x}) = 0 \tag{4.11}$$

erfüllt sein muss, können wir diese Gleichung numerisch lösen, etwa mit dem Newton-Verfahren. Dieses liefert aber "nur" eine Lösung dieser Gleichung, welche nicht notwendig ein Minimum ergibt. Deshalb interessiert man sich für numerische Verfahren, die (4.11) lösen und gleichzeitig die Minimierung in (PU) berücksichtigen. Dazu gehören **Abstiegsverfahren**. Das sind iterative Verfahren, die schrittweise den Funktionswert von f verkleinern.

Definition 4.1.1 *$f : \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar an der Stelle x. Ein Vektor $d \in \mathbb{R}^n$ heißt* **Abstiegsrichtung** *von f im Punkt x, wenn*

$$\nabla f(x) \cdot d < 0.$$



Der Sinn dieser Definition ist klar, er wird bestätigt durch

Lemma 4.1.1 *f sei differenzierbar an der Stelle x und d eine Abstiegsrichtung. Dann existiert $\bar{\sigma} > 0$ mit $f(x+\sigma d) < f(x) \quad \forall \sigma \in [0, \bar{\sigma}]$.*

Beweis: Wegen

$$\nabla f(x) \cdot d = \lim_{\sigma \to 0} \frac{f(x+\sigma d) - f(x)}{\sigma} < 0$$

muss für hinreichend kleines $\sigma > 0$ die Beziehung $f(x+\sigma d) < f(x)$ erfüllt sein. □

Wichtige Beispiele von Abstiegsrichtungen:

Beispiel 4.1.1

- *Gilt $\nabla f(x) \neq 0$, so ist der* **Antigradient** $-\nabla f(x)$ *Abstiegsrichtung, denn für $d = -\nabla f(x)$ gilt*
$$f'(x)d = \nabla f(x) \cdot (-\nabla f(x)) = -\|\nabla f(x)\|^2 < 0.$$
- *Ist A positiv definite (n,n)-Matrix, dann ist $-A^{-1}\nabla f(x)$ Abstiegsrichtung. Das liegt daran, dass mit A auch A^{-1} positiv definit ist.*

Verfahren 4.1.1 (Allgemeine Form von Abstiegsverfahren)

1. *Wähle Startpunkt $x^0 \in \mathbb{R}^n, k := 0$, Abbruchparameter $\varepsilon > 0$.*
2. *Abbruch, falls $\|\nabla f(x^k)\| < \varepsilon$.*
3. *Berechne Abstiegsrichtung $d = d^k$ und Schrittweite $\sigma = \sigma_k > 0$, so dass*
$$f\left(x^k + \sigma_k d^k\right) < f\left(x^k\right),$$
$$x^{k+1} := x^k + \sigma_k d^k$$
4. *$k := k+1$, gehe zu 1.*

Bemerkungen:

- Numerisch verwendet man auch als Abbruchkriterium
$$|f(x^{k+1}) - ft(x^k)| < \varepsilon_1 \wedge \|x^{k+1} - x^k\| < \varepsilon_2,$$
wobei $\varepsilon_1, \varepsilon_2$ positive Abbruchschranken sind.
- Alternativ:
$$|f(x^{k+1}) - f(x^k t)| \approx |\sigma_k f'(x^k) d^k| < \varepsilon_1$$
und
$$\|x^{k+1} - x^k\|_\infty = \sigma_k \|d^k\|_\infty < \varepsilon_2.$$
- In der Regel ist die Wahl der Schrittweite σ das Hauptproblem.

4.2 Das Newton-Verfahren

Das Newton-Verfahren zur Lösung der Gleichung $\nabla f(x) = 0$ ist ein gängiges Mittel zur Bestimmung lokaler Extrema. Setzen wir $F(x) := \nabla f(x)$, so ist $F : \mathbb{R}^n \to \mathbb{R}^n$ gegeben und das bekannte Newton-Verfahren auf die Gleichung

$$F(x) = 0$$

anzuwenden. Die Grundidee des Verfahrens ist schnell wiederholt. Ist x^k bereits bestimmt, so verhält sich $F(x)$ in erster Näherung wie $F\left(x^k\right)+F'\left(x^k\right)\left(x-x^k\right)$, so dass wir von dieser Funktion eine Nullstelle x suchen. Wir lösen also das lineare Gleichungssystem

$$\boxed{F\left(x^k\right)+F'\left(x^k\right)\left(x-x^k\right)=0} \tag{4.12}$$

und erhalten als neue Näherung $x =: x^{k+1}$. Ist $F'\left(x^k\right)$ invertierbar, so folgt

$$x^{k+1} = x^k - F'\left(x^k\right)^{-1} F\left(x^k\right). \tag{4.13}$$

Für die Konvergenzanalysis des Verfahrens benötigen wir folgende **Voraussetzungen:**

(i) $F : \mathbb{R}^n \supset D \to \mathbb{R}^n$ ist differenzierbar in D, D offen, und hat in D eine Nullstelle $\tilde{x}$.

(ii) F' ist Lipschitz-stetig in D, d. h. $\exists L > 0$:

$$\boxed{\| F'(x) - F'(y) \| \le L \| x - y \| \quad \forall x, y \in D,}$$

(iii) $\exists F'(\tilde{x})^{-1}$.

Der Konvergenzbeweis des Newton-Verfahrens beruht auf folgenden bekannten und mit relativ geringem Aufwand beweisbaren Fakten:

Lemma 4.2.1 *Es gilt*

$$\| F(x) - F(y) - F'(y)(x-y) \| \le \frac{L}{2} \| x - y \|^2 \quad \forall x, y \in D.$$

Das folgt aus dem Mittelwertsatz, angewendet auf $\varphi(t) = F\left(x + t(x-y)\right))$.

Lemma 4.2.2 *Ist A eine nichtsinguläre (n,n)-Matrix, S eine Matrix gleichen Typs und $\| A^{-1} \| \, \| S \| < 1$, dann existiert $(A+S)^{-1}$ und*

$$\| (A+S)^{-1} \| \le \frac{\| A^{-1} \|}{1 - \| A^{-1} \| \, \| S \|}.$$

Lemma 4.2.3 *Sei $G : \bar{B}(\tilde{x}, r) \to \mathbb{R}^n$ eine Kontraktion, d. h. in $\bar{B}(\tilde{x}, r)$ Lipschitz-stetig mit Konstante $L < 1$. Ist $\tilde{x}$ ein Fixpunkt von G, dann ist es der einzige in $\bar{B}(\tilde{x}, r)$. Ausgehend von jedem beliebigen Startpunkt x^0 in dieser Kugel konvergiert die Folge x^k*

$$x^{k+1} = G(x^k)$$

gegen $\tilde{x}$ und

$$\| x^k - \tilde{x} \| \le L^k \| x^0 - \tilde{x} \|.$$

Die Aussage folgt aus dem Banachschen Fixpunktsatz.

Satz 4.2.1 (Konvergenz des Newton-Verfahrens) *Unter den obigen Voraussetzungen (i) - (iii) gibt es $r > 0, c > 0$, so dass das Newton-Verfahren für jeden Startpunkt $x^0 \in B(\tilde{x}, r)$ eine gegen $\tilde{x}$ quadratisch konvergente Folge x^k erzeugt, d.h.*

$$\| x^{k+1} - \tilde{x} \| \leq c \| x^k - \tilde{x} \|^2. \tag{4.14}$$

Beweisskizze. Der Beweis ist aus dem Kurs Numerische Mathematik bekannt; er wird nur der Vollständigkeit halber angegeben.

a) Mit Lemma 4.2.2 zeigt man

$$\| F'(x)^{-1} \| \leq 2 \| F'(\tilde{x})^{-1} \| \quad \forall x \in B(\tilde{x}, r_1)$$

b) Das wendet man an und findet

$$\begin{aligned} \| F'(x)^{-1} - F'(y)^{-1} \| &= \| \underbrace{F'(x)^{-1}}_{\| \cdot \| \leq 2\|F(\tilde{x})^{-1}\|} \underbrace{(F'(y) - F'(x))}_{\leq L\|x-y\|} \underbrace{F'(y)^{-1}}_{\leq 2\|F(\tilde{x})^{-1}\|} \| \\ &\leq 4L \| F'(\tilde{x})^{-1} \|^2 \quad \| x - y \| \end{aligned}$$

c) Aus Formel (4.13) wird klar – das Newton-Verfahren ist Fixpunktiteration für

$$\boxed{G(x) := x - F'(x)^{-1} F(x).}$$

G ist Kontraktion in $B(\tilde{x}, r)$:

$$\begin{aligned} G(x) - G(y) &= \underbrace{x - y}_{= F'(x)^{-1}F'(x)(x-y)} - F'(x)^{-1}F(x) + F'(y)^{-1}F(y) \\ &= \underbrace{F'(x)^{-1}}_{\text{beschr. wegen a)}} \underbrace{\{F(y) - F(x) - F'(x)(x-y)\}}_{\leq \frac{L}{2}\|x-y\|\,\|x-y\|} \\ &\quad + \underbrace{\left(F'(y)^{-1} - F'(x)^{-1}\right)}_{\leq c\|x-y\| \text{ wegen b)}} \underbrace{\cdot F(y)}_{\text{klein, wenn } y \text{ nahe an } \tilde{x}} \end{aligned}$$

Man schätzt ab und findet z. B.

$$\| G(x) - G(y) \| \leq \frac{1}{2} \| x - y \| \quad \text{in} \quad B(\tilde{x}, r)$$

für ein $0 < r \leq r_1$ klein genug gewählt.

d) Nun wird das Kontraktionslemma 4.2.3 benutzt. Die Iterationsfolge konvergiert gegen $\tilde{x}$. Die quadratische Konvergenz folgt aus

$$\begin{aligned} \| x^{k+1} - \tilde{x} \| &= \| x^k - F'(x^k)^{-1} F(x^k) - \tilde{x} \| \\ &= \underbrace{\| F'(x^k)^{-1}}_{\leq 2\|F'(\tilde{x})^{-1}\|} \underbrace{\left\{ F(\tilde{x}) - F(x^k) - F'(x^k)(\tilde{x} - x^k) \right\}}_{\leq \frac{L}{2}\|\tilde{x} - x^k\|^2} \| \\ &\leq c \| x^k - \tilde{x} \|^2 \end{aligned}$$

mit $c = L \| F'(\tilde{x})^{-1} \|$ □

Anwendung auf Optimierungsprobleme

Das Newton-Verfahren konvergiert also lokal quadratisch. Im Fall des freien Optimierungsproblems (PU) hatten wir

$$F(x) := \nabla f(x)$$

gesetzt. Wir fordern daher als **Voraussetzungen**

- f'' ist Lipschitz-stetig in einer Umgebung eines lokalen Minimums $\tilde{x}$ von f.
  Damit ist implizit auch die Existenz einer Nullstelle von ∇f vorausgesetzt.
- $f''(\tilde{x})$ ist positiv definit.
  (Das sichert die Existenz von $F'(\tilde{x})^{-1} = f''(\tilde{x})^{-1}$ und passt zur Minimum-Eigenschaft der gesuchten Lösung.)

Satz 4.2.2 *Unter obigen Voraussetzungen konvergiert das Newton-Verfahren*

$$\boxed{x^{k+1} = x^k - f''\left(x^k\right)^{-1} \nabla f\left(x^k\right)} \tag{4.15}$$

lokal quadratisch gegen das lokale Minimum $\tilde{x}$, falls der Startwert hinreichend nahe an $\tilde{x}$ gewählt wird.

In der numerischen Umsetzung invertiert man natürlich $f''(x^k)$ nicht, sondern man löst das Gleichungssystem

$$f''(x^k)(\underbrace{x^{k+1} - x^k}_{d^k}) = -\nabla f(x^k),$$

d. h. man bestimmt eine Richtung d^k aus

$$f''(x^k)d^k = -\nabla f(x^k)$$

und setzt

$$x^{k+1} := x^k + d^k.$$

Für d^k gilt

$$d^k = \underbrace{f''(x^k)^{-1}}_{\text{pos. definit}} \quad \underbrace{(-\nabla f(x^k))}_{\text{Antigradient}}$$

Daher ist d^k nach Bsp 4.1.1 Abstiegsrichtung, die sogenannte **Newton-Richtung**. Damit wird das Newton-Verfahren aber nicht automatisch ein Abstiegsverfahren, denn es wählt immer die Schrittweite $\sigma = 1$, und die kann zu groß sein. Deshalb wendet man eine geänderte Verfahrensvorschrift an,

$$\boxed{x^{k+1} = x^k - \sigma_k f''(x^k)^{-1} \nabla f(x^k).}$$

Das ist das **gedämpfte Newton-Verfahren**, vgl. Abschnitt 4.6.

Bemerkung: Man kann das Newton-Verfahren auch anders interpretieren: Die Verfahrensvorschrift (4.15) bedeutet

$$f''(x^k)(x^{k+1}-x^k)+\nabla f(x^k)=0.$$

Das ist gerade die notwendige Optimalitätsbedingung für Lösungen der quadratischen Optimierungsaufgabe

$$(Q)_k \qquad \boxed{\min_{x\in\mathbb{R}^n} \quad \nabla f(x^k)^\top(x-x^k)+\frac{1}{2}(x-x^k)^\top f''(x^k)(x-x^k).}$$

Ist $f''(x^k)$ positiv definit, so besitzt diese, wie wir inzwischen wissen, genau eine Lösung. Diese ist gerade x^{k+1}. Damit ist das Newton-Verfahren äquivalent zur Lösung einer Folge quadratischer Optimierungsaufgaben, wenn $f''(\tilde{x})$ positiv definit ist. Es ist damit ein **sequentiell-quadratisches Optimierungsverfahren** – ein sogenanntes **SQP-Verfahren** (von Sequential Quadratic Programming).

Man schreibt die Iterationsvorschrift so auf: Bestimme

$$\min \nabla f(x^k)^\top z+\frac{1}{2}z^\top f''(x^k)z$$

und setze

$$x^{k+1}:=x^k+z^k.$$

4.3 Allgemeine Aussagen für Abstiegsverfahren

Bevor wir einzelne Abstiegsverfahren wie Gradientenverfahren, gedämpftes Newton- oder Quasi-Newtonverfahren weiter diskutieren, beweisen wir einige allgemeine Aussagen, die für alle Abstiegsverfahren gleichermaßen zutreffen.

4.3.1 Effiziente Schrittweiten

Ist d^k eine Abstiegsrichtung und σ_k hinreichend klein, so gilt $f(x^k+\sigma_k d^k)<f(x^k)$. Das muss aber keineswegs zur Konvergenz des Abstiegsverfahrens in ein lokales Minimum führen, wie folgendes einfache Beispiel zeigt:

Beispiel 4.3.1 *$f(x)=x^2$, $x^0=1$, $d^k=-1$ für alle $k\geq 0$ und*

$$\sigma_k=\left(\frac{1}{2}\right)^{k+2}, \qquad k=0,1,\ldots$$

Die Folge $\{x^k\}$ strebt gegen $\frac{1}{2}$ – die Schrittweiten waren zu klein gewählt.

Startet ein Abstiegsverfahren bei x^0, so entstehen nur noch kleinere Funktionswerte. Deshalb liegen die weiteren Iterierten stets in der Niveaumenge $N(f,f(x^0))$.

Definition 4.3.1 *Es sei x aus $N(f, f(x^0))$ und $d \in \mathbb{R}^n$ eine Abstiegsrichtung. Eine Schrittweitenstrategie $(x,d) \mapsto \sigma$ heißt* **effizient**, *falls*

$$\boxed{f(x+\sigma d) \leq f(x) - c\left(\frac{\nabla f(x)\cdot d}{\|d\|}\right)^2} \tag{4.16}$$

mit einer von $x \in N(f, f(x^0))$ und d unabhängigen Konstante $c > 0$ gilt.

Erläuterung: In der Nähe eines lokalen Minimums $\tilde{x}$ ist der Gradient $\nabla f(\tilde{x})$ fast null. Daher kann man nahe bei $\tilde{x}$ keinen wesentlichen Abstieg mehr erwarten. Die Effizienz bewertet daher den Abstieg in Relation zum Wert $\nabla f(x)\cdot d$ für Einheitsvektoren d. Beachte: $d/\|d\|$ ist Einheitsvektor.

Plausibilitätsbetrachtung: Wir betrachten die Funktion

$$\varphi(\sigma) := f(x+\sigma d)$$

und dazu die erste positive Nullstelle von φ', d.h. die exakte Schrittweite σ_E (wir werden diese später noch genau definieren). Dann haben wir

$$0 = \varphi'(\sigma) = \nabla f(x+\sigma d)\cdot d \approx [\nabla f(x) + f''(x)\sigma d]\, d$$

und deshalb

$$\sigma_E \approx \frac{-\nabla f(x)\cdot d}{\langle d, f''(x)\, d\rangle}.$$

Es sei nun $f''(x)$ positiv definit mit kleinstem Eigenwert $\lambda > 0$. Dann folgt

$$c\frac{-\nabla f(x)\cdot d}{\|d\|^2} \leq \sigma_E \leq \frac{1}{\lambda}\frac{-\nabla f(x)\cdot d}{\|d\|^2}.$$

Also verhält sich σ_E mit einer gewissen Konstanten c wie

$$\sigma = -c\,\frac{\nabla f(x)\cdot d}{\|d\|^2}.$$

Wir haben

$$f(x+\sigma d) = f(x) + \sigma\nabla f(x)\cdot d + o(\sigma) \approx f(x) + \sigma\nabla f(x)\cdot d.$$

Setzen wir nun die obige Faustformel für σ ein, so folgt

$$f(x+\sigma d) - f(x) \approx -c\,\frac{\nabla f(x)\cdot d}{\|d\|^2}\nabla f(x)\cdot d = -c\left(\frac{\nabla f(x)\cdot d}{\|d\|}\right)^2.$$

Das ist die Beziehung in der Definition der Effizienz.

Sind Folgen $\{x^k\}, \{d^k\}$ mit $\nabla f(x^k)^\top d^k < 0$ und effiziente Schrittweiten σ_k gegeben, dann ist (4.16) mit einer von k unabhängigen Konstante $c > 0$ erfüllt.

Eine spezielle Form der Effizienz ist das **Prinzip des hinreichenden Abstiegs:** Man verlangt von x und d unabhängige Konstanten $c_1, c_2 > 0$, so dass

$$f(x+\sigma d) \le f(x) + c_1 \sigma \nabla f(x) \cdot d \quad \text{(hinreichend schneller Abstieg)} \tag{4.17}$$

$$\sigma \ge -c_2 \frac{\nabla f(x) \cdot d}{\| d \|^2} \qquad \text{(Mindestschrittweite)} \tag{4.18}$$

gemeinsam erfüllt sind.

Aus diesen beiden Bedingungen folgt die Effizienzbedingung (4.16) mit $c = c_1 c_2$, denn

$$f(x+\sigma d) \le f(x) + c_1 \left(-c_2 \frac{\nabla f(x) \cdot d}{\| d \|^2} \right) \nabla f(x)^\top d = f(x) - c_1 c_2 \left(\frac{\nabla f(x) \cdot d}{\| d \|} \right)^2 .$$

Bemerkung: Man kann unter Voraussetzung der Lipschitz-Stetigkeit von f' auf $N(f, f(x^0))$ die Existenz effektiver Schrittweiten beweisen, vgl. Lemma 4.3.4 in [1].

Mindestschrittweite bei quadratischer Funktion: Zur Erläuterung betrachten wir die einfache quadratische Funktion

$$f(x) = \frac{1}{2}\|x\|^2 + b \cdot x$$

und setzen $\varphi(\sigma) = f(x+\sigma d)$. Eine einfache Rechnung ergibt

$$\varphi'(\sigma) = \sigma \|d\|^2 + (x+b) \cdot d.$$

Aus $\varphi'(\sigma) = 0$ erhalten wir die optimale (exakte) Schrittweite, vgl. mit (4.18),

$$\sigma = -\frac{\nabla f(x) \cdot d}{\|d\|^2}.$$

4.3.2 Gradientenbezogene Richtungen

Leider garantiert die Wahl effizienter Schrittweiten allein nicht, dass die Folge der Gradienten $\{\nabla f(x^k)\}$ gegen null strebt. Dazu benötigt man *gradientenbezogene Richtungen.*

Ist $N(f, f(x^0))$ kompakt, so ist die Folge der Funktionswerte $\{f(x^k)\}$ (nach unten) beschränkt. Ist die Schrittweitenfolge $\{\sigma_k\}$ effizient, dann gilt

$$f(x^{k+1}) \le f(x^k) - c \left(\frac{\nabla f\left(x^k\right) \cdot d^k}{\| d^k \|} \right)^2$$

bzw. nach Umstellung

$$c \left(\frac{\nabla f\left(x^k\right) \cdot d^k}{\| d^k \|} \right)^2 \le f(x^k) - f(x^{k+1}).$$

Aus der Monotonie folgt mit der Beschränktheit nach unten die Konvergenz der Funktionswerte, so dass die rechte Seite nach null streben muss.

$$\frac{\nabla f\left(x^k\right)\cdot d^k}{\| d^k \|} \to 0, \quad k \to \infty. \tag{4.19}$$

Man will nun die Richtungen so wählen, dass daraus auch

$$\nabla f(x^k) \to 0, \quad k \to \infty. \tag{4.20}$$

folgt.

Die Beziehung (4.19) kann ohne (4.20) gelten, wenn die Richtung d^k in der Grenze orthogonal zu $\nabla f(x^k)$ wird. Das muss man ausschließen und gleichmäßig größer als der rechte Winkel zu $\nabla f(x^k)$ bleiben. Nun gilt

$$\begin{aligned} \cos\left(\nabla f(x^k), d^k\right) &= \frac{\nabla f(x^k)\cdot d^k}{\| \nabla f\left(x^k\right) \| \, \| d^k \|} =: \beta_k \\ \Rightarrow \qquad \beta_k \| \nabla f(x^k) \| &= \underbrace{\frac{\nabla f\left(x^k\right)\cdot d^k}{\| d^k \|}}_{\to 0 \text{ bei Effizienz}} \end{aligned}$$

Daraus folgt $\nabla f(x^k) \to 0$, falls $-\beta_k \geq c > 0 \quad \forall k$.

Definition 4.3.2 *Seien $x \in N(f, f(x^0)), d \in \mathbb{R}^n$. Eine Abstiegsrichtungs-Strategie $x \mapsto d$ heißt* **gradientenbezogen** *in x, wenn die* **Winkelbedingung**

$$-\frac{\nabla f(x)\cdot d}{\| \nabla f(x) \| \, \| d \|} \geq c_3 \tag{4.21}$$

mit einer von x und d unabhängigen Konstanten $c_3 > 0$ erfüllt ist.

Sie heißt **streng gradientenbezogen**, *wenn zusätzlich*

$$\frac{1}{c_4} \| \nabla f(x) \| \leq \| d \| \leq c_4 \| \nabla f(x) \| \tag{4.22}$$

mit einer von x und d unabhängigen Konstante $c_4 > 0$ gilt.

Folgerung: *Ist die Niveaumenge $N(f, f(x^0))$ kompakt und sind die Suchrichtungen gradientenbezogen, dann gilt*

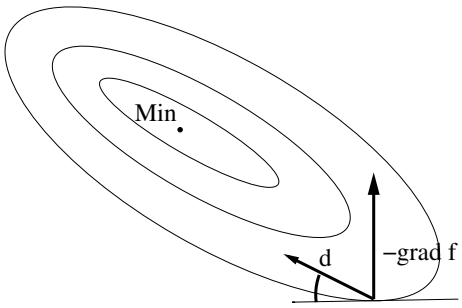
$$\lim_{k\to\infty} \nabla f(x^k) = 0.$$

In den Anwendungen dieses Begriffs arbeitet man mit von x abhängigen Richtungen $d = d(x)$. Dann sind die c_i unabhängig von x.

Beispiel 4.3.2 *Der Antigradient $d(x) = -\nabla f(x)$ ist streng gradientenbezogen, denn*

$$\begin{aligned} -\nabla f(x)\cdot d = \| \nabla f(x) \|^2 &= 1 \cdot \| \nabla f(x) \| \, \| d \| \\ \| \nabla f(x) \| \overset{(>)}{=} \| d \| &\overset{(>)}{=} \| \nabla f(x) \| \qquad d.h.\, c_3 = c_4 = 1. \end{aligned}$$

Veranschaulichung der Gradienten-Bezogenheit:



Mindestabstand zum rechten Winkel mit $-\nabla f(x)$ garantiert hinreichenden Abstieg relativ zu $\|d\|$.

Auch die **Newton-Richtung ist streng gradientenbezogen**. Wir wollen dies skizzieren und brauchen dazu folgende Voraussetzung:

(VLK) (Lokal gleichmäßige Konvexität)

Es sei $f : \mathbb{R}^n \to \mathbb{R}$; $N(f, f(x^0)) \subset D$ und D sei nichtleer, offen und konvex; $f \in C^2$ in D.

Mit einem $\alpha_1 > 0$ gelte

$$h^\top f''(x)h \geq \alpha_1 \| h \|^2 \quad \forall h \in \mathbb{R}^n, \forall x \in D,$$

d. h. gleichmäßige positive Definitheit von f'' auf D.

Die folgenden einfachen Schlüsse aus (VLK) kann man leicht selbst beweisen:

Lemma 4.3.1

- $N(f, f(x^0))$ *ist dann konvex und kompakt.*
- f *ist gleichmäßig konvex auf D.*

Für alle $x \in N(f, f(x^0))$ *und* $h \in R^n$ *gilt: Es existieren* $\alpha_2 > 0$, $\beta_1 > 0$, *so dass*

- $\| f''(x) \| \leq \alpha_2$
- $\| f''(x)^{-1} \| \leq \beta_2 := 1/\alpha_1$
- $\beta_1 \| h \|^2 \leq h^\top f''(x)^{-1} h \leq \beta_2 \| h \|^2$.

Beispiel 4.3.3 (Newton-Richtung) *Die lokale Konvexitätsvoraussetzung* (VLK) *sei erfüllt,* $x \in N(f, f(x^0))$, *und d die Newton-Richtung*

$$d = -f''(x)^{-1} \nabla f(x).$$

Diese ist dann **streng gradientenbezogen**, *denn zunächst folgt*

$$-\nabla f(x)^\top d = \nabla f(x)^\top f''(x)^{-1} \nabla f(x) \geq \beta_1 \| \nabla f(x) \|^2.$$

Die Winkelbedingung (4.21) folgt aus

$$-\nabla f \cdot d \overset{oben}{\geq} \beta_1 \|\nabla f\| \underbrace{\|\nabla f\|}_{\geq \frac{1}{\beta_2}\|d\|} \geq \frac{\beta_1}{\beta_2} \|\nabla f\| \|d\|.$$

Ferner

$$\|d\| \leq \beta_2 \|\nabla f(x)\| \;\; und \;\; \|\nabla f(x)\| = \|-f''(x)d\| \leq \alpha_2 \|d\|$$

$\Rightarrow$ (4.22), d.h. strenge Gradienten-Bezogenheit.

4.3.3 Allgemeine Konvergenzsätze

Folgende Voraussetzungen werden im Weiteren oft benötigt:

(VNK) **(Kompaktheit einer Niveaumenge)**

Für ein gegebenes $x^0 \in \mathbb{R}^n$ ist die Niveaumenge
$$N\left(f, f(x^0)\right) = \left\{x | f(x) \leq f(x^0)\right\}$$
kompakt.

(VFD) **(Stetige Differenzierbarkeit)**

$f \in C^1$ auf konvexer, offener Menge $D_0 \supset N(f, f(x^0))$.

Damit lässt sich zunächst zeigen:

Satz 4.3.1 *(VNK) und (VFD) seien erfüllt, die Suchrichtungen d^k des allgemeinen Abstiegsverfahrens 4.1.1 seien gradientenbezogen in x^k und die Schrittweiten σ_k effizient. Stoppt das Verfahren nicht nach endlich vielen Schritten, dann gilt $\nabla f(x^k) \to 0, k \to \infty$ und $\{x^k\}$ besitzt einen Häufungspunkt $\tilde{x}$. Für jeden solchen Häufungspunkt gilt*

$$\nabla f(\tilde{x}) = 0.$$

Beweis: Aus der vorausgesetzten Kompaktheit folgt sofort die behauptete Existenz mindestens eines Häufungspunkts von $\{x^k\}$. Wir wissen außerdem bereits, dass aus Effizienz und Gradientenbezogenheit die Konvergenz der Gradientenfolge gegen null folgt. □

Dass $\{x^k\}$ einen Häufungspunkt besitzt, bedeutet numerisch nicht viel. Man möchte $x^k \to \tilde{x}$ haben. In der Tat gilt

Satz 4.3.2 *Zusätzlich zu den Voraussetzungen von Satz 4.3.1 sei im allgemeinen Abstiegsverfahren 4.1.1*

- *die Folge der Richtungen d^k streng gradientenbezogen,*
- *die Schrittweitenfolge $\{\sigma_k\}$ beschränkt*
- *und die Menge aller Nullstellen von ∇f in $N(f, f(x^0))$ endlich.*

Stoppt das Verfahren nicht nach endlich vielen Schritten, dann konvergiert $\{x^k\}$ gegen eine Nullstelle von ∇f.

Beweis: Wir geben ein $\varepsilon > 0$ beliebig klein vor und zeigen die Existenz eines Punktes $\tilde{x}$ mit $\|x^k - \tilde{x}\| < \varepsilon$ für alle hinreichend großen k.

Wegen der Kompaktheitsvoraussetzung (VNK) ist H, die Menge aller Häufungspunkte (im Weiteren "HP") von x^k nichtleer. Jeder HP der Folge $\{x^k\}$ ist eine Nullstelle von $\nabla f(x)$, daher ist die Menge H endlich. Wir definieren

$$0 < \rho := \text{ kleinster Abstand zwischen verschiedenen Elementen von } H$$

und wählen

$$\varepsilon := \frac{\rho}{4}.$$

Der Abstand $d(x^k, H)$ strebt nach null (das überlegt man sich leicht durch einen indirekten Schluss). Daher existiert ein k_0 mit

$$d(x^k, H) < \varepsilon \quad \forall k > k_0. \tag{4.23}$$

Die Frage ist nun, ob sich das Verfahren einen festen Häufungspunkt "aussucht".

Strenge Gradientenbezogenheit, (4.22) $\Rightarrow$

$$\|x^{k+1} - x^k\| = \|\underbrace{\sigma_k}_{\leq \bar{\sigma}} d^k\| \leq c\bar{\sigma} \underbrace{\|\nabla f(x^k)\|}_{\to 0, \text{Satz 4.3.1}}. \tag{4.24}$$

Daher folgt

$$\|x^{k+1} - x^k\| < \varepsilon \quad \forall k \geq k_1.$$

Es sei $\tilde{x}$ irgendein *HP* von x^k. Dann existiert ein l, o.B.d.A. $l \geq \max\{k_0, k_1\}$, so dass

$$\|x^l - \tilde{x}\| < \varepsilon.$$

Nun folgt

$$\|x^{l+1} - \tilde{x}\| \leq \|x^{l+1} - x^l\| + \|x^l - \tilde{x}\| < 2\varepsilon = \frac{\rho}{2}.$$

Ist $\hat{x} \neq \tilde{x}$ ein anderer HP, so hat er mindestens den Abstand ρ zu $\tilde{x}$, also ist sein Abstand zu x^{l+1} größer oder gleich $\rho/2$. Laut (4.23) muss aber der Abstand kleiner als $\varepsilon = \rho/4$ sein, folglich kommt nur $\tilde{x}$ als nächster HP in Frage und wir erhalten

$$\|x^{l+1} - \tilde{x}\| < \varepsilon.$$

Induktiv folgt schließlich

$$\|x^k - \tilde{x}\| < \varepsilon \quad \forall k \geq l.$$

Wir folgern daraus $x^k \to \tilde{x}$. □

Diese bisherigen Resultate sind allgemein, aber schwach – sie sagen nichts über eine Konvergenzrate aus. Unter der Voraussetzung (VLK) hat man aber *gleichmäßige Konvexität* in $N(f, f(x^0))$ und man kann deshalb zeigen

$$\boxed{\frac{\alpha_1}{2}\|x - \tilde{x}\|^2 \leq f(x) - f(\tilde{x}) \leq \frac{1}{2\alpha_1}\|\nabla f(x)\|^2} \tag{4.25}$$

in $N(f, f(x^0))$, wobei $\tilde{x}$ das einzige lokale Minimum in $N(f, f(x^0))$ ist [1, Lemma 4.3.14]. Das ist automatisch das globale.

(Die linke Abschätzung, d.h. die quadratische Wachstumsbedingung, folgt aus der lokalen Konvexität (VLK), Taylorentwicklung und $\nabla f(\tilde{x}) = 0$ wie gehabt. Die rechte ist etwas komplizierter.)

Diese Eigenschaft ist die Grundlage für

Satz 4.3.3 *Die Konvexitätsvoraussetzung (VLK) sei erfüllt, die Richtungen d^k seien gradientenbezogen in x^k sowie die Schrittweiten $\{\sigma_k\}$ effizient.*

Stoppt das allgemeine Abstiegsverfahren nicht nach endlich vielen Schritten, dann konvergiert $\{x^k\}$ gegen das eindeutig bestimmte globale Minimum $\tilde{x}$ von f. Es gibt ein $q \in (0,1)$ mit

$$f(x^k t) - f(\tilde{x}) \leq q^k \left(f(x^0) - f(\tilde{x}) \right) \tag{4.26}$$

und

$$\| x^k - \tilde{x} \|^2 \leq \frac{2}{\alpha_1} q^k \left(f\left(x^0\right) - f(\tilde{x}) \right), \quad \forall k \geq 0. \tag{4.27}$$

Diesen Satz findet man in [1], 2. Auflage, als Satz 4.2.32.

Folgerung: $\| x^k - \tilde{x} \| \leq C \sqrt{q^k} = \tilde{C} \tilde{q}^k.$

Bemerkung: Damit verhält sich $\{x^k\}$ wie eine linear konvergente Folge. Eine Folge $\{x^k\}$ heißt *linear konvergent* wenn für alle $k = 0, 1, ...$

$$\| x^{k+1} - \tilde{x} \| \leq L \| x^k - \tilde{x} \|$$

mit einem $0 < L < 1$ erfüllt ist. Dann gilt

$$\| x^k - \tilde{x} \| \leq L^k \| x^0 - \tilde{x} \|.$$

4.4 Schrittweitenbestimmung

4.4.1 Exakte Schrittweite

Gegeben seien $x \in \mathbb{R}^n$ und eine Abstiegsrichtung $d \in \mathbb{R}^n$. Gesucht ist eine effiziente Schrittweitenstrategie σ. Wir definieren

$$\varphi(s) = f(x + s\,d).$$

Am besten wäre folgende Wahl von σ:

$$\min_{s \geq 0} f(x + sd) = \min_{s \geq 0} \varphi(s) = \varphi(\sigma).$$

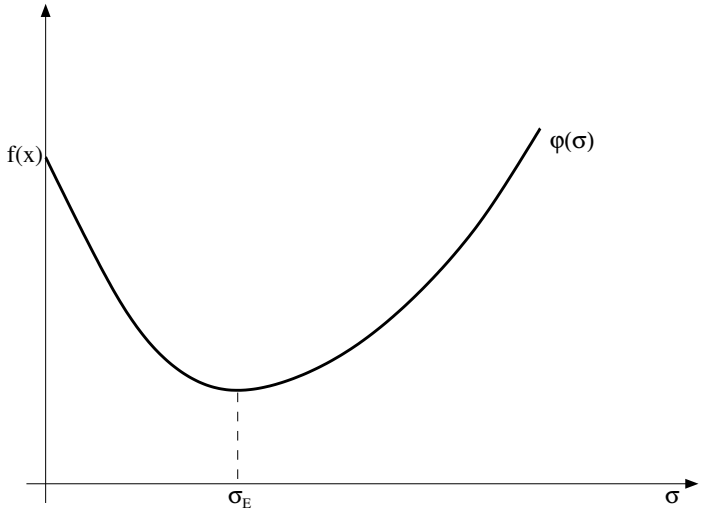
Das wird aber erstens nicht immer möglich sein (zum Beispiel bei $f(x) = e^{-x}$) und liefe zweitens auf globale Optimierung in $\mathbb{R}$ hinaus. Ist aber die Voraussetzung (VNK) erfüllt, die entsprechende Niveaumenge also kompakt, dann muss $\varphi(s)$ irgendwann größer als $\varphi(0)$ werden. Der Betrag des Vektors $x + s\,d$ wird ja für $s \to \infty$ beliebig groß, so dass $x + s\,d$ für hinreichend großes s die Menge $N(f, f(x^0))$ verlässt und dann einen größeren Wert als $f(x^0)$ ergibt.

Folglich hat $\varphi'(s) = \nabla f(x + sd) \cdot d$ eine kleinste positive Nullstelle σ_E.

Definition 4.4.1 *Die Zahl σ_E mit*

$$\varphi'(s)\begin{cases} = 0 & \text{für } s = \sigma_E, \\ < 0 & \text{für alle } s \in [0, \sigma_E) \end{cases}$$

heißt **exakte Schrittweite**.



Man kann sie nach unten wie folgt abschätzen:

$$\begin{aligned} 0 \underset{\text{Def von } \sigma_E}{\overset{\uparrow}{=}} \nabla f(x+\sigma_E d)\cdot d &\underset{\uparrow}{=} \nabla f(x)\cdot d + [\nabla f(x+\sigma_E d) - \nabla f(x)]\cdot d \\ &\underset{\text{Lipschitzbed.}}{\overset{\uparrow}{\leq}} \nabla f(x)\cdot d + \sigma_E L \,\|\, d \,\|^2 \end{aligned}$$

$$\Rightarrow \qquad \boxed{\sigma_E \geq \tilde{\sigma} = -\frac{\nabla f(x)\cdot d}{L\|d\|^2}.} \tag{4.28}$$

Außerdem bekommt man den Mindestabstieg

$$f(x+\sigma_E d) \leq f(x) + \frac{1}{2}\tilde{\sigma}\nabla f(x)\cdot d. \tag{4.29}$$

Damit sind σ_E und $\tilde{\sigma}$ effizient. Leider ist σ_E in der Regel schwer zu bestimmen. Eine Ausnahme bilden quadratische Funktionen

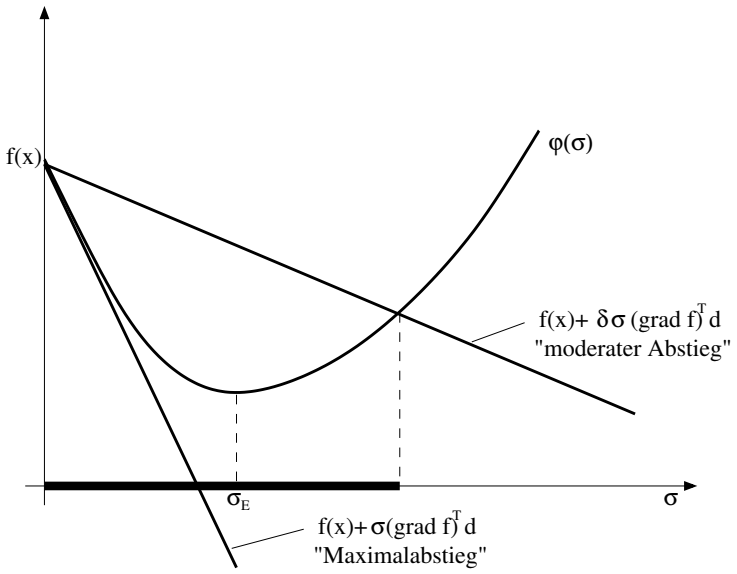
$$f(x) = \frac{1}{2}x^\top Hx + b^\top x$$

für die σ_E leicht explizit zu berechnen ist. Ansonsten muss man sich anders behelfen.

4.4.2 Schrittweite nach Armijo

Gegeben sei $x \in \mathbb{R}^n$ und eine Abstiegsrichtung d. Gesucht ist eine passende Schrittweite σ. Zur Bestimmung der **Armijo-Schrittweite** σ_A fordert man

- $f(x+\sigma_A d) \le f(x) + \delta\sigma_A \nabla f(x)\cdot d$ *Hinr. schneller Abstieg* (4.30)
- $\sigma_A \ge -c_2 \dfrac{\nabla f(x)\cdot d}{\|d\|^2}$ *Mindestschrittweite* (4.31)



Verfahren 4.4.1 (Armijo-Goldstein)

0. Fixiere Konstanten

$0 < \delta < 1$ *Abflachung*
$\gamma > 0$ *Effizienzkonstante*
$0 < \beta_1 \le \beta_2 < 1$

1. Startschrittweite: Wähle

$$\sigma_0 \ge -\gamma \frac{\nabla f(x)\cdot d}{\|d\|^2},$$

j : = 0

2. Wenn

$$f(x+\sigma_j d) \le f(x) + \delta\sigma_j \nabla f(x)\cdot d,$$

dann setze $\sigma_A := \sigma_j$*, fertig.*

3. Ansonsten verkleinere σ_j *so dass*

$$\sigma_{j+1} \in \left[\beta_1\sigma_j, \beta_2\sigma_j\right]$$

$j := j+1$*, gehe zu 2.*

Unter entsprechenden Voraussetzungen (d. h. Kompaktheit (VNK), Differenzierbarkeit (VFD), Lipschitzstetigkeit (VFL)) findet das Verfahren nach endlich vielen Schritten eine Schrittweite, die (4.30) – (4.31) erfüllt (vgl. [1, Satz 4.4.3]).

Die erste Beziehung ist klar, denn $\sigma_j \le \beta_2^j \sigma_0$ liegt irgendwann in diesem Bereich, siehe Skizze. Die zweite ist etwas kniffliger: l sei die Zahl der Iterationsschritte. Die Endlichkeit des Verfahrens beweisen wir hier nicht.

Gilt $l = 0$, so ist (4.31) mit $c_2 = \gamma$ erfüllt. Bei $l > 0$ liegt $s = \sigma_{l-1}$ noch außerhalb des akzeptablen Bereichs, also

$$\underbrace{f(x+sd) - f(x)}_{\substack{=\nabla f(x+\vartheta sd)\cdot s\,d \\ 0<\vartheta<1}} > \delta s \nabla f(x)^\top d.$$

$$\Rightarrow \nabla f(x+\vartheta sd)\cdot d = \frac{1}{s}[f(x+sd)-f(x)] > \delta \nabla f(x)\cdot d \quad \Big| \quad -\nabla f(x)\cdot d$$

$$\Rightarrow -(1-\delta)\nabla f(x)\cdot d < [\nabla f(x+\vartheta sd) - \nabla f(x)]\cdot d \underset{\substack{\uparrow \\ \text{Lipschitzst.}}}{\le} L\vartheta s \,\|d\|^2 \le sL\,\|d\|^2$$

$$\Rightarrow \boxed{s \ge -\frac{(1-\delta)}{L}\frac{\nabla f(x)\cdot d}{\|d\|^2}}$$

Wegen $\sigma_A \ge \beta_1 s$ (s ist die letzte Schrittweite und $\sigma_l \ge \beta_1 \sigma_{l-1}$ war gefordert) gilt am Ende die Beziehung

$$\sigma_A \ge -\underbrace{\frac{\beta_1(1-\delta)}{L}}_{c_2}\frac{\nabla f(x)\cdot d}{\|d\|^2},$$

$$c_2 = \min\left\{\gamma, \frac{\beta_1(1-\delta)}{L}\right\}$$

□

Bemerkung: Man kann z. B. $\beta_1 = \beta_2 = \frac{1}{2}$ wählen (Halbierung).

Zur Wahl der Verfahrensparameter siehe z.B. [1].

4.4.3 Schrittweite nach Powell

Dieses Verfahren wählt σ so, dass

$$f(x+\sigma d) \le f(x) + \delta\sigma\nabla f(x)\cdot d \qquad \text{(wie Armijo)} \tag{4.32}$$

und

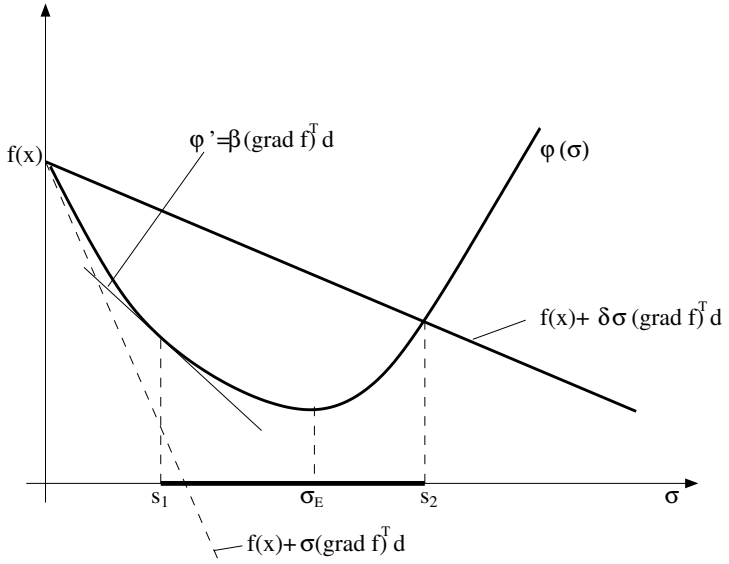
$$\nabla f(x+\sigma d)\cdot d \ge \beta \nabla f(x)\cdot d \quad \textit{Mindestschrittweite} \tag{4.33}$$

mit $0 < \delta < \beta < 1$ erfüllt ist.

Geometrische Interpretation: Sei $\varphi(s) := f(x+sd)$. Dann gilt

$$\varphi'(s) = \nabla f(x+sd)\cdot d$$

Demnach bestimmt das Verfahren die Powell-Schrittweite $\sigma = \sigma_p$ wie folgt:



Da β kleiner als 1 ist, kann nicht der für $\sigma = 0$ vorliegende Abstieg $\nabla f(x+sd)\cdot d$ gewählt werden, daher muss σ hinreichend positiv sein. So wird eine Mindestschrittweite garantiert.

Die Existenz einer solchen Schrittweite wird in [1, Satz 4.4.5] gezeigt. Die Bestimmung läuft über eine **Intervallschachtelung:**

Dazu definieren wir

$$G_1(\sigma) = \begin{cases} \frac{f(x+\sigma d)-f(x)}{\sigma \nabla f(x)\cdot d} & , \text{ für } \sigma > 0, \\ 1 & , \text{ für } \sigma = 0, \end{cases}$$

$$G_2(\sigma) = \frac{\nabla f(x+\sigma d)\cdot d}{\nabla f(x)\cdot d}.$$

Dann gilt

$$(4.32) \Leftrightarrow G_1(\sigma) \geq \delta$$

und

$$(4.33) \Leftrightarrow G_2(\sigma) \leq \beta.$$

Geometrisch bedeutet das in der Regel, dass sich $\mathbb{R}_+$ in 3 Intervalle $[0,s_1) \cup [s_1,s_2] \cup (s_2,\infty) =: I_1 \cup I_2 \cup I_3$ unterteilen lässt, mit

$$\begin{aligned} G_1(\sigma) \geq \delta \text{ und } G_2(\sigma) > \beta \quad & \text{in } I_1, \\ G_1(\sigma) \geq \delta \text{ und } G_2(\sigma) \leq \beta \quad & \text{in } I_2, \\ G_1(\sigma) < \delta \text{ und } G_2(\sigma) \leq \beta \quad & \text{in } I_3. \end{aligned}$$

Verfahren 4.4.2 (Powell)

1. *Wahl einer Startschrittweite* $\sigma_0 > 0$, $j := 0$.

   (i) Gilt $G_1(\sigma_0) \geq \delta$ *und* $G_2(\sigma_0) \leq \beta$*:* *Fertig!* $\sigma_p := \sigma_0$.

(ii) Liegt σ_0 in I_1, dann vergrößern wir das rechte Ende des Intervalls:

$$\begin{aligned} a_0 &:= \sigma_0 \\ b_0 &:= 2^l \sigma_0 \text{ mit minimalem } l \in \mathbb{N}, \text{ so dass} \\ &\qquad G_1(b_0) < \delta, \text{ (d.h. } b_0 \in I_3) \end{aligned}$$

Gehe zu 2.

(iii) Liegt σ_0 in I_3, also $G_1(\sigma_0) < \delta$, dann verkleinern wir das linke Ende des Intervalls:

$$\begin{aligned} b_0 &= \sigma_0, \\ a_0 &= 2^{-l} \sigma_0 \text{ mit minimalem } l \in \mathbb{N}, \text{ so dass} \\ &\qquad G_2(a_0) > \beta \text{ und } G_1(a_0) \geq \delta \\ &\qquad \text{(also } a_0 \in I_1). \end{aligned}$$

Damit ist das Intervall I_2 eingeschachtelt.

Gehe zu 2.

2. *Mittelwert $\sigma_j := \frac{1}{2}(a_j + b_j)$*

(i) Liegt σ_j in I_2: *Fertig,* $\sigma_p := \sigma_j$.

(ii) Liegt σ_j in I_1: *Dann* $a_{j+1} = \sigma_j, b_{j+1} = b_j$.

(iii) Liegt σ_j in I_3:

$$a_{j+1} = a_j, b_{j+1} = \sigma_j.$$

$j := j + 1$, *goto 2.*

Das Powell-Verfahren kann die Schrittweite auch vergrößern, ausgehend von der Startschrittweite σ_0, daher kann σ_0 an sich beliebig sein.

Typische Werte für β und δ sind z. B. $\delta = 0.1$ und $\beta = 0.9$.

Bemerkungen:

- σ_p wird (unter entsprechenden Voraussetzungen) in endlich vielen Schritten berechnet (vgl. [1, Satz 4.5.10])
- Unter den gleichen Voraussetzungen wie beim Armijo-Verfahren gilt folgendes allgemeines Konvergenzresultat:

  Wird die Schrittweite σ_k exakt, nach Armijo oder Powell gewählt, dann ist $\{\sigma_k\}$ Folge effizienter Schrittweiten.

4.5 Das Gradientenverfahren

Das Gradientenverfahren wird auch als **Verfahren des steilsten Abstiegs** bezeichnet. Als Richtung wählt man hier

$$d^k := -\nabla f(x^k).$$

Das Verfahren ist einfach zu implementieren aber in der Nähe des Optimums sehr langsam.

Verfahren 4.5.1 (Gradientenverfahren)

0. *Wähle einen Startvektor* $x^0, k := 0$ *und eine Abbruchschranke* $\varepsilon > 0$.
1. *Wenn* $\|\nabla f(x^k)\| < \varepsilon$: *Fertig.*
2. *Berechne*

$$d^k = -\nabla f(x^k)$$
$$\sigma_k \textit{ als effiziente Schrittweite (z. B. Armijo)}$$
$$x^{k+1} := x^k + \sigma_k d^k$$
$$k := k+1, \textit{goto } 1.$$

□

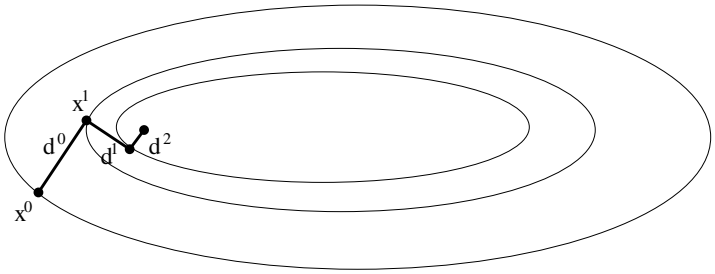
Unter den entsprechenden Voraussetzungen greifen unsere allgemeinen Konvergenzsätze.

Verfahrensnachteil: Die ersten Schritte sind noch schnell, aber "dann zieht sich's".

Begründung: Benutzt man die exakte Schrittweite, was ja an sich sinnvoll erscheint, dann hat man

$$0 = \frac{\partial}{\partial\sigma} f\big(\underbrace{x^k + \sigma d^k}_{x^{k+1}}\big)|_{\sigma=\sigma_E} = \nabla f\left(x^{k+1}\right) \cdot d^k$$
$$= -d^{k+1} \cdot d^k$$
$$\Rightarrow \quad d^{k+1} \perp d^k.$$

Die gewählten Richtungsvektoren sind daher zueinander orthogonal. In "schmalen Tälern" führt dieses *zig-zagging* zu sehr langsamer Konvergenz!



Ausweg: Berücksichtigung der Form der Niveaumengen, wie zum Beispiel im Zweidimensionalen unter Ausnutzung einer gewissen Elliptizität.

4.6 Gedämpftes Newton-Verfahren

4.6.1 Die Verfahrensvorschrift

Als Abstiegsrichtung wählt das Verfahren die Newton-Richtung

$$\boxed{d^k = -f''\left(x^k\right)^{-1} \nabla f\left(x^k\right).}$$

Verfahren 4.6.1 (Gedämpftes Newton-Verfahren)

0. Wähle einen Startpunkt $x^0 \in \mathbb{R}^n, k := 0$, *Abbruchschranke* $\varepsilon > 0$.

1. Wenn $\|\nabla f(x^k)\| \leq \varepsilon$: *Fertig.*

2. Berechne d^k *aus*

$$f''(x^k)d^k = -\nabla f(x^k).$$

Wähle eine effiziente Schrittweite σ_k *(z. B. Armijo o. Powell),*

$$x^{k+1} := x^k + \sigma_k d^k.$$

$k := k+1$, *goto 1.*

4.6.2 Interpretation der Newton-Richtung

Wir setzen $H = f''(x)$; H sei positiv definit. Außerdem definieren wir ein neues Skalarprodukt und eine zugehörige Norm in $\mathbb{R}^n$:

$$\langle x, y\rangle_H := x^\top H y.$$
$$\|x\|_H := \sqrt{\langle x, x\rangle_H} = \sqrt{x^\top H x}$$

Man kann nun zeigen:

Lemma 4.6.1 *Die Richtung*

$$\bar{d} = \frac{-H^{-1}\nabla f(x)}{\|H^{-1}\nabla f(x)\|_H}$$

löst unter der Voraussetzung $\nabla f(x) \neq 0$ *die Aufgabe*

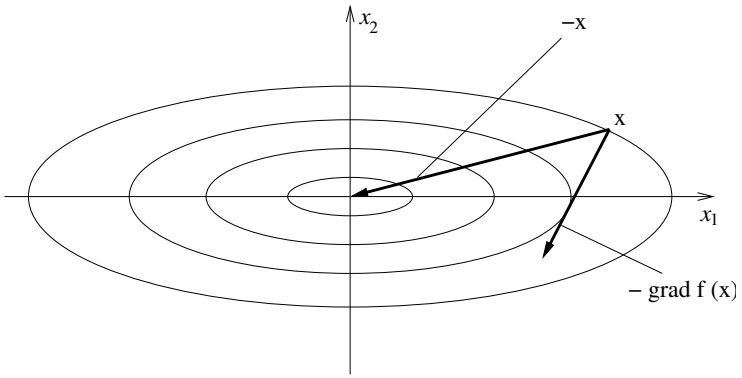
$$\min_{\|d\|_H = 1} \nabla f(x) \cdot d, \tag{4.34}$$

liefert also den steilsten Abstieg in der Norm $\|\cdot\|_H$.

Der Vorteil der Wahl von $-H^{-1}\nabla f$ anstelle der Gradientenrichtung $-\nabla f$ erschließt sich aus einer Betrachtung der quadratischen Funktion

$$f(x) = \frac{1}{2}x^\top H x,$$

z. B. bei $H = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ mit $a, b > 0$. Die Niveaulinien von $f(x)$ sind dann Ellipsen der Form $ax_1^2 + bx_2^2 = r^2$. Das Gradientenverfahren liefert eine Richtung, die am Nullpunkt – der Lösung der Aufgabe $\min f(x)$ – vorbeigeht.
Hingegen liefert $-f''(x)^{-1}\nabla f(x) = -H^{-1}Hx = -x$ genau die Richtung zur Lösung.



Folgerung: Bei der quadratischen Funktion würde das gedämpfte Newton-Verfahren bei exakter Wahl der Schrittweite in genau einem Schritt konvergieren.

4.6.3 Konvergenz des Verfahrens

Es sei die glm. Konvexitätsbedingung (VLK) sowie die glm. Lipschitzstetigkeit von f'' in $N(f, f(x^0))$ erfüllt, d. h.

$$\| f''(x) - f''(y) \| \leq L \| x - y \| \quad \forall x, y \in N(f, f(x^0)). \tag{4.35}$$

Damit sind die verwendeten Matrizen $f''(x^k)$ positiv definit, und das Verfahren ist durchführbar.

Nach dem allgemeinen Konvergenzsatz 4.3.3 ist das gedämpfte Newton-Verfahren linear konvergent. Aber es gilt mehr:

Satz 4.6.1 *Die Bedingung (VLK) sei erfüllt und die Schrittweiten σ_k nach Armijo oder Powell gewählt, wobei als Startschrittweite in jedem Schritt des gedämpften Newton-Verfahrens $\sigma_{k,0} = 1$ bestimmt werde. Weiter sei $0 < \delta < 1/2$ für die Konstante δ aus der Effizienzbedingung. Dann wählt das Verfahren für alle hinreichend großen k die volle Schrittweite $\sigma_k = 1$. Daher tritt in diesem Fall superlineare Konvergenz ein. Gilt zusätzlich (4.35), dann liegt quadratische Konvergenz vor.*

Der Beweis ist langwierig, siehe [1, Satz 4.7.4].

Folgerung: Nach endlich vielen Schritten geht das gedämpfte Newton-Verfahren in das ungedämpfte über. Von da ab konvergiert es wie dieses, nämlich quadratisch, falls (4.35) erfüllt ist, und sonst superlinear.

Definition 4.6.1 $\{x^k\}$ *konvergiert* **superlinear***gegen $\tilde{x}$, wenn*

$$\lim_{n\to\infty} \frac{\| x^{k+1} - \tilde{x} \|}{\| x^k - \tilde{x} \|} = 0.$$

Oft wird diese Eigenschaft auch q-superlineare Konvergenz genannt.

Beispiel: $x_k = q^k, |q| < 1$, konvergiert nur linear gegen Null, nicht superlinear. Aber $x_k = \frac{q^k}{k!}$ konvergiert superlinear, denn

$$\frac{q^{k+1}}{(k+1)!} / \frac{q^k}{k!} = \frac{q}{k+1} \to 0, k \to \infty.$$

Verwendet man die exakte Schrittweite, dann gilt $\sigma^k \to 1$, $k \to \infty$, und man kann quadratische Konvergenz zeigen (vgl. [1, Satz 4.6.4]).

Ein Vorteil superlinearer Konvergenz: *Bei superlinearer Konvergenz kann als Abbruchbedingung*

$$\|x^{k+1} - x^k\| \leq \varepsilon$$

benutzt werden.

Das sieht man so: Superlineare Konvergenz gegen $\tilde{x}$ ist äquivalent zur Existenz einer Nullfolge $\{\varepsilon_k\}$ mit

$$\frac{\|x^{k+1} - \tilde{x}\|}{\|x^k - \tilde{x}\|} = \varepsilon_k.$$

Wir schätzen ab und verwenden dabei die Beziehung $|\|a\| - \|b\|| \leq \|a - b\|$:

$$\left| \frac{\|x^{k+1} - x^k\|}{\|x^k - \tilde{x}\|} - 1 \right| = \left| \frac{\|x^{k+1} - x^k\|}{\|x^k - \tilde{x}\|} - \frac{\|x^k - \tilde{x}\|}{\|x^k - \tilde{x}\|} \right| \leq \frac{\|x^{k+1} - \tilde{x}\|}{\|x^k - \tilde{x}\|} \leq \varepsilon_k.$$

Damit verhält sich die Folge $\{x^{k+1} - \tilde{x}\}$ asymptotisch wie $\{x^{k+1} - x^k\}$.

Modifikation des Verfahrens:

- Wählt man $f''(x^0)$ anstelle von $f''(x^k)$ (*vereinfachtes Newton-Verfahren*), so ergibt sich globale aber nur lineare Konvergenz.
- Neuberechnung (einer Approximation) von $f''(x^k)$ nach jeweils n Schritten führt zu superlinearer Konvergenz.
- Verwendet man Differenzenquotienten zur Approximation der Ableitung, so ergibt sich superlineare Konvergenz, falls die Diskretisierung fein genug ist.

4.7 Variable Metrik- und Quasi-Newton-Verfahren

Die Grundidee dieser Verfahren ist folgende: Ausgehend von Informationen über $f''(x)$ (oder solchen, die dem nahekommen), wird eine Norm $\|\cdot\|_A$ benutzt, welche die Krümmung der Niveaulinien berücksichtigt – wie das bereits für quadratische Funktionen erläutert wurde.

4.7.1 Allgemeine Verfahrensvorschrift

Verfahren 4.7.1 (Variable Metrik)

0. *Startvektor* $x^0 \in \mathbb{R}^n$, $k := 0$, *Abbruchschranke* $\varepsilon > 0$.
1. *Wenn* $\|\nabla f(x^k)\| \leq \varepsilon$: *Fertig.*
2. *Berechne:*
   - *positiv definite symmetrische Matrix* A_k
   - $d^k = -A_k^{-1} \nabla f(x^k)$

- *effiziente Schrittweite* σ_k
- $x^{k+1} := x^k + \sigma_k d^k$
- $k := k+1$, *goto 1.*

In jedem Schritt wird die steilste Richtung bezüglich der Norm $\|\cdot\|_{A_k}$ gewählt.

Spezialfälle: $A_k \equiv I$: Gradientenverfahren
$A_k = f''(x^k)$: gedämpftes Newton-Verfahren

4.7.2 Globale Konvergenz von Variable-Metrik-Verfahren

Grundlegende Voraussetzung für das Verfahren ist die gleichmäßige positive Definitheit und Beschränktheit der Matrizen A_k.

Definition 4.7.1 *Eine Matrizenfolge* $\{A_k\}$ *symmetrischer* (n,n)*-Matrizen heißt gleichmäßig positiv definit und beschränkt, wenn Konstanten* $0 < \alpha_1 < \alpha_2$ *existieren, so dass*

$$\alpha_1 \|x\|^2 \leq x^\top A_k x \leq \alpha_2 \|x\|^2 \quad \forall x \in \mathbb{R}^n$$

für alle $k \in \mathbb{N}$ *gilt.*

Äquivalent dazu ist: Kleinster Eigenwert von $\lambda_1^{(k)} \geq \alpha_1$, größter $\leq \alpha_2$ oder : kleinster Eigenwert von $(A^k)^{-1} \geq \alpha_2^{-1}$, größter $\leq \alpha_1^{-1}$.
Im Vergleich mit den Sätzen über allgemeine Abstiegsverfahren ist es relativ plausibel, dass bei gleichmäßiger positiver Definitheit und Beschränktheit der gewählten Matrixfolge $\{A_k\}$ Folgendes gilt (beachte: Wegen gleichmäßiger positiver Definitheit ist (VLK) erfüllt):

- (VNK), (VFD) $\Rightarrow$ alle Richtungen d^k sind streng gradientenbezogen
- **Konvergenzaussagen** analog wie bei Satz 4.3.1 (Häufungspunkt von x^k mit $\nabla f = 0$), Satz 4.3.2 (Konvergenz gegen Nullstelle von ∇f) sowie Satz 4.3.3 (lineare Konvergenz)

4.7.3 Quasi-Newton-Methoden

In diesem Abschnitt behandeln wir zunächst die entsprechende Grundidee.

Ein Nachteil des gedämpften Newton-Verfahrens ist die aufwändige Berechnung von $f''(x^k)$ in jedem neuen Schritt. Im Gegensatz dazu möchte man an Stelle von $f''(x^k)$ eine Folge von Matrizen $\{A_k\}$ mit folgenden Eigenschaften aufbauen:

- Der Übergang von A_k zu A_{k+1} ist einfach zu bewerkstelligen und
- A_k approximiert $f''(x^k)$ in gewissem Sinne

und natürlich sollen alle Matrizen A_k positiv definit und symmetrisch sein.

Zur Motivation betrachten wir eine quadratische Funktion

$$f(x) = \frac{1}{2}x^\top Hx + b^\top x$$

und damit die quadratische unrestringierte Aufgabe

$$\text{(QU)} \qquad \min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Hx + b^\top x.$$

Hier gilt $\nabla f(x) = Hx + b$, $f''(x) = H$ und deshalb

$$\begin{aligned} f''(x^{k+1})(x^{k+1} - x^k) &= H(x^{k+1} - x^k) \pm b \\ &= \nabla f(x^{k+1}) - \nabla f(x^k), \end{aligned}$$

$k = 0, 1, 2 \ldots$.

Kennen wir H nicht, sondern nur die Gradienten von f und die Vektoren $x^0, \ldots, x^{n-1}$ und sind die Vektoren $x^{k+1} - x^k$ alle voneinander linear unabhängig, so bestimmen die n Gleichungssysteme

$$H(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k),\ k = 0, \ldots, n-1,$$

die Matrix H eindeutig.

Man beginnt mit einer positiv definiten symmetrischen Startmatrix A_0 und fordert dann von den A_k

$$\boxed{A_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k).} \tag{4.36}$$

(4.36) heißt **Quasi-Newton-Gleichung**.

4.7.4 BFGS-Update

Die Lösung der Quasi-Newton-Gleichung ist nicht eindeutig bestimmt. Man hat nun Formeln entwickelt, bei denen A_{k+1} recht einfach berechnet werden kann. Am bekanntesten ist die **BFGS-Formel** (nach Broyden, Fletcher, Goldfarb, Shanno).

Man definiert:

$$\boxed{\begin{aligned} &x^{k+1} - x^k =: s^k \\ &\nabla f(x^{k+1}) - \nabla f(x^k) =: y^k. \end{aligned}}$$

Die **Quasi-Newton-Gleichung** lautet mit diesen Bezeichnungen

$$A_{k+1}s^k = y^k.$$

Ausgehend von A_k wird dann A_{k+1} in zwei Schritten bestimmt:

- Zuerst berechnet man

$$\boxed{\tilde{A}_k := A_k - \frac{(A_k s^k)(A_k s^k)^\top}{(s^k, A_k s^k)}.} \tag{4.37}$$

Beachten Sie, dass in dieser Definition ein dyadisches Produkt auftritt.

Ist A_k bereits symmetrisch und positiv definit gewesen, so ist $\tilde{A}_k$ auch symmetrisch und zumindest positiv semidefinit. Das erhält man wie folgt:

Wir betrachten dazu die Matrix

$$\tilde{B}_k := A_k^{-1/2} \tilde{A}_k A_k^{-1/2}.$$

Dann ist positive Semidefinitheit von $\tilde{A}_k$ äquivalent zu der von $\tilde{B}_k$. Diese Matrix ist wohldefiniert, da A_k positiv definit ist. Wir setzen außerdem

$$\bar{s}^k := A_k^{1/2} s^k.$$

Dann folgt

$$\tilde{A}_k = A_k - \frac{(A_k s^k)(A_k s^k)^\top}{\langle s^k, A_k s^k \rangle} = A_k^{1/2} A_k^{1/2} - \frac{A_k s^k (s^k)^\top A_k}{\langle s^k, A_k s^k \rangle}$$

und damit

$$\tilde{B}_k = I - \frac{\bar{s}^k (\bar{s}^k)^\top}{\langle \bar{s}^k, \bar{s}^k rlangle e}.$$

Wir können jedes $x \in \mathbb{R}^n$ als orthogonale Summe $x = \lambda \bar{s}^k + d$ mit $\lambda \in \mathbb{R}$ und $(\bar{s}^k, d) = 0$ darstellen und finden

$$\langle x, \tilde{B}_k x \rangle = \langle \lambda \bar{s}^k + d, \lambda \bar{s}^k + d \rangle - \frac{\langle \lambda \bar{s}^k + d, s^k \rangle^2}{\langle \bar{s}^k, \bar{s}^k \rangle} = \|d\|^2 \geq 0 \quad \forall x \in \mathbb{R}^n,$$

also positive Semidefinitheit.

Außerdem gilt

$$\tilde{A}_k s^k = 0,$$

damit erfüllt $\tilde{A}_k$ allein nicht die Quasi-Newton-Gleichung. Es gilt offenbar

$$\operatorname{rang}(A_k s^k)(A_k s^k)^\top = 1,$$

deshalb heißt (4.37) **symmetrische Rang-1-Modifikation**.

- Durch eine zweite Rang-1-Modifikation versucht man, eine positiv definite Matrix zu bekommen:

$$A_{k+1} = \tilde{A}_k + \gamma_k w^k (w^k)^\top$$

und gleichzeitig die Quasi-Newton-Gleichung zu erfüllen.

Quasi-Newton-Gleichung:

$$A_{k+1} s^k = \underbrace{\tilde{A}_k s^k}_{=0} + \gamma_k w^k \overbrace{\langle w^k, s^k \rangle}^{\in \mathbb{R},\ =: \frac{1}{\gamma_k}} \overset{(!)}{=} y^k$$

$\Rightarrow w^k$ muss ein Vielfaches von y^k sein; wir wählen $w^k = y^k$ und

$$\gamma_k = \frac{1}{\langle y^k, s^k \rangle}.$$

Positive Definitheit: Zumindest muss dann für die spezielle Richtung s^k gelten

$$0 < \langle s^k, A_{k+1} s^k \rangle = \langle s^k, y^k \rangle. \tag{4.38}$$

Positive Definitheit: Man zeigt nun wie folgt, dass die Bedingung $\langle s^k, y^k \rangle > 0$ auch hinreichend für positive Definitheit von A_{k+1} ist, wenn A_k positiv definit war:

Wir übernehmen die Bezeichnungen von oben und definieren zusätzlich

$$\bar{y}^k := A_k^{-1/2} y^k \quad \text{sowie} \quad B_k = A_k^{-1/2} A_{k+1} A_k^{-1/2}.$$

Dann ist die positive Definitheit von A_{k+1} äquivalent zu der von B_k.

Wir haben nun

$$B_k = \tilde{B}_k + \frac{A_k^{-1/2} y^k \, (A_k^{-1/2} y^k)^\top}{\langle A_k^{-1/2} y^k, A_k^{1/2} s^k \rangle} = \tilde{B}_k + \frac{\bar{y}^k \, (\bar{y}^k)^\top}{\langle \bar{y}^k, \bar{s}^k \rangle}$$

Nach Multiplikation mit $x = \lambda \bar{s}_k + d$ wie oben ergibt sich

$$\langle x, B_k x \rangle = \langle x, \tilde{B}_k x \rangle + \frac{[\lambda \langle \bar{s}^k, \bar{y}^k \rangle + \langle d, \bar{y}^k \rangle]^2}{\langle \bar{y}^k, \bar{s}^k \rangle} = \|d\|^2 + \frac{[\lambda \langle \bar{s}^k, \bar{y}^k \rangle + \langle d, \bar{y}^k \rangle]^2}{\langle \bar{y}^k, \bar{s}^k \rangle}.$$

Die linke Seite kann deshalb nur dann null sein, wenn $d = 0$; daraus folgt aber auch, dass der zweite Summand verschwindet. Also folgt schließlich mit der positiven Semidefinitheit (wir haben oben gezeigt, dass $\tilde{B}_k$ positiv semidefinit ist) die positive Definitheit von B_k.

Bemerkung: Die Powell-Schrittweite sichert die Bedingung $\langle s^k, y^k \rangle > 0$: Wir haben offenbar $s^k = \sigma_k d^k$. Daher kann die Bedingung der Mindestschrittweite im Powell-Verfahren so aufgeschrieben werden:

$$\nabla f(x^{k+1}) \cdot s^k \geq \beta \, \nabla f(x^k) \cdot s^k.$$

Folglich können wir abschätzen

$$\begin{aligned} \langle y^k, s^k \rangle &= \langle \nabla f(x^{k+1}) - \nabla f(x^k), s^k \rangle \\ &= \underbrace{\langle \nabla f(x^{k+1}), s^k \rangle}_{\geq \beta \, \nabla f(x^k) \cdot s^k} - \langle \nabla f(x^k), s^k \rangle \\ &\geq \underbrace{(\beta - 1)}_{<0} \underbrace{\langle \nabla f(x^k), s^k \rangle}_{<0} > 0. \end{aligned}$$

Kurz: **Das Powell-Verfahren ist die passende Schrittweitenregel für BFGS.**

Insgesamt:

$$\boxed{A_{k+1} = A_k - \frac{A_k s^k \left(A_k s^k\right)^\top}{\langle s^k, A_k s^k \rangle} + \frac{y^k \left(y^k\right)^\top}{\langle y^k, s^k \rangle}} \tag{4.39}$$

Da die Summe von zwei Rang 1-Matrizen in der Regel vom Rang 2 ist, spricht man von einer **Rang-2-Modifikation**.

Bemerkung: (4.38) ist für eine quadratische Funktion erfüllt, falls H positiv definit ist:

$$f(x) = \frac{1}{2} x^\top H x, \qquad \nabla f = Hx$$

$$\begin{aligned}\Rightarrow y^k \cdot s^k &= \Big(\nabla f(x^{k+1}) - \nabla f(x^k)\Big) \cdot (x^{k+1} - x^k) \\ &= \Big(H(x^{k+1} - x^k)\Big) \cdot (x^{k+1} - x^k) \\ &\geq \alpha \,\| x^{k+1} - x^k \|^2 > 0.\end{aligned}$$

□

Nebenrechnungen:

- $\tilde{A}_k s^k = 0$: (ausgeschrieben ohne den Index k)

$$\tilde{A}s \;=\; As - \frac{(As)(As)^\top}{s^\top As}s = \frac{1}{s^\top As}[(s^\top As)As - As \quad \underbrace{s^\top A^\top s}_{= s^\top As,} \quad]$$

$$\text{da } A \text{ symmetrisch}$$

- Matrizen vom Typ $ss^\top$ haben Rang 1:

$$ss^\top = (s_i s_j) \;=\; \begin{pmatrix} s_1s_1 & s_1s_2 & \dots & s_1s_n \\ s_2s_1 & s_2s_2 & \dots & s_2s_n \\ \vdots & & & \\ s_ns_1 & s_ns_2 & \dots & s_ns_n \end{pmatrix}$$

Die Spalten sind Vielfache von s, deshalb liegt Rang 1 vor.

4.7.5 Das BFGS-Verfahren für quadratische Optimierungsprobleme

Wir wissen bereits: Bei quadratischer Funktion f ist

$$\sigma_E = \frac{-\nabla f(x) \cdot d}{d^\top H d}$$

die Formel für exakte Schrittweite. Die wenden wir an.

Verfahren 4.7.2 (BFGS für (QU))

0. *Wähle Startvektor x^0 und symmetrische positiv definite Startmatrix $A_0, k := 0$, Abbruchschranke $\varepsilon > 0$.*
1. *Wenn $\|\nabla f(x^0)\| < \varepsilon$: Fertig.*
2. *Berechne*

$$\begin{aligned} d^k &= -A_k^{-1} \nabla f(x^k) \\ \sigma_k &= \frac{-\langle \nabla f(x^k), d^k \rangle}{\langle d^k, Hd^k \rangle} \quad \textit{(exakte Schrittweite)} \\ x^{k+1} &= x^k + \sigma_k d^k \\ s^k &= x^{k+1} - x^k \\ y^k &= \nabla f(x^{k+1}) - \nabla f(x^k) \\ A_{k+1} &\textit{ als BFGS-Update} \\ k &:= k+1, \textit{ goto } 2. \end{aligned}$$

Dabei werden in der Praxis die Matrizen A_{k+1} etwas anders berechnet als nach der obigen Formel.

Von besonderer Bedeutung ist die folgende Definition:

Definition 4.7.2 (H-Orthogonalität) *Sei H eine symmetrische und positiv definite (n,n)-Matrix. Nicht verschwindende Vektoren $d^0,\dots,d^k, k<n$, heißen zueinander* **konjugiert** *bzw.* **orthogonal bezüglich** *H, wenn die folgenden Beziehungen erfüllt sind:*

$$\boxed{\langle d^i, Hd^j\rangle = 0, \qquad 0 \le i < j \le k.}$$

Genau das tritt beim BFGS-Verfahren ein!

Satz 4.7.1 *H sei symmetrisch und positiv definit. Dann berechnet das BFGS-Verfahren für (QU) in $m \le n$ Schritten das Minimum $\tilde{x}$ von f. Die vom Verfahren berechneten Richtungen sind zueinander konjugiert. Ist $m=n$, dann gilt $A_n = H$.*

Beweisidee: Die Konvergenz in $m \le n$ Schritten wird aus dem noch zu beweisenden Lemma 4.8.1 für das CG-Verfahren plausibel. Wir konzentrieren uns auf die H-Orthogonalität.

Es sei $\nabla f(x^0) \neq 0$, sonst ist die Lösung schon gefunden. Nun werden x^1, y^1, s^1 mit A_0 laut Verfahrensvorschrift berechnet. Dann wird A_1 bestimmt und ist nach [1, Lemma 4.8.5] positiv definit.

$$\begin{aligned}
\underbrace{x^1 - x^0}_{s^0} &= \sigma_0 d^0, \quad Hs^0 = \nabla f(x^1) - \nabla f(x^0) = y^0 \\
\Rightarrow \nabla f(x^1) &= \nabla f(x^0) + Hs^0 = \nabla f(x^0) + \sigma_0 H d^0 \\
\Rightarrow \nabla f(x^1)\cdot d^0 &= \nabla f(x^0)\cdot d^0 + \underbrace{\sigma_0 d^0 \cdot H}_{\substack{= -\nabla f(x^0) d^0 \\ \text{Def. von } \sigma_0}} \quad d^0 = 0
\end{aligned} \tag{4.40}$$

Wegen oben folgt $d^0 = \sigma_0^{-1} s^0$, also

$$\langle d^0, Hd^1\rangle = \frac{1}{\sigma_0} \underbrace{\left(Hs^0\right)^\top}_{y^{0\top} \text{ s. o.}} \underbrace{d^1}_{\substack{-A_1^{-1}\nabla f \\ \text{(BFGS-Verf.)}}} = -\frac{\overbrace{\langle y^0, A_1^{-1}}^{s^0:\text{Quasi-N.}} \nabla f(x^1)\rangle}{\sigma_0}$$

Nun kommt die Quasi-Newton-Gleichung für A_1 ins Spiel: $A_1^{-1} y^0 = s^0$, also

$$\langle d^0, Hd^1\rangle = -\frac{\langle s^0, \nabla f(x^1)\rangle}{\sigma_0} = -\langle \nabla f(x^1), d^0\rangle = 0 \quad \text{wegen (4.40).}$$

Damit haben wir für $k=1$ erhalten:

$$\left.\begin{array}{rlcl}
\text{(i)} & \nabla f(x^k)\cdot d^i & = & 0 \\
\text{(ii)} & A_k^{-1} y^i & = & s^i \\
\text{(iii)} & \langle d^i, Hd^k\rangle & = & 0
\end{array}\right\} \text{ für } \quad 0 \le i < k.$$

Induktion $k \to k+1$ liefert letztlich die Behauptung. □

Bemerkungen:

- Der Satz gilt nur bei exakter Rechnung und bei exakter Schrittweite.
- Alternativ könnte man sofort das Gleichungssystem

$$H\tilde{x} + b = 0$$

  mit dem Cholesky-Verfahren lösen.
- Ein Schritt BFGS entspricht im Rechenaufwand dem des ganzen Cholesky-Verfahrens, BFGS für quadratische Aufgaben lohnt sich also nicht. Es ist auch mehr für nichtlineare Probleme gedacht und so auch in MATLAB oder NAGLIB implementiert.

4.7.6 Das BFGS-Verfahren für nichtlineare Optimierungsaufgaben

Das Verfahren verläuft analog zum quadratischen Fall. Nur haben wir jetzt nicht mehr die exakte Schrittweite σ_E zur Verfügung, die sich im quadratischen Fall so gut berechnen lässt. Man kann beweisen:

- Lineare Konvergenz bei Verwendung effizienter Schrittweiten unter Voraussetzung (VLK) [1, Satz 4.8.12].
- Gilt zusätzlich noch (4.35) und werden die Schrittweiten nach Armijo oder Powell gewählt, dann tritt superlineare Konvergenz ein. Die erzeugten Matrizen A_k sind gleichmäßig positiv definit und beschränkt [1, Satz 4.8.13].

Es gilt nicht notwendig $\lim_{k\to\infty} A_k = f''(\tilde{x})$, sondern

$$\lim_{k\to\infty} \frac{\| (A_k - f''(\tilde{x}))\, d^k \|}{\| d^k \|} = 0.$$

4.8 Verfahren konjugierter Richtungen

4.8.1 CG-Verfahren für quadratische Optimierungsprobleme

Beim BFGS-Verfahren sind die erzeugten Richtungen zueinander H-orthogonal und man hat Konvergenz nach höchstens n Schritten. Nachteil: Die Matrizen A_k müssen abgespeichert werden. Bei hoher Dimension n ist das ein Problem: Bei 10000 Unbekannten müssen zum Beispiel $100 \cdot 10^6$ Elemente abgespeichert und berechnet werden.

Außerdem hat eine Matrix häufig eine besondere Struktur. Sie erlaubt es , Matrix-Vektor-Produkte effizient auszuführen, ohne dafür die Matrix aufbauen zu müssen. Die Hilbert-Matrix z.B. ist definiert als

$$H_{ij} = \frac{1}{i+j+1}$$

d.h., es gilt

$$[Hx]_i = \sum_{j=1}^{n} H_{ij}x_j = \sum_{j=1}^{n} \frac{x_j}{i+j+1}.$$

Die Koeffizienten sind bei Bedarf schnell berechnet, so dass man für eine Multiplikation mit einem Vektor nicht vorab die gesamte Matrix abspeichern muss, um das Produkt auszurechnen.

Das CG-Verfahren ist ein iteratives Verfahren, das für solche Zwecke zugeschnitten ist. Es ist eigentlich gedacht zur Lösung großdimensionierter Gleichungssysteme der Form

$$Hx + b = 0$$

mit positiv definiter und symmetrischer Matrix H. Alternativ, und darauf basiert das Verfahren, löst man die Aufgabe

$$\text{(QU)} \qquad \min_{\mathrm{x}\in\mathbb{R}^{\mathrm{n}}} \mathrm{f(x)} = \frac{1}{2}\mathrm{x}^{\top}\mathrm{Hx} + \mathrm{b}^{\top}\mathrm{x}.$$

Das Gradientenverfahren verwendet im Punkt x^k die Richtung $-\nabla f(x^k)$, die senkrecht auf den Niveauflächen der Funktion f steht und damit in der Regel nicht in Richtung des Minimums der quadratischen Funktion zeigt.

Die Idee der CG-Verfahren (Conjugate Gradient) ist es, statt dessen sukzessive H-orthogonale Richtungen zu generieren.

Konstruktion des Verfahrens. Die letzte Iterierte sei x^k, der letzte Gradient sei

$$g^k := \nabla f(x^k) = Hx^k + b.$$

Die neue Richtung sei d^k, die neue Schrittweite σ_k, also

$$x^{k+1} = x^k + \sigma_k d^k.$$

Die Kenntnis von d^k vorausgesetzt bestimmen wir zunächst die exakte Schrittweite aus

$$\frac{d}{d\sigma} f(x^k + \sigma d^k\rangle = 0 \tag{4.41}$$

und erhalten nach leichter Rechnung

$$\sigma_k = -\frac{\langle \nabla f(x^k), d^k\rangle}{\langle d^k, Hd^k\rangle} = -\frac{\langle g^k, d^k\rangle}{\langle d^k, Hd^k\rangle}. \tag{4.42}$$

Es bleibt die Bestimmung einer sinnvollen Richtung d^k. Ungünstig wäre, wie wir wissen, die Richtung des steilsten Abstiegs

$$d^k = -\nabla f(x^k) = -(b + Hx^k) = -g^k.$$

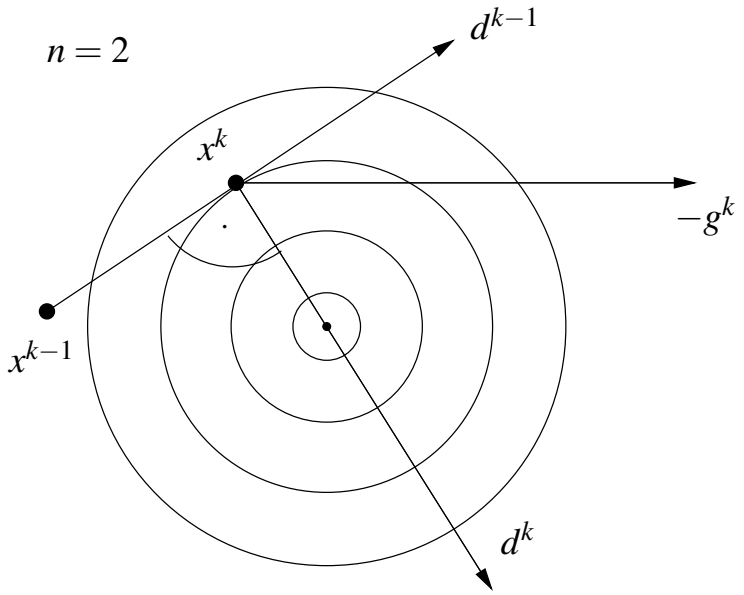
Ein Blick auf die Geometrie in der Energienorm zeigt die Lösung: Wir verwenden die bereits eingeführte Energienorm $\|x\|_H := \sqrt{(x, Hx)}$ und bezeichnen mit $\tilde{x} = -H^{-1}b$ die Lösung der Aufgabe. Dann ergibt die Taylorentwicklung bis zum Grad 2 (beachte $\nabla f(\tilde{x}) = 0$))

$$f(x) - f(\tilde{x}) = \frac{1}{2}\langle x - \tilde{x}, H(x - \tilde{x})\rangle,$$

also

$$f(x) = f(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_H^2.$$

Damit sind die Niveaulinien von f im Sinne der Energienorm Kreise um die Lösung $\tilde{x}$.



Die Iterierte $x^k = x^{k-1} + \sigma_{k-1} d^{k-1}$ ist bereits durch optimale Wahl von σ_{k-1} (exakte Schrittweite) entstanden. Deshalb wird (in Bezug auf die Energienorm) eine Niveaufläche von f an der Stelle x^k tangiert.

Weil in der Energienorm die Nivauflächen Sphären sind, sollte d^k H-orthogonal zu d^{k-1} sein, um schneller in den Mittelpunkt zu kommen. Um die bestehende Freiheit der Wahl von d^k einzugrenzen, bilden wir eine passende Linearkombination von d^{k-1} und g^k und bleiben damit in der von d^{k-1} und g^k aufgespannten Ebene (siehe obige Abbildung). Wir machen den Ansatz

$$d^k = -g^k + \beta_k d^{k-1}$$

und fordern $\langle d^k, d^{k-1} \rangle_H = 0$,

$$0 = \langle d^k, H\, d^{k-1} \rangle = \langle -g^k + \beta_k d^{k-1}, H\, d^{k-1} \rangle,$$

also

$$\beta_k = \frac{\langle g^k, H\, d^{k-1} \rangle}{\langle d^{k-1}, H\, d^{k-1} \rangle}.$$

Damit ist die Grundidee erklärt, aber die so gefundenen Formeln für die Parameter σ_k und β_k haben noch einen gewissen Nachteil: Sie enthalten zu viele, in der Regel teure Produkte mit der Matrix H. Wir leiten deshalb äquivalente, aber bessere Formeln her. Die folgenden Aussagen beschreiben die wichtigsten Eigenschaften des Verfahrens:

Lemma 4.8.1 *Es seien $d^0, d^1, ..., d^{n-1}$ gegebene konjugierte Richtungen. Dann liefert für jedes $x^0 \in \mathbb{R}^n$ die Vorschrift*

$$x^{k+1} = x^k + \sigma_k d^k$$
$$\sigma_k = -\frac{\langle \nabla f(x^k), d^k \rangle}{\langle d^k, Hd^k \rangle} \quad (\textit{exakte Schrittweite})$$

nach höchstens n Schritten die Lösung $\tilde{x} = -H^{-1}b$.

Beweis. Da die Richtungen d^i den ganzen Raum aufspannen, existieren Koeffizienten σ_i, $i = 1, \ldots, n-1$, so dass die Gleichung

$$\tilde{x} - x^0 = \sum_{i=0}^{n-1} \sigma_i d^i$$

erfüllt ist. Multiplizieren wir diese für festes k von links mit $(d^k)^\top H$, dann folgt

$$\sigma_k \langle d^k, Hd^k \rangle = \langle d^k, H(\tilde{x} - x^0) \rangle = \langle d^k, H(-H^{-1}b - x^0) \rangle = -\langle d^k, Hx^0 + b \rangle.$$

Weiter erhalten wir schrittweise

$$\langle d^k, Hx^0 + b \rangle = \langle d^k, Hx^1 + H \underbrace{(x^0 - x^1)}_{-\sigma_0 d^0} + b \rangle = \langle d^k, Hx^1 + b \rangle = \ldots = \langle d^k, Hx^k + b \rangle$$

wegen H-Orthogonalität. Aus beiden Beziehungen folgt

$$\sigma_k = -\frac{(d^k)^\top \nabla f(x^k)}{(d^k)^\top H d^k},$$

also stimmen die σ_k aus der Darstellung von $\tilde{x} - x^0$ mit den im Satz genannten Koeffizienten überein. Aus der Iterationsvorschrift ergibt sich aber ebenfalls

$$x^n - x^0 = \sum_{i=0}^{n-1} \sigma_i d^i$$

und daher haben wir $\tilde{x} = x^n$. □

Lemma 4.8.2 *Es sei $d^0 := -g^0$ definiert und $k \in \{1, ..., n\}$. Die Iterierte x^k minimiert f nicht nur auf der Geraden $\{x^{k-1} + \sigma d^{k-1} | \sigma \in \mathbb{R}\}$ sondern auch im Raum $x_0 + V_k$, mit $V_k = span \{d^0, ..., d^{k-1}\}$. Insbesondere gilt*

$$\langle g^k, d^i \rangle = 0 \quad \textit{für alle} \quad i < k. \tag{4.43}$$

Beweis. Es genügt, (4.43) zu zeigen, denn dann erfüllt x^k die notwendige Optimalitätsbedingung der entsprechenden Optimierungsaufgabe und ist wegen Konvexität auch optimal. Wir beweisen die Aussage induktiv. Für $k = 1$ stimmt (4.43) offenbar. Sie sei nun richtig für ein $k < n$. Wir zeigen zuerst $g^{k+1} \perp d^k$:

$$\langle d^k, g^{k+1} \rangle = \langle d^k, Hx^{k+1} + b \rangle = \langle d^k, \underbrace{Hx^k + b}_{=g^k} \rangle + \sigma_k \langle d^k, Hd^k \rangle = 0$$

wegen der Bildungsvorschrift der σ_k. Nun müssen wir noch die Orthogonalität zu $d^0,...,d^{k-1}$ zeigen. Wir haben

$$g^{k+1} - g^k = H(x^{k+1} - x^k) = \sigma_k H d^k. \tag{4.44}$$

Für alle $i \in \{0,\dots,k-1\}$ folgt daher wegen der H-Orthogonalität und der Induktionsvoraussetzung

$$0 = \langle d^i, \sigma_k H d^k \rangle = \langle d^i, g^{k+1} - g^k \rangle = \langle d^i, g^{k+1} \rangle.$$

□

Lemma 4.8.3 *Für die im obigen Verfahren erzeugten Unterräume gilt*

$$V_k = span\{d^0,...,d^{k-1}\} = span\{g^0,...,g^{k-1}\}, \quad k = 1,\dots,n.$$

Beweis. Wir zeigen die Aussage induktiv. Zunächst gilt $V_1 = span\{d^0\} = span\{g^0\}$. Die Aussage sei bereits für k bewiesen. Die Konstruktionsvorschrift für d^k besagt

$$d^k = g^k + \beta_k d^{k-1}.$$

So haben wir offenbar $g^k \in span\{d^0,...,d^k\} = V_{k+1}$ und deshalb

$$span\{g^0,...,g^k\} \subset V_{k+1}.$$

Wegen (4.43) gilt

$$\langle g^k, d \rangle = 0 \qquad \forall d \in V_k, \tag{4.45}$$

also ist g^k orthogonal zu V_k. Folglich hat $span\{g^0,...,g^k\}$ die Dimension $k+1$ und wegen der obigen Inklusion kann nur noch gelten

$$span\{g^0,...,g^k\} = V_{k+1}.$$

□

Diese Ergebnisse gestatten nun die Umformung der Koeffizienten, um am Ende auf die gängige Form des CG-Verfahrens zu kommen. Wegen des letzten Lemmas können wir $d = g^{k-1} \in V_k$ in (4.45) einsetzen und erhalten

$$\langle g^k, g^{k-1} \rangle = 0. \tag{4.46}$$

Außerdem haben wir nach (4.44)

$$Hd^k = \frac{1}{\sigma_k} H(x^{k+1} - x^k) = \frac{1}{\sigma_k}(g^{k+1} - g^k). \tag{4.47}$$

Schließlich folgt daraus noch

$$\langle d^k, g^k \rangle = \langle -g^k + \beta_k d^{k-1}, g^k \rangle = -\|g_k\|^2 + \beta_k \langle d^{k-1}, g^k \rangle = -\|g_k\|^2 \tag{4.48}$$

wegen $(d^{k-1}, g^k) = 0$. Deshalb können wir offenbar die Formel für die Schrittweite auch so aufschreiben:

$$\sigma_k = \frac{\langle g^k, g^k \rangle}{\langle d^k, Hd^k \rangle}. \tag{4.49}$$

Damit ergibt sich auch

$$\beta_{k+1} \overset{(4.47)}{=} \frac{\langle g^{k+1}, \sigma_k^{-1}(g^{k+1} - g^k) \rangle}{\langle d^k, g^{k+1} - g^k \rangle \sigma_k^{-1}} = \frac{\langle g^{k+1}, g^{k+1} \rangle}{-\langle d^k, g^k \rangle} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}.$$

Wir erinnern noch einmal an die Bezeichung $g^k = \nabla f(x^k)$.

Verfahren 4.8.1 (Quadratisches CG-Verfahren)

0. *Wähle x^0, berechne $d^0 = -g^0 = -(Hx^0 + b)$*
   Abbruchschranke $\varepsilon < 0$, $k := 0$

1. *Wenn $\|g^k\| < \varepsilon$:* *Fertig.*

2. *Berechne*

$$\sigma_k = \frac{\|g^k\|^2}{\langle d^k, Hd^k\rangle}$$
$$x^{k+1} = x^k + \sigma_k d^k$$
$$g^{k+1} = Hx^{k+1} + b = g^k + \sigma_k H d^k$$
$$\beta_{k+1} = \frac{\|g^{k+1}\|^2}{\|g^k\|^2}$$
$$d^{k+1} = -g^{k+1} + \beta_{k+1} d^k,$$

   $k := k+1$, *goto 1.*

Folgende weitere Eigenschaften des CG-Verfahrens sollte noch erwähnt werden:

$$V_k = \text{span}\{\mathrm{g}^0, \mathrm{Hg}^0, ..., \mathrm{H}^{\mathrm{k}-1}\mathrm{g}^0\}, \quad \mathrm{k} = 1, \ldots, \mathrm{n}, \tag{4.50}$$

denn aus (4.44) folgt

$$g^k = g^{k-1} + \sigma_{k-1} H d^{k-1}$$

und daraus schließlich induktiv die Aussage.

Bemerkung: Neben der oben angeführten Formel der Berechnung von β_k (das ist die Variante nach Fletcher und Reeves) gibt es auch andere Versionen. Bekannt sind zum Beispiel die von Hestenes und Stiefel, Polak und Ribiere sowie von Hager und Zhang.

4.8.2 Konvergenzgeschwindigkeit des CG-Verfahrens

Für symmetrische, positiv definite (n,n)-Matrizen H mit Eigenwerten $\lambda_1 < \ldots < \lambda_n$ ist die Kondition definiert durch

$$\kappa(H) = \frac{\lambda_n}{\lambda_1}.$$

Wendet man das **Gradientenverfahren mit exakter Schrittweite** auf unser quadratisches Optimierungsproblem an, so ergibt sich für den Fehler in der Energienorm $\|x\|_H := \sqrt{x^\top Hx}$ (vgl. z.B. [7]):

$$\|\tilde{x} - x^{k+1}\|_H \leq \left(\frac{\kappa(H)-1}{\kappa(H)+1}\right)^k \|\tilde{x} - x^0\|_H.$$

Für das CG-Verfahren ergibt sich folgende bessere Abschätzung:

Satz 4.8.1 *Der Approximationsfehler von $\tilde{x}-x^k$ im CG-Verfahren lässt sich in der Energienorm abschätzen durch*

$$\|\tilde{x}-x^k\|_H \le 2\left(\frac{\sqrt{\kappa(H)}-1}{\sqrt{\kappa(H)}+1}\right)^k \|\tilde{x}-x^0\|_H.$$

BEWEISSKIZZE. Nach Lemma 4.8.2 gilt

$$\|\tilde{x}-x^k\| \le \|\tilde{x}-y\| \quad \forall y \in V_k. \tag{$*$}$$

Wegen (4.50) lässt sich $y \in V_k$ als Linarkombination von Potenzen von H angewendet auf g^0 schreiben. D.h. es gibt ein Polynom P_{k-1} vom Grad $k-1$ so dass

$$\begin{aligned} y &= x^0 + P_{k-1}(H)g^0 = x^0 + P_{k-1}(H)(Hx^0+b) \\ &= x^0 + HP_{k-1}(H)(x^0-\tilde{x}) \end{aligned}$$

$$\begin{aligned} \Rightarrow \tilde{x}-y &= \tilde{x}-x_0 - HP_{k-1}(H)(x^0-\tilde{x}) \\ &= \underbrace{(I+HP_{k-1}(H))}_{=:Q_k(H)}(\tilde{x}-x^0) \end{aligned}$$

mit einem Polynom $Q_k \in \mathscr{P}_k$ vom Grad k mit $Q_k(0)=1$.

Ist $\{z_1,\ldots,z_n\}$ ein Orthonormalsystem aus Eigenvektoren von H, dann gilt

$$\tilde{x}-x^0 = \sum_{j=1}^n c_j z_j,$$

folglich

$$\tilde{x}-y = \sum_{j=1}^n c_j Q_k(H) z_j = \sum_{j=1}^n c_j Q_k(\lambda_j) z_j.$$

Daraus folgt

$$\begin{aligned} \|\tilde{x}-y\|_H^2 &= \left[\sum_{j=1}^n c_j Q_k(\lambda_j) z_j\right]^\top H \left(\sum_{j=1}^n c_j Q_k(\lambda_j) z_j\right) \\ &= \sum_{j=1}^n \lambda_j c_j^2 Q_k^2(\lambda_j) \\ &\le \min_{\substack{Q_k\in\mathscr{P}_k \\ Q_k(0)=1}} \max_\lambda |Q_k(\lambda)|^2 \underbrace{\sum_{j=1}^n \lambda_j c_j^2}_{=\|\tilde{x}-x^0\|_H^2}. \end{aligned}$$

Wählt man als Polynome die Tschebyschew-Polynome vom Grad $\le k$ so erhält man nach Transformation des Definitionsbereichs auf $[\lambda_1,\lambda_n]$ die Abschätzung

$$\alpha := \min_{\substack{Q_k\in\mathscr{P}_k \\ Q_k(0)=1}} \max_{1\le i\le n} |Q_k(\lambda_i)| \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$$

mit $\kappa = \kappa(H) = \frac{\lambda_n}{\lambda_1}$. □

4.8.3 Vorkonditionierung

Der letzte Satz besagt, dass für das CG-Verfahren gute Konvergenz zu erwarten ist, falls die Kondition der Matrix klein ist. Die Idee der Vorkonditionierung besteht in einer Modifikation des Problems, so dass die Kondition der modifizierten Systemmatrix klein ist. Die Niveaulinien von f sollen so gut wie möglich Kreise approximieren.

Im folgenden wählen wir eine **positiv definite und symmetrische Matrix** B, substituieren $x = B\bar{x}$ und erhalten mit $\bar{H} = HB$ das Problem

$$\bar{H}\bar{x} = -b.$$

$\bar{H} = HB$ ist i.a. nicht selbstadjungiert bzgl. des euklidischen Skalarprodukts aber bzgl. $(.,.)_B =$ denn

$$\langle x, HBy\rangle_B = \langle x, BHBy\rangle = \langle HBx, By\rangle = \langle HBx, y\rangle_B.$$

Die wesentliche Idee des vorkonditionierten CG - Verfahrens besteht in der Lösung des obigen Ersatzproblems, wobei das euklidische Skalarprodukt $(.,.)$ durch $(.,.)_B$ ersetzt wird. Details finden sich z. B. im Buch von Deuflhard und Hohmann.[1]

Für den Approximationsfehler hat man

$$\|\tilde{x} - x^k\|_H \leq 2\left(\frac{\sqrt{\kappa(H\cdot B)} - 1}{\sqrt{\kappa(H\cdot B)} + 1}\right)^k \|\tilde{x} - x^0\|_H .$$

Eine gute Vorkonditionierung ermittelt eine positiv definite symmetrische Matrix B, für die einerseits die Produkte By einfach auszuwerten sind und andererseits die Kondition $\kappa(HB)$ „klein“ ist.

Beispiele sind

- $B = D^{-1}$, wobei $D = diag(H)$ die Diagonalmatrix mit den Diagonalelementen von H ist.
- Unvollständige Cholesky-Zerlegung von H.

4.8.4 CG-Verfahren für nichtlineare Optimierungsprobleme

Das Verfahren wurde zuerst von Fletcher und Reeves untersucht, deshalb wird es auch als **Fletcher-Reeves-Verfahren** bezeichnet.

Es verläuft völlig analog zum letzten Verfahren, wobei g^k als Gradient $\nabla f(x^k)$ der gegebenen nichtlinearen Funktion f zu wählen ist und nicht mehr die einfache Form $Hx^k + b$ hat.

Unter der (theoretischen) Annahme der Verwendung *exakter* Schrittweiten $\sigma_k = \sigma_E$ kann man Konvergenz zeigen (vgl. z.B. [1, Satz 4.9.4]).

[1] Deuflhard/Hohmann: *Numerische Mathematik 1*. de Gruyter, Berlin 1993.

5 Probleme mit linearen Restriktionen – Theorie

5.1 Ein Beispiel

Wir nehmen als Beispiel ein Lagerhaltungsproblem

- Eine Firma verkauft (zunächst vereinfacht) ein Produkt.
- Verkauf und Lagerbestand werden zu diskreten Zeitpunkten $t_0 < t_1 < \ldots < t_N$ beobachtet.
- Betrieb eines Lagers, Belieferung am Anfang jeder Periode $[t_i, t_{i+1}]$

Ziel: Steuerung der Lagerhaltung des Produkts, um minimale Gesamtkosten zu haben.

Größen: z_i Lagerbestand bei t_i vor Neulieferung
r_i Nachfrage nach Produkt in $[t_i, t_{i+1}]$
u_i Liefermenge zum Zeitpunkt t_i
$z_0 = a \geq 0$ ist vorgegeben (Anfangsbestand)

Lagerbilanzgleichungen:

$$z_{i+1} = z_i - r_i + u_i \quad i = 0, \ldots, N-1.$$

Gegeben: $r_0, \ldots, r_N$.

Gesucht: $z = (z_1, \ldots, z_N)^\top$, $u = (u_0, \ldots, u_{N-1})^\top$, $x := \binom{z}{u}$.

Kosten: $f_i(z_i, u_i)$ (Liefern, Einkaufen, Lagern)

Zielfunktion:

$$f(z,u) \quad = \quad \rho z_N^2 + \sum_{i=0}^{N-1} f_i(z_i, u_i)$$

Der Endbestand wird mit einem Faktor ρ gewichtet.

Am Ende ergibt sich folgendes Problem:

$$\text{(LH1)} \left\{ \begin{array}{lll} \min f(z,u) := & \rho z_N^2 + \sum\limits_{i=0}^{N-1} f_i(z_i, u_i) & \\ \text{bei} & & \\ 0 \leq u_i \leq b_i & i = 0, \ldots, N-1 & \text{Kapazitätsschranken} \\ z_{i+1} = z_i - r_i + u_i & i = 0, \ldots, N-1 & \text{Lagerbilanz} \\ z_i \geq 0 & i = 1, \ldots, N & \text{Nichtnegativität} \end{array} \right.$$

Dabei ist $z_0 = a$ vorgegeben.

Das Problem hat wieder die allgemeine Form

$$\text{(P)} \quad \min f(x), \quad x \in \mathscr{F}$$

mit der zulässigen Menge

$$\mathscr{F} = \left\{ x = \binom{z}{u} \,|\, z \geq 0,\, 0 \leq u \leq b,\, z_{i+1} = z_i - r_i + u_i,\, i = 1,\ldots,N-1,\, z_0 = a. \right\}$$

Verallgemeinerung: Bisher waren z_i, u_i reelle Variablen. Bei mehreren Produkten können das Vektoren sein. Das ergibt schließlich die Aufgabe

$$\begin{array}{lll} \text{(LH2)} & \min & f(z,u) = \rho \,\|\, z_N \,\|^2 + \sum\limits_{i=0}^{N-1} f_i(z_i,u_i) \\ & & 0 \leq u_i \leq b_i \\ & & z_{i+1} = A_i z_i + B_i u_i - r_i \\ & & z_i \geq 0 \end{array}$$

mit $z_i \in \mathbb{R}^n, u_i \in \mathbb{R}^n, r_i \in \mathbb{R}^n$ und entsprechenden Matrizen A_i, B_i.

5.2 Optimalitätsbedingungen erster Ordnung

Wir brauchen zunächst einige Grundlagen und Hilfsmittel der konvexen Analysis.

Definition 5.2.1 (Kegel) *Eine nichtleere Teilmenge $K \subset \mathbb{R}^n$ heißt Kegel, wenn*

$$x \in K \Rightarrow \alpha x \in K \quad \forall \alpha > 0.$$

Beispiel 5.2.1 *Kegel sind*

- $\{x \in \mathbb{R}^n \,|\, x_i > 0 \quad \forall_i\}$
  Würden wir in der obigen Definition $\alpha = 0$ zulassen, dann wäre dies kein Kegel!
- $\{x \in \mathbb{R}^n \,|\, x_i \geq 0 \quad \forall_i\}$ **(Nichtnegativer Orthant)**
- $\{x \in \mathbb{R}^2 \,|\, x_1 \geq 0 \wedge x_2 = 0 \;\; \textit{oder} \;\; x_1 = 0 \wedge x_2 \geq 0\}$

Wir verwenden im Weiteren folgende Schreibweise:

$$x \geq 0 \Leftrightarrow x_i \geq 0 \quad \forall_i = 1,\ldots,n.$$

Für Mengen $A, B \subset \mathbb{R}^n$ und $\alpha, \beta \in \mathbb{R}$

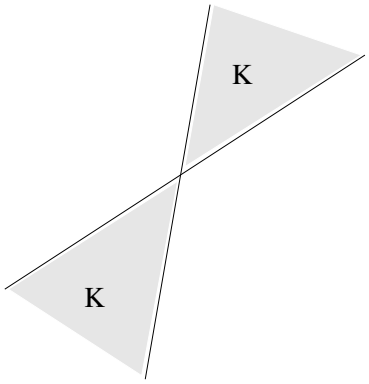
$$\alpha A + \beta B := \{\alpha a + \beta B \,|\, a \in A, b \in B\}.$$

Lemma 5.2.1 (Konvexitätskriterium für Kegel) *Ein Kegel $K \subset \mathbb{R}^n$ ist genau dann konvex, wenn*

$$K + K \subset K$$

Beweis: Ü.A.

Beispiel 5.2.2 *Ein nichtkonvexer Kegel aus $\mathbb{R}^2$ mit $K+K \not\subset K$:*



Nach dieser allgemeinen Kegelei kommen nun die Kegel, welche für die Optimierungstheorie von ausschlaggebender Bedeutung sind:

Definition 5.2.2 (Konische Hülle) *Es sei $S \subset \mathbb{R}^n$ und $x \in S$ fest. Dann heißt*

$$K(S,x) = \{\alpha(s-x) \,|\, s \in S, \alpha > 0\}$$

der von $S-\{x\}$ **erzeugte Kegel** *oder* **konische Hülle** *von $S-\{x\}$.*

Andere Schreibweise:

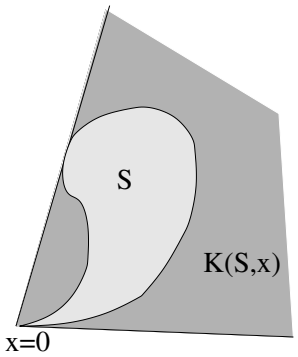
$$K(S,x) = \bigcup_{\alpha>0} \alpha(S-\{x\}).$$

Lemma 5.2.2 *Ist $C \subset \mathbb{R}^n$ konvex sowie $x \in C$, dann ist auch $K(C,x)$ konvex.*

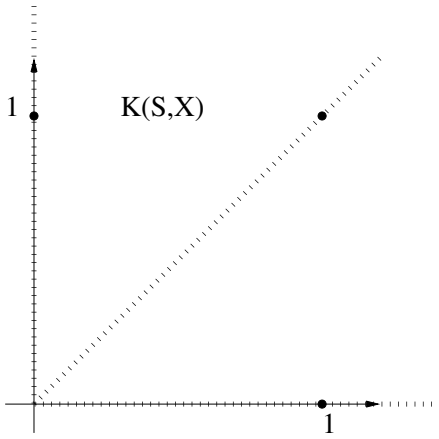
Beispiel 5.2.3

Die folgenden Bilder sind in der Regel nur korrekt, wenn x der Nullpunkt ist. Alle eingezeichneten Kegel $K(S,x)$ muss man sich ansonsten mit der Spitze in den Nullpunkt verschoben vorstellen.

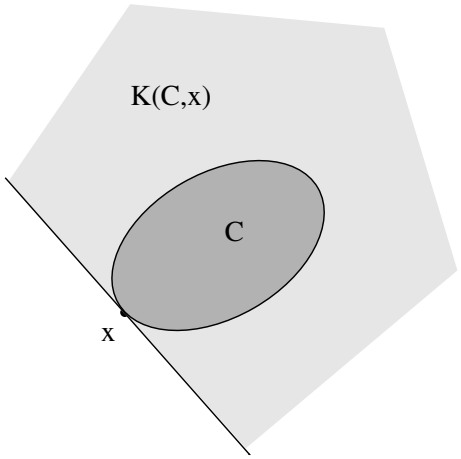
a)



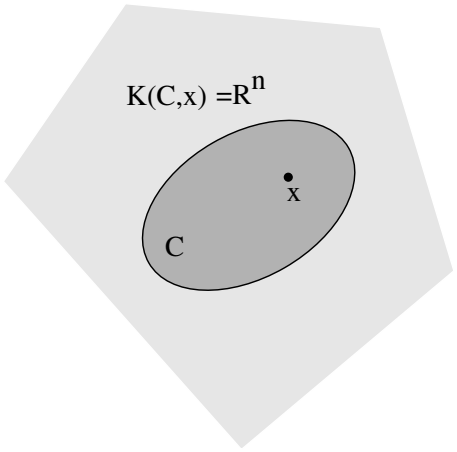
b) $S = \{(0,0)^\top, (1,0)^\top, (0,1)^\top, (1,1)^\top\}$*; das Bild zeigt den Kegel $K(S,0)$.*



c) Beachte hierzu auch e)

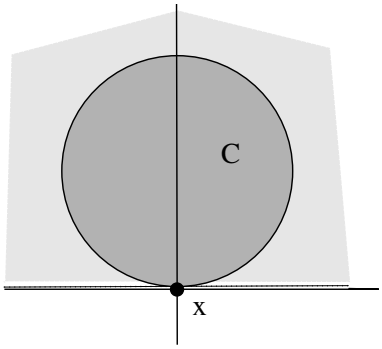


d)



e) Auch bei abgeschlossener Menge C muss $K(C,x)$ nicht abgeschlossen sein:
Es sei $C \subset \mathbb{R}^2$ die abgeschlossene Einheitskugel um $(0,1)^\top$ und $x = (0,0)^\top$. Dann

$$K(C,x) = \left\{ y \in \mathbb{R}^2 \,|\, y_2 > 0 \right\} \cup \{0\}.$$



f) Es sei $C = \mathbb{R}^n_+ = \{x \,|\, x \geq 0\}, \quad x \in C.$

$$K(C,x) = \{d \in \mathbb{R}^n \,|\, d_i \geq 0, \text{ wenn } x_i = 0\}$$

g) Es sei A eine $(m,n)-$*Matrix,* $b \in \mathbb{R}^m$ *fest gegeben und*

$$C = \{x \,|\, Ax = b\}$$

und $x \in C$*. Hier gilt*

$$K(C,x) = \{y \,|\, Ay = 0\}.$$

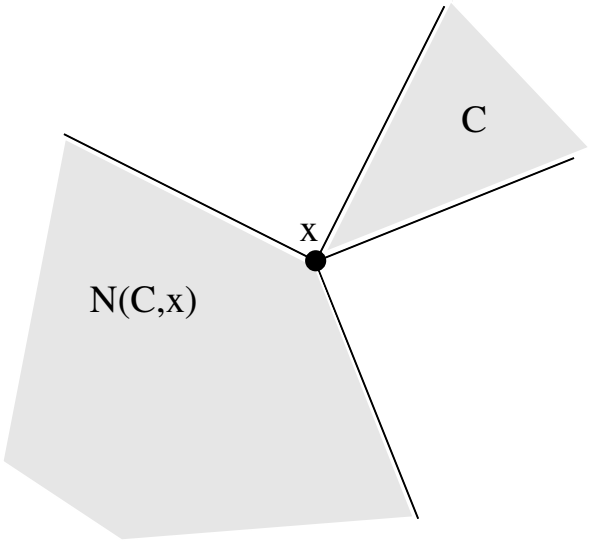
Definition 5.2.3 (Normalenkegel) *Sei* $C \subset \mathbb{R}^n$ *konvex,* $x \in C$*. Ein Vektor* $s \in \mathbb{R}^n$ *heißt Normalenrichtung von C in x, wenn*

$$\langle s, y - x \rangle \leq 0 \quad \forall y \in C.$$

Die Menge

$$N(C,x) = \{s \,|\, s \text{ ist Normalenrichtung}\}$$

heißt Normalenkegel *von C in x.*



Bemerkung:

- $N(C,x)$ ist stets abgeschlossen.
- $x \in int\, C \Rightarrow N(C,x) = \{0\}$

Definition 5.2.4 (Dualkegel) *$K \subset \mathbb{R}^n$ sei konvexer Kegel. Dann heißt*

$$K^* = \{s \in \mathbb{R}^n \,|\, \langle s,x\rangle \leq 0 \quad \forall x \in K\}$$

Dual- oder Polarkegel *zu K*.

Bemerkungen:

- Wenn $0 \in K$, dann gilt $K^* = N(K,0)$ (siehe oben)
- K_1, K_2 konvexe Kegel mit $K_1 \subset K_2 \Rightarrow K_1^* \supset K_2^*$
- K^* ist *immer* konvex und abgeschlossen

Satz 5.2.1 *Ist $C \subset \mathbb{R}^n$ konvex und $x \in C$, so gilt*

$$N(C,x) = K(C,x)^*.$$

Beweis: (i) $''\subset''$: Sei $s \in N(C,x)$, d. h. $\langle s, y-x\rangle \leq 0 \quad \forall y \in C$.

$$\begin{aligned} \Rightarrow \langle s, \alpha(y-x)\rangle &\leq 0 \quad \forall y \in C, \alpha > 0 \\ \Rightarrow \langle s,z\rangle &\leq 0 \quad \forall z \in K(C,x) \end{aligned}$$

und damit $s \in K(C,x)^*$.

(ii) $''\supset''$: Sei $s \in K(C,x)^*$, d. h. die letzte Ungleichung gilt. Sie gilt insbesondere für jedes $y \in C$ $z = 1 \cdot (y-x)$, also $s \in N(C,x)$. □

Wir kehren zurück zu unserem alten Problem

$$\min_{x \in \mathscr{F}} f(x). \tag{P}$$

Wir wissen bereits: Ist $\mathscr{F}$ nichtleer und konvex, f in $\tilde{x} \in \mathscr{F}$ differenzierbar und $\tilde{x}$ ein lokales Minimum von (P), dann gilt die Variationsungleichung

$$\nabla f(\tilde{x}) \cdot (x - \tilde{x}) \geq 0 \quad \forall x \in \mathscr{F}. \tag{5.51}$$

Umformulierung der Variationsungleichung:

Setze $s = \nabla f(\tilde{x})$. Dann gilt $\langle s, x-\tilde{x}\rangle \geq 0 \quad \forall x \in \mathscr{F}$, also $\langle -s, x-\tilde{x}\rangle \leq 0$. Damit

$$-\nabla f(\tilde{x}) \in N(\mathscr{F}, \tilde{x}) \tag{5.52}$$

oder

$$\boxed{0 \in \nabla f(\tilde{x}) + N(\mathscr{F}, \tilde{x}).} \tag{5.53}$$

Das ist eine Verallgemeinerung von $0 = \nabla f(\tilde{x})$.

Eine alternative Formulierung ist

$$\nabla f(\tilde{x}) \in -K(\mathscr{F}, \tilde{x})^*.$$

Ist f auch konvex, so ist (P) eine konvexe Optimierungsaufgabe. Dann ist (5.53) auch *hinreichend* für Optimalität.

Beispiel 5.2.4 (Nichtnegativitätsrestriktionen) *Wir betrachten das Problem*

$$\boxed{\min_{x\in\mathbb{R}^n} f(x), \quad x \geq 0.} \tag{PV}$$

Eine lokale Lösung sei $\tilde{x} \in \mathscr{F} = \mathbb{R}^+$.

$$\begin{aligned} \Rightarrow \quad K(\mathscr{F},\tilde{x}) = & \left\{ d \in \mathbb{R}^n \,|\, d_i \geq 0, \ \textit{falls} \ \tilde{x}_i = 0 \right\} \\ \Rightarrow \quad N(\mathscr{F},\tilde{x}) = & \left\{ s \in \mathbb{R}^n \,|\, s_i \leq 0, \textit{wenn} \quad \tilde{x}_i = 0 \ \textit{und} \ s_i = 0, \textit{wenn} \ \tilde{x}_i > 0 \right\}. \end{aligned}$$

Aus (5.52) *folgt also*

$$\begin{aligned} -\nabla f(\tilde{x})_i \leq 0, \quad \textit{falls} \quad \tilde{x}_i = 0 \\ \nabla f(\tilde{x})_i = 0, \quad \textit{falls} \quad \tilde{x}_i > 0. \end{aligned}$$

Alternativer Weg:

Das hätten wir aber auch ohne Verwendung von $N(\mathscr{F},\tilde{x})$ direkt aus der Variationsungleichung herleiten können:

$$\nabla f(\tilde{x})^\top (x - \tilde{x}) \geq 0 \Leftrightarrow \sum_{i=1}^{n} \nabla f(\tilde{x})_i (x_i - \tilde{x}_i) \geq 0$$

und wegen Unabhängigkeit der Komponenten

$$\nabla f(\tilde{x})_i (x_i - \tilde{x}_i) \geq 0 \quad \forall x_i \geq 0 \quad \forall i$$

d. h.

$$\nabla f(\tilde{x})_i \, \tilde{x}_i \leq \nabla f(\tilde{x})_i \, x \quad \forall x \geq 0$$

Damit muss $\tilde{x}_i$ das Minimum der rechten Seite unter allen $x \geq 0$ annehmen. Deshalb

$$\begin{aligned} \nabla f(\tilde{x})_i \geq 0, \quad \text{falls} \quad \tilde{x}_i = 0 \\ \nabla f(\tilde{x})_i = 0, \quad \text{falls} \quad \tilde{x}_i > 0. \end{aligned}$$

Diese Beziehungen können auch als Abbruchkriterium verwendet werden. Es folgt übrigens auch

$$\nabla f(\tilde{x})_i > 0 \ \Rightarrow \ \tilde{x}_i = 0.$$

Beispiel 5.2.5 (Lineare Gleichungsrestriktionen)

Wir betrachten das Problem

$$\boxed{\min_{x\in\mathbb{R}^n} f(x), \quad Ax = b} \tag{PLG}$$

mit einer (m,n)*-Matrix A; also* $\mathscr{F} = \{x \,|\, Ax = b\}$. *Wir wissen hier bereits*

$$K(\mathscr{F},\tilde{x}) = \{x \,|\, Ax = 0\} = \ \ker A =: U$$

mit dem linearen Unterraum $U = \ker A$. Offenbar gilt dann

$$K(\mathscr{F},\tilde{x})^* = N(\mathscr{F},\tilde{x}) = U^\perp,$$

überlegen Sie sich das bitte selbst!

$$U = \ker A \quad \Rightarrow \quad U^\perp = \operatorname{im} A^\top = \{A^\top \lambda \,|\, \lambda \in \mathbb{R}^m\}$$

Variationsungleichung (5.53) $\Rightarrow -\nabla f(\tilde{x}) \in N(\mathscr{F},\tilde{x}) \Rightarrow -\nabla f(\tilde{x}) \in \operatorname{im} A^\top$. *Folglich gilt*

$$-\nabla f(\tilde{x}) = A^\top \lambda$$

also

$$\boxed{\nabla f(\tilde{x}) + A^\top \lambda = 0.}$$

Das ist die klassische Lagrangesche Multiplikatorenregel mit Lagrange-Multiplikator $\lambda \in \mathbb{R}^m$. Diese wird später noch allgemeiner diskutiert.

5.3 Optimalitätsbedingungen zweiter Ordnung

5.3.1 Notwendige Bedingungen

Wir beweisen nun ein Analogon zum entsprechenden Satz für unrestringierte Aufgaben.

Satz 5.3.1 (Notwendige Bedingung zweiter Ordnung) *Es sei $\tilde{x} \in \mathscr{F}$ lokales Minimum von (P) und $\mathscr{F}$ konvex. Dann gilt neben der notwendigen Bedingung 1. Ordnung auch die notwendige Bedingung 2. Ordnung*

$$(x-\tilde{x})^\top f''(\tilde{x})(x-\tilde{x}) \geq 0 \quad \forall x \in \mathscr{F} \;\; mit \;\; \nabla f(\tilde{x})^\top (x-\tilde{x}) = 0. \tag{5.54}$$

Beweis: Wir setzen $d = x - \tilde{x}$ und fordern $\nabla f(\tilde{x})^\top d = 0$. Es gilt $\tilde{x} + td \in \mathscr{F}$ für alle $t \in [0,1]$ und

$$f(\tilde{x}+td) - f(\tilde{x}) \geq 0$$

für alle hinreichend kleinen $t > 0$. Mit einer Taylorentwicklung folgt

$$\begin{aligned} 0 &\leq \underbrace{\nabla f(\tilde{x})^\top td}_{=0} + \tfrac{1}{2}t^2 d^\top f''(\tilde{x})d + o(t^2) \quad | : \tfrac{1}{2}t^2 \\ 0 &\leq d^\top f''(\tilde{x})d + 2\underbrace{\frac{o(t^2)}{t^2}}_{\to 0,\, t\downarrow 0}. \end{aligned}$$

Grenzübergang $t \downarrow 0$ ergibt die Behauptung. □

Bemerkung: Man kann offenbar (5.54) auch so formulieren:

$$d^\top f''(\tilde{x})d \geq 0 \quad \forall d \in K(\mathscr{F},\tilde{x}) \text{ mit } \nabla f(\tilde{x})^\top d = 0. \tag{5.55}$$

(man multipliziere die quadratische Form in (5.54) mit α^2 durch).

Beispiel 5.3.1 (Gleichungsnebenbedingungen) *Hier wissen wir* $K(\mathscr{F},\tilde{x}) = \ker A$ *und deshalb*

$$d^\top f''(\tilde{x})d \geq 0 \quad \forall d \in U = \ker A \text{ mit } \nabla f(\tilde{x})^\top d = 0. \tag{5.56}$$

Außerdem haben wir

$$-\nabla f \in N(\mathscr{F},\tilde{x}) = U^\perp,$$

also gilt automatisch $\nabla f(\tilde{x})^\top d = 0$ *und somit*

$$\boxed{d^\top f''(\tilde{x})d \geq 0 \qquad \text{für alle } d \text{ mit } Ad = 0.}$$

$f''(\tilde{x})$ *muss auf dem linearen Unterraum* $U = \ker A$ *positiv semidefinit sein.*

5.3.2 Hinreichende Bedingungen

Satz 5.3.2 (Hinreichende Bedingung zweiter Ordnung) *Die Funktion* f *sei in* $\tilde{x} \in \mathscr{F}$ *zweimal stetig differenzierbar und die notwendige Bedingung 1. Ordnung sei erfüllt. Zusätzlich gelte mit einem* $\alpha > 0$

$$d^\top f''(\tilde{x})d \geq \alpha \,\|\, d \,\|^2 \tag{5.57}$$

für alle $d \in cl\, K(\mathscr{F},\tilde{x})$ *mit* $\nabla f(\tilde{x})^\top d = 0$*. Dann existiert zu jedem* $\beta \in (0,\alpha)$ *ein* $\rho > 0$*, so dass die* quadratische Wachstumsbedingung

$$f(x) \geq f(\tilde{x}) + \frac{\beta}{2}\|x-\tilde{x}\|^2 \quad \forall\, x \in \mathscr{F} \cap B(\tilde{x},\rho),$$

erfüllt ist. Damit ist $\tilde{x}$ *strenges lokales Minimum.*

Beweis (Indirekt): Die Behauptung sei falsch. Dann existiert ein $\beta \in (0,\alpha)$ und eine Folge $\{x^i\} \subset \mathscr{F}$ mit $x^i \to \tilde{x}$, $x^i \neq \tilde{x}\ \forall i \in \mathbb{N}$ und

$$f(\tilde{x}) + \frac{\beta}{2}\|x^i-\tilde{x}\|^2 > f(x^i) \quad \forall i \in \mathbb{N} \tag{5.58}$$

(man wähle $\rho_i = 1/i$). Wir zeigen, dass daraus $\beta \geq \alpha$ folgt, also ein Widerspruch.

- Die Elemente $d^i = \frac{x^i-\tilde{x}}{\|x^i-\tilde{x}\|}$ gehören zur (kompakten) Einheitskugeloberfläche. Daher existiert eine Teilfolge – o.B.d.A. sei das $\{d^i\}$ selbst – mit $d^i \to d$, $i \to \infty$. Dann gilt $\|d\| = 1$.
- Außerdem ist die Variationsungleichung für $\tilde{x}$ erfüllt, also

  $$\langle \nabla f(\tilde{x}), x^i - \tilde{x}\rangle \geq 0\,, \quad \text{daher } \langle \nabla f(\tilde{x}), d^i\rangle \geq 0$$

  und deshalb

  $$\nabla f(\tilde{x})^\top d \geq 0\,.$$

Ferner gilt $d \in cl\, K(\mathscr{F},\tilde{x})$, da $x^i - \tilde{x} \in K \;\; \forall i$.

Andererseits muss die Ungleichung

$$\nabla f(\tilde{x})^\top d \leq 0$$

gelten, denn mit Taylorentwicklung der rechten Seite von (5.58) folgt:

$$\begin{aligned} f(\tilde{x}) + \frac{\beta}{2}\|x^i - \tilde{x}\|^2 &\geq f(x^i) \\ &= f(\tilde{x}) + \nabla f(\tilde{x})^\top (x^i - \tilde{x}) + r_1(x^i - \tilde{x})\,. \end{aligned}$$

Wir teilen durch $\|x^i - \tilde{x}\|$, vorher streichen wir $f(\tilde{x})$ auf beiden Seiten. Dann

$$\frac{\beta}{2}\underbrace{\|x^i - \tilde{x}\|}_{\to 0} \geq \nabla f(\tilde{x})^\top \underbrace{\frac{x^i - \tilde{x}}{\|x^i - \tilde{x}\|}}_{\to d} + \underbrace{\frac{r_1(x^i - \tilde{x})}{\|x^i - \tilde{x}\|}}_{\to 0,\, x^i \to \tilde{x}}\,.$$

Daraus folgt $\nabla f(\tilde{x})^\top d \leq 0$ und somit

$$\nabla f(\tilde{x})^\top d = 0\,. \tag{5.59}$$

- Nun gehen wir nochmals in (5.58), entwickeln aber bis Ordnung 2,

$$\begin{aligned} f(\tilde{x}) + \frac{\beta}{2}\|x^i - \tilde{x}\|^2 &\geq f(x^i) \\ &= f(\tilde{x}) + \underbrace{\nabla f(\tilde{x})^\top (x^i - \tilde{x})}_{\geq 0,\ \text{Variations-ungleichung}} + \frac{1}{2}(x^i - \tilde{x})^\top f''(\tilde{x})(x^i - \tilde{x}) + r_2(x^i - \tilde{x}). \end{aligned}$$

Wir teilen durch $\|x^i - \tilde{x}\|^2$,

$$\frac{\beta}{2}\|d^i\|^2 \geq \frac{1}{2}(d^i)^\top f''(\tilde{x}) d^i + \underbrace{\frac{r_2(x^i - \tilde{x})}{\|x^i - \tilde{x}\|^2}}_{\to 0}$$

$i \to \infty \Rightarrow$

$$\frac{\beta}{2}\|d\|^2 \geq \frac{1}{2} d^\top f''(\tilde{x}) d \geq \frac{\alpha}{2}\|d\|^2$$

nach Voraussetzung der hinreichenden Bedingung, denn wir wissen bereits $d \in clK(\mathscr{F},\tilde{x})$ und $\nabla f(\tilde{x})^\top d = 0$ wegen (5.59). Außerdem gilt $\|d\| = 1$

$$\Rightarrow \beta \geq \alpha$$

ein Widerspruch! □

Beispiel 5.3.2 *Wir betrachten die Aufgabe*

$$\min f(x) = -(x_1x_2 + x_2x_3 + x_1x_3)$$

bei

$$x_1 + x_2 + x_3 - 3 = 0$$

Wir haben

$$\nabla f = -\begin{pmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_2 + x_1 \end{pmatrix}$$

Notwendige Bedingung:

$$\nabla f^\top d = 0 \quad \forall d \text{ mit } d_1 + d_2 + d_3 = 0.$$

Wir zeigen: $\tilde{x} = (1,1,1)^\top$ *ist lokales Minimum:*

- $\nabla f(1,1,1)^\top d = -2(1,1,1)\begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = -2(d_1 + d_2 + d_3) = 0$ *falls* $d_1 + d_2 + d_3 = 0$,

  die notwendige Bedingung ist erfüllt.

$$f''(1,1,1) = -\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

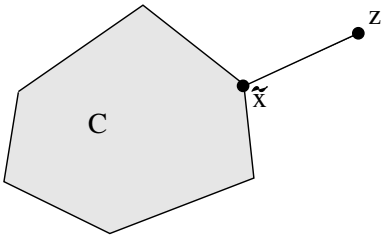
- f'' *ist offenbar* **nicht positiv definit**, *denn* $\det f'' = -2$. *Dennoch ist die hinreichende Bedingung 2. Ordnung erfüllt,*

$$f''(\tilde{x})\begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = -\begin{pmatrix} d_2 + d_3 \\ d_1 + d_3 \\ d_1 + d_2 \end{pmatrix} = -\underbrace{\begin{pmatrix} d_1 + d_2 + d_3 \\ d_1 + d_2 + d_3 \\ d_1 + d_2 + d_3 \end{pmatrix}}_{= 0 \quad \text{falls } d_1 + d_2 + d_3 = 0} + \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = d$$

  $\Rightarrow$

$$d^\top f''(\tilde{x}) d = d^\top d = \|d\|^2 \quad \text{falls } d_1 + d_2 + d_3 = 0.$$

  Positive Definitheit auf Unterraum $\Rightarrow \tilde{x}$ *ist lokales Minimum (vgl. auch den bewiesenen Satz über hinreichende Bedingungen bei Gleichungsrestriktionen).*



Beispiel 5.3.3 (Projektion auf eine konvexe Menge) *Sei $C \subset \mathbb{R}^n$ nicht leer, konvex und abgeschlossen und $y \notin C$ fest gewählt. Wir suchen das Element $\tilde{x} \in C$ kleinsten Abstands zu y, betrachten also die Aufgabe*

$$\boxed{\min_{x \in C} f(x) = \|x-y\|^2 .}$$

Offenbar existiert eine Lösung der Aufgabe, auch bei nicht kompaktem C, denn die Zielfunktion strebt gegen Unendlich, falls $\|x\|$ unbeschränkt wächst.

Es gilt

$$\begin{aligned} f(x) &= \langle x-y, x-y\rangle \\ \nabla f(x)^\top d &= 2\langle x-y, d\rangle \\ (d^1)^\top f''(x)(d^2) &= 2\langle d^1, d^2\rangle \end{aligned}$$

- *Notwendige (und wegen Konvexität auch hinreichende) Bedingung 1. Ordnung für Lösung $\tilde{x}$:*

$$\langle \tilde{x}-y, x-\tilde{x}\rangle \geq 0 \quad \forall x \in C \tag{5.60}$$

  (Charakterisierung der Projektion auf eine konvexe Menge)

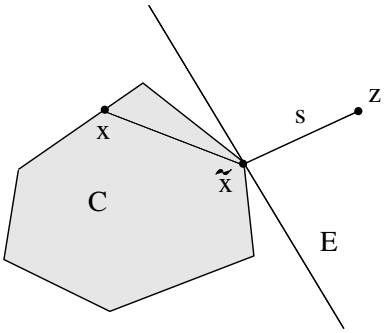
- *Hinreichende Bedingung 2. Ordnung: Wir erhalten ganz einfach*

$$d^\top f''(\tilde{x}) d = 2\langle d, d\rangle = 2\|d\|^2 .$$

  Positive Definitheit auf dem ganzen Raum!

Nach unserem letzten Satz ist damit jedes $\tilde{x}$, welches (5.60) *erfüllt, lokales* **strenges** *Minimum. Wegen Konvexität ist es sogar das globale.*

Als wichtige Grundlage für die Optimierung beweisen wir damit einen Trennungssatz.



Satz 5.3.3 (Trennungssatz) *Es sei $C \subset \mathbb{R}^n$ nichtleer, abgeschlossen und konvex und $z \notin C$. Dann gibt es eine Hyperebene, die C und $\{z\}$ strikt trennt, d. h., es existiert ein $s \neq 0 \in R^n$ mit*

$$\langle s, z\rangle > \sup_{x \in C} \langle s, x\rangle . \tag{5.61}$$

Bevor wir (5.61) beweisen, schreiben wir die Aussage etwas eingängiger auf:

$$\langle s,z\rangle \geq c + \langle s,x\rangle \quad \forall x \in C$$

mit einem gewissen $c > 0$, also

$$\boxed{\langle s, z-x\rangle \geq c \quad \forall x \in C\,.}$$

Damit trennt die Hyperebene $E = \{x | \langle s, z-x\rangle = c\}$ die Menge C und $\{z\}$.

Beweis: Geometrisch motiviert setzen wir $s = z - \tilde{x}$. Dann

$$\begin{aligned}
\langle \tilde{x} - z, x - \tilde{x}\rangle &\geq 0 \quad \forall x \in C \\
\langle z - \tilde{x}, x - \tilde{x}\rangle &\leq 0 \quad \forall x \in C \\
\langle \underbrace{z-\tilde{x}}_{s}, x - z + \underbrace{z-\tilde{x}}_{s}\rangle &\leq 0 \quad \forall x \in C \\
\langle s, x-z\rangle + \underbrace{\|s\|^2}_{C} &\leq 0
\end{aligned}$$

Das ist äquivalent zur Aussage des Satzes. □

5.4 Gleichungsnebenbedingungen

Wir untersuchen nun noch einmal detaillierter

$$\boxed{\min f(x)\,,\; Ax = b} \qquad \text{(PLG)}$$

mit einer (m,n)-Matrix A.

Hier ist $\mathscr{F} = \{x | Ax = b\}$ konvex und abgeschlossen.

5.4.1 Optimalitätsbedingungen erster Ordnung

Satz 5.4.1 (Multiplikatorenregel für Gleichungsrestriktionen) *Ist $\tilde{x}$ lokales Minimum von (PLG) und f an der Stelle $\tilde{x}$ differenzierbar, dann existiert ein $\lambda \in \mathbb{R}^m$, so dass*

$$\nabla f(\tilde{x}) + A^\top \lambda = 0\,. \tag{5.62}$$

Hat A vollen Rang, dann ist λ eindeutig bestimmt.

Beweis: (5.62) haben wir schon bewiesen (Beispiel 5.2.5).
Eindeutigkeit von λ: (5.62) heißt

$$\lambda_1 a^1 + \ldots + \lambda_m a^m = -\nabla f\,,$$

wobei $(a^i)^\top$ die Zeilenvektoren von A sind, also die Spalten von $A^\top$. Hat A vollen Rang, so ist λ eindeutig bestimmt. □

In dieser Form muss man sich sich den Satz aber nicht unbedingt einprägen. Sehr prägnant werden alle unsere weiteren Optimalitätsbedingungen durch Einführung einer *Lagrange-Funktion*.

Definition 5.4.1 (Lagrange-Funktion) $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$L(x, \lambda) := f(x) + \langle \lambda , Ax - b \rangle .$$

Der Vektor λ heißt **Lagrangescher Multiplikator** oder *Vektor der Lagrangeschen Multiplikatoren.*

Dann ergibt sich die Bedingung des letzten Satzes ganz einfach als

$$\nabla_x L(\tilde{x}, \lambda) = 0$$

denn

$$\nabla_x L(\tilde{x}, \lambda) = \nabla f(\tilde{x}) + A^\top \lambda .$$

Aber zu den notwendigen Bedingungen gehört natürlich auch die Nebenbedingung $Ax = b$ selbst. Diese bekommt man, wie man sofort sieht, aus

$$\nabla_\lambda L(\tilde{x}, \lambda) = 0 .$$

Insgesamt erhalten wir folgendes **Optimalitätssystem**:

$$\boxed{\begin{array}{ll} \nabla f(x) + A^\top \lambda = 0 \\ Ax - b \qquad\quad = 0 \end{array}} \quad \text{oder} \quad \begin{array}{ll} \nabla_x L(x, \lambda) = 0 \\ \nabla_\lambda L(x, \lambda) = 0. \end{array} \tag{5.63}$$

Definition 5.4.2 *Jedes $x \in \mathbb{R}^n$, welches mit einem $\lambda \in \mathbb{R}^m$ das Optimalitätssystem (5.63) erfüllt, heißt* **stationärer Punkt** *unserer Optimierungsaufgabe.*

Nicht jeder stationäre Punkt ist eine (lokale) Lösung, aber es gilt:

Satz 5.4.2 *Ist f konvex, dann ist jeder stationäre Punkt $\tilde{x}$ globale Lösung von* (PLG)*, d. h., hier ist (5.63) nicht nur notwendige sondern auch hinreichende Optimalitätsbedingung.*

Das ist klar, denn die Aufgabe ist hier konvex.

5.4.2 Bedingungen zweiter Ordnung bei Gleichungsrestriktionen

Aus den allgemeinen Betrachtungen vorher erhalten wir als Spezialfall

Satz 5.4.3 *Sei $f \in C^2$ in $\tilde{x}$, die notwendige Bedingung (5.63) erfüllt und es gelte*

$$d^\top f''(\tilde{x}) d \geq \alpha \|d\|^2 \tag{5.64}$$

für alle $d \in R^n$ mit $Ad = 0$. Dann ist $\tilde{x}$ striktes lokales Minimum für (PLG).

Beispiel 5.4.1 *Wir schauen uns nochmals die folgende Aufgabe an:*

$$\min f(x) = -(x_1x_2 + x_2x_3 + x_1x_3)$$

bei

$$x_1 + x_2 + x_3 = 3.$$

• Lagrange-Funktion:

$$L(x,\lambda) = -(x_1x_2 + x_2x_3 + x_1x_3) + \lambda(x_1 + x_2 + x_3 - 3)$$

• *Notwendige Bedingung 1. Ordnung*

$$\nabla_x L = -\begin{pmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0$$

3 Gleichungen für 4 Unbekannte. Die vierte ist die Nebenbedingung.

• *Hinreichende Bedingung zweiter Ordnung:*

$$d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \quad \forall d : d_1 + d_2 + d_3 = 0$$

• *Das hatten wir bereits alles für* $\tilde{x} = (1,1,1)^\top$ *nachgewiesen.*

Bemerkungen:

(i) Die Bedingung (5.64) is äquivalent zu

$$\boxed{d^\top L_x''(\tilde{x},\lambda)d = 0 \quad \text{für alle } d \in \mathbb{R}^n \text{ mit } Ad = 0,}$$

denn hier gilt wegen Linearität der Nebenbedingungen

$$L_x'' = f''.$$

In dieser Form sollte man sich (5.64) merken. Die Lagrange-Funktion ist also das passende Werkzeug, Optimalitätsbedingungen prägnant und einprägsam darzustellen.

(ii) Anstelle von $Ad = 0$ können wir natürlich eleganter schreiben $d \in \ker A$.

5.4.3 Nullraum-Matrizen

Die folgenden Beobachtungen werden später für numerische Verfahren gebraucht. Es gilt

$$\mathbb{R}^n = \ker A \oplus (\ker A)^\perp,$$

d. h. es gibt eine eindeutige Zerlegung

$$x = u + v,$$

wobei u die Projektion von x auf $\ker A$ und v diejenige auf $(\ker A)^\top$ ist. Es gilt

$$u = Px = \underbrace{(I - A^\top (AA^\top)^{-1} A)x}_{} \tag{5.65}$$

falls A vollen Rang hat (Übungsaufgabe).

Wir fassen nun ein System von l Vektoren, die $\ker A$ aufspannen, zu einer (n,l)-Matrix Z zusammen. Dann gilt $im Z = \ker A$, d. h.

$$\boxed{d \in \ker A \Leftrightarrow d = Zz \quad \text{mit } z \in \mathbb{R}^l.}$$

Durchläuft z als Parameter ganz $\mathbb{R}^l$, so durchläuft d ganz $\ker A$.

Definition 5.4.3 *Die eben eingeführte Matrix Z heißt* **Nullraum-Matrix**.

Aus der linearen Algebra ist bekannt: Die allgemeine Lösung von $Ax = b$ setzt sich zusammen aus einer speziellen Lösung w und der allgemeinen Lösung des homogenen Systems. Damit

$$\mathscr{F} = \{x | Ax = b\} = \{w\} + \ker A = \{w\} + \mathrm{im}\, Z,$$
$$\mathscr{F} = \{w\} + \{Zz \,|\, z \in \mathbb{R}^l\}.$$

Durchläuft z ganz $\mathbb{R}^l$, so durchläuft $x = w + Zz$ ganz $\mathscr{F}$. Damit ist Problem (PLG) äquivalent zu

$$\boxed{\min_{z \in \mathbb{R}^l} F(z) = f(w + Zz).} \tag{5.66}$$

Das ist ein unrestringiertes, auf z **reduziertes Problem** mit $l \leq n$ Variablen. Damit verbundene numerische Verfahren heißen **Reduktionsverfahren (variable-reduction methods)**.

Beispiel 5.4.2 *Die Projektionsmatrix P von oben ist eine Nullraum-Matrix. Allerdings ist dort $l = n$ (P enthält z.B. I_n als Anteil).*

$$(PLG) \Leftrightarrow \min_{z \in \mathbb{R}^n} F(z) = f(w + Pz).$$

Beispiel 5.4.3 *A habe vollen Rang. Dann sind m Spalten linear unabhängig, o.B.d.A. die ersten m, und*

$$A = (A_1, A_2) \qquad \begin{array}{l} A_1 : m \times m \quad \textit{invertierbar} \\ A_2 : \ \textit{“Rest”} \end{array}$$
$$Ad = 0 \Leftrightarrow A_1 d^1 + A_2 d^2 = 0 \quad \textit{mit} \quad d = \begin{pmatrix} d^1 \\ d^2 \end{pmatrix} \begin{array}{l} \to \in \mathbb{R}^m \\ \to \in \mathbb{R}^{n-m} \end{array}.$$

Nun kann man nach d^1 auflösen, und nur noch d^2 spielt als jetzt freie Variable eine Rolle:

$$d^1 = -A_1^{-1} A_2 d^2 \qquad \textit{(Elimination)}$$

und deshalb offenbar

$$d \in \ker A \Leftrightarrow d \begin{pmatrix} -A_1^{-1} A_2 d^2 \\ d_2 \end{pmatrix}.$$

Insgesamt ist

$$Z = \begin{pmatrix} -A_1^{-1}A_2 \\ I_{n-m} \end{pmatrix}$$

damit Nullraummatrix. (PLG) *ist äquivalent mit*

$$\min_{z\in\mathbb{R}^{n-m}} F(z) := \min f\left(w + \begin{pmatrix} -A_1^{-1}A_2 \\ I_{n-m} \end{pmatrix} z\right)$$

ein Problem mit nur $l = n-m$ Variablen.

Berechnung von $\nabla\mathbf{F}(\mathbf{z})$: Aus der Kettenregel folgt

$$\begin{aligned} F'(z) &= f(w+Zz)' = f'(w+Zz)\circ Z \\ \Rightarrow \nabla F(z) &= Z^\top \cdot \nabla f(w+Zx) = Z^\top \nabla f(x) \quad \text{(reduzierter Gradient).} \end{aligned}$$

Berechnung von $\mathbf{F''(z)}$:

$$\begin{aligned} F'(z)h_1 =: \varphi(z) &= f'(w+Zz)\,Zh_1 = \langle f'(w+Zz)^\top, Zh_1\rangle \\ \Rightarrow (F''(z)h_1)h_2 = \varphi'(z)h_2 &= \langle f''(w+Zz)^\top Zh_2, Zh_1\rangle = h_2^\top Z^\top f''(w+Zz)\,Zh_1. \end{aligned}$$

Somit folgt

$$F''(z) = Z^\top f''(x) Z \quad \text{(reduzierte Hesse-Matrix).}$$

Wir stellen noch einmal den reduzierten Gradienten und die reduzierte Hesse-Matrix zusammen:

$$\boxed{\begin{array}{rcll} x & = & w+Zz & \\ F(z) & = & f(x+Zz) & \\ \nabla F(z) & = & Z^\top \nabla f(x) & \textbf{reduzierter Gradient} \\ F''(z) & = & Z^\top f''(x) Z & \textbf{reduzierte Hesse-Matrix} \end{array}}$$

Man kann nun die Optimalitätsbedingungen auch in reduzierter Form aufschreiben, nämlich:

Satz 5.4.4 *Es gilt mit* $\tilde{x} := w + Z\tilde{z}$

$$\boxed{\nabla f(\tilde{x}) + A^\top \lambda = 0 \quad \Leftrightarrow \quad \nabla F(\tilde{z}) = 0.}$$

Beweis: (i) $\Rightarrow$: Sei mit $\lambda \in \mathbb{R}^m$

$$\nabla f(\tilde{x}) = -A^\top \lambda\,.$$

Dann folgt wegen obiger Darstellung

$$\begin{aligned} \nabla F(\tilde{z}) &= Z^\top \nabla f(\tilde{x}) = -Z^\top A^\top \lambda \\ \Rightarrow \nabla F(\tilde{z})^\top h &= (-Z^\top A^\top \lambda, h) = -(A^\top \lambda, \underbrace{Zh}_{\in \ker A}) = 0 \quad \forall h, \end{aligned}$$

also

$$\nabla F = 0.$$

(ii) Sei $\nabla F(\tilde{z}) = Z^\top \nabla f(\tilde{x}) = 0$. Wir wählen $d \in \ker A$ beliebig. Dann gilt $d = Zz$ und

$$\nabla f(\tilde{x})^\top d = \nabla f(\tilde{x})^\top Zz = 0$$

$\Rightarrow \nabla f(\tilde{x}) \perp \ker A$. Lineare Algebra: $\nabla f \in \mathrm{Im} A^\top$, also

$$\nabla f = A^\top \mu\,. \quad \text{Setzen} \quad \lambda := -\mu\,.$$

□

Analog vereinfachen sich die hinreichenden Bedingungen zweiter Ordnung:

Satz 5.4.5 *Es sei f zweimal stetig in $\tilde{x}$ differenzierbar, Z eine Nullraum-Matrix von A und $\tilde{x} = w + Z\tilde{z}$. Dann folgt aus der positiven Definitheit der reduzierten Hesse-Matrix*

$$F''(\tilde{z}) = Z^\top f''(\tilde{x}) Z$$

die Existenz von $\alpha > 0$ mit

$$d^\top f''(\tilde{x}) d \geq \alpha \|d\|^2 \quad \forall d \in \ker A\,. \tag{5.67}$$

Ist umgekehrt (5.67) *erfüllt und sind die Spaltenvektoren von Z linear unabhängig, dann ist $F''(\tilde{z})$ positiv definit.*

Beweis: (i) $\Rightarrow$: $F''(\tilde{z})$ sei positiv definit und $d \in \ker A$, also $d = Zz$. Dann

$$\begin{aligned} d^\top f''(\tilde{x}) d = (Zz)^\top f''(\tilde{x}) Zz &= z^\top \underbrace{Z^\top f''(\tilde{x}) Z}_{F''(\tilde{z})} z \\ &\geq \beta \|z\|^2 \geq \underbrace{\beta \|Z\|^{-2}}_{=\alpha} \|d\|^2 \end{aligned}$$

(wegen positiver Definitheit von F'' und $\|d\| \leq \|Z\|\,\|z\| \Rightarrow \|z\| \geq \|Z\|^{-1}\|d\|$)

(ii) $\Leftarrow$: (5.67) sei erfüllt, d.h. mit $d = Zz$

$$(Zz)^\top f''(\tilde{x}) Zz = z^\top \underbrace{Z^\top f''(\tilde{x}) Z}_{=F''(\tilde{z})} z \geq \alpha \|d\|^2.$$

Daraus folgt die positive Definitheit von $F''(\tilde{z})$. Ist nämlich $z \neq 0$, dann auch $d \neq 0$ wegen $d = Zz$ und der Voraussetzung an Z. □

Beispiel 5.4.4

$$\begin{aligned} &\min f(x) = x_1^2 + 2x_2^2 + 3x_3 \\ &\textit{bei} \quad x_1 + 2x_2 + 3x_3 = 6\,. \end{aligned}$$

Durch die Nebenbedingung können wir offenbar die Dimension um 1 *reduzieren.*

$$A = (\underset{\uparrow}{1}, \underbrace{2,3}) \qquad d_1 + 2d_2 + 3d_3 = 0$$
$$A_1\, A_2 \qquad \Updownarrow \quad d_1 = -2d_2 - 3d_3$$

$$d \in \ker A \Leftrightarrow d = \begin{pmatrix} -2d_2 & -3d_3 \\ d_2 & \\ d_3 & \end{pmatrix} = \underbrace{\begin{pmatrix} -2 & -3 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{Z} \begin{pmatrix} d_2 \\ d_3 \end{pmatrix}.$$

Eine spezielle Lösung der inhomogenen Gleichung ist $w = (6,0,0)^\top$.

Reduziertes unrestringiertes Problem:

$$\begin{aligned} \min F(z_1,z_2) &= f(w+Zz) = f(6-2z_1-3z_2, z_1, z_2) \\ &= (6-2z_1-3z_2)^2 + 2z_1^2 + z_2 \end{aligned}$$

(entstanden durch Elimination von x_1*).*

Stationärer Punkt aus $\nabla F(\tilde{z}) = 0$*,*

$$\tilde{z} = \begin{pmatrix} 1/2 \\ 3/2 \end{pmatrix}.$$

Aus Satz 5.4.4 folgt, dass

$$\tilde{x} = w + Zz = \begin{pmatrix} 1/2 \\ 1/2 \\ 3/2 \end{pmatrix}$$

den notwendigen Bedingungen der Lagrange-Multiplikatorenregel genügt. Konvexität von $f \Rightarrow$ $\tilde{x}$ *ist Lösung.*

Wir hätten aber auch einfach die hinreichende Bedingung nachprüfen können:

$$f''(x) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$F''(\hat{z}) = \begin{pmatrix} -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -2 & 3 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} -4 & 4 & 0 \\ -6 & 0 & 0 \end{pmatrix} \begin{pmatrix} -2 & -3 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 12 & 12 \\ 12 & 18 \end{pmatrix}$$

ist positiv definit.

5.4.4 Quadratische Optimierungsprobleme

Als Spezialfall von (PLG) betrachten wir

$$\begin{aligned} &\min f(x) = \frac{1}{2}\langle Qx,x\rangle + \langle q,x\rangle \\ &\text{bei} \quad Ax = b \end{aligned} \tag{QG}$$

mit einer (n,n)-Matrix Q, $q \in \mathbb{R}^n$. Klar ist:

Satz 5.4.6 *Ist Q positiv definit auf* $\ker A$*, d. h.*

$$d^\top Qd \geq \alpha \|d\|^2 \quad \forall d \in \ker A\,,$$

und ist die Gleichung $Ax = b$ lösbar, dann hat (QG) *genau eine Lösung.*

Beweisidee: Das liegt an der strikten Konvexität von f auf $\mathscr{F}$! Wir nutzen einfach die bereits diskutierte Darstellung $x = w + d$ mit einer speziellen Lösung w des Gleichungssystems und $d = Zz \in \ker A$. Auf $\ker A$ ist f aber streng konvex. □

Bemerkung: Die Gleichung $Ax = b$ ist lösbar für $m \leq n$, wenn A vollen Rang hat.

Besonders schön ist hier das Optimalitätssystem: Wir haben für symmetrisches Q

$$\nabla f = Qx + q$$

Daraus folgen die notwendigen Bedingungen

$$Qx + q + A^\top \lambda = 0\,,\ Ax = b.$$

Anders aufgeschrieben erhalten wir das **Optimalitätssystem**

$$\boxed{\begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}.} \tag{5.68}$$

Jede Lösung dieses Systems ist bei positiv definiter Matrix Q auf $\ker A$ eine Lösung von (QG) (Konvexität!). Hat A vollen Rang, dann gibt es wegen Satz 5.4.6 höchstens eine. Umgekehrt existiert dann genau eine Lösung des Optimalitätssystems, insbesondere für $q = 0$, $b = 0$. Also ist der Kern der Matrix in (5.68) gleich $\{0\}$, die obige Matrix invertierbar und folglich gilt

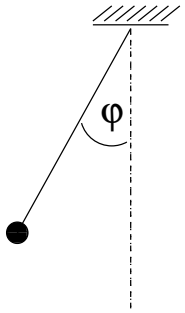
Lemma 5.4.1 *Wenn $m \leq n$, rang $A = m$ und Q positiv definit auf* $\ker A$*, dann ist die Matrix* $\begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix}$ *invertierbar.*

Die eindeutige Bestimmtheit von λ ergibt sich aus dem vollen Rang von A.

5.4.5 Dynamische Optimierungsprobleme

Probleme dieser Art sind sehr wichtig für viele Probleme in Wissenschaft und Technik. Eigentlich kommen sie her von Optimalsteuerungsproblemen bei gewöhnlichen Differentialgleichungen.

Beispiel 5.4.5 *Optimale Steuerung eines schwingenden Pendels in die Ruhelage:*



Zeitraum:	$t \in [0,T]$		
Auslenkung:	$\varphi = \varphi(t),$	$\varphi(0) = \varphi_0$	
Winkelgeschwindigkeit:	$\omega = \omega(t),$	$\omega(0) = \omega_0,$	$\omega = \dot{\varphi}$

Angreifende, steuerbare Kraft: $u = u(t)$

Bewegungsgleichung:
$$\begin{aligned} &\ddot{\varphi}(t) = -c\sin(\varphi(t)) + u(t), \quad t \in (0,T) \\ &\varphi(0) = \varphi_0 \\ &\dot{\varphi}(0) = \omega_0. \end{aligned}$$

c steht für das Verhältnis von Masse und Gravitation. Umformung als System 1. Ordnung: Wir setzen $\omega = \dot{\varphi}$

$$\begin{aligned} \dot{\omega} &= -c\sin(\varphi) + u \\ \dot{\varphi} &= \omega. \end{aligned}$$

Der zu steuernde Zustand des Systems ist

$$z(t) := \begin{pmatrix} \varphi(t) \\ \omega(t) \end{pmatrix}.$$

Optimierungsziel: Ruhelage bei $t = T$*, d.h. wir erhalten das* **Optimalsteuerungsproblem**

$$\min \frac{1}{2}\|z(T)\|^2 + \nu \int_0^\top u^2(t)dt$$

bei

$$\begin{aligned} \dot{z}(t) &= \begin{pmatrix} \omega(t) \\ -c\sin(\varphi(t)) \end{pmatrix} + \begin{pmatrix} 0 \\ u(t) \end{pmatrix} \\ z(0) &= \begin{pmatrix} \varphi_0 \\ \omega_0 \end{pmatrix} = z_0\,. \end{aligned}$$

Der zweite Term in der Zielfunktion misst die Energiekosten.

Linearisierung bei Annahme kleinen Schwingungen:

Für kleine Schwingungen haben wir $\sin\varphi \approx \varphi$. *Dann ergibt sich folgendes Optimalsteuerungsproblem:*

$$\begin{aligned} &\min \frac{1}{2}\|z(T)\|^2 + \frac{\nu}{2}\int_0^\top u(t)^2 dt \\ \dot{z}(t) &= \underbrace{\begin{pmatrix} 0 & 1 \\ -c & 0 \end{pmatrix}}_{A} z(t) + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{B} u(t) \\ z(0) &= z_0. \end{aligned}$$

Dieses Optimalsteuerungsproblem kann theoretisch im Funktionenraum gut behandelt werden. Dies wollen wir hier nicht tun, sondern betrachten statt dessen eine

Diskretisierte Variante:

Sei dazu $0 = t_0 < t_1 < \ldots < t_N = T$ eine äquidistante Zerlegung von $[0,T]$ mit $t_i = \tau \cdot i$, $0 \le i \le N$ und $\tau = T/N$. Wir setzen $z(t)$ als vektorwertige stückweise lineare Funktion an, $u(t)$ als Treppenfunktion,

$$\begin{aligned} u(t) &\equiv u_i \quad \text{auf } [t_i, t_{i+1}] \\ z(t_i) &= z_i \quad \text{in } t_i,\ i = 0,\ldots,N. \end{aligned}$$

Zur Approximation der Ableitung verwenden wir der Einfachheit halber das explizite Euler-Verfahren,

$$\dot{z}(t_i) \approx \frac{z(t_{i+1}) - z(t_i)}{\tau} \qquad (z \in \mathbb{R}^2)$$

und nach Ersetzen von $z(t_i)$ durch z_i

$$\begin{aligned} z_{i+1} &= z_i + \tau A z_i + \tau B u_i \qquad i = 0,\ldots,N-1 \\ &= \underbrace{(I + \tau A)}_{\tilde{A}} z_i + \underbrace{\tau B}_{\tilde{B}} u_i. \end{aligned}$$

Insgesamt ergibt sich die diskretisierte Schwingungsgleichung

$$\boxed{z_{i+1} = \tilde{A} z_i + \tilde{B} u_i \quad i = 0,\ldots,N-1.}$$

Wegen

$$\int_0^\top u(t)^2 dt = \sum_0^{N-1} \int_{ti}^{t_{i+1}} u_i^2 dt = \tau \sum_0^{n-1} u_i^2$$

erhalten wir als zu minimierendes Zielfunktional

$$\boxed{f(z,u) = \frac{1}{2}\|z_N\|^2 + \frac{\nu}{2}\sum_{i=0}^{N-1} \tau u_i^2 .}$$

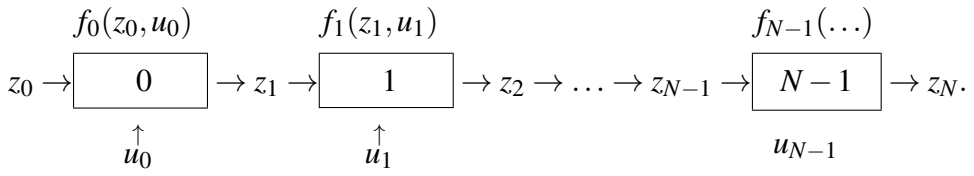
Diese Aufgabe gehört zum Typ der dynamischen Optimierungsprobleme, deren allgemeine Version folgende Form hat (vgl. [1,S. 202ff]):

Dynamisches Optimierungsproblem:

$$\boxed{\begin{aligned} &\min f(z,u) = \frac{1}{2} z_N^\top Q z_N + \sum_{i=0}^{N-1} f_i(z_i,u_i) \\ &\text{bei} \quad z_{i+1} = A_i z_i + B_i u_i + c_i, \quad i = 0,\ldots,N-1\,. \end{aligned}} \tag{DLG}$$

Hier ist $z_0 \in \mathbb{R}^n$ fest vorgegeben, Q eine symmetrische (n,n)-Matrix, A_i sind (n,n)-Matrizen, B_i sind (n,m)-Matrizen, $c_i \in \mathbb{R}^n$, $u_i \in \mathbb{R}^m$ und $x^\top = (z^\top, u^\top)$, $x \in \mathbb{R}^{N(m+n)}$ und $f_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ sind gegebene differenzierbare Funktionen.

Wir können uns das als ***N*-stufigen Entscheidungsprozess** vorstellen:



Entscheidungen: u_i.

Herleitung der notwendigen Optimalitätsbedingungen: Wir haben dazu folgende zwei Möglichkeiten: Erstens können wir die Nebenbedingungen kompakt in Form eines linearen Gleichungssystems aufschreiben und dann die Regel der Lagrangeschen Multiplikatoren anwenden. Dies wird im Folgenden beschrieben (Selbststudium). Zweitens können wir die Gleichungssysteme der einzelnen Zeitschritte getrennt in die Lagrange-Funktion aufnehmen und die notwendigen Bedingungen direkt aus der Lagrange-Funktion durch Ableiten und Nullsetzen herleiten. Das ist kürzer und folgt nach der ersten Variante.

Variante 1:

Wir schreiben die Nebenbedingungen von (DLG) etwas anders auf:

$$-A_i z_i + z_{i+1} - B_i u_i = c_i, \quad i = 0,\ldots,N-1$$

und beachten, dass für $i = 0$ die Variable z_0 fest vorgegeben ist, also spielt diese Gleichung eine Sonderrolle:

$$z_1 - B_0 u_0 = c_0 + A_0 z_0$$

Die Nebenbedingungen lauten dann

$$\mathscr{A}\mathbf{x} = \mathbf{b}$$

für $x \in \mathbb{R}^{N(n+m)}$ mit

$$x = (z^\top, u^\top)^\top, \quad b = ((A_0 z_0 + c_0)^\top, c_1^\top, \ldots, c_{N-1}^\top)^\top \in \mathbb{R}^{nN}, \qquad \mathscr{A} = (\mathscr{A}_1, \mathscr{A}_2),$$

$$\mathscr{A}_1 = \begin{pmatrix} I & & & & \\ -A_1 I & & & & \\ & \ddots & & & \\ & & -A_2 I & & \\ & & & \ddots\ddots & \\ & & & -A_{N-1} I & \end{pmatrix}, \quad \mathscr{A}_2 = \begin{pmatrix} -B_0 & & \\ & \ddots & \\ & & -B_{N-1} \end{pmatrix}$$

wobei $I = I_n$. $\mathscr{A}_1$ ist auf alle Fälle invertierbar, weil untere Dreiecksmatrix mit I in Hauptdiagonale (lässt sich von Anfang zum Ende durchlösen). Damit hat $\mathscr{A}$ vollen Rang. Offenbar gilt daher $\mathscr{F} = \{x | \mathscr{A}x = b\} \neq \phi$.

Nun sei $\tilde{x} = (\tilde{z}, \tilde{u})^\top$ eine (lokale) Lösung von (DLG). Weil $\mathscr{A}$ vollen Rang hat, existiert genau ein **Lagrangescher Multiplikator** $\tilde{\lambda} \in \mathbb{R}^{Nn}$, $\tilde{\lambda} = (\tilde{\lambda}_1^\top, \ldots, \tilde{\lambda}_N^\top)^\top$ mit $\nabla f(\tilde{x}) + \mathscr{A}^\top \tilde{\lambda} = 0$, d. h.

$$\boxed{\nabla f(\tilde{z}, \tilde{u}) + \mathscr{A}^\top \tilde{\lambda} = 0\,.}$$

Leider passt dieses Pluszeichen oben nicht in die allgemeine Theorie der Optimalsteuerung. Deshalb setzen wir

$$\lambda := -\tilde{\lambda}\,,$$

um die Gleichung

$$\nabla f(\tilde{z}, \tilde{u}) - \mathscr{A}^\top \lambda = 0$$

zu erhalten.

Wir transponieren die Gleichung und betrachten

$$f_z(\tilde{z}, \tilde{u}) z + f_u(\tilde{z}, \tilde{u}) u - \lambda^\top \mathscr{A} \begin{pmatrix} z \\ u \end{pmatrix} = 0 \quad \forall \begin{pmatrix} z \\ u \end{pmatrix} \in R^{N(n+m)}$$

(f_z bezeichnet die partielle Ableitung von f nach z, nicht den partiellen Gradienten!)

Jetzt heißt es, kräftig zu rechnen, um alles in eine anwendbare Form zu bringen. Wir schreiben dabei f_z kurz für $f_z(\tilde{z}, \tilde{u})$, analog bei f_u.

$$f_z z = \sum_{i=1}^{N-1} f_{i,z} \cdot z_i + \tilde{z}_N^\top Q z_N \quad \text{(keine Abhängigkeit von } z_0\text{!)}$$

$$f_u u = \sum_{0}^{N-1} f_{i,u} \cdot u_i$$

$$\lambda^\top \mathscr{A} \begin{pmatrix} z \\ u \end{pmatrix} = \lambda^\top (\mathscr{A}_1 z + \mathscr{A}_2 u) = (\lambda_1^\top, \ldots, \lambda_N^\top) \begin{pmatrix} z_1 & -B_0 u_0 \\ z_2 - A_1 z_1 & -B_1 u_1 \\ \vdots & \\ z_N - A_{N-1} z_{N-1} & -B_{N-1} u_{N-1} \end{pmatrix} \begin{matrix} \\ \leftarrow \text{ Sonderrolle} \\ \leftarrow \text{ ab 2. Komp.} \\ \text{ normal} \end{matrix}$$

Einsetzen in die notwendige Bedingung

$$\begin{aligned} &\tilde{z}_N^\top Q z_N + \sum_{i=1}^{N-1} f_{i,z} z_i + \sum_{i=0}^{N-1} f_{i,u} u_i \\ &- \sum_{i=1}^{N-1} \lambda_{i+1}^\top [z_{i+1} - A_i z_i - B_i u_i] - \lambda_1^\top [z_1 - B_0 u_0] = 0 \quad \text{für alle } \begin{pmatrix} z \\ u \end{pmatrix}. \end{aligned} \tag{5.69}$$

Nun kann man z, u beliebig passend einsetzen, um Gleichungen für λ zu erhalten.

• $u = 0$, $z_1, \ldots, z_{N-1} = 0$, nur z_N frei, also (Q ist symmetrisch!)

$$(\tilde{z}_N^\top Q - \lambda_N^\top) z_N = 0 \quad \forall z_N,$$

$\Rightarrow$

$$\lambda_N = Q\tilde{z}_N,$$

Das ist eine **Endbedingung** für λ. Damit sind jetzt alle mit z_N verbundenen Terme Null.

• Wählen jetzt z_i beliebig, die restlichen Null, auch die u's

$\Rightarrow f_{i,z} - \lambda_i^\top z_i + \lambda_{i+1}^\top A_i z_i = 0 \quad \forall z_i \quad i \in \{1, \ldots, N-1\}$

also $f_{i,z} - \lambda_i^\top + \lambda_{i+1}^\top A_i = 0$ oder $\lambda_i^\top = \lambda_{i+1}^\top A + f_{i,z}$

oder, noch schöner,

$$\begin{aligned} \lambda_i &= A_i^\top \lambda_{i+1} + \nabla_{z_i} f_i(\tilde{z}_i, \tilde{u}_i), \quad i = N-1, \ldots, 1 \\ \lambda_N &= Q\tilde{z}_N. \end{aligned} \tag{5.70}$$

Definition 5.4.4 *Die obere Gleichung (5.70) heißt* **adjungierte Gleichung***, die untere ist eine zugehörige Endbedingung.*

Nun werten wir noch die u_i's aus: Setzen alle z_j Null und alle u_j außer u_i. Dann

$$[f_{i,u} + \lambda_{i+1}^\top B_i] u_i = 0 \quad \forall u_i\,,$$

also die **Minimumbedingung**

$$B_i^\top \lambda_{i+1} + \nabla_{u_i} f_i(\tilde{z}_i, \tilde{u}_i) = 0 \quad \forall i = 0, \ldots, N-1.$$

Folgerung: Die notwendigen Optimalitätsbedingungen erster Ordnung für die diskretisierte dynamische Optimierungsaufgabe lauten

$\lambda_i = A_i^\top \lambda_{i+1} + \nabla_{z_i} f_i(\tilde{z}_i, \tilde{u}_i), \quad i = N-1, \ldots, 1$	(adjungierte Gleichung)
$\lambda_N = Q\tilde{z}_N,$	(Endbedingung)
$B_i^\top \lambda_{i+1} + \nabla_{u_i} f_i(\tilde{z}_i, \tilde{u}_i) = 0 \quad \forall i = 0, \ldots, N-1.$	(Minimumbedingung)

Interpretation der Minimumbedingung: Wir betrachten $g_i(u) = f_i(\tilde{z}, u) + B_i^\top \lambda \cdot u$. Dann steht oben, dass $\tilde{u}_i$ die *notwendige Bedingung* für die Aufgabe $\min\limits_u g_i(u)$ erfüllt (was nicht heißt, dass $\tilde{u}_i$ diese Minimumaufgabe auch wirklich löst. Die braucht nicht einmal lösbar zu sein!).

Variante 2. Wir haben *adjungierte Gleichung* und *Minimumprinzip* aus der Lagrangeschen Multiplikatorenregel $\nabla f(\tilde{x}) + A^\top \lambda = 0$ hergeleitet. Diese Herangehensweise ist zu wenig praktikabel. Äquivalent ist das Arbeiten mit der Lagrange-Funktion. Definiere dazu

$$\mathscr{L}(z, u, \lambda) := f(z, u) \quad \text{"minus herangehangene Gleichungen"}$$

(Minuszeichen wegen $\lambda := -\tilde{\lambda}$) , also

$$\mathscr{L}(z,u,\lambda) = \frac{1}{2} z_N^\top Q z_N + \sum_{i=0}^{N-1} f_i(z_i,u_i) \\ - \lambda_1^\top [z_1 - Bu_0 - Az_0 - c_0] - \sum_{i=1}^{N-1} \lambda_{i+1}^\top [z_{i+1} - A_i z_i - B_i u_i - c_i].$$

Dann gilt: Die notwendigen Bedingungen sind äquivalent mit

$\nabla_{z_i}\mathscr{L}(\tilde{z},\tilde{u},\lambda) = 0\,,$ $\quad i = 1,\ldots,N$	$\rightarrow$	adjungierte Gleichung
$\nabla_{u_i}\mathscr{L}(\tilde{z},\tilde{u},\lambda) = 0\,,$ $\quad i = 0,\ldots,N-1$	$\rightarrow$	Minimumbedingung

So sollte man sich das merken. Im Übrigen stellt (5.69) genau dieses System dar; d. h., ausgehend von $\mathscr{L}(z,u)$, kommt man direkt zu (5.69).

5.5 Affine Ungleichungsnebenbedingungen

5.5.1 Problemdefinition

Wir betrachten nun die etwas allgemeinere, nämlich durch Ungleichungsrestriktionen erweiterte Aufgabe

$$\min f(x) \tag{PLU}$$

bei

$$\langle a^i, x\rangle = b_i\,,\ i = 1,\ldots,m$$
$$\langle g^j, x\rangle \le r_j\,,\ j = 1,\ldots,p.$$

Abgekürzte Schreibweise: Sei

$$\mathscr{F} = \{x \,|\, Ax = b\,,\ Gx \le r\},$$

mit Matrizen

$$A := \left(a^1,\ldots,a^m\right)^\top, \quad G := \left(g^1,\ldots,g^m\right)^\top$$

somit

$$\boxed{\begin{array}{l}\min f(x)\\ \text{bei} \quad Ax = b\,,\ Gx \le r.\end{array}} \tag{PLU}$$

Beispiel 5.5.1 (Konvexe Linearkombination kleinster Norm)

Gegeben: Vektoren $s^j \in \mathbb{R}^n$, $j = 1,\dots,p$. Löse

$$\min_{\alpha\in\mathbb{R}^p} f(\alpha) = \frac{1}{2}\|\sum_{j=1}^{p} \alpha_j s^j\|^2$$

$$bei \quad \sum_{j=1}^{p} \alpha_j = 1\,, \alpha_j \geq 0,\ j = 1,\dots,p.$$

Das ist eine Aufgabe mit Nichtnegativitätsforderungen.

5.5.2 Notwendige Optimalitätsbedingungen

Bisher waren die notwendigen Optimalitätsbedingungen für eine Lösung $\tilde{x}$ aus der Analysis bekannt. Das ist jetzt anders, jetzt wird wirklich Neuland betreten!

Definition 5.5.1 (Aktive Ungleichungen) *Zu gegebenem $x \in \mathscr{F}$ heißt*

$$J(x) := \{j \in \{1,\dots,p\} | \langle g^j, x\rangle = r_j\}$$

Indexmenge der aktiven Ungleichungs-Restriktionen.

Nun sei $\tilde{x}$ eine Lösung der Aufgabe (PLU).

Definition 5.5.2 (Linearisierender Kegel) *Es sei $\tilde{x} \in \mathscr{F}$;*

$$L(\mathscr{F},\tilde{x}) := \{d \in \mathbb{R}^n \,|\, Ad = 0\,,\ \langle g^j, d\rangle \leq 0 \quad \forall j \in J(\tilde{x})\}$$

heißt linearisierender Kegel *von $\mathscr{F}$ an der Stelle $\tilde{x}$.*

Lemma 5.5.1 *Es sei $\mathscr{F}$ die oben durch affine Gleichungen und Ungleichungen definierte zulässige Menge. Dann gilt für $\tilde{x} \in \mathscr{F}$*

$$K(\mathscr{F},\tilde{x}) = L(\mathscr{F},\tilde{x})\,.$$

Insbesondere ist damit $K(\mathscr{F},\tilde{x})$ abgeschlossen, weil das für L offensichtlich gilt.

Beweis:

(i) $\mathbf{K} \subset \mathbf{L}$: Sei $d \in K(\mathscr{F},\tilde{x})$, d. h. $d = \alpha(x-\tilde{x})$ mit $x \in \mathscr{F}$, $\alpha \geq 0$.
$\Rightarrow Ax = A\tilde{x} = b \Rightarrow$

$$Ad = 0.$$

Das war einfach. Außerdem gilt für alle $j \in J(\tilde{x})$

$$\begin{aligned} &\langle g^j, x\rangle \leq r_j = \langle g^j, \tilde{x}\rangle \\ \Rightarrow \quad &\langle g^j, d\rangle = \alpha\langle g^j, x-\tilde{x}\rangle \leq 0 \quad \forall j \in J(\tilde{x})\,. \end{aligned}$$

Insgesamt also $d \in L(\mathscr{F},\tilde{x})$.

(ii) $\mathbf{L} \subset \mathbf{K}$: Wir zeigen $\tilde{x} + \bar{t}\, d \in \mathscr{F}$ mit einem $\bar{t} > 0$. Daraus folgt dann $\tilde{t}d \in \mathscr{F} - \{\tilde{x}\}$ und deshalb $d \in K(\mathscr{F}, \tilde{x})$.

Sei dazu $d \in L(\mathscr{F}, \tilde{x})$. Für $j \neq J(\tilde{x})$ gilt $\langle g^j, \tilde{x} \rangle < r_j$ und damit

$$\langle g^j, \tilde{x} + td \rangle < r_j \quad \forall t \in [0, \bar{t}] \ , \quad \forall j \notin J(\tilde{x})$$

mit einem $\bar{t} > 0$. Von d wissen wir wegen $d \in L$:

$$\begin{aligned} & Ad = 0 \quad \text{und} \quad \langle g^j, d \rangle \leq 0 \quad j \in J(\tilde{x}) \\ \Rightarrow \quad & A(\tilde{x} + td) = A\tilde{x} + t \underbrace{Ad}_{0} = A\tilde{x} = b \\ & \langle g^j, \tilde{x} + \bar{t}d \rangle = \underbrace{\langle g^j, \tilde{x} \rangle}_{= r_j} + \bar{t} \underbrace{\langle g^j, d \rangle}_{\leq 0} \leq r_j \quad \forall j \in J(\tilde{x}) \, . \end{aligned}$$

Deshalb gilt wie behauptet $\tilde{x} + \bar{t}d \in \mathscr{F}$. □

Nun brauchen wir aber auch noch den **Normalenkegel** N. Dazu benötigen wir:

Lemma 5.5.2 *Sei $K \subset \mathbb{R}^n$ konvexer abgeschlossener Kegel und $x \notin K$. Dann existiert ein $s \in \mathbb{R}^n$ mit*

$$\langle s, x \rangle > 0 = \max_{y \in K} \langle s, y \rangle \, .$$

Beweis: Aus dem Trennungssatz 5.3.3 folgt: $\exists s \in \mathbb{R}^n$ mit

$$\langle s, x \rangle > \sup_{y \in K} \langle s, y \rangle. \tag{5.71}$$

Daraus folgt

Fall 1: $\langle s, y \rangle \leq 0 \quad \forall y \in K$.

Dann ist das Supremum Null (wegen $0 \in K$).

Fall 2: $\exists z \in K : \quad \langle s, z \rangle > 0$

Dann ist das Supremum unendlich (man nehme $\alpha \cdot z$, mit $\alpha \to \infty$).

Fall 2 kann wegen (5.71) nicht eintreten, nur Fall 1, d. h.

$$\langle s, x \rangle > 0 = \max_{y \in K} \langle s, y \rangle \, .$$

□

Außerdem haben wir:

Ist $K \subset \mathbb{R}^n$ abgeschlossener konvexer Kegel und $K \neq \emptyset$, dann gilt $(K^*)^* = K$ (Übungsaufgabe).

Bemerkung: Hierzu benötigen wir die Aussage von Lemma 5.5.2

Nun können wir $N(\mathscr{F}, \tilde{x})$ angeben:

Lemma 5.5.3 *Für (PLU) gilt*

$$N(\mathscr{F}, \tilde{x}) = \left\{ \sum_{i=1}^{m} \lambda_i a^i + \sum_{j \in J(\tilde{x})} \mu_j g^j \,\middle|\, \lambda_i \in \mathbb{R}, \, i = 1, \ldots, m, \; \mu_j \geq 0 \right\}$$

Beweis: Wir wollen die rechte Seite, d. h. $\{\ldots\}$, zunächst mit $N(A,G,\tilde{x})$ bezeichnen.

(i) "$\supset$": Sei $s \in N(A,G,\tilde{x})$, d. h.

$$s = \sum \lambda_i a^i + \sum_J \mu_j g^j\,,\ \mu_j \geq 0\,.$$

Wegen $N(\mathscr{F},\tilde{x}) = K(\mathscr{F},\tilde{x})^*$ ist $s \in K(\mathscr{F},\tilde{x})^*$ zu zeigen, d.h.

$$\langle s,d\rangle \leq 0 \quad \forall d \in K(\mathscr{F},\tilde{x}).$$

Multiplizieren skalar mit $d \in K(\mathscr{F},\tilde{x})$ durch. Dann gilt wegen $K = L$ (Form von K!)

$$\langle s,d\rangle = \sum \lambda_i \underbrace{\langle a^i,d\rangle}_{=0} + \sum_J \mu_j \underbrace{\langle g^j,d\rangle}_{\leq 0} \leq 0 \quad \forall d \in K(\mathscr{F},\tilde{x})\,.$$

Damit nach Definition

$$s \in K(\mathscr{F},\tilde{x})^* = N(\mathscr{F},\tilde{x}).$$

(ii) "$\subset$", d.h. $K(\mathscr{F},\tilde{x})^* \subset N(A,G,\tilde{x})$:

Das zeigen wir in einer dualisierten Form, nämlich zuerst

$$N(A,G,\tilde{x})^* \subset K(\mathscr{F},\tilde{x}).$$

Das sieht man wie folgt ein: Sei $s \in N(A,G,\tilde{x})^*$, also

$$\langle s\,, \sum_{i=1}^{m} \lambda_i a^i + \sum_{j\in J(\tilde{x})} \mu_j g^j\rangle \leq 0 \qquad \forall \lambda_i \in \mathbb{R},\ \forall \mu_j \geq 0.$$

Setzen wir speziell alle μ_j null und auch alle λ_i mit Ausnahme von λ_k, so folgt

$$\lambda_k \langle s\,, a^k\rangle \leq 0 \qquad \forall \lambda_k \in \mathbb{R}$$

und deshalb $\langle s\,, a^k\rangle = 0$. Analog folgert man für variables μ_ℓ auch $\langle s\,, \mu^\ell\rangle \leq 0$. Insgesamt erhalten wir so $s \in L(\mathscr{F},\tilde{x})$ und wegen Lemma 5.5.1 auch $s \in K(\mathscr{F},\tilde{x})$.

Aus diesem Zwischenresultat folgt sofort

$$N(\mathscr{F},\tilde{x}) = K(\mathscr{F},\tilde{x})^* \subset (N(A,G,\tilde{x}))^{**} = N(A,G,\tilde{x}).\square$$

Nun haben wir alles bereit zum Aufstellen der notwendigen Bedingungen für $\tilde{x}$: Wir kennen die allgemeine notwendige Bedingung, die aus der Variationsungleichung gefolgert wurde,

$$-\nabla f(\tilde{x}) \in N(\mathscr{F},\tilde{x}).$$

Aus dem eben bewiesenen Lemma folgt

$$-\nabla f(\tilde{x}) = \sum_{i=1}^{m} \lambda_i a^i + \sum_{j\in J(\tilde{x})} \mu_j g^j\,,\ \mu_j \geq 0$$

oder

$$\boxed{0 = \nabla f(\tilde{x}) + \sum_{i=1}^{m} \lambda_i a^i + \sum_{j \in J(\tilde{x})} \mu_j g^j \,,\ \mu_j \geq 0\,.} \tag{5.72}$$

Das können wir noch verschönern. Wir setzen für $j \notin J(\tilde{x})$ einfach $\mu_j = 0$. Dann gilt

- $\mu_j \geq 0 \quad \forall j \in \{1,...,p\}$.
- $\mu_j \underbrace{(\langle g^j, \tilde{x}\rangle - r_j)}_{\text{denn das ist Null für } j \in J(\tilde{x}).} = 0 \quad \forall j \in \{1,...,p\}$

Außerdem nimmt (5.72) die leicht zu merkende Form

$$\nabla f(\tilde{x}) + A^\top \lambda + G^\top \mu = 0$$

an.

Definition 5.5.3 (Lagrangesche Multiplikatoren) *Vektoren $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$ heißen Lagrangesche Multiplikatoren zu $\tilde{x} \in \mathscr{F}$, wenn die Gradientengleichung*

$$\nabla f(\tilde{x}) + A^\top \lambda + G^\top \mu = 0 \tag{5.73}$$

und die Komplementaritätsbedingungen *(auch komplementäre Schlupfbedingungen)*

$$\mu \geq 0, \quad G\tilde{x} - r \leq 0, \quad \langle \mu, G\tilde{x} - r\rangle = 0 \tag{5.74}$$

erfüllt sind.

Wir haben somit bewiesen:

Satz 5.5.1 (Karush-Kuhn-Tucker-Satz für affine Restriktionen) *Ist $\tilde{x}$ lokales Minimum von (PLU) und f in $\tilde{x}$ differenzierbar, dann existieren Lagrangesche Multiplikatoren $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$ zu $\tilde{x}$. Sind die Vektoren a^i, $i = 1,...,m$, sowie g^j, $j \in J(\tilde{x})$, linear unabhängig, so sind λ und μ eindeutig bestimmt.*

Damit bestimmen sich x, λ und μ aus dem Optimalitätssystem

$$\boxed{\begin{aligned} &Ax = b\,,\ Gx \leq r\,,\ \nabla f + A^\top \lambda + G^\top \mu = 0\,,\\ &\mu \geq 0\,,\ \langle \mu, Gx - r\rangle = 0\,. \end{aligned}}$$

Bemerkungen zum Satz:

- *Zur Historie: Die Multiplikatorenregel für Gleichungsrestriktionen stammt von Lagrange. Verallgemeinerung auf Ungleichungsrestriktionen: 1951, Harold W. Kuhn und Albert W. Tucker; später wurde erkannt, dass die Aussage bereits 1939 von William Karush in seiner Masterarbeit bewiesen worden war. Dieses Resultat hat er aber nicht publiziert, weil es für ihn ein Nebenresultat war.*

  Deshalb verwendet man heute die Bezeichnung **Karush-Kuhn-Tucker-System** *oder kurz* **KKT-System** *für das obige Optimalitätssystem.*

- Verwendung der $\mathscr{L}$-Funktion:

$$\mathscr{L}(x,\lambda,\mu) := f(x) + \langle \lambda, Ax-b\rangle + \langle \mu, Gx-r\rangle .$$

Dann bedeutet (5.73) wieder

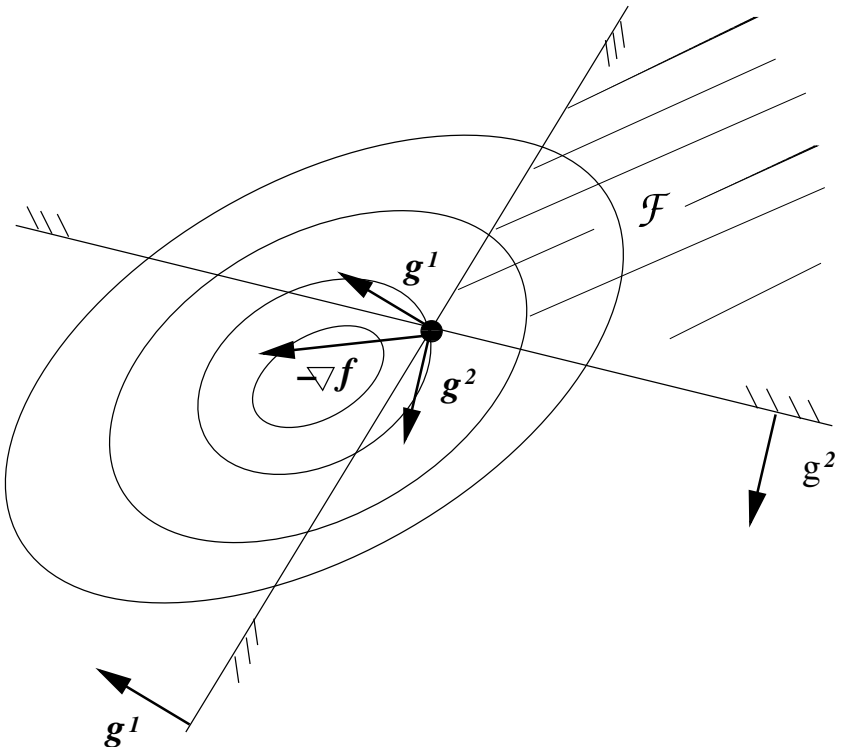
$$\nabla_x \mathscr{L}(\tilde{x},\lambda,\mu) = 0 .$$

Die Nichtnegativitätsforderungen an μ und die komplementäre Schlupfbedingung muss man sich zusätzlich merken.

- **Geometrische Interpretation** (für zwei aktive Ungleichungsrestriktionen): Unsere Optimalitätsbedingung fordert

$$-\nabla f = \mu_1 g^1 + \mu_2 g^2,$$

d.h. $-\nabla f$ muss positive Linearkombination von g^1 und g^2 sein, liegt also im von g^1 und g^2 aufgespannten Kegel. Man sieht schnell ein, dass diese Beziehung im Optimum erfüllt sein muss, denn sonst würde im zulässigen Bereich eine Abstiegsrichtung existieren.



- Bei Konvexität sind die Bedingungen wieder hinreichend für (globale) Optimalität von $\tilde{x}$.

Beispiel 5.5.2 *Wir betrachten die Aufgabe*

$$\begin{array}{rl} \min f(x) = -\sqrt{x_1} - x_2 & \\ \textit{bei} \quad x_i \geq 0\,,\ i=1,2 & \Leftrightarrow \\ x_1 + x_2 \leq 1 & \end{array} \left| \begin{array}{r} -x_1 \leq 0 \\ -x_2 \leq 0 \\ x_1 + x_2 - 1 \leq 0. \end{array} \right.$$

Diese Aufgabe passt nicht ganz in unser Schema, denn im Punkt $x_1 = 0$ ist die Zielfunktion nicht differenzierbar. Sie überlegen sich aber leicht, dass dieser Punkt nicht in Betracht kommt (vergleichen Sie $\sqrt{x}$ mit x für $0 < x < 1$).

Aufstellen des Optimalitätssystems:

$$\mathscr{L}(x,\lambda,\mu) = \mathscr{L}(x,\mu) = -\sqrt{x_1} - x_2 - \mu_1 x_1 - \mu_2 x_2 + \mu_3(x_1 + x_2 - 1),$$

$$\nabla_x \mathscr{L} = \begin{pmatrix} -\dfrac{1}{2\sqrt{x_1}} - \mu_1 + \mu_3 \\ -1 - \mu_2 + \mu_3 \end{pmatrix} \overset{(!)}{=} 0$$

⇒ **Optimalitätssystem:**

$$\begin{array}{rr} -\dfrac{1}{2\sqrt{x_1}} - \mu_1 + \mu_3 = 0 & x_i \geq 0\,, \quad i = 1,2 \\ -\mu_2 + \mu_3 = 1 & \mu_j \geq 0\,, \quad j = 1,2,3 \\ & \mu_1 x_1 = 0 \\ & \mu_2 x_2 = 0 \\ & \mu_3(x_1 + x_2 - 1) = 0\,. \end{array}$$

Eine Lösung ist

$$\begin{aligned} \tilde{x}_1 &= 1/4, \quad \tilde{\mu}_1 = \tilde{\mu}_2 = 0 \\ \tilde{x}_2 &= 3/4, \quad \tilde{\mu}_3 = 1\,. \end{aligned}$$

Wie kommt man drauf? Graphisch oder durch die Eliminationsmethode (es ist klar, dass im Optimum $x_1 + x_2 = 1$ erfüllt sein muss).

Graphische Lösung: *$t := \sqrt{x_1} + x_2$ ist zu maximieren. Man stelle die Funktion*

$$x_2 = t - \sqrt{x_1}$$

graphisch dar und verschiebe den Graphen so lange nach oben, bis er mit dem zulässigen Bereich genau einen Punkt gemeinsam hat. Dort muss er den gleichen Anstieg haben wie die Funktion $x_2 = 1 - x_1$, also -1. Daraus ermittelt man $x_1 = 1/4$.

5.5.3 Hinreichende Optimalitätsbedingungen

Wir wissen wir schon, wie diese in allgemeiner Form aussehen (Satz 5.3.2): Es existiert ein $\alpha > 0$ mit

$$d^\top f''(\tilde{x}) d \geq \alpha \|d\|^2 \quad \forall d \in K(\mathscr{F}, \tilde{x}) \quad \text{mit} \quad \nabla f(\tilde{x})^\top d = 0\,.$$

Für (PLU) haben wir den Kegel K bereits berechnet,

$$K = \{d \,|\, Ad = 0\,,\ \langle g^j, d\rangle \leq 0 \quad \forall j \in J(\tilde{x})\}\,.$$

Nun erfüllt aber $\tilde{x}$ auch die notwendigen Bedingungen, also

$$\nabla f(\tilde{x}) = -A^\top \lambda - \sum_{j \in J(\tilde{x})} \mu_j g^j.$$

Multiplikation mit $d \in K(\mathscr{F},\tilde{x})$ ergibt

$$0 = \langle \nabla f(\tilde{x}), d\rangle = \underbrace{\langle -\lambda, Ad\rangle}_{=0 \text{ wenn } d\in K} - \sum_{j\in J(\tilde{x})} \mu_j \underbrace{\langle g^j, d\rangle}_{\leq 0 \text{ wenn } d\in K},$$

also ist obige Gleichung äquivalent zu

$$0 = \sum_{j\in J(\tilde{x})} \mu_j \langle g^j, d\rangle .$$

Da alle Skalarprodukte $\langle g^j, d\rangle$ nichtpositiv und alle μ_j nichtnegativ sind enthält die Summe nur Summanden gleichen Vorzeichens. Daher verschwindet sie genau dann, wenn alle Summanden verschwinden. Also:

Falls $\mu_j > 0$, so muss $\langle\cdot,\cdot\rangle$ verschwinden, d. h., wir erhalten als zusätzliche Bedingung zu $d \in \mathscr{K}(\mathscr{F},\hat{x})$:

$$\langle g^j, d\rangle = 0 \quad \text{falls } \mu_j > 0 .$$

Die vorher geforderte Bedingung $\nabla f(\tilde{x})^\top d = 0$ ist hierin schon enthalten.

Damit ergibt sich als *hinreichende Optimalitätsbedingung:*

$$\boxed{\begin{array}{rlll} d^\top f''(\tilde{x}) d & \geq \alpha \|d\|^2 & & \\ \text{für alle } d\in\mathbb{R}^n & \text{mit} & & \\ Ad & = 0 & & \\ \langle g^j, d\rangle & \leq 0 & \text{für } j\in J(\tilde{x}) & \text{mit } \mu_j = 0 \\ \langle g^j, d\rangle & = 0 & \text{für } j\in J(\tilde{x}) & \text{mit } \mu_j > 0. \end{array}} \tag{5.75}$$

Insgesamt haben wir folgende Aussage bewiesen:

Satz 5.5.2 (Hinreichende Bedingung 2. Ordnung) *Es sei f in einer Umgebung von $\tilde{x}$ zweimal stetig differenzierbar und $\tilde{x}$ erfülle die notwendigen Optimalitätsbedingungen erster Ordnung aus Satz 5.5.1. Weiter seien die hinreichenden Bedingungen (5.75) erfüllt. Dann ist $\tilde{x}$ striktes lokales Minimum von (PLG).*

Die Restriktionen mit $\mu_j > 0$ sind besonders wichtig:

Definition 5.5.4 (Streng aktive Restriktionen) *Die für $\tilde{x}$ aktiven Restriktionen mit $\tilde{\mu}_j > 0$ heißen* streng aktive Restriktionen.

Beispiel 5.5.3 (Fortsetzung Beispiel 5.5.2)

$$\begin{aligned} \min f(x) &= -\sqrt{x_1} - x_2 \\ x_1 \geq 0,\ x_1 + x_2 &\leq 1. \end{aligned}$$

Wir wissen bereits:

$$\tilde{x}_1 = \frac{1}{4}\,,\ \tilde{x}_2 = \frac{3}{4}\,,\ \mu_1 = \mu_2 = 0\,,\ \mu_3 = 1$$

erfüllen die notwendigen Kuhn-Tucker-Bedingungen.

Hinreichende Bedingung: *Letzte Ungleichung ist streng aktiv, daher lauten die hinreichenden Bedingungen 2. Ordnung*

$$d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \quad \forall d : d_1 + d_2 = 0\,.$$

$$f''(x):\ \textit{Nur}\ f_{x_1x_1} \neq 0; \quad f_{x_1} = -\frac{1}{2}x_1^{-1/2},$$
$$f_{x_1x_1} = \frac{1}{4}\tilde{x}_1^{-3/2} = \frac{1}{4}\left(\frac{1}{4}\right)^{-3/2} = 2.$$

Die Hessematrix

$$f''(\tilde{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

ist nicht positiv definit auf $\mathbb{R}^2$ aber auf dem linearen Unterraum der den streng aktiven Ungleichungen zugeordnet ist, denn

$$\begin{aligned} d_1 + d_2 = 0 &\Rightarrow d_1^2 = d_2^2 \\ \Rightarrow d^\top \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} d &= 2d_1^2 = d_1^2 + d_1^2 \\ &= d_1^2 + d_2^2 = 1 \cdot \|d\|^2\,. \end{aligned}$$

Die hinreichende Bedingung ist mit $\alpha = 1$ erfüllt.

Bemerkung: Man sieht, dass die Verifikation der hinreichenden Bedingungen schwierig sein kann; (5.75) ist i.A. nur numerisch nachprüfbar.

Insbesondere enthält die hinreichende Bedingung 2. Ordnung die Forderung $\langle g^j, d\rangle \leq 0$ für $j \in J$ mit $\mu_j = 0$. Diese Bedingung lässt sich numerisch schwer nachprüfen. Man lässt diese Einschränkung deshalb gern weg und hofft, dass die **strenge hinreichende Bedingung**

$$d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \tag{5.76}$$

bei

$$Ad = 0\,,\ \langle g^j, d\rangle = 0 \quad \forall j \in J(\tilde{x}) \text{ mit } \mu_j > 0$$

verifizieren zu können. Dazu bilden wir die Matrix

$$B = B(\tilde{x}) = \begin{pmatrix} (a^1)^\top \\ \ddots \\ (a^m)^\top \\ \ddots \\ (g^j)^\top \\ \ddots \end{pmatrix} \quad \text{mit allen} \quad g^j \in J(\tilde{x}) \text{ mit } \mu_j > 0\,.$$

Die **strenge Bedingung zweiter Ordnung** lautet dann

$$\boxed{d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \quad \forall d \in \ker B\,.} \tag{5.77}$$

Nun bestimmt man eine Nullraum-Matrix zu $B(\tilde{x})$, dies sei $Z(\tilde{x})$. Ist diese von vollem Rang, dann ist (5.77) äquivalent zur positiven Definitheit von $Z(\tilde{x})^\top f''(\tilde{x})Z(\tilde{x})$, also zur positiven Definitheit der reduzierten Hesse-Matrix.

Beispiel 5.5.4 (Nichtkonvexe quadratische Optimierung) *Wir untersuchen*

$$\begin{aligned} &\min f(x) := -x^2, \\ &x \geq -1\,,\; x \leq \frac{1}{2}. \end{aligned}$$

Offenbar hat die Aufgabe das **globale Minimum** *bei* $x = -1$ *und ein* **lokales Minimum** *bei* $x = \frac{1}{2}$.

Wir untersuchen $\tilde{x} = \frac{1}{2}$ *näher:*

$$\mathscr{L}(x,\mu_1,\mu_2) = -x^2 + \mu_1(-1-x) + \mu_2\left(x - \frac{1}{2}\right).$$

Offenbar muss $\mu_1 = 0$ *sein bei* $\tilde{x} = \frac{1}{2}$. *Kuhn-Tucker-Bedingung* $\Rightarrow$

$$\mathscr{L}_x = -2\tilde{x} + \mu_2 = 0 \quad \Rightarrow \boxed{\mu_2 = 1}\,.$$

Hinreichende Bedingung: *Offenbar ist*

$$d\mathscr{L}_{xx}d \geq \alpha d^2 \quad \forall d \in R \text{ mit } d = 0$$

erfüllt für jedes α. *Damit gilt (5.75). Wie wir gesehen haben, ist aber* $\tilde{x} = \frac{1}{2}$ *nur lokal optimal.*

5.5.4 Strikte Komplementarität

Manche numerischen Verfahren und manch theoretische Aussage für (PLU) bleibt nur richtig bei strikter Komplementarität. Was heißt das?

Definition 5.5.5 (Strikte Komplementarität) *In* $\tilde{x} \in \mathscr{F}$ *ist die Bedingung der strikten Komplementarität erfüllt, wenn*

$$j \in J(\tilde{x}) \Rightarrow \mu_j > 0$$

gilt, wenn also alle aktiven Ungleichungsrestriktionen streng aktiv *sind.*

Folgerung: Hier gilt

$$\begin{cases} \langle g^j, \tilde{x}\rangle = r_j & \Rightarrow \mu_j > 0 \\ \mu_j = 0 & \Rightarrow \langle g^j, x\rangle < r_j \quad j = 1,\ldots,p\,. \end{cases}$$

In diesem Falle sind dann die hinreichenden Bedingungen (5.75) und (5.76) äquivalent, denn der Fall $\mu_j = 0$ und $\langle g^j, \tilde{x}\rangle = r_j$ tritt nicht auf. Wir bilden

$$B(\tilde{x}) = \begin{pmatrix} (a^i)^\top \\ (g^j)^\top \end{pmatrix} \begin{matrix} i \in 1,...,m \\ j \in J(\tilde{x}) \end{matrix}$$

und fordern

$$d^\top f''(\tilde{x}) d \geq \alpha \|d\|^2 \quad \forall d \in \ker B(\tilde{x}) \,. \tag{5.78}$$

Eine weitere Bedingung für spätere Zwecke ist die folgende Voraussetzung:

Lineare Unabhängigkeit der Gradienten der aktiven Restriktionen: *Die Vektoren $a^1,...,a^m$ sowie $\{g^j\}_{j\in J(\tilde{x})}$ seien linear unabhängig.*

Unter dieser Voraussetzung ist

$$\mathscr{A} = \begin{pmatrix} f''(\tilde{x}) & A^\top & G(\tilde{x})^\top \\ A & 0 & 0 \\ G(\tilde{x}) & 0 & 0 \end{pmatrix}$$

invertierbar, wobei die Matrix G durch

$$G(\tilde{x}) := ((g^j)^\top)_{j\in J(\tilde{x})}$$

definiert ist.

5.5.5 Probleme mit oberen und unteren Schranken (box constraints)

In diesem Fall, der besonders einfach diskutiert werden kann, ist der zulässige Bereich $\mathscr{F}$ ein Quader in $\mathbb{R}^n$,

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, v_i \leq x_i \leq w_i,\, i = 1,...,n\} \,.$$

Wir untersuchen also die oft auftretende Aufgabe

$$\boxed{\begin{array}{l} \min_{x\in\mathbb{R}^n} f(x) \\ v \leq x \leq w \,. \end{array}} \tag{PB}$$

Dabei sind $v < w$ gegebene Vektoren unterer und oberer Schranken.

Die allgemeine, bisher entwickelte Theorie liefert hier sehr einfach Beziehungen.

• **Umformung in die allgemeine Form $Gx \leq r$:**

$$v \leq x \leq w \quad \Leftrightarrow \quad \begin{matrix} -x & \leq -v \\ x & \leq w \end{matrix} \quad \Leftrightarrow \underbrace{\begin{pmatrix} -I \\ I \end{pmatrix}}_{G} x \leq \underbrace{\begin{pmatrix} -v \\ w \end{pmatrix}}_{r}$$

Also

$$G = \begin{pmatrix} -1 & & \\ & \vdots & \\ & & -1 \\ 1 & & \\ & \vdots & \\ & & 1 \end{pmatrix}$$

• **Lineare Unabhängigkeit der aktiven Gradienten:**

Ist erfüllt, denn obere und untere Restriktionen können wegen $v < w$ nie gleichzeitig aktiv sein (siehe auch die Form von G). Deshalb sind die Multiplikatoren μ_j eindeutig bestimmt. Aber das bekommen wir alles noch direkter:

• **Notwendige Optimalitätsbedingungen**

$$L = L(x,\mu) = f(x) + \sum_{i=1}^{n} \mu_i^u(-x_i + v_i) + \sum_{i=1}^{n} \mu_i^o(x_i - w_i)$$

$$\frac{\partial L}{\partial x_i} = 0 \quad \Leftrightarrow \quad \boxed{\frac{\partial f}{\partial x_i} - \mu_i^u + \mu_i^o = 0\,,\ \mu_i^u, \mu_i^o \geq 0}$$

μ_i^u: Multiplikatoren zu den unteren Schranken, μ_i^o: Multiplikatoren zu den oberen Schranken

Zusatzinformation: Für jedes i darf genau einer der Multiplikatoren positiv sein (muss aber nicht). Einer ist *immer* Null. Wir erhalten damit folgende möglichen Fälle:

$$\begin{aligned}
\Rightarrow \frac{\partial f}{\partial x_i} = 0 \quad &\Rightarrow \quad \mu_i^u = \mu_i^o = 0 \\
\frac{\partial f}{\partial x_i} > 0 \quad &\Rightarrow \quad \mu_i^u = \frac{\partial f}{\partial x_i}\,,\ \mu_i^o = 0 \\
\frac{\partial f}{\partial x_i} < 0 \quad &\Rightarrow \quad \mu_i^u = 0\,,\ \mu_i^o = -\frac{\partial f}{\partial x_i}.
\end{aligned}$$

Folgerung: Bei Box-Restriktionen gilt

$$\boxed{\begin{aligned}\mu_i^u &= \left(\tfrac{\partial f}{\partial x_i}\right)^+ \\ \mu_i^o &= \left(\tfrac{\partial f}{\partial x_i}\right)^-\end{aligned}}$$

• **Hinreichende Bedingung 2. Ordnung:**

Betrachten wir die Sache aus einem anderen Blickwinkel als bisher in der allgemeinen Theorie. Die *notwendigen Bedingungen* schreiben wir zur Abwechslung nicht in Kuhn-Tucker-Form auf, sondern als *Variationsungleichung*, d. h.

$$\langle \nabla f(\tilde{x}), x - \tilde{x} \rangle \geq 0 \quad \forall x \text{ mit } v \leq x \leq w\,.$$

Das heißt

$$\begin{aligned}
&\langle \nabla f(\tilde{x}), \tilde{x} \rangle \leq \langle \nabla f(\tilde{x}), x \rangle \quad \forall x \text{ mit } v \leq x \leq w \\
&\Updownarrow \\
&\langle \nabla f(\tilde{x}), \tilde{x} \rangle = \min_{v \leq x \leq w} \langle \nabla f(\tilde{x}), x \rangle \\
&\Updownarrow \\
&\frac{\partial f}{\partial x_i} \cdot \tilde{x}_i = \min_{v_i \leq x \leq w_i} \frac{\partial f}{\partial x_i} x\,.
\end{aligned}$$

Folglich

$$\left.\begin{array}{ll} \dfrac{\partial f}{\partial x_i} > 0 & \Rightarrow \tilde{x}_i = v_i \\ \dfrac{\partial f}{\partial x_i} < 0 & \Rightarrow \tilde{x}_i = w_i \end{array}\right\} \begin{array}{l}\text{In diesen Fällen ist } \tilde{x}_i \text{ durch die notwendigen} \\ \text{Bedingungen bei Kenntnis von } \nabla f \text{ festgelegt!}\end{array}$$

$$\frac{\partial f}{\partial x_i} = 0 \quad : \text{keine Aussage (außer eben } \frac{\partial f}{\partial x_i} = 0).$$

Bei den Komponenten mit $\frac{\partial f}{\partial x_i} = 0$ brauchen wir also zusätzliche Informationen. Die holen wir uns aus hinreichenden Bedingungen 2. Ordnung. Wir fordern die strenge hinreichende Bedingung

$$d^\top f''(\tilde{x}) d \geq \alpha \|d\|^2 \quad \text{für alle } d \in \mathbb{R}^n \text{ mit } d_i = 0 \text{ falls } \left|\frac{\partial f}{\partial x_i}\right| \neq 0\,. \tag{5.79}$$

Numerisch kann das so bewerkstelligt werden:

$$Z := \operatorname{diag}(c_i)_{i=1,\ldots,n} \quad \text{mit} \quad c_i = \begin{cases} 1 & \left|\frac{\partial f}{\partial x_i}\right| = 0 \\ 0 & \text{sonst.} \end{cases}$$

Dann ist obige Bedingung (5.79) äquivalent zur positiven Definitheit von

$$Z^\top f''(\tilde{x}) Z\,.$$

5.6 Lineare Optimierungsprobleme

Wir betrachten die Aufgabe, ein *lineares* Funktional f bei linearen Restriktionen zu minimieren

$$\boxed{\begin{array}{l} \min f(x) = c^\top x \\ \text{bei} \quad Ax = b \\ \qquad\quad x \geq 0\,. \end{array}} \tag{LP}$$

(LP) heißt **lineare Optimierungsaufgabe** in kanonischer Form. Zur Herleitung der Kuhn-Tucker-Bedingungen schreiben wir $\mathscr{F}$, die zulässige Menge, wieder um:

$$\mathscr{F} = \{x | Ax = b\,,\ Gx \leq 0\} \quad \text{mit} \quad G = -I\,.$$

Wegen $\nabla f = c$ erhalten wir die notwendigen Bedingungen

$$\begin{aligned} c + A^\top \lambda + G^\top \mu &= 0 \\ \mu \geq 0\,,\ \langle Gx, \mu \rangle &= 0 \end{aligned}$$

also $c = \mu - A^\top \lambda$, und wenn wir $y = -\lambda$ setzen

$$\boxed{A^\top y + \mu = c\,,\ \mu \geq 0\,,\ \langle \mu\,, x \rangle = 0\,.} \tag{5.80}$$

Beachte, dass (LP) eine konvexe Aufgabe ist. Aus unserer bisherigen Theorie folgt

Satz 5.6.1 *$\tilde{x}$ ist genau dann (globales) Minimum von* (LP)*, wenn $\mu \geq 0$ und $y \in \mathbb{R}^m$ mit* (5.80) *existieren.*

Bemerkung: Hier kann man die Lagrangeschen Multiplikatoren als Lösung einer dualen Optimierungsaufgabe erhalten. Diese konstruiert man leicht mit folgendem Trick als Lagrange-duale Aufgabe:

$$\begin{aligned} L(x,\lambda) &:= f(x) + \langle \lambda, Ax-b\rangle \\ &= \langle c,x\rangle + \langle \lambda, Ax-b\rangle . \end{aligned}$$

Es ist sinnvoll, die "einfachen" Restriktionen $x \geq 0$ nicht in die Lagrange-Funktion aufzunehmen. Dann folgt

$$(\mathrm{LP}) \quad \Leftrightarrow \quad \min_{x\geq 0} \Big(\underbrace{\max_{\lambda\in\mathbb{R}^m} (f(x) + \langle \lambda, Ax-b\rangle)}_{\text{falls } Ax-b\neq 0, \text{ so bekommt man max} = +\infty \text{ als Ergebnis,}} \Big)$$

$$\text{und diese } x \text{ fallen bei der Minimierung heraus.}$$

Das **Dualproblem** (DP) erhält man durch Vertauschen von min und max:

$$\begin{aligned} &\max_{\lambda\in\mathbb{R}^m} \Big(\min_{x\geq 0} f(x) + \langle \lambda, Ax-b\rangle \Big) \\ &= \max_{\lambda\in\mathbb{R}^m} \Big(\underbrace{\min_{x\geq 0} \langle c + A^\top \lambda, x\rangle}_{\begin{array}{l} = -\infty, \text{ falls } c + A^\top\lambda \not\geq 0 \\ = 0, \quad \text{falls } c + A^\top \lambda \geq 0. \end{array}} - \langle b, \lambda\rangle \Big) \end{aligned}$$

$\Rightarrow$

$$(\mathrm{DP}) \quad \Leftrightarrow \quad \boxed{\begin{array}{l} \max_{\lambda\in\mathbb{R}^m} -\langle b,\lambda\rangle \\ \text{bei} \quad A^\top\lambda + c \geq 0 \end{array}}$$

Umformulierung: Setze $A^\top\lambda + c = \mu \geq 0$, $\lambda := -y$, dann erhält man – wie man beweisen kann – den Multiplikator y als Lösung der Aufgabe

$$(\mathrm{DP}) \quad \begin{cases} \max \langle b, y\rangle \\ \text{bei} \quad A^\top y + \mu = c, \;\; \mu \geq 0. \end{cases}$$

Das ist genau die Form von Satz 5.6.1. Somit kann $y = -\lambda$ also Lösung der dualen Aufgabe (DP) bestimmt werden.

6 Probleme mit nichtlinearen Restriktionen – Theorie

6.1 Grundlagen

Die vorliegenden Optimalitätsbedingungen erster Ordnung gestatten uns nun die Behandlung des allgemeinsten Fall von Aufgaben unserer Vorlesung, nämlich voll nichtlinearer Aufgaben

des Typs

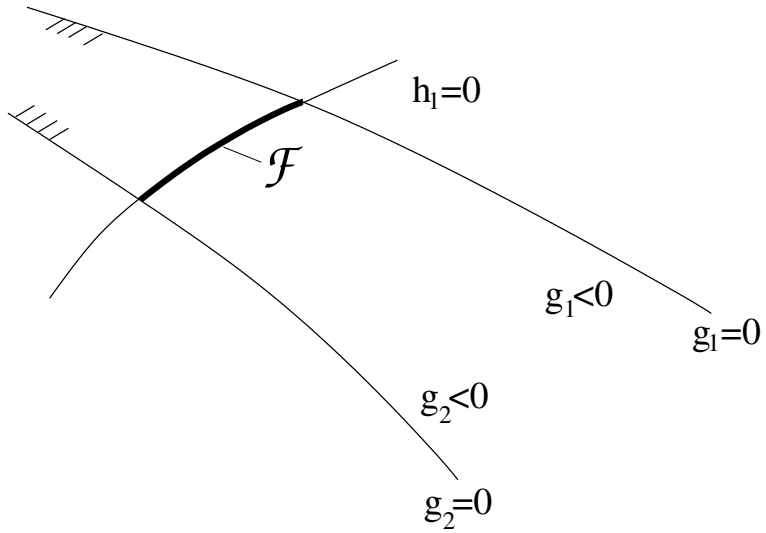
$$\begin{aligned} &\min f(x) \\ &h_i(x) = 0 \quad i = 1,...,m \\ &g_j(x) \le 0 \quad j = 1,...,p. \end{aligned} \tag{PNU}$$

Dabei seien $f : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$, $g_j : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. In vektorieller Form lautet die Aufgabe mit $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^p$,

$$\boxed{\min f(x) \;\text{ bei }\; h(x) = 0,\; g(x) \le 0.}$$

Hier gilt $\mathscr{F} = \{x \in \mathbb{R}^n \,|\, h(x) = 0\,,\; g(x) \le 0\}$.

Illustration für $n = 2$:



6.2 Notwendige Optimalitätsbedingungen erster Ordnung

Es sei $\tilde{x}$ eine lokale Lösung von (PNU). Wir wollen nun wieder K-K-T-Sätze beweisen. Schlüsselidee ist die folgende:

Löst $\tilde{x}$ (PNU), so sollte auch $\tilde{x}$ das Problem lösen, das durch *Linearisierung an der Stelle* $\tilde{x}$ entsteht:

$$\boxed{\begin{aligned} &\min f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) \\ &\text{bei } \; h(\tilde{x}) + h'(\tilde{x})(x - \tilde{x}) = 0 \\ &\qquad g(\tilde{x}) + g'(\tilde{x})(x - \tilde{x}) \le 0. \end{aligned}} \quad \textbf{(Linearisiertes Problem)} \tag{6.1}$$

Sind h, g affin-linear, dann gilt das wirklich! Denn:

$$\begin{aligned} h(x) = Ax - b \qquad \Rightarrow \quad h(\tilde{x}) + h'(\tilde{x})(x - \tilde{x}) \quad &= A\tilde{x} - b + A(x - \tilde{x}) \\ &= Ax - b \end{aligned}$$

Analog folgt mit $g(x) = Gx - r$

$$g(\tilde{x}) + g'(\tilde{x})(x - \tilde{x}) = Gx - r.$$

Obige Nebenbedingungen sind dann äquivalent zu $Ax = b$, $Gx \leq r$, d. h. $x \in \mathscr{F}$.

Für $x \in \mathscr{F}$ galt aber die Variationsungleichung

$$\begin{aligned} & f'(\tilde{x})(x - \tilde{x}) \geq 0 && \forall x \in \mathscr{F}, \\ \Leftrightarrow \quad & f'(\tilde{x})\tilde{x} \leq f'(\tilde{x})x && \forall x \in \mathscr{F} \\ \Leftrightarrow \quad & \min_{x \in \mathscr{F}} f'(\tilde{x})x = f'(\tilde{x})\tilde{x}. \end{aligned}$$

Im nichtlinearen Fall löst $\tilde{x}$ leider nicht immer die linearisierte Aufgabe! Deshalb wird eine Zusatzvoraussetzung nötig sein.

Beispiel 6.2.1 *Ein Gegenbeispiel ist die Aufgabe*

$$\begin{aligned} \min f(x_1, x_2) &= x_1 \\ g_1(x) &= -x_1^3 + x_2 \leq 0 \\ g_2(x) &= -x_2 \leq 0. \end{aligned}$$

Der obige zulässige Bereich ist in ≤ 0-Form geschrieben. Einprägsamer ist $x_2 \geq 0$, $x_2 \leq x_1^3$.

Die Lösung ist offenbar: $\tilde{x}_1 = \tilde{x}_2 = 0$

Linearisierte Aufgabe (Tangente an den Graphen von $x_2 = x_1^3$ anlegen!)

$$\min x_1 \quad \textit{bei} \quad x_2 \leq 0, \ -x_2 \leq 0.$$

also

$$\boxed{\min x_1 \quad \textit{bei} \quad x_2 = 0.}$$

Das Infimum ist offenbar $-\infty$, die Aufgabe hat **keine Lösung!** *Ausgangsproblem und linearisiertes Problem verhalten sich völlig unterschiedlich.*

Solche Phänomene müssen ausgeschlossen werden, denn Linearisierung spielt sowohl in der theoretischen Begründung als auch für die numerischen Verfahren eine wichtige Rolle. Deshalb hat man sich ausführlich mit Voraussetzungen an $\mathscr{F}$ befasst, welche so etwas verhindern. Das sind **Regularitätsbedingungen** (*Constraint qualifications*). Auf Grund ihrer Wichtigkeit wollen wir diese ausführlich behandeln.

Zunächst sieht man, dass man (bei gegebenem $\tilde{x}$) im linearisierten Problem (6.1) die inaktiven Restriktionen weglassen kann – sie spielen lokal (d. h. in einer Umgebung von $\tilde{x}$) keine Rolle

Wir betrachten die aktiven Restriktionen $J(\tilde{x}) = \{1 \leq j \leq p \,|\, g_j(\tilde{x}) = 0\}$ und die linearisierte Aufgabe

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} \langle \nabla f(\tilde{x}), x - \tilde{x} \rangle \\ & h'(\tilde{x})(x - \tilde{x}) = 0 \\ & \langle \nabla g_j(\tilde{x}), x - \tilde{x} \rangle \leq 0 \quad \forall j \in J(\tilde{x}). \end{aligned} \tag{6.2}$$

Lemma 6.2.1 *$\tilde{x}$ löst (6.1) genau dann, wenn $\tilde{x}$ das Problem (6.2) löst.*

Das ist eigentlich klar, weil die inaktiven Restriktionen keine Rolle spielen. Der Beweis findet sich z.B in [1, Lemma 7.2.6].

Folgerung 6.2.1 *Die Richtung $d = 0$ aus (6.2) löst das Problem*

$$\begin{aligned} &\min_{d\in\mathbb{R}^n} \langle \nabla f(\tilde{x}), d\rangle \\ &h'(\tilde{x})d = 0 \\ &\langle \nabla g_j(\tilde{x}), d\rangle \le 0 \quad \forall\, j \in J(\tilde{x})\,. \end{aligned} \tag{6.3}$$

Der Beweis ist trivial, denn es kann keine zulässige Abstiegsrichtung geben, sonst gälte in (6.2) $\inf = -\infty$.

Definition 6.2.1 (Linearisierungskegel) *Die Menge*

$$L(\mathscr{F},\tilde{x}) = \{d \,|\, h'(\tilde{x})d = 0, \quad \langle \nabla g_j(\tilde{x}), d\rangle \le 0 \quad \forall j \in J(\tilde{x})\}$$

heißt Linearisierungskegel von $\mathscr{F}$ in $\tilde{x}$.

Nun wollen wir zunächst annehmen, dass die lokale Lösung $\tilde{x}$ von (PNU) auch Lösung der linearisierten Aufgabe ist (was, wie wir wissen, schiefgehen kann). Dann folgt also

$$\langle \nabla f(\tilde{x}), d\rangle \ge 0 \quad \forall d \in L(\mathscr{F},\tilde{x})$$

bzw.

$$-\nabla f(\tilde{x}) \in L(\mathscr{F},\tilde{x})^*.$$

Wir haben bereits früher bewiesen, dass daraus folgt

$$-\nabla f(\tilde{x}) = \sum_{i=1}^{m} \lambda_i a^i + \sum_{j\in J(\tilde{x})} \mu_j g^j\,,\ \mu_j \ge 0$$

mit (angepasst an unseren Fall)

$$a^i := \nabla h_i(\tilde{x})\,,\ g^j := \nabla g_j(\tilde{x})\,.$$

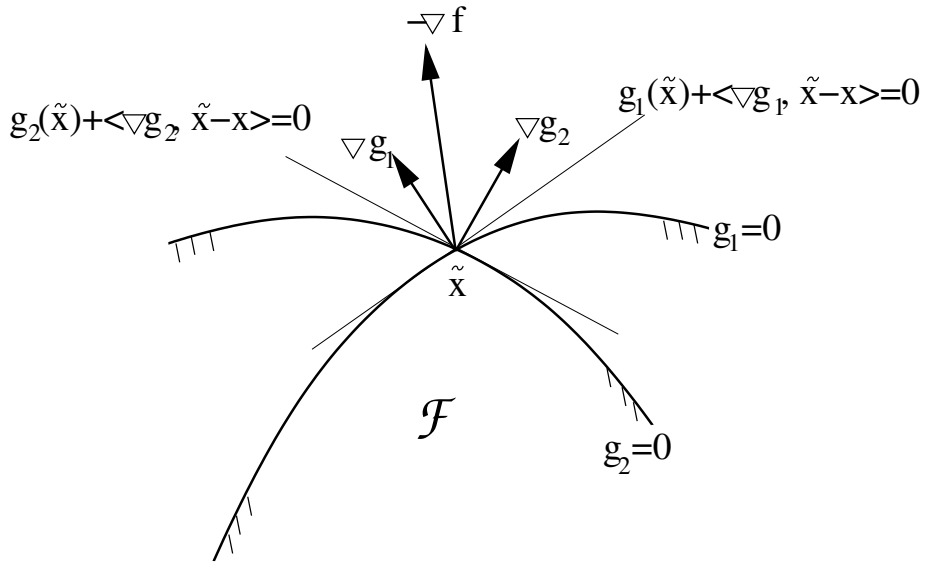
Folgerung 6.2.2 *Ist die lokale Lösung $\tilde{x}$ von (PNU) auch Lösung der linearisierten Aufgabe, dann existieren Vektoren $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$, $\mu \ge 0$ mit*

$$0 = \nabla f(\tilde{x}) + h'(\tilde{x})^\top \lambda + g'(\tilde{x})^\top \mu \tag{6.4}$$

$$\mu \ge 0 \quad \textit{und} \quad \langle \mu, g(\tilde{x})\rangle = 0 \quad \forall j. \tag{6.5}$$

Das sind bereits die sogenannten Karush-Kuhn-Tucker-Bedingungen, wie wir später noch in Form eines Satzes einführen werden.

Geometrische Illustration im Fall von reinen Ungleichungsrestriktionen:



Ist $\tilde{x}$ lokale Lösung, dann muss $-\nabla f$ in dem von $\nabla g_1, \nabla g_2$ aufgespannten Kegel liegen.

Definition 6.2.2 *λ und μ mit den Eigenschaften (6.4-6.5) heißen Lagrangesche Multiplikatoren zu $\tilde{x} \in \mathscr{F}$.*

Somit kann unsere obige Folgerung auch so formuliert werden:

Satz 6.2.1 *Die Funktionen f, g, h seien differenzierbar an der Stelle $\tilde{x}$. Löst $\tilde{x} \in \mathscr{F}$ die linearisierte Aufgabe (6.1), dann existieren Lagrangesche Multiplikatoren λ, μ zu $\tilde{x}$.*

Bemerkungen:

a) Dieser Satz war insofern schon klar, weil (6.1) eine differenzierbare Aufgabe mit linearen Restriktionen ist, und dafür kennen wir ja schon die Lagrangesche Multiplikatorenregel. Wie bisher, können wir diese Regel wie folgt aufschreiben:

$$\mathscr{L} = \mathscr{L}(x, \lambda, \mu) := f(x) + \langle h(x), \lambda \rangle + \langle g(x), \mu \rangle .$$

Diese Funktion nennen wir wieder *Lagrange-Funktion*.

Dann gilt

$$\boxed{\nabla_x \mathscr{L}(\tilde{x}, \lambda, \mu) = 0 , \ \mu \geq 0 , \ \langle g(\tilde{x}), \mu \rangle = 0 .}$$

b) Sind f und $\mathscr{F}$ konvex, dann ist diese Optimalitätsbedingung auch hinreichend für Optimalität.

Wir haben Satz 6.2.1 aber nur unter der Bedingung gezeigt, dass $\tilde{x}$ die linearisierte Aufgabe löst. Wann gilt das? Dazu müssen wir etwas weiter ausholen, vorher aber rechnen wir zur Auflockerung ein weiteres Beispiel (noch mit linearen Restriktionen).

Beispiel 6.2.2

$$\begin{aligned} & \min f(x) = \|x\|^2,\ x \in \mathbb{R}^3 \\ \textit{bei} \quad & 2x_1 - x_2 + x_3 \le 5 \\ & x_1 + x_2 + x_3 = 3\,. \end{aligned}$$

Es existieren Lagrangesche Multiplikatoren, da die Restriktionen linear sind.

$$\mathscr{L} = x_1^2 + x_2^2 + x_3^3 + \mu(2x_1 - x_2 + x_3 - 5) + \lambda(x_1 + x_2 + x_3 - 3)$$

$$\begin{aligned} \mathscr{L}x_1 = 0 \quad &\Rightarrow \quad 2x_1 + 2\mu + \lambda = 0 \\ \mathscr{L}x_2 = 0 \quad &\Rightarrow \quad 2x_2 - \mu + \lambda = 0 \\ \mathscr{L}x_3 = 0 \quad &\Rightarrow \quad 2x_3 + \mu + \lambda = 0\,. \end{aligned} \tag{6.6}$$

Außerdem muss gelten $\mu(2x_1 - x_2 + x_3 - 5) = 0$, $\mu \ge 0$.

Annahme $\mu > 0$**:** *Dann muss gelten* $2x_1 - x_2 + x_3 = 5$. *Aus (6.6) folgt*

$$\begin{aligned} x_1 &= -\mu - \frac{\lambda}{2} \\ x_2 &= +\frac{\mu}{2} - \frac{\lambda}{2} \\ x_3 &= -\frac{\mu}{2} - \frac{\lambda}{2}\,. \end{aligned} \tag{6.7}$$

Einsetzen in die aktiven Nebenbedingungen

$$5 = 2\left(-\mu - \frac{\lambda}{2}\right) - \left(\frac{\mu}{2} - \frac{\lambda}{2}\right) + \left(-\frac{\mu}{2} - \frac{\lambda}{2}\right) = -3\mu - \lambda$$

$$3 = \left(-\mu - \frac{\lambda}{2}\right) + \left(\frac{\mu}{2} - \frac{\lambda}{2}\right) + \left(-\frac{\mu}{2} - \frac{\lambda}{2}\right) = -\mu - \frac{3}{2}\lambda$$

$$\left.\begin{aligned} -\frac{15}{2} &= \frac{9}{2}\mu + \frac{3}{2}\lambda \\ 3 &= -\mu - \frac{3}{2}\lambda \end{aligned}\right\} \quad \Rightarrow -\frac{9}{2} = \frac{7}{2}\mu$$

$$\mu = -\frac{9}{7} \quad \textit{Widerspruch zu} \quad \mu > 0\,.$$

Also nehmen wir $\mu = 0$ **an.** *Dann müssen wegen (6.7) alle* x_i *gleich sein und aus* $x_1 + x_2 + x_3 = 3$ *folgt* $x_1 = x_2 = x_3 = 1$, $\lambda = -2$.

$\Rightarrow$ $x = (1,1,1)^\top$ *erfüllt die Optimalitätsbedingungen "kritischer Punkt".*

Dieser Vektor x ist auch eine Lösung, denn

- $f(x) \to \infty$, $\|x\| \to \infty$.
- *Daher können wir das Minimum in der beschränkten Menge*

$$\{x \in \mathscr{F} \mid \|x\|^2 \le f(1,1,1)\}$$

suchen.

- *Nach dem Satz von Weierstraß existiert das Minimum unserer Aufgabe.*
- *Dieses* **muss** *die notwendigen Bedingungen erfüllen, also ist obiges x die Lösung.*

Wir setzen nun unsere theoretischen Untersuchungen fort.

Definition 6.2.3 (Tangentialkegel) *Es sei $S \subset \mathbb{R}^n$ eine beliebige Menge, wir denken dabei an $\mathscr{F}$. Ein Vektor $d \in \mathbb{R}^n$ heißt* Tangentialrichtung *an S in x, $x \in S$, wenn es Folgen $\{x^k\} \subset S$ mit $x^k \to x$, $k \to \infty$, und $\{t_k\} \subset \mathbb{R}$, $t_k > 0$, mit $t_k \downarrow 0$ gibt, so dass*

$$\lim_{k\to\infty} \frac{x^k - x}{t_k} = d\,.$$

Die Menge aller Tangentialrichtungen an S in x heißt Tangentialkegel $T(S,x)$.

Es gilt immer $0 \in T(S,x)$. Außerdem ist $T(S,x)$ ein Kegel. Man kann folgende Eigenschaften zeigen (vgl. [1, Lemma 7.2.10,11,13]), dabei sei $x \in S$:

- $T(S,x)$ ist abgeschlossen
- $T(S,x) \subset clK(S,x)$
- S konvex $\Rightarrow\ T(S,x) = clK(S,x)$.
- Im Falle linearer Gleichungen und Ungleichungen ist $K(\mathscr{F},x)$ abgeschlossen und konvex, d. h. es gilt

$$T(\mathscr{F},x) = K(\mathscr{F},x) = L(\mathscr{F},x).$$

Im nichtlinearen Fall gilt das nicht immer, aber es gilt zumindest immer bei Differenzierbarkeit ohne weitere Voraussetzungen die folgende Aussage:

Lemma 6.2.2 *Es sei $\mathscr{F}$ die zulässige Menge von (PNU) und die Funktionen g und h seien an der Stelle x differenzierbar. Dann gilt*

$$\boxed{T(\mathscr{F},x) \subset L(\mathscr{F},x).}$$

Beweis: Siehe [1, Lemma 7.2.15]. Man zeigt: Gilt $d \in T(\mathscr{F},x)$ dann

$$h'(x)d = 0 \quad \text{sowie} \quad \nabla g^i(x)d \le 0 \quad \forall i \in J(x)\,.$$

Das heißt aber gerade $d \in L(\mathscr{F},x)$. □

Beispiel 6.2.3 *Wir betrachten als zulässige Menge den Einheitskreis,*

$$\mathscr{F} = \{\mathbf{x} \in \mathbb{R}^2 \,|\, \mathbf{g(x)} = \mathbf{x_1^2} + \mathbf{x_2^2} - \mathbf{1} \le \mathbf{0}\}$$

a) $\mathbf{x} = \mathbf{0}$: $\quad \Rightarrow g(x) < 0$ *ist inaktiv* $\Rightarrow$

$$L(\mathscr{F},0) = \mathbb{R}^2.$$

Außerdem gilt auch

$$T(\mathscr{F},0) = \mathbb{R}^2,$$

da $x = 0$ ein innerer Punkt von $\mathscr{F}$ ist (man setze $x^k = x + t_k d$; dann folgt $x^k \in \mathscr{F}$ für kleine t_k und damit $\frac{x^k - x}{t_k} = d \quad \forall k > k_0$).

Für $x = 0$ gilt also $L(\mathscr{F},0) = T(\mathscr{F},0)$.

b) $\mathbf{x} = (\mathbf{0}, -\mathbf{1})^\top$:

$$\begin{aligned} L(\mathscr{F},x) &= \{d \,|\, g'(x)d \le 0\} \\ &= \left\{ d \,\middle|\, 2\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \le 0 \right\} \\ &= \{d \,|\, -2d_2 \le 0\} \\ &= \{d \,|\, d_2 \ge 0\}. \end{aligned}$$

Analog erkennt man aus der Geometrie von $\mathscr{F}$ die Beziehung $T(\mathscr{F},x) = \{d \,|\, d_2 \ge 0\}$ also auch hier $T(\mathscr{F},x) = L(\mathscr{F},x)$.

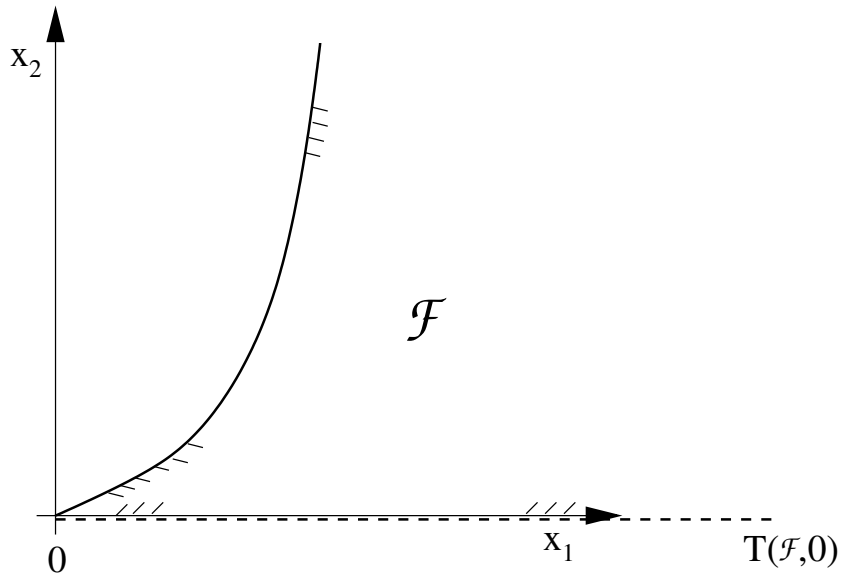
Beispiel 6.2.4 $\mathscr{F} = \{\mathbf{x} \in \mathbb{R}^2 \,|\, -\mathbf{x_1^3} + \mathbf{x_2} \le \mathbf{0},\ -\mathbf{x_2} \le \mathbf{0}\}$ *also unser Gegenbeispiel.*

Im Nullpunkt gilt mit $g_1 = -x_1^3 + x_2$, $g_2 = -x_2$,

$$\begin{aligned} L(\mathscr{F},0) &= \{d \,|\, -3 \cdot 0^2 d_1 + d_2 \le 0,\ -d_2 \le 0\} \\ &= \{d \,|\, d_2 = 0\}. \end{aligned}$$

Für den Tangentialkegel ergibt sich (Bild unten)

$$T(\mathscr{F},0) = \{d \,|\, d_1 \ge 0,\ d_2 = 0\}.$$



Der negative Teil von $\mathbb{R}$ fällt bei der Bildung von $T(\mathscr{F},0)$ weg!

Hier haben wir wirklich nur die echte Inklusion

$$T(\mathscr{F},0) \subset L(\mathscr{F},0)\,.$$

Der Vorteil des Tangentialkegels ist, dass man mit ihm einen vernünftigen Ersatz unserer nur bei Konvexität gültigen Variationsungleichung bekommt. Es gilt nämlich für beliebige Mengen S die folgende Aussage:

Satz 6.2.2 *Sei $S \subset \mathbb{R}^n$ nichtleer, $f : S \to \mathbb{R}$, $\tilde{x} \in S$ ein lokales Minimum von f in S und f in $\tilde{x}$ differenzierbar. Dann gilt*

$$\langle \nabla f(\tilde{x}), d\rangle \geq 0 \quad \forall d \in T(S,\tilde{x})\,.$$

Beweis: Wir wählen $d \in T(S,\tilde{x})$ beliebig aus. Dann gilt

$$d = \lim_{k\to\infty} \frac{x^k - \tilde{x}}{t_k}$$

mit $x^k \in S$, $x^k \to \tilde{x}$, $t_k \to 0$. Also mit $r^k \to 0$

$$\begin{aligned} t_k(d + r^k) &= x^k - \tilde{x} \\ x^k &= \tilde{x} + t_k(d + r^k)\,. \end{aligned}$$

Da $\tilde{x}$ lokales Minimum und $x^k \to \tilde{x}$, folgt für hinreichend großes k

$$\begin{aligned} & 0 \leq f(x^k) - f(\tilde{x}) = \langle \nabla f(\tilde{x}), t_k(d + r^k)\rangle + o(t_k(d + r^k))\,|\,: t_k \\ \Rightarrow\ & 0 \leq \langle \nabla f(\tilde{x}), d + r^k\rangle + \underbrace{\frac{o(t_k(d + r^k))}{t_k}} \\ & \qquad = \frac{o(t_k(d + r^k))}{|t_k(d + r^k)|} \cdot |d + r^k| \end{aligned}$$

für $k \to \infty$ strebt r^k gegen Null und auch $t_k(d + r^k)$ gegen Null. Insgesamt

$$0 \leq \langle \nabla f(\tilde{x}), d\rangle\,.$$

□

Wir kehren nun zurück zu unserer oben definierten Menge $\mathscr{F}$. Wir wissen bereits, dass

$$\langle \nabla f(\tilde{x}), d\rangle \geq 0 \quad \forall d \in T(\mathscr{F},\tilde{x})$$

gilt und hätten gern

$$\langle \nabla f(\tilde{x}), d\rangle \geq 0 \quad \forall d \in L(\mathscr{F},\tilde{x})\,,$$

denn Letzteres ergibt – wie wir gesehen haben – die Lagrangesche Multiplikatorenregel. Dazu bräuchten wir aber, dass $\tilde{x}$ die linearisierte Aufgabe löst...

Deshalb fordern wir einfach

$$\boxed{T(\mathscr{F},\tilde{x}) = L(\mathscr{F},\tilde{x}).}$$

Definition 6.2.4 (Regularität) *Ein Vektor $\tilde{x} \in \mathscr{F}$ heißt* regulär, *wenn $T(\mathscr{F},\tilde{x}) = L(\mathscr{F},\tilde{x})$ gilt. Diese Regularitätsbedingung wird auch* Bedingung von Abadie *genannt.*

Wir folgern aus dem Bisherigen:

Satz 6.2.3 (Karush-Kuhn-Tucker-Satz) *Ist $\tilde{x} \in \mathscr{F}$ regulär und lokales Minimum für (PNU), dann existieren Lagrangesche Multiplikatoren λ und μ, es gilt also*

$$\nabla f(\tilde{x}) + h'(\tilde{x})^\top \lambda + g'(\tilde{x})^\top \mu = 0,$$
$$\langle \mu, g(\tilde{x}) \rangle = 0,\ \mu \geq 0.$$

Sind zusätzlich die Gradienten der aktiven Restriktionen, d. h.

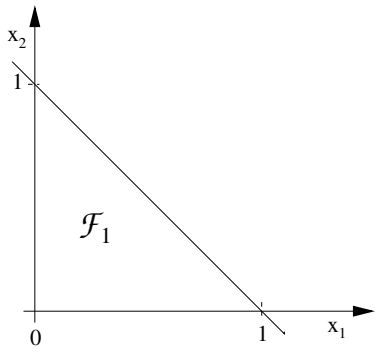
$$\nabla h_i(\tilde{x}),\ i = 1,...,m \quad \textit{sowie} \quad \nabla g_j(\tilde{x}),\ j \in J(\tilde{x}),$$

linear unabhängig, dann sind λ und μ eindeutig bestimmt.

Bemerkung: $T(\mathscr{F},x)$ hängt `nicht` von der konkreten Darstellung von $\mathscr{F}$ ab, während $L(\mathscr{F},x)$ davon abhängen kann!

Beispiel 6.2.5 *Es sei*

$$\begin{aligned} \mathscr{F}_1 = \{x | g_1(x) = x_1 + x_2 - 1 &\leq 0 \\ g_2(x) = -x_1 &\leq 0 \\ g_3(x) = -x_2 &\leq 0\ \} \end{aligned}$$



Andere Darstellung von $\mathscr{F}_1$:

$$\mathscr{F}_2 = \{x | (x_1 + x_2 - 1)^3 \leq 0,\ -x_1 \leq 0,\ -x_2 \leq 0\ \}.$$

Offenbar gilt $\mathscr{F}_1 = \mathscr{F}_2$.

Wir betrachten den Punkt $\tilde{x} = (1/2, 1/2)^\top$. Hier gilt:

$$T(\mathscr{F}_1,\tilde{x}) = T(\mathscr{F}_1,\tilde{x}) = \{d \,|\, d_1 + d_2 \leq 0\}$$
(*geometrisch klar*)

$$L(\mathscr{F}_1,\tilde{x}) = \{d \,|\, d_1 + d_2 \leq 0\} = T(\mathscr{F},\tilde{x})$$
(g_1 *war affin linear,* g_2, g_3 *inaktiv*).

$$L(\mathscr{F}_2,\tilde{x}) = \left\{ d \;\middle|\; \langle \nabla g_1(\tilde{x}), d \rangle = \left\langle \underbrace{3(\tilde{x}_1 + \tilde{x}_2 - 1)^2}_{=0} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, d \right\rangle = \langle 0, d \rangle \leq 0 \right\} = \mathbb{R}^2$$

$$\Rightarrow \quad L(\mathscr{F}_2,\tilde{x}) = \mathbb{R}^2 \supset T(\mathscr{F}_2,\tilde{x}).$$

Man kann zeigen, dass für entsprechende Optimierungsaufgaben im zweiten Fall keine Lagrangeschen Multiplikatoren existieren müssen (vgl. [1]).

Wir widmen uns nun dem Problem der Regularität. Wann gilt

$$T(\mathscr{F},\tilde{x}) = L(\mathscr{F},\tilde{x})?$$

Fall reiner Ungleichungsrestriktionen

Wir betrachten

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, g(x) \leq 0\},$$

$g : \mathbb{R}^n \to \mathbb{R}^p$ sei differenzierbar. Ein $\tilde{x} \in \mathscr{F}$ erfülle folgende **Regularitätsbedingung:**

$$\boxed{\exists \bar{d} \in \mathbb{R}^n : \langle \nabla g_j(\tilde{x}), \bar{d} \rangle < 0 \quad \forall j \in J(\tilde{x}).}$$

Die Bedingung fordert, dass eine Richtung existiert, die ins Innere der zulässigen Menge weist. Dann ist $\tilde{x}$ regulär, d. h. $T(\mathscr{F},\tilde{x}) = L(\mathscr{F},\tilde{x})$.

Beweis: Da stets $T \subset L$ gilt, ist nur die andere Inklusion zu zeigen. Dazu sei $d \in L(\mathscr{F},\tilde{x})$ beliebig gegeben. Wir setzen

$$x(t) = \tilde{x} + t(d + \alpha\bar{d})$$

mit $\alpha > 0$. Dann folgt $x(t) \in \mathscr{F}$ *für hinreichend kleine t:*

Für die inaktiven Restriktionen $j \notin J(\tilde{x})$ ist das klar, denn $g_j(\tilde{x} + t(d + \alpha\bar{d})) \to g_j(\tilde{x}) < 0, t \downarrow 0$. Bei den aktiven gilt

$$\begin{aligned} g_j(x(t)) &= \underbrace{g_j(\tilde{x})}_{=0} + \langle \nabla g_j(\tilde{x}), t(d + \alpha\bar{d}) \rangle + o(t) \\ &= t \underbrace{\langle \nabla g_j(\tilde{x}), d \rangle}_{\leq 0, \text{ da } d \in L(\mathscr{F},\tilde{x})} + t\Big\{ \alpha \underbrace{\langle \nabla g_j(\tilde{x}), \bar{d} \rangle}_{<0} + \underbrace{o(t)t^{-1}}_{\to 0} \Big\} \end{aligned}$$

Für hinreichend kleines t wird $\{...\}$ negativ, daher (es gibt nur endlich viele $j \in J(\tilde{x})$)

$$g_j(x(t)) \leq 0 \quad \forall t \in [0,\bar{t}] \quad \forall j \in J(\tilde{x}).$$

Nun ist klar, was passieren muss: Wir setzen

$$t_k = \frac{1}{k},\ x^k = x(t_k).$$

Dann gilt $x^k \in \mathscr{F}\ \forall k > k_0$ und nach Konstruktion

$$\frac{x^k - \tilde{x}}{t_k} = d + \alpha\bar{d}.$$

Daraus folgt

$$d + \alpha\bar{d} \in T(\mathscr{F},\tilde{x}) \quad \forall \alpha > 0.$$

$T(\mathscr{F},\tilde{x})$ ist abgeschlossen. Damit gilt auch $d = \lim\limits_{\alpha \downarrow 0} d + \alpha\bar{d} \in T(\mathscr{F},\tilde{x})$. □

Man kann zeigen, dass für die affin-linearen Restriktionen sogar $\langle \nabla g_j, \bar{d} \rangle \leq 0$ ausreicht. Insgesamt folgt dann

Satz 6.2.4 *Sei $\tilde{x} \in \mathscr{F}$ und g differenzierbar in $\tilde{x}$. Gibt es ein $\bar{d} \in \mathbb{R}^n$ mit*

$$\left.\begin{array}{ll} \langle \nabla g_j(\tilde{x}), \bar{d} \rangle \leq 0 & \text{für alle affin-linearen aktiven Restriktionen} \\ \langle \nabla g_j(\tilde{x}), \bar{d} \rangle < 0 & \text{für alle anderen aktiven Restriktionen,} \end{array}\right\} \tag{6.8}$$

dann ist $\tilde{x}$ regulär.

Hinreichend für (6.8) ist die folgende Forderung:

Linearisierte Slater-Bedingung:

$$\boxed{\exists \bar{d} \in \mathbb{R}^n : g(\tilde{x}) + g'(\tilde{x})\bar{d} < 0.} \tag{6.9}$$

Diese Bedingung heißt auch *lokale Slater-Bedingung* (lokal, weil von $\tilde{x}$ abhängig).

Im Falle konvexer Funktionen g_j ist folgende Bedingung hinreichend für (6.8):

Slater-Bedingung:

$$\boxed{\exists \bar{v} \in \mathscr{F} : g_j(\bar{v}) < 0 \quad \forall j \in J(\tilde{x}), \quad \text{falls } g_j \text{ nichtlinear}}$$

Beweis: Falls $\bar{v} = \tilde{x}$, dann gibt es unter Voraussetzung der obigen Slater-Bedingung keine aktiven und gleichzeitig nichtlinearen Restriktionen. Deshalb ist (6.8) mit $\bar{d} = 0$ erfüllt. Anderenfalls gilt $\bar{v} - \tilde{x} \neq 0$ und aus der Konvexität folgt

$$\begin{aligned} g_j(\bar{v}) - g_j(\tilde{x}) &\geq \langle \nabla g_j(\tilde{x}), \underbrace{\bar{v} - \tilde{x}}_{d} \rangle \\ \text{also} \quad \langle \nabla g_j(\tilde{x}), d \rangle &\leq \underbrace{g_j(\bar{v})}_{<0} - \underbrace{g_j(\tilde{x})}_{=0 \; \forall j \in J(\tilde{x})} < 0 \quad \forall j \in J(\tilde{x}). \end{aligned}$$

.

Auch diese Bedingung hängt noch von der unbekannten Lösung $\tilde{x}$ ab und kann daher nicht a priori nachgeprüft werden. Folgende Bedingung ist von $\tilde{x}$ unabhängig und ist deshalb am praktikabelsten: Es ist *die*

Slater-Bedingung:

$$\boxed{\exists \bar{v} : \quad g_j(\bar{v}) < 0 \quad \forall j.}$$

Diese ist hinreichend für Regularität im Falle konvexer Funktionen g_j.

Korollar 6.1 *Sind die Gradienten der aktiven Restriktionen, $\nabla g_j(\tilde{x})$, $j \in J(\tilde{x})$, linear unabhängig, dann ist $\tilde{x}$ regulär.*

Beweis: Das System $\langle \nabla g_j(\tilde{x}), d \rangle = b_j$, $j \in J(\tilde{x})$, hat dann für alle b_j eine Lösung. Insbesondere für $b_j < 0$. □

Regularität bei Gleichungsrestriktionen

Nun sei

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, h(x) = 0\}.$$

Satz 6.2.5 *Die Gradienten $\nabla h_i(\tilde{x})$, $i = 1,...,m$, seien linear unabhängig. Dann ist $\tilde{x}$ regulär.*

Der Beweis ist eine Anwendung des Satzes über implizite Funktionen. Er ist in vielen Büchern zu finden, z.B. in [1, Satz 7.2.26], deshalb lassen wir ihn weg. Dieser Satz wurde auch im Kurs Analysis II bewiesen.

Andere Formulierung dieser Bedingung:

$$\boxed{h'(\tilde{x}) \quad \text{sei surjektiv}}$$

Regularität bei Gleichungs- und Ungleichungsrestriktionen

Jetzt ist

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, h(x) = 0\,,\; g(x) \leq 0\}\,.$$

Satz 6.2.6 *Sei $\tilde{x} \in \mathscr{F}$, h und g stetig differenzierbar. Die Gradienten $\nabla h_i(\tilde{x})$ seien linear unabhängig, und es existiere ein $\bar{d} \in \mathbb{R}^n$ mit*

$$\boxed{h'(\tilde{x})\bar{d} = 0 \quad \textit{und} \quad \nabla g_j(\tilde{x})\bar{d} < 0 \quad \forall j \in J(\tilde{x})\,.} \tag{6.10}$$

Dann ist $\tilde{x}$ regulär.

Wir verzichten auf den Beweis. Diese Regularitätsbedingung nennt man **Mangasarian-Fromovitz-Bedingung***, kurz MFCQ für "Mangasarian Fromovitz constraint qualification".*

Bemerkung: Wie oben ist dafür wiederum hinreichend:

$$\nabla h_i(\tilde{x})\,,\; i = 1,...,m\,,\; \nabla g_j(\tilde{x})\,,\; j \in J(\tilde{x}), \quad \text{sind linear unabhängig}\,. \tag{6.11}$$

Einige kleine Beispiele sollen die Anwendung von Regularitätsbedingungen illustrieren:

Beispiel 6.2.6

a) $\min \mathbf{f}(\mathbf{x}), |\mathbf{x}|^{\mathbf{2}} \leq \mathbf{1}$.

Der zulässige Bereich ist

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, g(x) = |x|^2 - 1 \leq 0\}.$$

Die Funktion g ist konvex und differenzierbar und $\bar{v} = 0$ erfüllt

$$g(\bar{v}) = -1 < 0.$$

Damit ist die Slater-Bedingung erfüllt und jedes $x \in \mathscr{F}$ regulär.

b) $\min \mathbf{f(x)}, \quad \mathbf{e^{\sum_1^n x_i} = 3}, \quad \mathbf{|x|^2 = 4}$.

Das ist ein Problem mit Gleichungsrestriktionen,

$$h_1(x) = e^{\sum x_i} - 3\,,\ h_2(x) = |x|^2 - 4$$

$$\nabla h_1(x) = \left(e^{\sum x_i}\right) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\,,\ \nabla h_2(x) = 2x\,.$$

Diese Gradienten sind – unabhängig von x – stets linear unabhängig, wenn x zulässig ist; denn dann gilt wegen $|x|^2 = 4$

$$x_1^2 + \ldots + x_n^2 = 4;$$

x müsste bei linearer Abhängigkeit die Darstellung $x = \alpha(1,\ldots,1)^\top$ *haben, also*

$$\alpha^2(1+\ldots+1) = 4\,,\ \alpha^2 = \frac{4}{n} \ \Rightarrow\ \alpha = \frac{2}{\sqrt{n}}$$

$$\Rightarrow \quad x = \frac{2}{\sqrt{n}}(1,\ldots,1)^\top \Rightarrow \quad e^{\sum_1^n x_i} = e^{\frac{2n}{\sqrt{n}}} = e^{2\sqrt{n}} = 3$$

$$\Rightarrow 2\sqrt{n} = \ln 3$$

$$\underbrace{n}_{\in\mathbb{N}} = \left(\frac{1}{2}\ln 3\right)^2,$$

ein Widerspruch. Also ist jedes zulässige x regulär.

Wir halten als Zusammenfassung dieses Abschnitts fest: Ist $\tilde{x}$ eine reguläre Lösung von (PNU), dann existieren als Folgerung des Karush-Kuhn-Tucker-Satzes zugehörige Lagrangesche Multiplikatoren.

Man kann einen sehr ähnlichen, aber meist nicht so praktikablen Satz beweisen, der aber *ohne* Regularität auskommt.

Satz 6.2.7 (Satz von Fritz John) *Die Funktionen* f,g,h *seien stetig differenzierbar und* $\tilde{x}$ *eine lokale Lösung von* (PNU)*. Dann existieren Multiplikatoren* $\mu_0 \geq 0$, $\lambda \in \mathbb{R}^m$, $\mu \geq 0 \in \mathbb{R}^p$, *so dass* $(\mu_0, \lambda^\top, \mu^\top) \neq 0$ *und die Beziehungen*

$$\mu_0 \nabla f(\tilde{x}) + h'(\tilde{x})^\top \lambda + g'(\tilde{x})^\top \mu = 0,$$

$$\langle \mu, g(\tilde{x}) \rangle = 0$$

erfüllt sind.

Dieser Satz ist also immer anwendbar, nur ist seine Aussage schwach, denn im Fall $\mu_0 = 0$ tritt die zu minimierende Zielfunktion gar nicht in der Optimalitätsbedingung auf.

Bemerkung: Wüsste man $\mu_0 \neq 0$, dann könnte man durch μ_0 teilen und hätte mit $\tilde{\lambda} := \frac{1}{\mu_0}\lambda$ und $\tilde{\mu} = \frac{1}{\mu_0}\mu$ "richtige" Lagrangesche Multiplikatoren.

Oft baut man die Theorie so auf: Man beweist zuerst mit einem Trennungssatz den Fritz-John-Satz und zeigt dann: $\tilde{x}$ regulär $\Rightarrow \mu_0 \neq 0$.

Beispiel 6.2.7 (Der KKT-Satz ist nicht anwendbar, aber dafür der Fritz-John-Satz)

$$\begin{aligned} \min f(x_1,x_2) &= x_1 + x_2^2 \\ h_1(x) &= -x_1^2 + x_2 = 0 \\ h_2(x) &= x_1^2 + x_2 = 0. \end{aligned}$$

Das ist ein typisches Beispiel für einen ungeschickt formulierten zulässigen Bereich.

Offenbar gilt $\mathscr{F} = \{0\}$, also ist $\tilde{x} = 0$ die Lösung. Würde die KKT-Aussage gelten, dann hätten wir

$$\nabla f(0) + \nabla h_1(0)\lambda_1 + \nabla h_2(0)\lambda_2 = 0\,,$$

also

$$\begin{aligned} \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\lambda_1 + \begin{pmatrix}0\\1\end{pmatrix}\lambda_2 &= 0\,, \\ \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}(\lambda_1+\lambda_2) &= 0. \end{aligned}$$

Das ist wegen linearer Unabhängigkeit unmöglich. Die Behauptung des Fritz-John-Satzes gilt natürlich, denn

$$\mu_0\begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}(\lambda_1+\lambda_2) = 0$$

ist mit $\mu_0 = 0$, $\lambda_1 = -\lambda_2$ zu erfüllen.

Beispiel 6.2.8

$$x_1^2 + x_2^2 \to \textit{ extr.} \qquad \textit{bei } x_1^4 + x_2^4 = 1.$$

Der zulässige Bereich ist beschränkt, also existiert eine Lösung des Problems. Wir wenden die notwendigen Optimalitätsbedingungen an:

Erweiterte Lagrange-Funktion:

$$\begin{aligned} \mathscr{L} &= \mu_0(x_1^2 + x_2^2) + \lambda(x_1^4 + x_2^4 - 1) \\ \frac{\partial \mathscr{L}}{\partial x_1} = 0 \quad &\Leftrightarrow \quad 2\mu_0 x_1 + 4\lambda x_1^3 = 0 \\ \frac{\partial \mathscr{L}}{\partial x_2} = 0 \quad &\Leftrightarrow \quad 2\mu_0 x_2 + 4\lambda x_2^3 = 0. \end{aligned}$$

Falls $\mu_0 = 0$, so gilt laut Fritz-John-Satz $\lambda \neq 0$ und wir erhalten $x_1^3 = 0$ sowie $x_2^3 = 0$, also keine Zulässigkeit. Daraus folgt $\mu_0 \neq 0$. Wir können o.B.d.A. $\mu_0 = \frac{1}{2}$ annehmen $\Rightarrow$

$$\begin{aligned} x_1(1 + 4\lambda x_1^2) &= 0 \\ x_2(1 + 4\lambda x_2^2) &= 0 \end{aligned}$$

Möglichkeiten:

$$\begin{aligned} x_1 = 0 \quad &\Rightarrow \quad x_2 = \pm 1 \qquad\qquad (\textit{aus Nebenbedingung}) \\ x_2 = 0 \quad &\Rightarrow \quad x_1 = \pm 1 \\ 1 + 4\lambda x_1^2 = 1 + 4\lambda x_2^2 &= 0. \end{aligned}$$

Die letzte Beziehung ergibt mit passendem λ

$$x_1^2 = x_2^2 = -\frac{1}{4\lambda} \quad \Rightarrow \quad |x_1| = |x_2|.$$

Eingesetzt in $x_1^4 + x_2^4 = 1$ *folgt* $2|x|^4 = 1$*, also*

$$|x_1| = |x_2| = 2^{-1/4}.$$

Was sind all diese Punkte wert? Nach dem Satz von Weierstraß existieren das Minimum und das Maximum der Zielfunktion im zulässigen Bereich. Diese müssen *die notwendigen Bedingungen erfüllen. Andere Lösungen gibt es nicht. Bei den beiden* $0, \pm 1$*-Varianten gilt*

$$f(x) = 1\,.$$

Für $|x_i| = 2^{-1/4}$ *folgt* $f(x) = 2 \cdot \dfrac{1}{\sqrt{2}} = \sqrt{2} > 1. \Rightarrow$

Maximum bei	$\lvert x_1\rvert = \lvert x_2\rvert = 2^{-1/4}$	*(4 Punkte)*
Minimum bei	$\binom{0}{\pm 1}$ *und* $\binom{\pm 1}{0}$	*(4 Punkte)*

6.3 Optimalitätsbedingungen zweiter Ordnung

Analog zu den Aufgaben mit oder ohne lineare Restriktionen kann man nun wieder mit hinreichenden Bedingungen 2. Ordnung überprüfen, ob wirklich ein lokales Minimum vorliegt. Dazu braucht man die zweite Ableitung der Lagrange-Funktion,

$$\mathscr{L}_{xx}(x, \lambda, \mu) = f''(x) + \sum_{i=1}^{m} \lambda_i h_i''(x) + \sum_{j=1}^{m} \mu_j g_j''(x)\,.$$

Wir beweisen den folgenden Satz nicht. Der Beweis findet sich zum Beispiel im Buch von W. Alt.

Satz 6.3.1 *Es gelte* $f, h, g \in C^2$*,* $\tilde{x}$ *sei regulär, und* $\lambda, \mu \geq 0$ *seien die entsprechenden Lagrangeschen Multiplikatoren. Es existiere ein* $\alpha > 0$*, so dass für alle* $d \in L(\mathscr{F}, \tilde{x})$ *mit der zusätzlichen Eigenschaft* $\langle \nabla f(\tilde{x}), d \rangle = 0$ *die Beziehung*

$$d^\top \mathscr{L}_{xx}(\tilde{x}, \lambda, \mu)\, d \geq \alpha \|d\|^2 \tag{6.12}$$

mit einem positiven α *gilt. Dann existieren Konstanten* $\rho, \beta > 0$*, so dass die* quadratische Wachstumsbedingung

$$f(x) \geq f(\tilde{x}) + \beta \|x - \tilde{x}\|^2$$

für alle $x \in \mathscr{F} \cap B(\tilde{x}, \rho)$ *erfüllt ist. Damit ist* $\tilde{x}$ *striktes lokales Minimum von* (PNU).

Diese Bedingung kann man etwas anders aufschreiben: $\tilde{x}$ erfüllt nach Voraussetzung die Kuhn-Tucker-Bedingungen mit Multiplikatoren λ und μ. Folglich

$$\begin{aligned} \nabla f(\tilde{x}) &= -h'(\tilde{x})^\top \lambda - g'(\tilde{x})^\top \mu\,, \\ \text{also} \quad \langle \nabla f(\tilde{x}), d\rangle = 0 \quad &\Leftrightarrow \quad \langle -h'(\tilde{x})^\top \lambda, d\rangle - \langle g'(\tilde{x})^\top \mu, d\rangle \qquad\qquad = 0 \end{aligned}$$

Also

$$\langle \lambda, h'(\tilde{x})d\rangle + \langle g'(\tilde{x})d, \mu\rangle = 0\,. \tag{6.13}$$

Für $d \in L(\mathscr{F},\tilde{x})$ haben wir automatisch $h'(\tilde{x})d = 0$. Außerdem gilt $\langle \nabla g_j(\tilde{x}), d\rangle \le 0$ für alle $j \in J(\tilde{x})$ (aktive Restriktionen). Wegen (6.13) ist noch

$$\langle \nabla g_j(\tilde{x}), d\rangle \mu_j = 0 \quad \forall j = 1, ..., p$$

zu fordern. Für die inaktiven Restriktionen gilt das wegen $\mu_j = 0$ automatisch; für die aktiven mit $\mu_j = 0$ auch (also keine zusätzliche Bedingung). Somit bleiben noch die aktiven mit $\mu_j > 0$. Hier muss deshalb gelten $\langle \nabla g_j(\tilde{x}), d\rangle = 0$.

$\Rightarrow$

Äquivalente Form der hinreichenden Bedingungen:

Satz 6.3.2 *Der Vektor $\tilde{x} \in \mathscr{F}$ genüge den Karush-Kuhn-Tucker-Bedingungen und der Definitheitsbedingung (6.12) für alle d mit*

$$\begin{aligned} h'(\tilde{x})d &= 0 \\ \langle \nabla g_j(\tilde{x}), d\rangle &= 0 \quad \forall j \in J(\tilde{x}) \quad \textit{mit} \quad \mu_j > 0 \quad \text{(streng aktive Restriktionen)} \\ \langle \nabla g_j(\tilde{x}), d\rangle &\le 0 \quad \forall j \in J(\tilde{x}) \quad \textit{mit} \quad \mu_j = 0. \end{aligned}$$

Dann ist die quadratische Wachstumsbedingung von Satz 6.3.1 erfüllt und $\tilde{x}$ striktes lokales Minimum.

7 Probleme mit linearen Restriktionen-Verfahren

7.1 Quadratische Optimierungsprobleme

Am einfachsten laufen die Dinge (wie generell) bei reinen Gleichungsrestriktionen.

7.1.1 Aufgaben mit Gleichungsrestriktionen

Wir betrachten wieder die Aufgabe

$$\boxed{\begin{array}{l} \min\limits_{x\in\mathbb{R}^n} \dfrac{1}{2}\langle x, Qx\rangle + \langle q, x\rangle \\ \text{bei} \quad Ax = b. \end{array}} \tag{QG}$$

Q : symmetrisch, (n,n), $A : (m,n)$, $m \leq n$.

Voraussetzungen: A habe vollen Rang m und es sei

$$d^\top Qd \geq \alpha \|d\|^2 \qquad \forall d \in \ker A.$$

Damit hat (QG) genau eine Lösung $\tilde{x}$ und genau einen zugehörigen Lagrangeschen Multiplikator $\tilde{\lambda}$ (Satz 5.4.6). Beide erfüllen zusammen das System

$$\mathscr{A}\begin{pmatrix}\tilde{x}\\ \lambda\end{pmatrix} = \begin{pmatrix}-q\\ b\end{pmatrix} \quad \text{mit} \quad \mathscr{A} = \begin{pmatrix}Q & A^\top\\ A & 0\end{pmatrix}. \tag{7.14}$$

$\mathscr{A}$ ist invertierbar (Lemma 5.4.1). Damit braucht man „nur“ $\mathscr{A}^{-1}$ zu berechnen. *Das ist aber in der Regel zu teuer!*

Bessere Idee: Bestimmen einer Nullraum-Matrix zu A und nachfolgende Elimination der Nebenbedingung $Ax = b$. Dazu werden 3 Schritte ausgeführt:

1. Nullraum-Matrix bestimmen
2. $\tilde{x}$ als Lösung eines unrestringierten Optimierungsproblems berechnen
3. λ ausrechnen.

Schritt 1: *QR-Zerlegung von* $A^\top$

Finde unitäre Matrix H und obere Dreiecksmatrix R mit

$$\boxed{HA^\top = \begin{pmatrix}R\\ 0\end{pmatrix}} \qquad \begin{array}{l}H : (n,n)\\ R : (m,m).\end{array}$$

Wir spalten H wie folgt auf:

$$H = \begin{pmatrix}Y^\top\\ Z^\top\end{pmatrix} \begin{array}{l}\}\, m \text{ Zeilen}\\ \}\, n-m \text{ Zeilen.}\end{array}$$

Wir zeigen, dass Z eine Nullraummatrix ist. Die Spalten von Y bilden die ersten m Zeilen von H, die von Z die letzten $(n-m)$ Zeilen.

Sowohl Y als auch Z müssen wegen rang $H = n$ Vollrang haben. Damit spannen die Spalten von Y und Z gemeinsam den ganzen $\mathbb{R}^n$ auf. Jedes $x \in \mathbb{R}^n$ hat damit die eindeutige Darstellung

$$x = H^\top \begin{pmatrix}x_y\\ x_z\end{pmatrix} = Yx_y + Zx_z. \tag{7.15}$$

Das gilt speziell für alle $x = d \in \ker A$, und deshalb

$$0 = Ad = A(Yd_y + Zd_z) = AH^\top \begin{pmatrix}d_y\\ d_z\end{pmatrix} = (R^\top, 0)\begin{pmatrix}d_y\\ d_z\end{pmatrix} = R^\top d_y\,.$$

Daraus folgt sofort

$$d_y = 0.$$

Deshalb erhält man alle $d \in \ker A$ durch $d = Zd_z$ mit beliebigem $d_z \in \mathbb{R}^{n-m}$.

Folgerung: Das durch die QR-Zerlegung konstruierte Z **ist eine Nullraum-Matrix.**

Wir haben x in folgender Form dargestellt:

$$x = \underbrace{Yx_y}_{\Rightarrow \in (\ker A)^\perp} + \underbrace{Zx_z}_{\in \ker A}.$$

Schritt 2a: $\tilde{x}$ sei die unbekannte Lösung. Einen Teil davon können wir nun sofort berechnen:

$$\begin{aligned} & \tilde{x} = Y\tilde{x}_y + Z\tilde{x}_z \\ \Rightarrow \quad & b = A\tilde{x} = AY\tilde{x}_y + 0 \\ & b = AY\tilde{x}_y = R^\top \tilde{x}_y \end{aligned}$$

$$\boxed{R^\top \tilde{x}_y = b}$$

Somit bestimmen wir $\tilde{x}_y = (R^\top)^{-1} b$ durch einfaches Auflösen des Gleichungssystems $R^\top \tilde{x}_y = b$. Diese Lösung $\tilde{x}_y$ ist aber noch keine spezielle Lösung von $Ax = b$.

Schritt 2b: "$\mathscr{F}$ = spezielle Lösung von $Ax = b$ plus allgemeine von $Ax = 0$"

Spezielle Lösung: $Y\tilde{x}_y =: w$; wir haben $Aw = b$.

Somit gilt $x \in \mathscr{F}$ genau dann wenn

$$x = w + Zz,\ z \in \mathbb{R}^{n-m}.$$

Allgemeine Lösung: Zz, $z \in \mathbb{R}^{n-m}$.

$\Rightarrow$ **Reduziertes Problem:**

$$\min_{z \in \mathbb{R}^{n-m}} f(w+Zz) = \frac{1}{2}(w+Zz)^\top Q(w+Zz) + q^\top (Zz+w).$$

Durch Ausmultiplizieren vereinfacht sich das, wobei wir den konstanten Term $\frac{1}{2} w^\top Q w$ weglassen können:

$$f = \frac{1}{2} z^\top \underbrace{Z^\top QZ}_{\tilde{Q}} z + \langle \underbrace{Z^\top Qw, z\rangle}_{=\tilde{q}^\top z} + \langle Z^\top q, z.\rangle$$

Damit ergibt sich das unrestringierte Problem

$$\boxed{\min_{z \in \mathbb{R}^{n-m}} F(z) = \frac{1}{2} z^\top \tilde{Q} z + \tilde{q}^\top z} \qquad (\mathrm{QG})_{ur}$$

$$\begin{aligned} \text{mit} \quad & \tilde{Q} := Z^\top QZ \\ & \tilde{q} := Z^\top Qw + Z^\top q. \end{aligned}$$

Unter unseren Voraussetzungen ist $\tilde{Q}$ positiv definit, damit ist das Problem eindeutig lösbar. Notwendige Bedingung für $\tilde{z}$:

$$\nabla F(\tilde{z}) = 0 \quad \Leftrightarrow \quad \tilde{Q}\tilde{z} = -\tilde{q}\,,$$

dabei spielt $\tilde{z}$ die Rolle von $\tilde{x}_z$ oben, also

$$Z^\top QZ\tilde{x}_z = -q = -Z^\top q - Z^\top Qw = -Z^\top q - Z^\top QY\tilde{x}_y\,. \tag{7.16}$$

Wie berechnet man günstig die Lösung dieses Systems? Anstelle von $Z^\top QZ$ und $Z^\top q$ betrachten wir zunächst das Ganze für die größere, $Z^\top$ enthaltene Matrix H:

- **Ähnlichkeitstransformation**

$$\boxed{-Hq =: \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \begin{matrix} \leftarrow \mathbb{R}^m \\ \leftarrow \mathbb{R}^{m-n} \end{matrix} =: h}$$

- Berechne

$$\boxed{B := HQH^\top = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}}$$

Das ergibt

$$\begin{aligned} B_{11} &= Y^\top QY \quad B_{12} = Y^\top QZ & h_1 &= -Y^\top q \\ B_{21} &= Z^\top QY \quad B_{22} = Z^\top QZ, & h_2 &= -Z^\top q\,. \end{aligned}$$

Somit stecken in H alle in Gleichung (7.16) für $\tilde{x}_z$ interessanten Terme:

$$\underbrace{Z^\top QZ}_{B_{22}}\tilde{x}_z = \underbrace{-Z^\top q}_{h_2}\underbrace{-Z^\top QY}_{B_{21}}\tilde{x}_y$$

und liest sich nun als

$$\boxed{B_{22}\tilde{x}_z = h_2 - B_{21}\tilde{x}_y.}$$

Lösung z.B. mit Cholesky-Zerlegung.

Am Ende haben wir $\tilde{x}_z$, $\tilde{x}_y$ und damit

$$\boxed{\tilde{x} := Y\tilde{x}_y + Z\tilde{x}_z.}$$

Schritt 3: Bestimmung des Multiplikators λ

$$Q\tilde{x} + A^\top\lambda = -q\,.$$

Wir setzen $\tilde{x}$ in die Zerlegung (7.15) ein, nämlich

$$\tilde{x} = H^\top \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix}$$

und multiplizieren die Gleichung von links mit H.

$\Rightarrow$

$$\underbrace{HQH^\top}_{B} \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix} + \underbrace{HA^\top}_{\binom{R}{0}} \lambda = \underbrace{-Hq}_{h}$$

d. h.

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \tilde{x}_y \\ x_z \end{pmatrix} + \begin{pmatrix} R \\ 0 \end{pmatrix} \lambda = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$\Rightarrow \qquad \boxed{R\lambda = h_1 - B_{11}\tilde{x}_y - B_{12}\tilde{x}_z}.$$

Das ist einfach aufzulösen, da R eine obere Dreiecksmatrix ist. Damit ist alles gelöst!

7.1.2 Aufgaben mit Ungleichungsrestriktionen

Wir betrachten

$$\boxed{\begin{array}{l} \min\limits_{x\in\mathbb{R}^n} \dfrac{1}{2}\langle x, Qx\rangle + \langle q, x\rangle \\ \text{bei} \quad Ax = b \text{ und } Gx \le r \end{array}} \tag{QLU}$$

mit Q : symmetrisch, (n,n), $A : (m,n)$, $m \le n$, $G : (p,n)$. Die zulässige Menge ist

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, Ax = b, Gx \le r\}.$$

Definition 7.1.1 *Es sei $\tilde{x}$ ein gegebener zulässiger Punkt. Ein Vektor d heißt dann zulässige Richtung im Punkt $\tilde{x}$, wenn $Ad = 0$ und $\langle g^j, d\rangle \le 0 \quad \forall j \in J(\tilde{x})$ gilt. Die Menge aller zulässigen Richtungen wird mit $K(\mathscr{F}, \tilde{x})$ bezeichnet.*

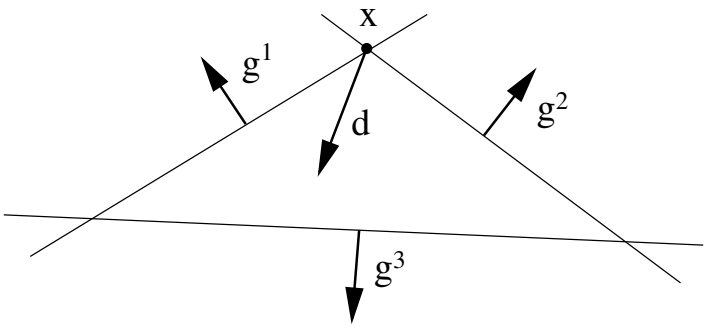
Nach Lemma 5.5.1 gilt für den Kegel der zulässigen Richtungen

$$K(\mathscr{F}, \tilde{x}) = L(\mathscr{F}, \tilde{x}) = \{d \in \mathbb{R}^n | Ad = 0\,,\ \langle g^j, d\rangle \le 0 \quad \forall j \in J(\tilde{x})\}\,.$$

Wir schreiben das noch etwas anders auf: Wie früher fassen wir alle Vektoren der aktiven Ungleichungen, g^i, $i \in J(x)$, zu einer Matrix $G(x)$ zusammen. Diese enthält als Zeilen die Vektoren $(g^i)^\top$. Dann ist d genau dann zulässige Richtung im Punkt x, wenn

$$Ad = 0, \qquad G(x)d \le 0.$$

Geometrische Illustration: (3 Ungleichungsrestriktionen $\langle g^i, x\rangle \le r^i$, $i = 1,2,3$)



$$\left.\begin{array}{l}\langle g^1,x\rangle = r^1\\ \langle g^2,x\rangle = r^2\end{array}\right\}\text{aktiv} \qquad J(x)=\{1,2\}$$
$$\left.\langle g^3,x\rangle < r^3\right\}\text{inaktiv} \qquad G(x)=\begin{pmatrix}(g^1)^\top\\(g^2)^\top\end{pmatrix}.$$

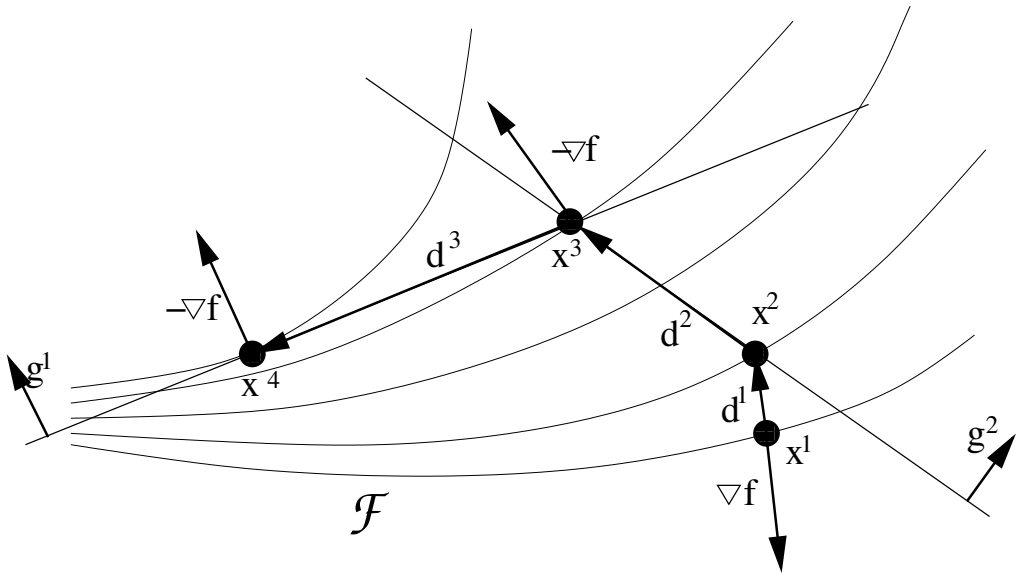
Anschaulich ist klar: Damit $x+td$ für kleine t zulässig bleibt, muss gelten

$$\langle g^1,d\rangle \le 0 \quad \wedge \quad \langle g^2,d\rangle \le 0\,.$$

Die dritte Nebenbedingung – die inaktive – hat darauf keinen Einfluss.

Bevor wir nun zur mathematischen Beschreibung des numerischen Verfahrens – Verfahren zulässiger Richtungen mit Aktive-Mengen-Strategie – kommen, wollen wir dessen grundlegenden Ideen geometrisch motivieren.

Wir betrachten dazu folgende Konstellation:



Schritt 1:
Startpunkt x^1 liegt im Inneren von $\mathscr{F}$. Es können zuerst beide Restriktionen ignoriert und mit einem Verfahren der freien Optimierung gestartet werden, bis eine (oder mehrere) Restriktionen aktiv werden.

Schritt 2:
Unser Verfahren hat im Punkt x^2 den Rand von $\mathscr{F}$ erreicht – die Restriktion Nr. 2 ist aktiv geworden.

- Wäre g^2 parallel zu ∇f, dann würde gelten

$$\nabla f + \mu g^2 = 0\,,$$

in diesem Falle, falls $\mu \geq 0$: **Fertig.**

In unserem Bild gilt das nicht, ∇f ist kein Vielfaches von g^2.

- Wir suchen daher im Unterraum mit $\langle g^2, d\rangle = 0$ weiter, d. h. in

$$\{x^2 + d \mid \langle g^2, d\rangle = 0\}.$$

Schritt 3:
In unserem Fall gelangt das Verfahren schließlich in x^3 zu einem Punkt, in dem eine weitere Restriktion aktiv wird, nämlich $\langle g^1, x\rangle$.

- Offenbar ist x^3 noch nicht optimal, denn

$$-\nabla f = \mu^1 g^1 + \mu^2 g^2 \tag{7.17}$$

mit $\mu^1 > 0$ aber $\mu^2 < 0$, also $0 = \nabla f + \mu^1 g^1 + \mu^2 g^2$.

- In welcher Richtung sollte weiter gesucht werden?
Die Richtung d muss für die Zulässigkeit

$$\begin{aligned}\langle g^1, d\rangle &\leq 0\\ \langle g^2, d\rangle &\leq 0\end{aligned}$$

erfüllen und für einen Abstieg

$$\begin{array}{c}\langle \nabla f, d\rangle < 0\\ \updownarrow\\ \langle -\nabla f, d\rangle > 0.\end{array}$$

Aus (7.17) folgt

$$\langle -\nabla f, d\rangle = \underbrace{\mu^1}_{>0}\underbrace{\langle g^1, d\rangle}_{\leq 0} + \underbrace{\mu^2}_{<0}\underbrace{\langle g^2, d\rangle}_{\leq 0}.$$

Weil $\langle -\nabla f, d\rangle$ positiv sein soll und möglichst groß, besteht die Strategie darin, $\langle g^1, d\rangle = 0$ zu wählen, unter Beachtung von

$$\langle g^2, d\rangle < 0.$$

$\Rightarrow$ **Wir deaktivieren Restriktion 2** und halten Nr. 1 aktiv.

- Als Resultat gelangt das Verfahren hier zu x^4. Das ist die Lösung, denn hier gilt $-\nabla f = \mu^1 g^1$ mit $\mu^1 > 0$, und die Optimalitätsbedingungen sind erfüllt

$$\begin{aligned}0 &= \nabla f + \mu^1 g^1 + 0 \cdot g^2\\ \langle g^1, x^4\rangle &= r^1\,,\ \langle g^2, x^4\rangle < r^2\\ \Big(\langle g^2, x^4\rangle - r^1\big) \cdot 0 &= 0\,.\end{aligned}$$

Fazit:

- Stop, wenn die notwendigen Bedingungen mit $\mu^i \geq 0$ erfüllt sind
- Aktivierung von Nebenbedingungen, auf die das Verfahren trifft
- Deaktivierung, wenn Multiplikatoren negativ werden (wird noch präzisiert).
- Ansonsten Suche in affin-linearen Unterräumen, d. h. Optimierung unter Gleichungsnebenbedingungen.

Wir kommen nun zur mathematischen präzisen Formulierung des Verfahrens.

- Aktueller Iterationspunkt sei x^k.

  $J_k := J(x^k)$: Menge der aktiven Indizes (hier gilt $\langle g^j, x^i \rangle = r^i$, $i \in J_k$)
  $p_k = |J(x^k)|$: Zahl der aktiven Indizes
  $G_k = G(x^k)$: Matrix der g^i, $i \in J_k$ (genauer: mit Zeilen $g_i^\top$)
  $B_k = \binom{A}{G_k}$: Beschreibt das zur Zeit aktive lineare Gleichungssystem

- Ausgehend von x^k wird ein im nächsten Schritt zu lösendes *quadratisches Optimierungsproblem* mit Gleichungsrestriktionen aufgestellt:

$$\boxed{\begin{array}{l} \min\limits_{d \in \mathbb{R}^n} \dfrac{1}{2}\langle Qd, d\rangle + \langle Qx^k + q, d\rangle \\ \text{bei} \quad B_k d = 0. \end{array}} \tag{Q_k}$$

Bemerkung: Bis auf eine von x^k abhängige Konstante ist die Zielfunktion von (Q_k) gerade $f(x^k + d)$.

Ergebnis der Optimierung:

- Richtung d^k
- Multiplikatoren μ_j^k, $j \in J_k$; λ_i^k (für Gleichungsrestriktionen)
- Wir ergänzen diese durch $\mu_j^k = 0$, $j \notin J_k$.

$$\begin{array}{lll} \text{Insgesamt:} & \lambda^k \in \mathbb{R}^m & (m \text{ Gln.}) \\ & \mu^k \in \mathbb{R}^{p_k} & \text{bzw. } \tilde{\mu}^k \in \mathbb{R}^p \quad \text{(durch Nullen aufgefüllt).} \end{array}$$

Voraussetzungen für das Verfahren:

- B_k hat immer vollen Rang (Lineare Unabhängigkeit des Systems $a^i, i = 1, \ldots, m$; g^j, $j \in J_k$).
- Positive Definitheit von Q auf $\ker B_k$ für alle k.

Durch diese Voraussetzungen ist (Q_k) eindeutig lösbar und die Multiplikatoren λ^k, μ^k sind eindeutig bestimmt.

Zulässige Menge von (Q_k):

$$\mathscr{F}_k = \{d \in \mathbb{R}^n | Ad = 0,\ G_k d = 0\} \subset L(\mathscr{F}, x^k)\,.$$

Damit ist jedes $d \in \mathscr{F}_k$ automatisch eine zulässige Richtung.

Der nächste Verfahrensschritt ergibt sich nun durch Auswertung der **notwendigen Optimalitätsbedingungen für** (Q_k):

$$\nabla f(x^k + d^k\rangle + A^\top \lambda^k + G_k^\top \mu^k = 0$$

d. h.

$$\boxed{Q(d^k + x^k) + q + A^\top \lambda^k + G_k^\top \mu^k = 0\,.} \tag{7.18}$$

Aus diesem System ergeben sich λ^k, μ^k wegen linearer Unabhängigkeit eindeutig.

Nun Fallunterscheidung:

Fall 1 $\boxed{d^k = 0 \quad \text{und} \quad \mu^k \geq 0\,.}$

Dann gilt $\nabla f(x^k) + A^\top \lambda^k + G_k^\top \mu^k = 0$, und x^k erfüllt die Kuhn-Tucker-Bedingungen. Wegen Konvexität sind diese hinreichend für Optimalität. Damit ist x^k die Lösung der Aufgabe (QU): **Stop**

Fall 2 $\boxed{d^k = 0 \quad \text{aber} \quad \mu^k \ngeq 0\,.}$

Hier gibt es *mindestens ein* $j \in J_k$ mit $\mu_j^k < 0$. Wie in unserem Illustrationsbeispiel sollte dann eine Nebenbedingung deaktiviert werden. Am lohnendsten: Wähle ein $j \in J_k$ mit

$$\mu_j^k = \min\{\mu_i^k\,,\ i \in J_k\}.$$

Die *Deaktivierung* erfolgt durch Neufestsetzung der aktiven Menge:

$$\boxed{\tilde{J}_k := J_k \backslash \{j\}\,.}$$

Nun wird entsprechend $(\tilde{Q}_k)$ aufgestellt und gelöst. Das Verfahren wird dabei sichern, dass die deaktivierte Restriktion nicht verletzt (d. h. in der falschen Richtung verlassen) wird. Dieser Zwischenschritt garantiert eine Lösung mit $\tilde{d}^k \neq 0$

Ergebnis: $\tilde{d}^k$, $\tilde{\mu}^k$, $\tilde{\lambda}^k$

Damit Q positiv definit auf ker $\tilde{B}_k$ bleibt, obwohl man ja B_k nicht kennt, wird der Einfachheit halber **positive Definitheit auf ker** A (dem größtmöglichen Unterraum) vorausgesetzt.

Wir zeigen: $\boxed{\tilde{d}^k \neq 0}$

Denn: Es galt $d^k = 0$,

$\Rightarrow \quad Qx^k + q + A^\top \lambda^k + G_k^\top \mu^k = 0$

$Q\tilde{d}^k + Qx^k + q + A^\top \tilde{\lambda}^k + \tilde{G}_k^\top \tilde{\mu}^k = 0$

Wäre $\tilde{d}^k = 0$, so

$$A^\top \lambda^k + G_k^\top \mu^k = A^\top \tilde{\lambda}^k + \tilde{G}_k^\top \tilde{\mu}^k$$

Damit ist $\mu_j g^j$ Linearkombination der anderen auftretenden a^i, g^i. Wegen $\mu_j \neq 0$ gilt Gleiches für g^j, ein Widerspruch zur vorausgesetzten linearen Unabhängigkeit (Vollrang von B_k).

Somit haben wir noch folgenden Fall zu diskutieren:

Fall 3 $\boxed{d^k \neq 0}$

In diesem Fall ist d^k **Abstiegsrichtung**.

Beweis: Aus den notwendigen Bedingungen (7.18) für x^k folgt

$$\begin{aligned} \nabla f(x^k) &= Q^k x^k + q = -Qd^k - A^\top \lambda^k - G_k^\top \mu^k \qquad | \cdot d^k \\ \langle \nabla f(x^k), d^k \rangle &= - \underbrace{\langle d^k, Qd^k \rangle}_{>0 \text{ wegen Definitheit}} - 0 < 0. \end{aligned} \tag{7.19}$$

Bemerkung. Im Fall 2 muss noch gesichert werden, dass die Richtung $\tilde{d}^k$ zulässige Richtung ist, d. h. dass auch für die eine deaktivierte Restriktion Nr. j gilt

$$\langle g^j, \tilde{d}^k \rangle < 0\,.$$

Das gilt auch wirklich, denn: Wegen $d^k = 0$ und (7.18) gilt ausgeschrieben:

$$0 = \nabla f(x^k) + \sum_{i=1}^{m} a^i \lambda_i^k + \sum_{\substack{i \neq j \\ i \in \tilde{J}_k}} g^i \mu_i^k + \mu_j^k g^j \qquad | \cdot \tilde{d}^k\,.$$

Wir multiplizieren skalar mit $\tilde{d}^k$ durch. Wir wissen aus Fall 3, dass $\tilde{d}^k$ eine Abstiegsrichtung ist. Außerdem war $A\tilde{d}^k = 0$ und $\langle g^i, \tilde{d}^k \rangle = 0$, $i \in \tilde{J}_k$ gefordert. Daher

$$\underbrace{\langle \nabla f(x^k), \tilde{d}^k \rangle}_{\substack{<0 \\ \text{(Abstieg)}}} + \underbrace{\mu_j^k}_{\substack{<0 \\ \text{(Fall 2)}}} \langle g^j, \tilde{d}^k \rangle = 0$$

$\Rightarrow \langle g^j, \tilde{d}^k \rangle < 0$.

Das entspricht auch der Erkenntnis aus unserem geometrischen Beispiel – $\tilde{d}^k$ zeigt aus der Sicht der j-ten Restriktion in das Innere von $\mathscr{F}$.

Zusammengefasst ergibt sich folgendes

Verfahren 7.1.1 (Aktive-Mengen-Strategie für (QU))

1. *Berechne Startpunkt $x^0 \in \mathscr{F}$, setze $k := 0$.*
2. *Stelle (Q_k) auf und bestimme daraus die Richtung d^k, Multiplikatoren λ^k, μ^k.*
3. *Wenn $d^k = 0$ und $\mu^k \geq 0$:* **Stop***; x^k ist die gesuchte Lösung.*

4. *Wenn $d^k = 0$ und $\mu^k \not\geq 0$, dann führe einen Inaktivierungsschritt durch:*

   - *Bestimme $\mu_j^k = \min\{\mu_i^k,\ i \in J_k\}$*
   - *$J_k := J_k \backslash \{j\}$*
   - *streiche in G_k die zu j gehörige Zeile*
   - *Löse das entsprechende neue Problem $(\tilde{Q}_k)$.*
     Das Ergebnis ist auf jeden Fall $d^k := \tilde{d}^k \neq 0$.

5. *Es gilt jetzt $d^k \neq 0$. Berechne eine Schrittweite σ_k (Erklärung unten) und setze*

$$x^{k+1} = x^k + \sigma_k d^k.$$

$k := k+1$, goto 2.

Schrittweitenbestimmung
Wir gehen von einer zulässigen Abstiegsrichtung d^k aus. Die neue Lösung ist dann

$$x^{k+1} = x^k + \tau d^k$$

mit einem gewissen $\tau > 0$. Wie sollte τ gewählt werden?

- **Maximaler Abstieg**

$$\begin{aligned} f(x^k + t d^k\rangle &= f(x^k) + \underbrace{t\nabla f(x^k) d^k}_{= -t\langle d^k, Q d^k\rangle \quad \text{siehe (7.19),}} + \frac{1}{2} t^2 \langle d^k, Q d^k\rangle \\ &= f(x^k) + \underbrace{\left(\frac{1}{2} t^2 - t\right)}_{\text{minimal bei } t=1.} \langle d^k, Q d^k\rangle \end{aligned}$$

Aus Sicht des maximalen Abstiegs wäre also $\tau = 1$ zu setzen, also Ausführung eines vollen Schritts.

- **Zulässigkeit von x^{k+1}**

Für die letzte Iterierte x^k galt

$$\begin{aligned} &\langle a^i, x^k\rangle = b_i \quad i = 1, \ldots, m \\ &\langle g^i, x^k\rangle = r_i \quad i \in J_k \quad \text{(aktive Ungln.)} \\ &\langle g^i, x^k\rangle < r_i \quad i \notin J_k \quad \text{(inaktive Ungln.)}. \end{aligned}$$

Durch die Wahl von d^k ist *für alle $t \geq 0$* Folgendes gesichert:

$$\begin{aligned} &\langle a^i, x^k + t d^k\rangle = b_i, \quad \text{denn } \langle a^i, d^k\rangle = 0 \\ &\langle g^i, x^k + t d^k\rangle = r_i, \quad \forall i \in J_k \backslash \{j\}, \quad \text{aus gleichem Grund} \\ &\langle g^j, x^k + t d^k\rangle < r_j, \quad \text{denn bei der inaktivierten Restriktion gilt } \langle g^j, d^k\rangle < 0. \end{aligned}$$

Damit brauchen wir uns nur um die inaktiven Restriktionen zu kümmern! Offenbar gibt es nur dann eine Schranke für t, wenn mindestens ein $i \notin J_k$ existiert mit

$$\langle g^i, d^k \rangle > 0 .$$

Es muss dann gefordert werden

$$\langle g^i, d^k \rangle + t \langle g^i, d^k \rangle \leq r_i .$$

Das Maximum von t ergibt sich bei Gleichheit, also für dieses spezielle i durch

$$t = \frac{r_i - \langle g^i, x^k \rangle}{\langle g^i, d^k \rangle} .$$

$\Rightarrow$

Maximal zulässige Schrittweite: Es sei

$$I_k = \{ i \,|\, \langle g^i, d^k \rangle > 0 \}.$$

$$\tau_k = \min_{i \in I_k} \left\{ \frac{r_i - \langle g^i, x^k \rangle}{\langle g^i, d^k \rangle} \right\}, \quad \text{falls} \quad I_k \neq \phi$$
$$\tau_k := \infty, \quad \text{falls} \quad I_k = \phi \quad \text{(keine Beschränkung nötig)} .$$

Insgesamt: $\boxed{\sigma_k = \min(1, \tau_k) \, .}$

Damit ist das Verfahren vollständig beschrieben. Es gilt

Satz 7.1.1 *Es sei Q positiv definit auf $\ker A$ und für alle $x \in \mathscr{F}$ habe die Matrix $B(x) = \binom{A}{G(x)}$ vollen Rang. Dann berechnet das Verfahren die Lösung des Problems (QU) in endlich vielen Schritten.*

Beweis: Die Durchführbarkeit des Verfahrens haben wir bereits diskutiert.

Im Verlauf des Verfahrens sind jeweils quadratische Optimierungsprobleme der Form

$$\min f(x) \quad \text{bei} \quad Ax = b \,, \; \langle g^i, x \rangle = r_i \quad \forall i \in J \tag{7.20}$$

von Bedeutung, wobei J eine beliebige Teilmenge von $\{1, ..., p\}$ ist, die für mögliche aktive Ungleichungsrestriktionen steht. Wir lösen im Verfahren nicht direkt (7.20), aber es treten Optimalitätssysteme von (7.20) auf, nämlich die Gleichungen

$$Qx + q + A^\top \lambda + G(x)^\top \mu = 0 , \tag{7.21}$$

wobei x jeweils für die eindeutig bestimmte Lösung von (7.20) steht. Es gibt nur endlich viele mögliche Teilmengen J, damit nur endlich viele verschiedene Probleme (7.20), also auch nur endlich viele solche Lösungen x und damit nur endlich viele Möglichkeiten, solche Systeme (7.21) zu erzeugen.

Das bedeutet noch nicht die Endlichkeit des Verfahrens, es könnten noch Zyklen auftreten.

Wir nehmen nun an, dass das Verfahren im Schritt k noch nicht zu Ende ist, d. h., wir führen den **Schritt 5** mit $d^k \neq 0$ durch. Dann:

(i) $I_k = \emptyset$. Hier gilt $\sigma_k = 1$, also

$$x^{k+1} = x^k + d^k$$

und wir wissen dann wegen (7.18)

$$Q(\underbrace{x^k + d^k}_{x^{k+1}}) + q + A^\top \lambda^k + G_k^\top \mu^k = 0\,,$$

so dass $x^{k+1} =: x$ das System (7.21) erfüllt (λ^k und μ^k ergeben sich daraus eindeutig). Das heißt nicht, dass x^{k+1} optimal ist, denn $\mu^k \not\geq 0$ kann eintreten. Damit ist x^{k+1} *eine der endlich vielen Lösungen von* (7.20).

(ii) $I_k \neq \emptyset$. Hier betrachten wir 2 Unterfälle:

- $\tau_k \geq 1$: Dann gilt $\sigma_k = \min(1, \tau_k) = 1$. Fall wie eben – d. h. x^{k+1} ist eine der Lösungen von (7.20)
- $\tau_k < 1$: Alle aktiven Restriktionen bleiben aktiv, aber es kommt mindestens eine neue hinzu, so dass die Kardinalzahl der aktiven Restriktionen wächst.
  Voraussetzung war: Das System der $\{a^i\}_{i=1,\ldots,m} \cup \{g^j\}$, $j \in J_k$, ist stets linear unabhängig. Es können also höchstens $n-m$ solcher Zuwachsfälle hintereinander auftreten (am Anfang waren es mindestens m linear unabhängige Vektoren, und jedes Mal kommt ein neuer hinzu).
  Nach maximal $n-m$ Iterationen gilt also $I_{k+i} = \phi$ oder $\tau_{k+i} \geq 1 \quad \Rightarrow$ *neue* Lösung von (7.21). Dann ist x^{k+1} Lösung des Systems (7.21)

Außerdem ist d^k eine *Abstiegsrichtung*, also gilt auf alle Fälle

$$f(x^{k+j}) < f(x^k)\,.$$

Damit sind die auftretenden x^{k+j} alle verschieden, und wegen Endlichkeit der Möglichkeiten für (7.20) bzw. (7.21) muss das Verfahren nach endlich vielen Schritten stoppen.

□

7.2 Gleichungsnebenbedingungen und nichtquadratische Zielfunktion

Hier wird die gleiche Grundidee wie bei quadratischer Zielfunktion angewendet: Man "eliminiert" die Gleichungsrestriktion mit Hilfe einer Nullraummatrix. Wir betrachten die Aufgabe

$$\boxed{\begin{array}{l}\min f(x)\\ Ax = b\end{array}} \qquad \text{(PLG)}$$

mit $f : \mathbb{R}^n \to \mathbb{R}$, jetzt nicht mehr notwendig quadratisch. Die Aufgabe ist also

$$\min_{x \in \mathscr{F}} f(x)$$

mit $\mathscr{F} = \{x \in \mathbb{R}^n \,|\, Ax = b\}$.

Es sei wieder $w \in \mathscr{F}$ eine spezielle Lösung von $Ax = b$ und $Z : \mathbb{R}^l \to \ker A$ eine Nullmatrix. Dann wird die unrestringierte Aufgabe

$$\boxed{\min_{z\in\mathbb{R}^l} F(z) := f(w+Zz)} \tag{7.22}$$

gelöst. Die Bestimmung einer Nullraummatrix hängt nicht von f ab, nur von $\mathscr{F}$, erfolgt also genauso, wie bereits beschrieben (QR-Zerlegung etc.).

Die freie Optimierungsaufgabe (7.22) kann nun (bei entsprechender Glattheit von f) mit jedem Verfahren der unrestringierten Optimierung behandelt werden. Damit könnten wir diesen Abschnitt abschließen, wenn es nicht noch einige interessante Nebenaspekte gäbe! Diese bestehen in der Parallelität der Minimierung von f und der von F.

Nehmen wir an, wir untersuchen ein normales Abstiegsverfahren.

$$\begin{aligned} \textit{Für} \quad f: \qquad & x^{k+1} = x^k + \sigma_k d^k \\ \textit{Für} \quad F: \qquad & z^{k+1} = z^k + \sigma_k v^k. \end{aligned}$$

Ist z.B. v^k eine Abstiegsrichtung für F in z^k, dann gilt für das Bild $d^k := Zv^k$

$$\begin{aligned} \nabla f(x^k)^\top d^k &= \nabla f(x^k)^\top Z v^k \\ &= (Z^\top \nabla f(x^k))^\top v^k \\ &= \nabla F(z^k)^\top v^k < 0, \end{aligned}$$

damit ist auch d^k eine Abstiegsrichtung, aber für f. Ferner

$$x^{k+1} = w + Zz^{k+1} = \underbrace{w + Zz^k}_{x^k} + \sigma_k \underbrace{Zv^k}_{d^k} = x^k + \sigma_k d^k .$$

Folgerung: Man kann das Verfahren im Raum der x-Variablen durchführen und muss die z-Variablen eigentlich gar nicht verwenden. Selbst v^k wird nicht benötigt.

Verfahren 7.2.1 (Reduziertes Abstiegsverfahren)

1. *Berechne $x^0 \in \mathscr{F}$, Nullraum-Matrix Z, $k := 0$.*
2. *Falls* $\underbrace{Z^\top \nabla f(x^k)}_{\text{reduzierter Gradient}} = 0$*:* **Stop**.
3. *Ansonsten berechne Abstiegsrichtung $d^k := Zv^k$, eine effiziente Schrittweite σ_k und*

$$x^{k+1} := x^k + \sigma_k d^k;$$

$k := k+1$, goto 2.

Man braucht dazu Z, v^k sowie die Korrespondenzen

$$\begin{array}{cccc} F(z^k), & F(z^k+\sigma_k v^k), & \nabla F(z^k), & \nabla F(z^k+\sigma_k v^k), \\ \Updownarrow & \Updownarrow & \Updownarrow & \Updownarrow \\ f(x^k) & f(x^k+\sigma_k d^k\rangle & Z^\top \nabla f(x^k) & Z^\top \nabla f(x^k+\sigma_k d^k\rangle \end{array}$$

Noch sieht es so aus, als würde man bei der Berechnung von $d^k = Zv^k$ zumindest den Vektor v^k brauchen und nicht nur im x-Raum arbeiten können.
Bei konkreten Verfahren sieht das aber anders aus!

Beispiel 7.2.1
Reduziertes Gradientenverfahren:

$$v^k := -\nabla F(z^k) = -Z^\top \nabla f(x^k)$$

$$\Rightarrow \quad \boxed{d^k = Zv^k = -ZZ^\top \nabla f(x^k)}$$

Spezialfall: $Z = P$, Projektionsmatrix auf $\ker A$ $\Rightarrow$

Projiziertes Gradientenverfahren: $Z = P$

$$d^k = -ZZ^\top \nabla f(x^k) = -ZZ\nabla f(x^k)$$

$$\boxed{d^k = -Z\nabla f(x^k)}$$

Variable-Metrik-Verfahren (reduziert):

Folge $\{A_k\}$ positiv definiter Matrizen;

$$v^k = -(A_k)^{-1}\nabla F(z^k) = -(A_k)^{-1}Z^\top \nabla f(x^k)$$

$$\Rightarrow \quad \boxed{d^k = Zv^k = -Z(A_k)^{-1}Z^\top \nabla f(x^k)}$$

Speziell: reduziertes Newton-Verfahren:

$$A_k := F''(z^k) = \underbrace{Z^\top f''(x^k)Z}_{\text{reduzierte Hesse-Matrix}}$$

$$\Rightarrow \quad \boxed{d^k = -Z(Z^\top f''(x^k)Z)^{-1}Z^\top \nabla f(x^k)\,.}$$

Analoge Betrachtungen gibt es für das reduzierte BFGS-Verfahren.

Eine schöne Anwendung der nichtlinearen Optimierung mit linearen Gleichungsrestriktionen:

Nichtlineare Regression mit Splines 3. Ordnung

Messwerte (ξ_i, η_i), $i = 1, ..., m$
Ansatz $\eta(\xi) = g(x, \xi)$
Gesucht *Vektor x und vorher aber ein geeigneter Ansatz g.*
Idee: Splines mit Koeffizienten x.

Sei $\xi_1 < \xi_2 < ... < \xi_m$. Wir überdecken das Intervall $[\xi_1, \xi_m]$ durch Knotenpunkte

$$\tau_0 < \tau_1 < ... < \tau_N\,,\ \textit{mit}\ \tau_0 \leq \xi_1, \xi_m \leq \tau_N\,.$$

Forderungen:

- Auf $[\tau_i, \tau_{i+1}]$ ist $g =: g_i(x, \xi)$ Polynom dritten Grades in ξ
- $g(x, \cdot) \in C^2[\tau_0, \tau_N] \quad \Rightarrow g, g', g''$ müssen in den Knotenpunkten stetig sein.

Man definiert auf $[\tau_i, \tau_{i+1}]$

$$g_i(x,\tau) = \frac{1}{\tau_{i+1}-\tau_i}(\gamma_{i+1}(\tau-\tau_i)^3 + \gamma_i(\tau_{i+1}-\tau)^3) + \beta_i(\tau-\tau_i) + \alpha_i$$
$$\gamma_0 = \gamma_N = 0$$
$$\gamma_i = g_i''(x,\tau_i)/6 \quad \text{“Momente”}$$
$$i = 1, ..., N-1\,.$$

Durch diese Wahl ist g'' automatisch stetig. Das heißt nicht, dass auch g und g' stetig sein müssen. Diese Forderung ergibt zusätzliche Bedingungen an

$$x = (\alpha_0, ..., \alpha_{N-1}, \beta_0, ..., \beta_{N-1}, ..., \gamma_1, ..., \gamma_{N-1})\,.$$

Nämlich: Stetigkeit von g': $\quad \Delta\tau_i := \tau_{i+1} - \tau_i$

$$g_i'(x,\tau_i) = g_{i-1}'(x,\tau_i)$$
$$3\gamma_i\Delta\tau_i + \beta_i = \beta_{i-1} + 3\gamma_i\Delta\tau_{i-1}$$
$$\Rightarrow \qquad \beta_i = \beta_{i-1} + 3\gamma_i(\Delta\tau_{i-1} - \Delta\tau_i)$$

Eigentlich ist damit nur β_0 wirklich frei.

Stetigkeit von g:

$$g_i(x,\tau_i) = g_{i-1}(x,\tau_i)$$
$$\gamma_i(\Delta\tau_i)^2 + \alpha_i = \gamma_i(\Delta\tau_{i-1})^2 + \alpha_{i-1} + \beta_{i-1}\Delta\tau_{i-1}$$
$$\Rightarrow \alpha_i = \alpha_{i-1} + \gamma_i((\Delta\tau_{i-1})^2 - (\Delta\tau_i)^2) + \beta_{i-1}\Delta\tau_{i-1}\,.$$

Auch hier ist eigentlich wieder nur α_0 frei. Insgesamt ergibt sich die Aufgabe

$$\min f(x) = \sum_{i=1}^{m} (\eta_i - g(x,\xi_i))^2$$
$$\text{bei} \quad \beta_i = \beta_{i-1} + 3\gamma_i(\Delta\tau_i - \Delta\tau_{i-1})$$
$$\alpha_i = \alpha_{i-1} + \beta_{i-1}\Delta\tau_{i-1} + \gamma_i(\Delta\tau_{i-1}^2 - \Delta\tau_i^2)$$
$$i = 1, ..., N-1\,.$$

Diese Aufgabe kann man auf eine unrestringierte Optimierungsaufgabe in den Variablen $(\alpha_0, \beta_0, \gamma_1, ..., \gamma_{N-1})^\top$ reduzieren.

7.3 Ungleichungsnebenbedingungen – nichtquadratische Zielfunktionen

Jetzt ist eine Aufgabe der Bauart

$$\boxed{\begin{array}{l} \min f(x) \\ Ax = b\,,\ Gx \le r \end{array}}$$

gegeben.

Grundidee: Taylorapproximation von f bis zur zweiten Ordnung $\rightsquigarrow$ quadratische Zielfunktion, Lösung der Aufgabe mit der vorn eingeführten Methode und dann neue Approximation von f: *SQP-Verfahren*.

Zur Einstimmung ein kurzer Exkurs zur einfachsten **Idee des SQP-Verfahrens**:
Wir betrachten dazu parallel zwei einfache Optimierungsaufgaben

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{sowie} \qquad \min_{x\in\mathscr{F}} f(x)\,.$$

$\mathscr{F}$ ist dabei eine konvexe Menge.

Die notwendigen Bedingungen 1. Ordnung für eine Lösung x^* lauten

$$\nabla f(x^*) = 0 \qquad \text{bzw.} \qquad \langle \nabla f(x^*), x - x^*\rangle \geq 0 \quad \forall x \in \mathscr{F}\,.$$

Die linke Beziehung ist ein nichtlineares Gleichungssystem. Wir wissen, wie wir so etwas lösen können – z.B. mit dem Newton-Verfahren. Rechts steht eine Variationsungleichung – da haben wir erst einmal keine Idee. Schreiben wir deshalb zunächst das Newton-Verfahren für die linke Gleichung auf:

$$\nabla f(x^k) + f''(x^k)(x - x^k) = 0\,. \tag{7.23}$$

Die Lösung ist $x = x^{k+1}$. Rechts haben wir noch keine Entsprechung. Offenbar ist aber (7.23) gerade die notwendige Bedingung 1. Ordnung für die Optimierungsaufgabe

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\langle x - x^k, f''(x^k)(x - x^k)\rangle + \langle \nabla f(x^k), x - x^k\rangle\,. \tag{7.24}$$

Es ist also egal, ob wir die Aufgabe (7.24) oder die Gleichung (7.23) lösen. Während aber (7.23) keine Entsprechung für die beschränkte Optimierungsaufgabe hat, ist das bei (7.24) kein Problem. Wir berücksichtigen einfach nur die Beschränkung $x \in \mathscr{F}$:

$$\min_{x\in\mathscr{F}} \langle \nabla f(x^k), x - x^k\rangle + \frac{1}{2}\langle x - x^k, f''(x^k)(x - x^k)\rangle\,.$$

Die Lösung ist $x = x^{k+1}$.

Genau diese Idee wendet man auf (PLU) an. Hier ist

$$\mathscr{F} = \{x \in \mathbb{R}^n \,|\, Ax = b\,,\ Gx \leq r\}$$

konvex. Wir iterieren wie folgt: x^k sei berechnet. Man stellt dann

$$\boxed{\begin{array}{ll} \min & \langle \nabla f(x^k), x - x^k\rangle + \frac{1}{2}\langle x - x^k, f''(x^k)(x - x^k)\rangle \\ \text{bei} & Ax = b\,,\ Gx \leq r \end{array}} \tag{QP_k}$$

auf und löst diese Aufgabe. Die Lösung sei x^{k+1}. Dann $k := k + 1$ und neue Iteration.

Ganz umsonst gibt es aber wie auch beim Newton-Verfahren für (7.23) auch hier die Konvergenz nicht: Dort muss $f''(x^k)$ stets invertierbar sein, was man durch $f \in C^2$ und $f''(x^*)$ regulär sichert. Da es aber um ein Minimum geht, muss $f''(x^*)$ noch dazu positiv semidefinit sein. Zusammen mit der Regularität muss somit $f''(x^*)$ positiv definit sein! Genau das aber hilft in (QP_K) auch: Wenn $f''(x^*)$ positiv definit ist, so auch $f''(x^k)$ für x^k nahe bei x^*, und damit hat (QP_k) genau eine Lösung!

Bei der numerischen Umsetzung schreibt man das Verfahren ein wenig anders auf: Man setzt

$$d = x - x^k \quad \Leftrightarrow \quad x = x^k + d\,.$$

Wegen $Ax^k = b$ muss $Ad = 0$ gelten und insgesamt

$$(QP_k) \quad \Leftrightarrow \quad \boxed{\begin{array}{l} \min \langle \nabla f(x^k), d\rangle + \frac{1}{2}\langle d, f''(x^k)d\rangle \\ \text{bei} \quad Ad = 0\,,\; Gx^k + Gd \le r\,. \end{array}}$$

Diese Aufgabe dient also der Berechnung einer Suchrichtung d. Man kann nun "voll" in die Richtung d gehen, d. h. $x^{k+1} = x^k + d^k$ setzen (Newton-Verfahren) oder eine Schrittweitensteuerung verwenden.

Bemerkung: Anstelle von $f''(x^k)$ kann man wie beim Variable-Metrik-Verfahren auch entsprechende Matrizen A_k nutzen – siehe z.B. [1]. Wir bleiben bei f''.

Voraussetzungen für die Durchführbarkeit des Verfahrens:

- $f''(x^k)$ soll jeweils positiv definit auf $\ker A$ sein
  $\Rightarrow$ Existenz genau einer Lösung von (QP_k).
- $B = \binom{A}{G}$ habe vollen Rang
  $\Rightarrow$ Multiplikatoren λ, μ sind eindeutig bestimmt.

Optimalitätsbedingung für (QP_k):

$$f''(x^k)d^k + \nabla f(x^k) + A^\top \lambda^{k+1} + G^\top \mu^{k+1} = 0 \tag{7.25}$$

$$\mu^{k+1} \ge 0\,,\; \langle \mu^{k+1}, Gx^k + Gd^k - r\rangle = 0\,. \tag{7.26}$$

Bei der Lösung von (QP_k) gibt es nun zwei Fälle:

Fall 1 $d^k = 0$.
Keine Änderung, x^k müsste optimal gewesen sein. In der Tat, (7.25–7.26) ergeben dann

$$\nabla f(x^k) + A^\top \lambda^{k+1} + G^\top \mu^{k+1} = 0\,,\; \mu^{k+1} \ge 0$$
$$\langle \mu^{k+1}, Gx^k - r\rangle = 0\,.$$

$\Rightarrow x^k$ erfüllt die Optimalitätsbedingungen $\Rightarrow$ STOP
(Optimalität folgt aus den hinreichenden Bedingungen).

Fall 2 $d^k \neq 0$.

Dann ist d^k Abstiegsrichtung, denn

$$\begin{aligned}\nabla f(x^k) &= -f''(x^k)d^k - A^\top \lambda^{k+1} - G^\top \mu^{k+1} \mid \cdot d^k \\ \langle \nabla f, d^k\rangle &= -\underbrace{\langle d^k, f''d^k\rangle}_{>0} - \underbrace{\langle Ad^k}_{=0}, \lambda^{k+1}\rangle - \underbrace{\langle Gd^k, \mu^{k+1}\rangle}_{\geq 0} \\ &< 0 \quad \Rightarrow \quad \text{Abstiegsrichtung} \qquad \text{s.unten.}\end{aligned}$$

Zur Diskussion von $\langle Gd^k, \mu^{k+1}\rangle$: Die Beschränkungen lauten ausgeschrieben

$$\langle g^i, x^k\rangle + \langle g^i, d^k\rangle \leq r_i\,.$$

Für die inaktiven Indizes $i \notin J(d^k)$ gilt $\mu_i^{k+1} = 0$. Für die aktiven gilt $\mu_i^{k+1} \geq 0$ sowie

$$\langle g^i, d^k\rangle = \underbrace{r_i - \langle g^i, x^k\rangle}_{\geq 0,\ \text{weil}\ x^k \text{ zulässig war}}$$

also $\langle g^i, d^k\rangle \geq 0$. Insgesamt folgt daraus leicht $\langle Gd^k, \mu^{k+1}\rangle \geq 0$.

Schrittweitenbestimmung:
Da d^k zulässig ist, kann mindestens $x^{k+1} = x^k + 1 \cdot d^k$ gewählt werden. Die maximale Schrittweite ist daher $\tau_k \geq 1$. Gängig: $\sigma_k = 1$ ("Newton-Verfahren") oder aber: Schrittweitensteuerung.

Wie wir gesehen haben, ist bei Wahl von $\sigma_k \equiv 1$ das SQP-Verfahren eine Verallgemeinerung des Newton-Verfahrens, und es wird deshalb auch so genannt. Unter natürlichen Voraussetzungen ist es wie dieses lokal quadratisch konvergent.

Voraussetzungen: Es sei $\tilde{x}$ ein lokales Minimum von (PLU).

(i) $f \in C^{2,1}$ in einer Kugel $B(\tilde{x}, \delta)$ um $\tilde{x}$

(ii) f'' ist auf $B(\tilde{x}, \delta)$ Lipschitz, d. h.

$$\|f''(x) - f''(y)\| \leq L\|x - y\| \quad \forall x, y \in B(\tilde{x}, \delta)$$

(iii) B hat vollen Rang

(iv) Positive Definitheit:

$$d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \quad \forall d : Ad = 0,\ G(\tilde{x})d = 0$$

(Hinreichende Optimalitätsbedingung 2. Ordnung)

(v) **strenge Komplementarität**: Gilt $\langle g^i, \tilde{x}\rangle = r_i$, so gilt auch $\tilde{\mu}_i > 0$ für den entsprechenden Multiplikator

Satz 7.3.1 *Unter den Voraussetzungen (i)-(v) konvergiert das beschriebene SQP-Verfahren* lokal quadratisch*, d. h. es gilt mit einer Konstanten $c > 0$ für alle $k = 0, 1, \ldots$*

$$\begin{aligned}&\|x^{k+1} - \tilde{x}\| + \|\lambda^{k+1} - \tilde{\lambda}\| + \|\mu^{k+1} - \tilde{\mu}\| \\ &\qquad \leq c\big(\|x^k - \tilde{x}\|^2 + \|\lambda^k - \tilde{\lambda}\|^2 + \|\mu^k - \tilde{\mu}\|^2\big).\end{aligned}$$

8 Probleme mit nichtlinearen Restriktionen-Verfahren

8.1 Das Lagrange-Newton-Verfahren

Wir betrachten nun die allgemeinste bereits diskutierte Aufgabe

$$\boxed{\begin{array}{ll} \min & f(x) \\ h(x) & = 0 \\ g(x) & \leq 0. \end{array}} \tag{PNU}$$

Dabei setzen wir jetzt generell voraus:

- $f, g, h \in C^2$
- $\tilde{x}$ ist eine *reguläre* lokale Lösung
- $\tilde{\lambda}, \tilde{\mu}$ sind zugehörige Lagrangesche Multiplikatoren.

Wir finden $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\mu}$ aus den Karush-Kuhn-Tucker-Bedingungen, die unter anderem fordern:

$$\boxed{\begin{aligned} \nabla f(\tilde{x}) + h'(\tilde{x})^\top \tilde{\lambda} + g'(\tilde{x})^\top \tilde{\mu} &= 0 \\ h(\tilde{x}) &= 0 \\ \tilde{g}(\tilde{x}) &= 0\,. \end{aligned}}$$

Dabei sind in $\tilde{g}$ nur die aktiven Ungleichungen berücksichtigt, d. h.

$$\tilde{g}(x) = (g_j(x))_{j \in J(\tilde{x})}\,.$$

Die nicht aktiven interessieren in einer Umgebung von $\tilde{x}$ nicht, sie können lokal als Nebenbedingungen vernachlässigt werden. Durch Lösen dieses Systems sollten $\tilde{x}, \tilde{\lambda}, \tilde{\mu}$ bestimmbar sein. Um das System kürzer zu formulieren, definieren wir

$$z = \begin{pmatrix} x \\ \lambda \\ \nu \end{pmatrix}, \quad F(z) = \begin{pmatrix} \nabla f(x) + h'(x)^\top \lambda + \tilde{g}'(x)^\top \nu \\ h(x) \\ \tilde{g}(x) \end{pmatrix}$$

mit $\nu = (\mu_j)_{j \in J(x)}$, $\tilde{\nu} := (\tilde{\mu}_j)_{j \in J(\tilde{x})}$. Dann haben wir mit $\tilde{z} = (\tilde{x}^\top, \tilde{\lambda}^\top, \tilde{\nu}^\top)$

$$F(\tilde{z}) = 0\,.$$

Was liegt näher, als darauf das Newton-Verfahren anzuwenden, um $\tilde{z}$ numerisch zu bestimmen? Als Vorabinformation muss man aber wissen, welche Indizes zur Menge $J(\tilde{x})$ gehören, also die aktiven Ungleichungen kennen! Für die Konvergenz des Verfahrens brauchen wir folgende Voraussetzungen:

- **Voraussetzung 1** *$f, g, h \in C^{2,1}$ und $F'(\tilde{z})$ ist nichtsingulär.*

  Die Matrix $F'(z)$ hat die Form

$$F'(z) = \begin{pmatrix} \mathscr{L}_{xx}(x, \lambda, \nu) & h'(x)^\top & \tilde{g}'(x)^\top \\ h'(x) & 0 & 0 \\ \tilde{g}'(x) & 0 & 0 \end{pmatrix}.$$

Dies ist eine Matrix vom Typ

$$\mathscr{A} = \begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix},$$

für welche in Lemma 5.4.1 gezeigt wurde: Ist Q positiv definit auf $\ker A$ und hat A vollen Rang, dann ist $\mathscr{A}$ invertierbar. Hier gilt

$$Q = \mathscr{L}_{xx},\ A = \begin{pmatrix} h'(x) \\ \tilde{g}'(x) \end{pmatrix},$$

also ist $F'(\tilde{z})$ unter folgenden zwei Voraussetzungen nichtsingulär:

- **Voraussetzung 2** *Die Gradienten $\nabla h_i(\tilde{x})$, $\nabla \tilde{g}_j(\tilde{x})$ der aktiven Restriktionen sind linear unabhängig.*
- **Voraussetzung 3** *Die strenge hinreichende Optimalitätsbedingung 2. Ordnung ist erfüllt:*

  $$d^\top \mathscr{L}_{xx}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) d \geq \alpha \|d\|^2$$

  für alle d mit $h'(\tilde{x})d = 0$ und $\tilde{g}'(\tilde{x})d = 0$.

Folglich hat A vollen Rang und Q ist positiv definit auf $ker A$.

Weiter brauchen wir später noch:

- **Voraussetzung 4** *Es liegt* strenge Komplementarität *vor, d.h. für alle $j = 1, \ldots, p$ gilt*

  $$g_j(\tilde{x}) = 0 \quad \Rightarrow \quad \mu_j > 0.$$

  Dahinter steht die Idee, dass aktive Ungleichungen in einer Umgebung von $\tilde{x}$ aktiv bleiben, falls die μ_j stetig von $\tilde{x}$ abhängen.

Damit sind die Voraussetzungen erfüllt, welche die lokal-quadratische Konvergenz des Newton-Verfahrens garantieren:

Ausgehend vom Startvektor $z^0 = (x^0, \lambda^0, \nu^0)$ berechnet man

$$z^{k+1} = z^k - F'(z^k)^{-1} F(z^k),$$

d. h., man löst das lineare System

$$F'(z^k)(z - z^k) = -F(z^k)$$

für z^{k+1}. Das sieht so aus:

Es sei $\tilde{\mathscr{L}}$ die Lagrange-Funktion ohne die inaktiven Restriktionen,

$$\tilde{\mathscr{L}} = f + \langle h, \lambda \rangle + \langle \tilde{g}, \mu \rangle.$$

Dann folgt

$$\begin{pmatrix} \tilde{\mathscr{L}}_{xx}(x^k, \lambda^k, \nu^k) & h'(x^k)^\top & \tilde{g}'(x^k)^\top \\ h'(x^k) & 0 & 0 \\ \tilde{g}'(x^k) & 0 & 0 \end{pmatrix} \begin{pmatrix} x - x^k \\ \lambda - \lambda^k \\ \nu - \nu^k \end{pmatrix}$$

$$= - \begin{pmatrix} \nabla f(x^k) + h'(x^k)^\top \lambda^k + \tilde{g}'(x^k)^\top \nu^k \\ h(x^k) \\ \tilde{g}(x^k) \end{pmatrix}.$$

Einiges hebt sich hier auf, so dass schließlich folgendes System zu lösen ist:

$$\begin{aligned} \nabla f(x^k) + \tilde{\mathscr{L}}_{xx}(x^k,\lambda^k,\nu^k)(x-x^k) + h'(x^k)^\top \lambda + \tilde{g}'(x^k)^\top \nu &= 0 \\ h(x^k) + h'(x^k)(x-x^k) &= 0 \\ \tilde{g}(x^k) + \tilde{g}'(x^k)(x-x^k) &= 0\,. \end{aligned} \tag{8.27}$$

Im Prinzip sind das die notwendigen Optimalitätsbedingungen für eine Lösung der linear-quadratischen Aufgabe

$$\begin{aligned} &\min_x \langle \nabla f(x^k), (x-x^k)\rangle + \frac{1}{2}(x-x^k)^\top \mathscr{L}_{xx}(x^k,\lambda^k,\nu^k)(x-x^k) && (\mathrm{Q1})_k \\ \text{bei}\quad & h(x^k) + h'(x^k)(x-x^k) = 0 \\ & g(x^k) + g'(x^k)(x-x^k) \le 0\,, \end{aligned}$$

wenn wir zeigen können, dass die inaktiven Restriktionen $j \notin J(\tilde{x})$ auch für $(Q1_k)$ nicht aktiv sind, also bedeutungslos bleiben. In gewissem Sinne sind also das Newton-Verfahren (8.27) und das durch $(\mathrm{Q1})_k$ beschriebene SQP-Verfahren äquivalent.

Nun führen wir wieder die Richtung $d = x - x^k$ ein und lösen

$$\boxed{\begin{aligned} \min_{d\in\mathbb{R}^n} \tfrac{1}{2}\langle d, \mathscr{L}_{xx}(x^k,\lambda^k,\mu^k)d\rangle &+ \langle \nabla f(x^k), d\rangle \\ \text{bei}\quad h(x^k) + h'(x^k)d &= 0 \\ g(x^k) + g'(x^k)d &\le 0. \end{aligned}} \tag{Q2}_k$$

Aus der Lösung d^k ergeben sich

$$x^{k+1} := x^k + d^k$$

und neue Multiplikatoren μ^{k+1}, λ^{k+1}. So erhalten wir

Verfahren 8.1.1 (Lagrange-Newton-Verfahren für (PNU)**)**

1. $k := 0$, $z^0 := (x^0, \lambda^0, \mu^0)$.
2. *Stelle* $(\mathrm{Q2})_k$ *auf und löse diese quadratische Optimierungsaufgabe; Ergebnis:* d^k; λ^{k+1}, μ^{k+1}.
3. *Falls* $d^k = 0$, *STOP.*
4. $x^{k+1} := x^k + d^k$, *goto 2.*

Bemerkungen:

1. Das Verfahren konvergiert lokal quadratisch! Das ist plausibel, weil es eigentlich unter entsprechenden Voraussetzungen äquivalent zum Newton-Verfahren ist (wenn man die aktiven Restriktionen kennt und strenge Komplementarität gilt. Strenge Komplementarität: Aktive Restriktionen bleiben in der entsprechenden Umgebung aktiv, weil die Multiplikatoren dort positiv bleiben).
2. Die Iterierten dieses Verfahrens sind in der Regel wegen der Nichtlinearität der Restriktionen unzulässig, d. h. $x^k \notin \mathscr{F}$. Außerdem haben wir nur lokale Konvergenz. Deshalb sind Modifikationen angebracht!

8.2 Sequentielle quadratische Optimierung

Das reine Lagrange-Newton-Verfahren ist nur lokal konvergent. Außerdem kann die Berechnung von $\mathscr{L}_{xx}$ teuer werden. Deshalb modifiziert man das Verfahren.

• **Verwendung von Approximationen $\mathbf{A_k}$ für $\mathscr{L}_{\mathbf{xx}}(\mathbf{x^k}, \lambda^{\mathbf{k}}, \mu^{\mathbf{k}})$**

Man verwendet wie bei Variable-Metrik-Verfahren symmetrische und positiv definite Matrizen A_k und bestimmt d^k aus

$$\boxed{\begin{array}{rll} & \min_d \langle \nabla f(x^k), d\rangle & +\frac{1}{2}\langle d, A_k d\rangle \\ \text{bei} & h(x^k) + h'(x^k)d & = 0 \\ & g(x^k) + g'(x^k)d & \le 0\,. \end{array}} \qquad (\mathrm{QP})_k$$

Wir wollen annehmen, dass die entsprechende zulässige Menge $\mathscr{F}_k$ nicht leer ist. Eine hinreichende Bedingung gibt [1,Satz8.2.1] an (h_i affin-linear, lineare Unabhängigkeit der Gradienten, Slater-Typ-Bedingung).

Unter natürlichen Voraussetzungen kann dann die Existenz genau einer Lösung von $(\mathrm{QP})_k$ sowie die gleichmäßige Beschränktheit der Folgen d^k, α^k, μ^k bewiesen werden [1,Satz8.2.2].

• **Schrittweitensteuerung**

Wir gehen nicht den gesamten Newtonschritt, sondern setzen

$$x^{k+1} = x^k + \sigma_k d^k\,.$$

Man muss dabei Folgendes beachten:

a) Die erzeugten x^k brauchen für (PNU) nicht zulässig zu sein.

b) d^k ist nicht notwendig eine Abstiegsrichtung.
Die Lösung x^k könnte wegen Unzulässigkeit einen zu kleinen Wert ergeben haben.

Deshalb benutzt man zur Schrittweitensteuerung sogenannte

• **Merit-Funktionen**

Definition 8.2.1 *Eine Funktion ϕ heißt Merit-Funktion für* (PNU)*, wenn sie folgende Eigenschaften hat:*

(i) Ist $\tilde{x} \in \mathscr{F}$ lokale Lösung von (PNU)*, dann ist $\tilde{x}$ auch lokales (freies) Minimum von ϕ.*

(ii) Die Richtung d^k ist Abstiegsrichtung für ϕ.

Praktisch bewährt haben sich *exakte Penalty-Funktionen* als Merit-Funktionen,

$$\boxed{\phi = \phi(x;\beta,\gamma) := f(x) + \sum_{j=1}^{p} \beta_j g_j(x)_+ + \sum_{i=1}^{m} \gamma_i |h_i(x)|}$$

mit Konstanten $\beta_j \geq 0$, $\gamma_i \geq 0$ und

$$g_j(x)_+ := \max\{0, g_j(x)\} = \frac{g_j(x) + |g_j(x)|}{2}.$$

Die Funktion ϕ ist nicht differenzierbar!

Für $x \in \mathscr{F}$ gilt $g_j(x) \leq 0$, also $g_j(x)_+ = 0$ sowie $h_i(x) = 0$, also $|h_i(x)| = 0$. Damit

$$f(x) = \phi(x; \beta, \gamma) \quad \forall x \in \mathscr{F}.$$

Für $x \notin \mathscr{F}$ haben wir $\phi > f$. In diesem Sinne sind $\Sigma\beta_j(g_j)_+$ und $\Sigma\gamma_i|h_i|$ Strafterme, die eine Verletzung der Nebenbedingungen bestrafen:

$$\underbrace{\sum_{i=1}^{m} \gamma_i |h_i(x)| + \sum_{j=1}^{p} \beta_j g_j(x)_+}_{\textit{Penalty-Term}} > 0 \quad \forall x \notin \mathscr{F}.$$

Die Koeffizienten γ_i, β_j heißen *Penalty-Parameter*. Die Merit-Funktion ϕ heißt auch **exakte Penalty-Funktion**.

Es gilt der wichtige

Satz 8.2.1 *Die obigen Voraussetzungen 1 - 4 seien erfüllt. Ist dann $\tilde{x} \in \mathscr{F}$ lokales Minimum von* (PNU) *mit Lagrangeschen Multiplikatoren λ sowie μ und gilt*

$$\beta_j > \mu_j, \ \gamma_i > |\lambda_i| \quad \forall\ j = 1, ..., p,\ i = 1, ..., m,$$

dann ist $\tilde{x}$ striktes lokales Minimum von $x \mapsto \phi(x; \beta, \gamma)$.

Durch relativ aufwändige Abschätzungen kann man letztlich folgendes zeigen:
Man gibt $\varepsilon > 0$ und $\delta \in (0, 1)$ vor. Sind die Parameter β_j und γ_i hinreichend groß gewählt, dann gilt

$$\beta_j \geq \mu_j^{(k+1)} + \varepsilon, \ \gamma_i \geq |\lambda_i^{(k+1)}| + \varepsilon,$$

und für hinreichend kleines $\sigma > 0$ erhält man

$$\phi(x^k + \sigma d^k, \beta, \gamma) \leq \phi(x^k; \beta, \gamma) - \sigma\delta \left[\langle d^k, A_k d^k \rangle + \varepsilon \|g(x^k)_+\|_1 + \varepsilon \|h(x^k)\|_1 \right],$$

d. h., die Merit-Funktion kann wirklich verkleinert werden. Außerdem gilt

$$\begin{aligned} |h_i(x^k + \sigma d^k\rangle| &\leq (1 - \delta\sigma)|h_i(x^k)| \quad i = 1, ..., m \\ g_j(x^k + \sigma d^k\rangle &\leq (1 - \delta\sigma) g_j(x^k) \quad j = 1, ..., p, \end{aligned}$$

d. h., die Unzulässigkeit wird in jedem Schritt geringer. Darauf basiert eine Grundversion des SQP-Verfahrens, auf deren ausführliche Darstellung wir verzichten. Wir verweisen auf [1, Abschnitt 8.2.3].