Inhaltsverzeichnis

1. Parameterabhängige Anfangswertprobleme

Parameterabhängige gewöhnliche Differentialgleichung mit Parametervektor p und parameterabhängiger Anfangsbedingung:

$$\dot{y}(t) = f(t, y(t), p) \tag{1.1}$$
$$y(t_0) = y_0(p) \tag{1.2}$$

wobei

- $t \in [t_0, t_{end}]$ „Zeit“
- $y\colon [t_0, t_{end}] \to \mathbb{R}^{n_y}$ „Zustände“
- $p \in \mathbb{R}^{n_p}$ „Parameter“
- $f\colon [t_0, t_{end}] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}$ sei hinreichend oft in t, y (stückweise) stetig differenzierbar, damit numerische Integrationsverfahren mit Fehlerkontrolle funktionieren, außerdem einmal in p stetig differenzierbar.
- $y_0\colon \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}$ einmal stetig differenzierbar.

Variante: Differentiell-algebraische Gleichungssysteme (DAE):

$$\begin{aligned}
&f\colon [t_0, t_{end}] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y} \\
&g\colon [t_0, t_{end}] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_z} \\
&\dot{y} = t(t, y, z, p) && (1.3) \\
&0 = g(t, y, z, p) \\
&y\colon [t_0, t_{end}] \to \mathbb{R}^{n_y} && \text{differentielle Zustände} \\
&z\colon [t_0, t_{end}] \to \mathbb{R}^{n_z} && \text{algebraische Zustände}
\end{aligned}$$

Das DAE benötigt nur Anfangswerte für y:

$$y(t_0) = y_0(p) \tag{1.4}$$

Die Anfangswerte der z sind durch die *Konsistenzbedingung* gegeben:

$$g(t_0, y(t_0), z(t_0), p) = 0 \tag{1.5}$$

Wir betrachten in dieser Vorlesung nur den *Index-1-Fall*, d. h.

$$\frac{\partial_g}{\partial_z} \quad \text{hat den Rang } n_z \tag{1.6}$$

Daraus folgt, dass die algebraischen Gleichungen lokal eindeutig nach $\dot{z}$ auflösbar sind:

$$\begin{aligned} 0 &= g(t, y, z, p) \\ 0 &= \frac{\partial}{\partial t} g(t, y, z, p) \\ &= \frac{\partial g}{\partial t} + \frac{\partial g}{\partial y}\dot{y} + \frac{\partial g}{\partial z}\dot{z} \\ \dot{z} &= -\left(\frac{\partial g}{\partial z}\right)^{-1} \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial y}\dot{y}\right) \end{aligned}$$

Man erhält zusammen mit $\dot{y} = f$ eine ODE für y und z, deren Anfangswerte aber die Konsistenzbedingung (1.5) erfüllen müssen. Diese ODE- bzw. DAE-Anfangswertprobleme können auch aus im Ort diskretisierten Anfangs-(Rand-)Wertproblemen von instationären partiellen Differentialgleichungen kommen. In dieser Vorlesung behandeln wir nicht die Diskretisierungsmethoden für PDEs.

Satz 1.1: Lokaler Stabilitätssatz

Seien die beiden Anfangswertprobleme

$$\dot{u}(t) = f(t, u, p_1), \quad u(t_0) = u_0 \tag{1.7}$$

$$\dot{v}(t) = f(t, v, p_1), \quad v(t_0) = v_0 \tag{1.8}$$

auf $[t_0, t_{end}]$ gegeben. Die Funktion $f(t, y, p)$ sei stetig in (t, y) und genüge einer Lipschitz-Bedingung in y mit Konstante $L < \infty$. Dann gilt für die Lösungen u und v von (1.7) und (1.8):

$$\|u(t) - v(t)\| \leq e^{L(t-t_u)} \left(\|u_0 - v_0\| + \int_{t_0}^{t} \sup_{t,y} \|f(\tau, y(\tau), p_1) - f(\tau, y(\tau), p_2)\| \, \mathrm{d}\tau \right)$$

Der Beweis wurde bereits in Numerik 1 gegeben, er benutzt das Gronwall-Lemma.

Korollar 1.2: Trompetenabschätzung

Sei f wie in Satz 1.1 mit $L < \infty$. Für die Abweichung $\delta y := v - y$ der Lösungen y und v von

$$\dot{y}(t) = f(t, y, p), \quad y(t_0) = y_0(p)$$
$$\dot{v}(t) = f(t, v, p_1), \quad v(t_0) = y_0(p + \delta p)$$

gilt:

$$\|\delta y(t)\| \leq \varepsilon_1 e^{L(t-t_0)} + \varepsilon_2 e^{L(t-t_0)}(t - t_0) \quad (1.10)$$

wobei

$$\|\delta y_0\| := \|\frac{\partial y_0}{\partial p}(p)\delta p\| \leq \varepsilon_1$$
$$\|\delta f\| := \|\frac{\partial f}{\partial p}(t, y, p)\partial p\| \leq \varepsilon_2$$

Beweis:
Siehe Numerik 1, folgt aus Satz 1.1.
Für Änderungen von p können die Lösungen der Anfangswertprobleme exponentiell auseinanderlaufen. Kleine Störungen von p in Anfangsbedingungen und Rechter-Seite-Funktion können sehr große Unterschiede in $y(t_{end})$ zur Folge haben.

Differentiation der Lösung des AWPs

Schreibweise:
Betrachte das Anfangswertproblem

$$\dot{y} = f(t, y, p) \quad y(0) = y_0 \quad (1.11)$$

Die Lösung y hängt von t, t_0, y_0 und p ab. Wir schreiben:

$$y(t) = y(t; t_0, y_0, p)$$

Satz 1.3:
Sei $f \in \mathbb{C}^m$, $m \geq 1$. Dann ist $y(t; t_0, y_0, p)$

- (m+1) mal stetig differenzierbar in t
- m mal stetig differenzierbar in t_0, y_0, p

Beweis:
Integraldarstellung der Lösung:

$$y(t) = y_0 + \int_{t_0}^{t} f(\tau, y(\tau), y)\,\mathrm{d}\tau$$

Definition 1.4: Ableitungen der Lösung des AWP

- Ableitung nach Anfangswerten: $G(t; t_0, y_0, p) \colon = \frac{\partial}{\partial y_0} y(t; t_0, y_0, p)$
- Ableitung nach Anfangszeitpunkt: $G_{t_0}(t; t_0, y_0, p) \colon = \frac{\partial}{\partial t_0} y(t; t_0, y_0, p)$
- Ableitung nach Parametern: $G_p(t; t_0, y_0, p) \colon = \frac{\partial}{\partial p} y(t; t_0, y_0, p)$

Berechnung der Ableitungen:

Integralform des AWP:

$$y(t) = y_0 + \int_{t_0}^{t} f(\tau, y(\tau), p)\,\mathrm{d}\tau \quad (1.12)$$

a) Differenziere nach Anfangswerten:

$$\frac{\partial}{\partial y_0} y(t) = \frac{\partial y_0}{\partial y_0} + \int_{t_0}^{t} \frac{\partial}{\partial y} f(\tau, y(\tau), p) \frac{\partial y}{\partial y_0}(\tau)\,\mathrm{d}\tau$$

$$G(t) = I + \int_{t_0}^{t} \frac{\partial}{\partial y} f(\tau, y(\tau), p) G(\tau)\,\mathrm{d}\tau \quad (1.13)$$

Dies ist äquivalent zum Anfangswertproblem

$$\dot{G}(t) = \frac{\partial f}{\partial y}(t, y(t), p) G(t) \quad (1.14)$$
$$G(t_0) = I$$

zur sogenannten Variationsdifferentialgleichung (VDE) nach Anfangswerten.

Differenziere nach dem Anfangszeitpunkt:

$$\frac{\partial}{\partial t_0} y(t) = \frac{\partial y_0}{\partial t_0} + \frac{\partial}{\partial t_0} \int_{t_0}^{t} f(\tau, y(\tau), p)\, \mathrm{d}\tau$$

$$G_{t_0}(t) = 0 - f(t_0, y(t_0), p) + \int_{t_0}^{t} \frac{\partial f}{\partial y} G_{t_0}(\tau)\, \mathrm{d}\tau \tag{1.15}$$

Dies ist äquivalent zum Anfangswertproblem

$$\begin{aligned} \dot{G}(t_0)(t) &= \frac{\partial f}{\partial y} G_{t_0}(t) \\ G_{t_0}(t_0) &= -f(t_0 y(t_0), p) \end{aligned} \tag{1.16}$$

zur VDE nach dem Anfangszeitpunkt.

c) Differenziere nach den Parametern:

$$\frac{\partial}{\partial p} y(t) = \frac{\partial y_0}{\partial p} + \int_{t_0}^{t} \frac{\partial f}{\partial y}(\tau, y(\tau), p) \frac{\partial y}{\partial p}(\tau) + \frac{\partial f}{\partial p}(\tau, y(\tau), p)\, \mathrm{d}\tau$$

$$G_p(t) = \frac{\partial y_0}{\partial p} + \int_{t_0}^{t} \frac{\partial f}{\partial y} G_p(\tau) + \frac{\partial f}{\partial p}\, \mathrm{d}\tau \tag{1.17}$$

Dies ist äquivalent zum Anfangswertproblem

$$\begin{aligned} \dot{G}_p(t) &= \frac{\partial f}{\partial y}(t, y(t), p) G_p(t) + \frac{\partial f}{\partial p}(t, y(t), p) \\ G_p(t_0) &= \frac{\partial y_0}{\partial p} \end{aligned} \tag{1.18}$$

zur VDE nach den Parametern.

Bemerkung 1.5:

Die Anfangswertprobleme der VDEs (1.14), (1.16) und (1,18) hängen von der Lösung $y(t)$ des „nominalen“ Anfangswertproblems (1.11) ab. Sie müssen also jeweils mit (1.11) in einem gemeinsamen System gelöst werden.

Satz 1.6

Es gilt:

$$G_{t_0}(t;t_0,y_0,p) = -G(t;t_0,y_0,p)f(t_0,y_0,p) \quad (1.19)$$

Beweis:
Multipliziere $\dot{G} = \frac{\partial f}{\partial y}G$, $G(t_0) = I$ von rechts mit $-f(t_0,y_0,p)$:

$$-\dot{G}f(t_0,y_0,p) = -\frac{\partial f}{\partial y}Gf(t_0,y_0,p)$$
$$\text{und } -G(t_0)f(t_0,y_0,p) = -f(t_0,y_0,p)$$

Also erfüllt $y\colon = -Gf(t_0,y_0,p)$ die Differentialgleichung $\dot{y} = \frac{\partial f}{\partial y}y$ und die Anfangsbedingung $y(t_0) = -f(t_0,y_0,p)$. Das ist die VDE für G_{t_0}, also ist $y = -Gf(t_0,y_0,p) = G_{t_0}$

Variationsdifferentialgleichungen für Richtungsableitungen

Gegeben sei eine Richtung $\Delta y_0 \in \mathbb{R}^{n_y}$. Die Richtungsableitung von y nach y_0 in der Richtung Δy_0 ist gegeben durch

$$\frac{\partial y}{\partial y_0}(t;t_0,y_0,p)\Delta y_0\colon = \frac{\partial}{\partial h}y(t;t_0,y_0+h\Delta y_0,p)\Big|_{h=0}$$
$$= \lim_{h\to 0}\frac{y(t;t_0,y_0+h\Delta y_0,p) - y(t;t_0,y_0,p)}{h} \in \mathbb{R}^{n_y} \quad (1.19a)$$

Die Richtungsableitung erfüllt

$$\frac{\partial}{\partial t}(G(t;t_0,y_0,p)\Delta y_0) = \frac{\partial f}{\partial y}(t,y,p)(G(t;t_0,y_0,p)\Delta y_0)$$
$$G(t;t_0,y_0,p)\Delta y_0 = I\Delta y_0 = \Delta y_0 \quad (1.20)$$

Dies ist ein (n_y-dimensionales) VDE-Anfangswertproblem für jede Richtung Δy_0. hat man mehrere Richtungen, kann man diese spaltenweise in einer Richtungsmatrix S zusammenfassen:

$$S = (\Delta y_{0,1},\cdots,\Delta y_{y,n_s})$$

Zur Berechnung aller zugehörigen Richtungsableitungen muss man also lösen:

$$(\dot{GS}) = \frac{\partial f}{\partial y}(GS), \quad (GS)(t_0) = S \quad (1.21)$$

mit dem n_s-fachen Aufwand wie für eine Richtung. Für die gesamte Ableitung $\frac{\partial y}{\partial y_0} = \frac{\partial y}{\partial y_0}I_{n_y\times n_y}$ braucht man n_y Richtungen, hat also den n_y-fachen Aufwand. Deshalb berechnet man Richtungsableitungen nicht durch Berechnen von $\frac{\partial y}{\partial y_0}$ und anschließende Multiplikation mit den Richtungen, sondern durch Lösen von Richtungs-VDE.

Für Richtungsableitungen nach p analog:

$$\frac{\partial}{\partial t}(G_p(t;t_0,y_0,p)\Delta p) = \frac{\partial f}{\partial y}(t,y,p)G_p(t;t_0,y_0,p)\Delta p + \frac{\partial f}{\partial p}(t,y,p)$$
$$G_p(t_0;t_0,y_0,p)\Delta p = \frac{\partial y_0}{\partial p}\Delta p \tag{1.22}$$

Eigenschaften der Ableitungsmatrizen

Füge an der Stelle $t_0 < s < t_e nd$ einen „Haltepunkt“ ein:

$$\dot{y} = f(t,y,p), \quad t \in [s, t_{end}], \quad y(t_0) = y_0$$

und weiter

$$\dot{y} = f(t,y,p), \quad t \in [s, t_{end}], \quad y(s) = y(s;t_0,y_0,p)$$

Nach dem Satz von Picard-Lindelöff ist die abschnittsweise Lösung die selbe wie für

$$\dot{y} = f(t,y,t), t \in [t_0, t_{end}], y(t_0) = y_0$$

Definition 1.7: Wronski-Matrix

$$W(t,s) \colon = G(t;s,y_s,p)$$

Satz 1.8: Eigenschaften der Wronski-Matrizen

- $W(t,s)$ erfüllt $\partial_t W(t,s) = \partial_y f(t,y,p)W(t,s)$, $W(s,s) = I$
- $W(t,s)W(s,r) = W(t,r)$ (1.23)
- Für alle $t,s,r \in [t_0, t_{end}]$ ist $W(t,s)$ invertierbar und es gilt $W(t,s)^{-1} = W(s,t)$ (1.24)
- Für beliebige $t_1, \cdots, t_n$ ist $W(t_1,t_2)W(t_2,t_3)\cdots W(t_{n-1},t_n) = W(t_1,t_n)$

Wronski-Matrizen für Ableitungen nach Parametern:
Definition 1.9:

$$W_p(t,s) \colon = G_p(t;s,y_s,p)$$

Satz 1.10

$$W_p(t_{end},t_0) = W(t_{end},s)W_p(s,t_0) + W_p(t_{end},s) \tag{1.25}$$

Beweis:

$$
\begin{aligned}
y(t_{end} = y(t_{end}; t_0, y_0, p) &= y(t_{end}; s, y_s, p) \\
y_s = y(s) &= y(s; t_0, y_0, p) \\
W_p(t_{end}, t_0) &= \frac{\partial y}{\partial y_s}(t_{end}; s, y_s, p) \cdot \frac{\partial y}{\partial p}(s; t_0, y_0, p) + \frac{\partial y}{\partial p}(t_{end}; s, y_s, p) \\
&= W(t_{end}, s) W_p(s, t_0) + w_p(t_{end}, s)
\end{aligned}
$$

Anwendung:
Das System ODE+VDE wird abschnittsweise auf den Teilintervallen $[t_0, s]$, $[s; t_{end}]$ gelöst, liefert

$$y(s), \quad W(s, t_0), \quad W_p(s, t_0)$$

und

$$y(t_{end}), \quad W(t_{end}, s), \quad W_p(t_{end}, s)$$

Dann kann man zusammensetzen:

$$
\begin{aligned}
W(t_{end}, t_0) &= W(t_{end}, s) W(s, t_0) \\
W_p(t_{end}, t_0) &= W(t_{end}, s) W_p(s, t_0) + W_p(t_{end}, s)
\end{aligned}
$$

Ableitungsmatrizen bei abschnittsweise definierter rechter Seite

$$
\dot{y}(t) = \begin{cases} f_1(t, y, p) & \text{für } t < t_s \\ f_2(t, y, p) & \text{für } t > t_s \end{cases} \qquad (1.26) y(t_0) = y_0
$$

Wobei der Umschaltpunkt $t_0 < t_s < t_{end}$ als eindeutige einfache Nullstelle der Schaltbedingung

$$Q(t_s, y(t_s), p) = 0$$

gegeben sei.
Satz 1.11:
Dann gilt:

$$
\begin{aligned}
\frac{\partial y}{\partial y_0}(t_{end}; t_0, y_0, p) = W(t_{end}, t_s) \Big[I - (f_1(t_s, y(t_s), p) - f_2(t_s, y(t_s), p)) \\
\left(\frac{\partial Q}{\partial t}(t_s, y(t_s)) \right)^{-1} \frac{\partial Q}{\partial y}(t_s, y(t_s)) \Big] W(t_s, t_0) \qquad (1.27)
\end{aligned}
$$

Beweis:

$$
\begin{aligned}
y(t_{end} &= y(t_{end}; t_s, y_s, p) \\
y(t_s) &= y(t_s; t_0, y_0, p) \\
\frac{\partial y}{\partial y_0} &= \frac{\partial y}{\partial t_s}(t_{end}; t_s, y_s, p)\frac{\partial t_s}{\partial y_s}\frac{\partial y}{\partial y_0}(t_s; t_0, y_0, p) \\
&\quad + \frac{\partial y}{\partial y_s}(t_{end})\left(\frac{\partial y(t_s)}{\partial y_0} + \left(\frac{\partial y(t_s)}{\partial t_s}\right)\frac{\partial t_s}{\partial y_s}\frac{\partial y}{\partial y_0}(t_s)\right)
\end{aligned}
$$

t_s ist Endzeitpunkt des ersten und Anfangszeitpunkt des zweiten Intervalls.

$$
\begin{aligned}
0 &= Q(t_s, y(t_s), p) \\
\Rightarrow 0 &= \frac{\partial Q}{\partial t}\frac{\partial t}{\partial y} + \frac{\partial Q}{\partial y}\bigg|_{y=y_s} \\
\Rightarrow \frac{\partial t_s}{\partial y_s} &= -\left(\frac{\partial Q}{\partial t}\right)^{-1}\frac{\partial Q}{\partial y_s}\bigg|_{t=t_s, y=y_s} && \text{(Satz für implizite Funktionen)} \\
\frac{\partial y}{\partial t_s}(t) &= -W(t, t-s)f_2(t_s, y_s, p) && \text{(Satz 1.6)} \\
\frac{\partial y}{\partial y_0}(t_{end}) &= W(t_{end}, t_s)\,[I - (f_1(t_s, y_s, p) \\
&\quad -f_2(t_s, y_s, p))\left(\frac{\partial Q}{\partial t}\right)^{-1}\frac{\partial Q}{\partial y}\bigg|_{t=t_s, y=y_s}\Bigg] W(t_s, t_0)
\end{aligned}
$$

Adjungierte Differentialgleichung

Will man Matrix-Matrix-Produkte von links an $\frac{\partial y}{\partial y_0}$ oder $\frac{\partial y}{\partial p}$ berechnen:

$$
\begin{aligned}
u^T\frac{\partial y}{\partial y_0} &\qquad \text{mit } u \in \mathbb{R}^{n_y} \text{ oder} \\
U^T\frac{\partial y}{\partial y_0} &\qquad \text{mit } U \in \mathbb{R}^{n_y \times n}
\end{aligned}
$$

Berechnet man ebenfalls nicht zuerst $\frac{\partial y}{\partial y_0}$ und multipliziert dann, sondern löst die adjungierte Differentialgleichung.

Satz 1.12

Gegeben sei das ODE-AWP

$$\dot{y} = f(t, y, p), \quad y(t_0) = y_0 \quad (1.28)$$

Dann gilt: Integriert man die adjungierte Differentialgleichung (ADE)

$$\dot{\Lambda}(t)^T = -\Lambda(t)^T \frac{\partial f}{\partial y}(t, y, p) \quad (1.29)$$

rückwärts, d. h. ausgehend von $T > t_0$ mit dem Anfangswert $\Lambda(T) = I$, dann gilt:

$$\frac{\partial y}{\partial y_0}(T) = \Lambda(t_0)^T \quad (1.30)$$

$$\frac{\partial y}{\partial p}(T) = \int_{t_0}^{T} \Lambda(t)^T \frac{\partial f}{\partial p}(t, y, p)\, \mathrm{d}t \quad (1.31)$$

Beweis:
Es gilt:

$$\dot{y} - f(t, y, p) = 0$$

$$\Rightarrow \int_{t_0}^{T} \Lambda T(\dot{y} - f(t, y, p))\, \mathrm{d}t = 0 \quad (1.32)$$

Leite (1.32) nach y_0 ab:

$$\begin{aligned} 0 &= \frac{\partial}{\partial y_0} \int_{t_0}^{T} \Lambda^T (\dot{y} - f(t, y, p))\, \mathrm{d}t \\ &= \int_{t_0}^{T} \Lambda^T \left(\frac{\partial \dot{y}}{\partial y_0} - \frac{\partial f}{\partial y} \frac{\partial y}{\partial y_0} \right) \mathrm{d}t \end{aligned}$$

Partielle Integration:

$$\int_{t_0}^{T} \Lambda^T \frac{\partial \dot{y}}{\partial y_0}\, \mathrm{d}t = \left[\Lambda^T \frac{\partial y}{\partial y_0} \right]_{t_0}^{T} - \int_{t_0}^{T} \dot{\Lambda}^T \frac{\partial y}{\partial y_0}\, \mathrm{d}t$$

Daraus folgt:

$$
\begin{aligned}
0 &= \int_{t_0}^{T} \Lambda^T \left(\frac{\partial \dot{y}}{\partial y_0} - \frac{\partial f}{\partial y_0} \frac{\partial y}{\partial y_0} \right) \mathrm{d}t \\
&= \int_{t_0}^{T} \left(-\dot{\Lambda} - \Lambda^T \frac{\partial f}{\partial y} \frac{\partial y}{\partial y_0} \right) \mathrm{d}t + \left[\Lambda^T \frac{\partial y}{\partial y_0} \right]_{t_0}^{T} \\
&= \int_{t_0}^{T} \underbrace{\left(-\dot{\Lambda}^T - \Lambda^T \frac{\partial f}{\partial y} \right)}_{=0 \text{ adj. DGL}} \frac{\partial y}{\partial y_0} \mathrm{d}t + \underbrace{\Lambda(T)^T}_{=I \text{ (AW)}} \frac{\partial y}{\partial y_0}(T) - \Lambda(t_0)^T \underbrace{\frac{\partial y}{\partial y_0}}_{=I} \\
&= \frac{\partial y}{\partial y_0}(T) - \Lambda(t_0)^T \\
\Rightarrow \Lambda^T(t_0) &= \frac{\partial y}{\partial y_0}(T)
\end{aligned}
$$

Leite (1.32) nach p ab mit analoger partieller Integration:

$$
\begin{aligned}
0 &= \int_{t_0}^{T} \Lambda^T \left(\frac{\partial \dot{y}}{\partial p} - \frac{\partial f}{\partial y} \frac{\partial y}{\partial p} - \frac{\partial f}{\partial p} \right) \mathrm{d}t \\
&= \int_{t_0}^{T} \underbrace{\left(-\dot{\Lambda}^T - \Lambda^T \frac{\partial f}{\partial y} \right)}_{=0} \frac{\partial y}{\partial p} \mathrm{d}t - \int_{t_0}^{T} \Lambda^T \frac{\partial f}{\partial p} \mathrm{d}t + \underbrace{\Lambda^T}_{=I} \frac{\partial y}{\partial p}(T) - \Lambda^T(t_0) \underbrace{\frac{\partial y}{\partial p}(t_0)}_{=0} \\
&= - \int_{t_0}^{T} \Lambda^T \frac{\partial f}{\partial p} \mathrm{d}t + \frac{\partial y}{\partial p}(T) \\
\Rightarrow \frac{\partial y}{\partial p}(T) &= \int_{t_0}^{T} \Lambda^T \frac{\partial f}{\partial p} \mathrm{d}t
\end{aligned}
$$

Bemerkung 1.13

Zur Berechnung von Λ^T muss man erst vorwärts (1.28) lösen und y berechnen, die Werte von y dabei zwischenspeichern und dann rückwärts (1.29) lösen, um Λ^T zu berechnen.

Ableitung von Linearkombinationen von $\frac{\partial y}{\partial y_0}$

Gegeben: $u \in \mathbb{R}^{n_y}$ (adjungierte Richtung). Die Linearkombinations-ADE

$$
\partial_t (u^T \Lambda(t)^T) = -(u^T \Lambda(t)^T) \frac{\partial f}{\partial y}(t, y, p) \quad (1.23)
$$

mit dem Anfangswert

$$(u^T \Lambda(T)^T) = u^T \quad (1.34)$$

hat die Lösung

$$u^T \Lambda(t_0)^T = u^T \frac{\partial y}{\partial y_0}(T) \quad (1.35)$$

$$\text{und} \int_{t_o}^{T} (u^T \Lambda(t)^T) \frac{\partial f}{\partial p}(t, y, p)\, \mathrm{d}t = u^T \frac{\partial y}{\partial p}(T) \quad (1.36)$$

D. h. pro Linearkombination der Zeilen von $\frac{\partial y}{\partial y_0}$ und $\frac{\partial y}{\partial p}$ muss nur eine Rückwärts-AWP gelöst werden.

Anwendung:

$\Phi(y(T; t_0, y_0, p))$ „Zielfunktion“
$\Phi : \mathbb{R}^{n_y} \to \mathbb{R}$

Gradient:

$$\nabla_p \Phi(y(T; t_0, y_0, p)) = \frac{\partial \Phi}{\partial y} \frac{\partial y}{\partial p}(T; t_0, y_0, p)$$

benötigt nur eine adjungierte Richtung $\frac{\partial \Phi}{\partial y} \in \mathbb{R}^{n_y}$

Zusammenfassung:

- Variationsdifferentialgleichung
  - Vorwärtsdifferentiation
  - Richtungsableitungen
- Adjungierte Differentialgleichung
  - Rückwärtsdifferentiation
  - Linearkombinationen

Simulation und Optimierungsprobleme bei Differentialgleichungen

- *Simulation:* Löse die Mathematischen Modellgleichungen, in dieser Vorlesung: Integration von ODE/DAE-AWPn

- *Optimierungsprobleme:*
  - *Parameterschätzung:* Bestimme die Modellparameter p so, dass Modell und Realität möglichst gut übereinstimmen.
  - *Modelldiskriminierung:* Bestimme durch Experimente, welche Modellvariante die Realität besser beschreibt.
  - *Optimale Versuchsplanung:* Bestimme Experimente, aus denen die Parameter mögliuchst signifikant geschätzt werden können.
  - *Optimales Design:* Berechne, wie ein System, Gerät etc. nach einem bestimmten Ziel optimal gebaut werden soll.
  - *Optimale Steuerung:* Berechne, wie ein Prozess nach einem bestimmten Ziel optimal durchgeführt werden soll, typischerweise: Minimale Kosten oder maximale Ausbeute.
  - *Optimale modellbasierte Regelung:* Wie ändert sich die optimale Steuerung bei Störungen des Systems?

2. Formulierung von Parameterschätzproblemen

englisch: Parameter Estimation

Beispiel: Chemische Reaktion („Bimolekulare Katalyse“), zwei Stoffe reagieren zu einem dritten:

$$A + B \to C$$

Es gibt zwei sogenannte Reaktionspfade, einen mit Katalysator und einen ohne Katalysator, beide Pfade laufen parallel ab und die Konzentrationsänderungen durch die beiden Pfade werden addiert. Wir nennen den Reaktionsgeschwindigkeitskoeffizienten k_1 für die Reaktion ohne Katalysator bzw. k_2 für die Reaktion mit Katalysator.

$$k_1 = f_1 \exp\left(-\frac{E_1}{RT}\right)$$

Die Wirkung des Katalysators nimmt mit der Zeit ab, das wird durch eine Exponentialfunktion beschrieben:

$$k_2 = c_{kat} \exp(-\lambda t) f_2 \exp\left(-\frac{E_2}{RT}\right)$$

Reaktionsgeschwindigkeit: $r = k_1 c_1 c_2 + k_2 c_1 c_2$

Daraus ergibt sich ein DGL-System mit Anfangsbedingungen:

$$\begin{aligned}
\dot{n}_1 &= -V(k_1 + k_2)\frac{n_1}{V}\frac{n_2}{V} \\
\dot{n}_2 &= -V(k_1 + k_2)\frac{n_1}{V}\frac{n_2}{V} \\
\dot{n}_3 &= -V(k_1 + k_2)\frac{n_1}{V}\frac{n_2}{V} \\
n_1(0) &= n_{1,0} \\
n_2(0) &= n_{2,0} \\
n_3(0) &= 0
\end{aligned}$$

Wobei n_1 die Anzahl Moleküle des ersten Stoffes beschreibt und V das Gesamt-Volumen aller an dem Experiment beteiligter Stoffe.

Der Prozessverlauf wird bestimmt von Größen, die von dem Experminentator eingestellt werden: $T, V, n_{1,0}, n_{2,0}$. Diese nennen wir „Steuergrößen“ Außerdem hängt der

Verlauf auch von den Größen $f_1, E_1, f_2, E_2, \lambda$ ab. Diese Größen sind durch Naturgesetze, Stoffeigenschaften etc. bestimmte Größen, diese nennen wir "Parameter", sie sind nicht zeitabhängig.

Experiment: Wähle Steuerungen, führe Prozess durch, erhebe Messdaten und Messfehler. Die Messdaten hier liefert eine Apparatur zur Messung von C.

Parameterschätzung: Bestimme die Werte der unbekannten Modellparameter so, dass die Simulation die Messwerte möglichst gut beschreibt.

Problemformulierung

Prozessmodell

$$\dot{y} = f(t, y(t), \overline{p}) \quad y(t_0) = y_0(\overline{p}) \quad (2.1)$$

Modell für Beobachtungen

- Nicht notwendigerweise verschiedene Messzeitpunkte t_i, $i = 1, \cdots, M$
- Messwerte $\eta_i \in \mathbb{R}$ und zugehörige Modellantworten $h_i(t_i, y(t_i), \overline{p}) \in \mathbb{R}$
- Messfehler $\varepsilon_i \in \mathbb{R}$ mit Standardabweichung $\sigma_i \in \mathbb{R}_+$

$\overline{p}$ seien die wahren, aber unbekannten Werte der Parameter.

Annahme: Die Messfehler seien unabhängige, additive und normalverteilte Zufallsgrößen mit bekannter Standardabweichung:

$$\eta_i = h_i(t_i, y(t_i), \overline{p}) + \varepsilon_i \quad (2.2)$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2) \quad (2.3)$$

Erwartungswert 0 bedeutet, dass es keinen systematischen Messfehler gibt.

Bemerkung: Da die Messzeitpunkte nicht notwendig verschieden sind, kann es an einem Zeitpunkt t auch mehrere Messungen geben. Auch die Messfunktionen h_i sind nicht notwendig verschieden.

Beispiele für die Standardabweichung σ :

- absolute Messfehler: $\sigma_i \equiv \sigma$
- relative Messfehler: $\sigma_i = \frac{x}{100} |h_i(t_i, y(t_i), \overline{p})| \sim \frac{x}{100} |\eta_i|$

Wichtig: Zur Beschreibung von Messdaten benötigt man Messwert η und „Genauigkeit“ (Standardabweichung) σ. Wenn man die wahren Parameter $\overline{p}$ nicht kennt kann man trotzdem immernoch die Residuen $\eta_i - h_i(t_i, y(t_i), \overline{p})$, $i = 1, \cdots, M$ oder die gewichteten Residuen $\sigma_i^{-1}(\eta_i - h_i(t_i, y(t_i), \overline{p}))$, $i = 1, \cdots, M$ (2.4) betrachten.

Ziel der Parameterschätzung: passe Modell an die Daten an.

Möglicher Ansatz: Maximum Likelihood, d. h. suche die Parameter, unter denen die beobachteten Daten die höchste Wahrscheinlichkeit haben. Das ist der ML-Schätzer $\hat{p}$:

$$\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$$

Mit der Likelihood-Funktion $L(p)$, die die bedingte Wahrscheinlichkeit der Daten in Abhängigkeit von den Parametern angibt. Oft betrachtet man die sogenannte log-Likelihood-Funktion $\log L$. Für unabhängige, normalverteilte Messfehler gilt:

$$\begin{aligned}
\hat{p} &= \underset{p}{\operatorname{argmax}} \log L(p) \\
&= \underset{p}{\operatorname{argmax}} \log \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2\right) \\
&= \underset{p}{\operatorname{argmax}} \sum_{i=1}^{M} \log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2\right)\right] \\
&= \underset{p}{\operatorname{argmax}} \sum_{i=1}^{M} -\log\sqrt{2\pi\sigma_i^2} + \log\exp\left(-\frac{1}{2}\left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2\right) \\
&= \underset{p}{\operatorname{argmax}} \sum_{i=1}^{M} -\frac{1}{2}\left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2 - \sum_{i=1}^{M} \log\sqrt{2\pi\sigma_i^2} \\
&= \underset{p}{\operatorname{argmax}} -\frac{1}{2} \sum_{i=1}^{M} \left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2 \\
&= \underset{p}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{M} \left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2 \qquad (2.5)
\end{aligned}$$

Also minimiert $\hat{p}$ die $\|\cdot\|_2$-Norm des Residuen-Vektors. Gleichzeitig müssen y und p das AWP (2.1) erfüllen.

Parameterschätzproblem:

$$\min_{p} \frac{1}{2} \sum_{i=1}^{M} \left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i}\right)^2$$

So dass gilt: $\dot{y} = f(t, y(t), p)$, $y(t_0) = y_0(p)$ (2.6).
Wichtig: Es werden immer Gewichte verwendet.

- Wenn man keine hat, macht man die Annahme, dass alle $\sigma_1 \equiv 1$
- Ein gemeinsamer (positiver) Faktor an den verwendeten Gewichten ändert das Problen nicht.

- Das Problem (2.6) und seine Lösung hängen sehr von den Gewichten ab.
- Jede Wahl der Gewichte macht eine spezielle Annahme über die Statistik der Messfehler: $N(0, \sigma_i^2)$-verteilt.

Schreibweisen:

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_M \end{pmatrix}$$

$$h = h(y,p) = \begin{pmatrix} h_1(t_1, y(t_1), p) \\ \vdots \\ h_M(t_M, y(t_M), p) \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_M \end{pmatrix}$$

$$S = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_M^2 \end{pmatrix}$$

Dann gilt: $\varepsilon \sim N(0,S) \Rightarrow S^{-1/2}\varepsilon \sim N(0,I)$
Das Zielfunktional von (2.6) minimiert

$$\frac{1}{2}\|S^{-\frac{1}{2}}(\eta - h)\|_2^2 = \frac{1}{2}(\eta - h)^T S^{-1}(\eta - h) \quad (2.7)$$

S heißt Kovarianzmatrix der Messfehler

$$S = E(\varepsilon\varepsilon^T) \quad \text{„Erwartungswert von } \varepsilon\varepsilon^T\text{“}$$

Für unabhängige normalverteilte Messfehler ist S diagonal und positiv definit. Allgemeiner für unabhängige (korrelierte) multinomialverteile Messfehler ist S symmetrisch und positiv definit.

$$S = \begin{pmatrix} \sigma_1^2 & s_{ij} \\ s_{ji} & \sigma_M^2 \end{pmatrix}$$

mit Varianzen σ_i^2 und Kovarianzen s_{ij}. Zusätzlich definieren wir Korrelationskoeffizienten:

$$r_{ij} = \frac{s_{ij}}{\sigma_i \sigma_j} \quad \text{Korrelation } -1 \leq r_{ij} \leq 1$$

$$R = \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_M^{-1} \end{pmatrix} S \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_M^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & r_{ij} \\ & \ddots & \\ r_{ji} & & 1 \end{pmatrix}$$

R heißt Korrelationsmatrix. Für nicht-diagonales S ist die ML-Schätzung gegeben durch

$$\min(\eta - h(y,p))^T S^{-1} (\eta - h(y,p)) \quad (2.8)$$

Transformation: Zerlege $S = LDL^T$ und minimiere folgendes:

$$\begin{aligned} \frac{1}{2}(\eta - h)^T S^{-1} (\eta - h) &= \frac{1}{2}(\eta - h) L^{-T} D^{-\frac{1}{2}} D^{-\frac{1}{2}} L^{-1} (\eta - h) \\ &= \frac{1}{2} \| D^{-\frac{1}{2}} L^{-1} (\eta - h) \|_2^2 \quad (2.9) \\ D^{-\frac{1}{2}} L^{-1} (\eta - h(y, \overline{p})) &\sim \mathcal{N}(0, I) \end{aligned}$$

Andere Messverteilungen:
Dichte:

$$\exp\left(- \left(\frac{|\varepsilon|}{\sigma} \right)^q \right) \quad (2.10)$$

Schätzer:

$$\min \| S^{-\frac{1}{2}} (\eta - h(y,p)) \|_q \quad (2.11)$$

Beispiele:
„Robuste Parameterschätzung“

$$q = 1$$

$$\min \sum_{i=1}^{M} \left| \frac{\eta_i - h_i(y,p)}{\sigma_i} \right| \quad (2.12)$$

„Worst-Case-Parameterschätzung“

$$q = \infty$$
$$\min \max_{i=1,\cdots,M} \left\{ \left| \frac{\eta_i - h_i(y,p)}{\sigma_i} \right| \right\} \quad (2.13)$$

Nebenbedingung des Parameterschätzproblems:

- y, p müssen die ODE-Modellgleichungen erfüllen: $\dot{y} = f(t, y, p)$
  Variante: y, z, p müssen ein DAE erfüllen $\dot{y} = f(t, y, z, p)$, $0 = g(t, y, z, p)$
- Anfangsbedingung: Konsistenzbedingung $y(t_0) = y_0(p)$, im DAE-Fall zusätzlich $0 = g(t_0, y_0, z(t_0), p)$
- Randbedingungen $r(y(t_0), y(t_{end}), p) = 0$ oder auch $r(y(t_0), y(t_1), \cdots, y(t_{K-1}, y(t_{end}), p) = 0$

Beispiele und Spezialfälle:

- *Periodizität:* $y(t_0) = y(t_{end})$.
- *Innere-Punkt-Bedingung:* $r(y(t_i), p) = 0$.
- *Linear gekoppelte Randbedingungen:* $r(y(t_0), p) + r(y(t_1), p) + \cdots + r(y(t_e nd), p) = 0$.
- *Ungleichsbendingungen:* $s(y(t_0), \cdots, y(t_k), p) \geq 0$, z. B. Vorzeichenbedingung $p_i \geq 0$ für einige i, Schranken $a_i \leq p_i \leq b_i$, Zustandsbeschränkungen $\overline{y}_i \geq y(t_i) \geq \underline{y}_i$

In einem effizienten numerischen Code sollten solche speziellen Strukturen berücksichtigt werden.

Bemerkung zu Ungleichsbedingungen: Wir betrachten PS-Probleme, bei denen man ein physikalisches Modell an experimentelle Daten fitten will. Typischerweise wird das Modell durch Gleichungen beschrieben: DGL-System, Anfangs- und Rand-Bedingungen. Ungleichungen werden dagegen vom Modellierer formuliert mit irgendwie willkürlichen Grenzen. Wenn wir ein PS-Problem mit Ungleichsbedingungen gelöst haben kann folgendes passieren:

- Im Lösungspunkt sind die Ungleichungen inaktiv, d. h. der Lösungspunkt erfüllt die echte Ungleichung $s(y(t_0), \cdots, y(t_k), p) > 0$. Dann können wir die Ungleichung weglassen.
- Im Lösungspunkt sind Ungleichungen aktiv, aber die zugehörigen Lagrange-Multiplikatoren sind Null. Dann können wir die Ungleichung ebenfalls weglassen.
- Eine Ungleichung ist aktiv mit Lagrange-Multiplikator ungleich Null, d. h. wenn man die Ungleichung weglässt wird die Zielfunktion besser, d. h. die Daten werden besser gefittet. Der Schätzer wird dann nicht nur durch die Daten sondern auch durch die vom Modellierer festgelegten Grenzen bestimmt. Das ist im allgemeinen physikalisch nicht sinnvoll.

Es kann sinnvoll sein, während der Algorithmus noch nicht terminiert ist, Grenzen zu fordern, um Auswertbarkeit des Modells zu gewährleisten. Im Lösungspunkt sollten diese Grenzen nicht aktiv sein. Wir behandeln bis auf weiteres Gleichungsbeschränkte PS-Probleme.

Allgemeine Problemformulierung:

$$\min_{p,y} \frac{1}{2} \sum \left(\frac{\eta_i - h_i(t_i, y(t_i), p)}{\sigma_i^2} \right)^2 \quad (2.14)$$

so dass

$$\begin{aligned} \dot{y}(t) &= f(t, y(t), p), \quad y(t_0) = y_0(p) \\ 0 &= r(y(t_0), \cdots, y(t_k), p) \end{aligned}$$

oder mit DAE

$$\begin{aligned} \dot{y}(t) &= f(t, y, z, p), \quad y(t_0) = y_0(p) \\ 0 &= g(t, y, z, p) \end{aligned}$$

Lösungsmethoden

Parametrisierung der Lösung des AWP durch Single Shooting, Multiple Shooting oder Kollokation

Dadurch wird (2.14) endlichdimensional:

$$\min \frac{1}{2} \|F_1(x)\|_2^2 \quad (2.15)$$

so dass $F_2(x) = 0$ mit $x \in \mathbb{R}^n$ geeignet.

Löse (2.15) mit verallgemeinertem Gauß-Newton-Verfahren

3. Shooting-Verfahren und Kollokation

3.1 Single-Shooting: Einfachschießverfahren

Vorgehensweise:

- Wähle Werte für die Parameter p („Initial Guess“)
- Wir lösen das AWP $\dot{y} = f(t, y, p)$, $y(t_0) = y_0(p)$ (3.1) mit einem numerischen Verfahren und erhalten eine Darstellung der Lösung $y(t; t_0 y_0, p)$
- Setze die Lösung an den Messzeitpunkten in die Modellantwortsfunktionen ein und berechne $F_{1,i}(p) = \sigma_i^{-1}(\eta_i - h_i(t_i, y(t_i; t_0, y_0, p), p)$, $i = 1, \cdots, M$ (3.2). Setze die Lösung außerdem an den Randbedingungspunkten in die Randbedingung ein: $F_2(p) = r(y(t_0; t_0, y_0, p), \cdots, y(t_k; t_0, y_o, p), p)$ (3.3). Halte dazu den Integrator an den Punkten t_i an oder benutze die fehlerkontrollierte kontinuierliche Ausgabe.
- Das ergibt ein endlichdimensionales nichtlineares Ausgleichsproblem, nämlich $\min \frac{1}{2}\|F_1(p)\|_2^2$ so dass $F_2(p) = 0$ (3.4).
- Löse diese mit einer geeigneten Methode. Kriterien: (3.5)
  - iterativ, da das Problem nichtlinear ist
  - sollte unzulässige Iterierte erlauben, d. h. Zwischenwerte, bei denen $F_2(p) \neq 0$
  - sollte für Least-Squares-Zielfunktion geeignet sein

Wir benutzen daher das verallgemeinerte Gauß-Newton-Verfahren. Dieses benötigt folgende Ableitungen:

$$J_1(p) := \frac{\partial}{\partial p} F_1(p)$$
$$J_2(p) := \frac{\partial}{\partial p} F_2(p)$$

In jeder Iteration des Gauß-Newton-Verfahrens muss also das AWP gelöst und F_1 und F_2 ausgewertet werden.

Bemerkung 3.1 Berechnung von J_1 und J_2.

$$
\begin{aligned}
J_1(p) :&= \frac{\partial F_1}{\partial p}(p) \\
&= \frac{\partial}{\partial p}\left(\frac{\eta_i - h_i(\cdots)}{\sigma_i}\right)_{i=1,\cdots,M} \\
&= \left(-\frac{1}{\sigma_i}\left(\frac{\partial h_i}{\partial_y}(\cdots)\frac{\partial y}{\partial p}(t_i;t_0,y_0,p) + \frac{\partial h_i}{\partial p}(t_i;t_0,y_0,p)\right)\right)_{i=1,\cdots,M} \qquad (3.6) \\
J_2(p) &= \frac{\partial F_2}{\partial p}(p) \\
&= \sum_{i=0}^{k} \frac{\partial r}{\partial y_i}\frac{\partial y}{\partial p}(t_i;t_0,y_0,p) + \frac{\partial r}{\partial p} \qquad (3.7)
\end{aligned}
$$

Dazu berechnet man $\frac{\partial y}{\partial p} =: G_p$ als Lösung der VDE

$$\dot{G}_p = \frac{\partial f}{\partial y}G_p + \frac{\partial f}{\partial p} \qquad (3.8)$$

sowie die Ableitungen

$$\frac{\partial f}{\partial y}, \frac{\partial f}{\partial p}, \frac{\partial h_i}{\partial y}, \frac{\partial h_i}{\partial p}, \frac{\partial r}{\partial y_i}, \frac{\partial r}{\partial p}$$

der Modellfunktion per Hand, durch numerische Differentiation, oder durch automatische Differentiation.

Algorithmus 3.2: Verallgemeinertes Gauß-Newton-Verfahren

Zur Lösung von $\min_p \frac{1}{2}\|F_1(p)\|_2^2$ s. t. $F_2(p) = 0$

- Start mit einer Startschätzung p^0, $k = 0$ („Initial Guess“)
- Solange ein Abbruchkriterium verletzt ist:
  - Berechne δp^k durch Lösung des linearisierten Ausgleichsproblems $\min \frac{1}{2}\|F_1(p^k) + J_1(p^k)\delta p\|_2^2$ s. t. $F_2(p^k) + J_2(p^k)\delta p = 0$ (3.9)
  - Bestimme eine Schrittweite α^k, z. B. durch Linesearch
  - Iteriere $p^{k+1} = p^k + \alpha^k \delta p^k$ (3.10)

Mögliches Abbruchkriterium: $\|\delta p^k\| \leq \varepsilon$. Mehr zu Gauß-Newton-Verfahren in Kapitel 4.

Algorithmus 3.3: Single-Shooting Gauß-Newton

- Start mit einer Startschätzung p^0, $k = 0$

- Solange ein Abbruchkriterium verletzt ist:
  - Integriere das AWP (3.1) zusammen mit der VDE (3.8) für $p = p^k$
  - Halte an den Punkten t_i an und werte $F_1(p^k)$ und $F_2(p^k)$ gemäß (3.2) und (3.3) aus. Berechne $J_1(p^k)$ und $J_2(p^k)$ gemäß (3.6) und (3.7)
  - Berechne δp^k durch Lösen des linearen Ausgleichsproblems (3.9).
  - Berechne eine Schrittweite δp^k und iteriere gemäß (3.10)

Implementierung: Praktische Aufgabe 1
Benötigte Bestandteile für die Implementierung:

- Integrator für AWP/VDE: entweder durch Integrator mit Interner Numerischer Integration („IND“), siehe Kapitel 5. Oder durch Integrieren des Systems

$$\begin{pmatrix} \dot{y} \\ \dot{G}_p \end{pmatrix} = \begin{pmatrix} f \\ \frac{\partial f}{\partial y} G_p + \frac{\partial f}{\partial p} \end{pmatrix}, \quad y(t_0) = \begin{pmatrix} y_0(p) \\ \frac{\partial y_0}{\partial p}(p) \end{pmatrix}$$

- Löser für lineare Ausgleichsprobleme $\min \frac{1}{2}\|F_1 + J_1\delta x\|_2^2$ s. t. $F_2 + J_2\delta x = 0$. KKT-Bedingung:

$$\exists \lambda \colon \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \begin{pmatrix} \delta x \\ \lambda \end{pmatrix} = - \begin{pmatrix} J_1^T F_1 \\ F_2 \end{pmatrix}$$

- Globalisierungs-Strategie: BT-Linesearch oder $\alpha^j = 1$

Bemerkung 3.4

Wen wir keine Randbedingung $r(\cdots) = 0$ haben und die Differentialgleichung eine ODE und keine DAE ist dann lösen wir in jeder Iteration des Gauß-Newton-Verfahrens („GN-Verfahren“) für die aktuellen Paramter das komplette Simulationsproblem. Das Ausgleichsproblem ist dann unbeschränkt: $\min_p \frac{1}{2}\|F_1(p)\|_2^2$. Das linearisierte Ausgleichsproblem $\min_{\Delta p} \frac{1}{2}\|F_1 + \partial_1 \Delta p\|_2^2$ hat die Lösung $\Delta p = -(J_1^T J_1)^{-1} J_1^T F_1$. Berechnung über QR-Zerlegung von J_1.

Schwierigkeiten beim Single-Shooting:

- In Satz 1.2 (Trompetenabschätzung) haben wir gesehen, dass kleine Störungen der Parameter sehr große Änderungen der Lösung y des AWP und damit auch große Änderungen des Zielfunktions-Wertes und Nebenbedingungen des Parameter-Schätz-Problems zur Folge haben können. Wir sind dann „weit weg“ von der Lösung und das GN-Verfahren konvergiert nicht, siehe auch Kapitel 4.
- Wenn die Ableitungen $\frac{\partial f}{\partial y}$ nicht auf ganz $[t_0, t_{end}] \times R^{n_y} \times \mathbb{R}^{n_p}$ beschränkt bleiben kann es sein, dass für Parameter-Werte weit weg von den wahren Parametern die Lösung des AWP nicht auf dem ganzen Intervall $[t_0, t_{end}]$ existiert. Dann ist das PS-Problem nicht auswertbar.
- Die Vorinformation über die Trajektorien, die durch die Messwerte gegeben ist wird nicht genutzt.

3.2 Multiple Shooting, die Mehrzielmethode

Zerlege das Integrationsintervall in Teilintervalle $t_0 = \tau_0 < \tau_1 < \cdots < \tau_m = t_{end}$. Die τ_i nennt man Mehrzielknoten. Löse das AWP nur auf diesen Intervallen $i = 0, \cdots, m-1$: $\dot{y} = f(t, y, p)$, $t \in [\tau_i, \tau_{i+1}]$ (3.10) mit den Anfangsbedingungen $y(\tau_i) = s_i$ (3.11). Erhalte die Lösung $y(t; \tau_i, s_i, p)$, $t \in [\tau_i, \tau_{i+1}]$ (3.12). Setze diese y in die Messfunktion und Randbedingungs-Funktion an den entsprechenden Zeitpunkten t_i ein. Die s_i, $i = 0, \cdots, m-1$ sind zusätzliche Variablen des Problems. An den Mehrzielknoten τ_i formulieren wir zusätzliche Stetigkeitsbedingungen, sog. „Anschlussbedingungen“:

$$s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) = 0 \quad i = 0, \cdots, m-2 \quad (3.13)$$

Für DAEs ist $s_i = (s_i^y, s_i^z)$ und im Punkt τ_i müssen die Konsistenzbedingungen $g(\tau_i, s_i^y, s_i^z, p) = 0$, $i = 0, \cdots, m-1$ erfüllt sein (3.14). Anschlussbedingungen braucht man dann nur für die s_i^y Das ergibt insgesamt das beschränkte, nichtlineare Ausgleichsproblem

$$\min_x \frac{1}{2} \|F_1(x)\|_2^2 \quad (3.15)$$

so dass $F_2(x) = 0$

- mit den Variablen $x = (s_0, \cdots, s_{m-1}, p)$ (3.16)
- den Zielfunktionstermen

  $$F_{1,j}(x) = \frac{1}{\sigma_j} \left(\eta_j - h_j(t_j, y(t_j; \tau_{i_j}, s_{i_j}, p), p) \right) \quad (3.17)$$

  mit $t_j \in [\tau_{i_j}, \tau_{i_j}]$, $j = 1, \cdots, M$
- den Nebenbedingungen $F_2(x)$, die bestehen aus:
  - den Randbedingungen: $r(y(t_0; \tau_0, s_0, p), \cdots, y(t_k; \tau_{i_k}, s_{i_k}, p), p) = 0$ (3.18) mit $t_j \in [\tau_{i_j}, \tau_{i_j+1}]$, $i = 1, \cdots, k$
  - den Anschlussbedingungen $s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) = 0$, $i = 0, \cdots, m-2$ (3.19)
  - und für DAEs den Konsistenzbedingungen $g(\tau_i, s_i^y, s_i^z, p) = 0$, $i = 0, \cdots, m-1$ (3.20).

Die Lösungen der Probleme (3.4) (Single-Shooting) und (3.15) (Multiple-Shooting) sind identisch. Das Problem (3.15) hat aber mehr Variablen.

Nachteile von Multiple-Shooting:

- Wesentlich höherer Programmieraufwand.
- Mehraufwand bei der Bestimmung von Startwerten für die Variablen
- Höherdimensionales Optimierungsproblem

Vorteile von Multiple-Shooting:

- Die Existenz einer (unstetigen) „Start-Trajektorie“ auf dem gesamten Intervall $[t_0, t_{end}]$ ist gesichert.
- Oft können die Multiple-Shooting-Variablen s_i durch Messwerte initialisiert werden.
- Man kann viel näher an der Lösung $(s_0^*, s_1^*, \cdots, s_{m-1}^*, p)$ des Problems starten. Der Einfluss schlechter Startwerte für die p wird abgemildert. Die Chance auf Konvergenz des GN-Verfahrens ist höher.
- Die Nichtlinearität des Problems wird reduziert: Bei Single-Shooting $y(t_{end}; t_0, y_0, p)$ (3.21), bei Multiple-Shooting $y(t_{i+1}; \tau_i, s_i, p)$ (3.22). Nach der Trompetenabschätzung ist die Nichtlinearität in (3.21) i. A. größer als in (3.22).

Zur Lösung von (3.15) verwenden wir wieder das verallgemeinerte GN-Verfahren und benötigen dazu die Ableitungen $J_1 = \partial_x F_1$ und $J_2 = \partial_x F_2$ (3.23). Diese haben die folgende Gestalt:

$$\begin{pmatrix} J_1 \\ J_2 \end{pmatrix} = \ldots\ldots\ldots \quad (3.22a)$$

mit

$$D_1^i = \frac{\partial F_1}{\partial s_i} \quad i = 0, \cdots, m-1$$
$$D_1^p = \frac{\partial F_1}{\partial p} \qquad \text{Ableitung der ZF-Terme}$$
$$D_2^o = \frac{\partial r}{\partial s_i} \quad i = 0, \cdots, m-1$$
$$D_2^p = \frac{\partial r}{\partial p} \qquad \text{Ableitung der Randbedingungen}$$

Ableitung der Anschlussbedingungen:

$$-s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) = 0 \quad i = 0, \cdots, m-2$$
$$\text{nach } s_i : \frac{\partial}{\partial s_i} y(\tau_{i+1}; \tau_i, s_i, p) =: G_i$$
$$\text{nach } s_{i+1} : \; -I$$
$$\text{nach } p : \frac{\partial}{\partial p} y(\tau_{i+1}; \tau_i, s_i, p) =: G_i^p \qquad \text{(sonst 0)}$$

Ableitung der Konsistenzbedingungen:

$$H_i := \frac{\partial}{\partial s_i} g(\tau; i, s_i^y, s_i^z, p) \quad i = 0, \cdots, m-1$$
$$H_i^p = \frac{\partial}{\partial p} g(\tau_i, s_i^y, s_i^z, p) \quad i = 0, \cdots, m-1$$

Lösungsverfahren müssen diese spezielle Struktur ausnutzen („Kondensierung“, siehe Kapitel 4). Die D-Matrizen können mit der Kettenregel berechnet werden, z. B.

$$\frac{\partial F_1}{\partial p} = \sum_{j=1}^{M} \frac{\partial F_1}{\partial y(t_j)} \frac{\partial y(t_j)}{\partial p} + \frac{\partial F_1}{\partial p}$$
$$\frac{\partial F_1}{\partial s_i} = \sum_{j=1}^{M} \frac{\partial F_1}{\partial y(t_j)} \frac{\partial y(t_j)}{\partial s_i}$$

Bemerkung 3.5

Der Aufwand zur Integration und zur Berechnung der G_i ist bei Single-Shooting und Multiple-Shooting im Wesentlichen gleich.

3.3 Kollokation

Literatur: Ascher, Mettheij, Russel: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations

Biegler: Nonlinear Programming

Kollokations-Diskretisierung

Approximiere die Lösung der ODE $\dot{y} = f(t, y, p)$ (3.24) durch stückweise Polynome vom Grad k auf einem Gitter $t_0 = \tau_0 < \tau_1 < \cdots < \tau_{m-1} < \tau_m = t_{end}$ (3.25). Auf jedem dieser Teilintervalle wird die Lösung dargestellt durch

$$y_j^\tau(t; s_j) : = \sum_{l=0}^{k} s_j l \chi_l \left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j} \right) \tag{3.2}$$
$$t \in [\tau_j, \tau_{j+1}], \quad j = 0, \cdots, m-1$$

Wobei $\{\chi_l\}_{l=0,\cdots,k}$ eine Basis des Polynomraums $P_k([0,1])$ der Polynome vom Grad $\leq k$ ist

Die Koeffizienten $s_{jl} \in \mathbb{R}^{n_y}$, $j = 0, \cdots, m-1$, $l = 0, \cdots, k$, $s_j = (s_{j0}, \cdots, s_{jk})$, $j = 0, \cdots, m-$ (3.27) sind Variablen und bestimmt durch diese Bedingungen:

- Die approximierte Lösung y_j^τ soll auf einer Unterteilung des Gitters die ODE erfüllen:

$$t_{jl} := \overline{c}_j + \rho_l h_j \quad l = 1, \cdots, k, \quad j = 0, \cdots, m-1 \tag{3.28}$$

(Kollokationspunkte) $\rho_l \in [0,1]$, $h_j := \tau_{j+1} - \tau_j$. Also:

$$\dot{y}_j^\tau(t_{jl}; s_j) = f(t_{jl}, y^\tau(t_{jl}, s_j), p), \quad j = 0, \cdots, m-1 \tag{3.29}$$

- Die approximierte Lösung y^τ soll an den Gitterpunkten τ_j stetig sein:

$$y_j^\tau(\tau_{j+1}; s_j) = y_{j+1}^\tau(\tau_{j+1}, s_{j+1}) \quad j = 0, \cdots, m-2 \tag{3.30}$$

- Anfangsbedingung:

$$y^\tau(\tau_0; s_0) = y_0(p) \tag{3.31}$$

Das sind mk Kollokationsbedingungen, $m-1$ Stetigkeitsbedingungen und eine Anfangsbedingung, also $m(k+1)$ Bedingungen für die $m(k+1)$ Variablen s_{jl}.

Wahl der Polynombasis und der Kollokationspunkte

z. B. B-Splines, Hermite-Splines. Wir benutzen die Runge-Kutta-Basis:

$$y_j^\tau(t; s_j) = s_{jl} + h_j \sum_{l=1}^{k} s_{jl} \Psi_l \left(\frac{t - \tau_j}{\tau_{j+1} - \tau_j} \right) \tag{3.32}$$

$$\text{mit } \Psi_l(0) = 0$$

$$\text{und } \dot{\Psi}_l(\rho_i) = \begin{cases} 1 & i = l \\ 0 & \text{sonst} \end{cases}$$

$$\Psi_l \in \mathcal{P}_k([0,1])$$

Dann lauten die Kollokationsbedingungen:

$$\begin{aligned} \dot{y}_j^\tau(t_{jl}; s_j) &= \dot{y}_j^\tau(\tau_j + \rho_l h_j; s_j) \\ &= f(t_{jl} y^\tau(t_{jl}; s_j), p) \quad l = 1, \cdots, k \end{aligned} \tag{3.33}$$

und die Stetigkeitsbedingunng:

$$y_{j-1}^\tau(\tau_j; s_{j-1} - s_{j0} = 0 \tag{3.34}$$

Die Basis $\{1, \Psi_1, \cdots, \Psi_k\}$ heißt Runge-Kutta-Basis, weil die s_{jl}, $l = 1, \cdots, k$ dabei die Stufen des Runge-Kutta-Schemas

$$\begin{pmatrix} \rho_1| & \Psi_1(\rho_1) & \cdots & \Psi_k(\rho_1) \\ \vdots| & \vdots & & \vdots \\ \rho_k| & \Psi_1(\rho_k) & \cdots & \Psi_k(\rho_k) \\ & \Psi_1(1) & \cdots & \Psi_k(1) \end{pmatrix} \tag{3.35}$$

$$\text{mit } \Psi_l(\rho) = \int_0^\rho L_l(\overline{\rho})\,\mathrm{d}\overline{\rho}$$

$$\Psi_l(1) = \int_0^1 L_l(\rho)\,\mathrm{d}\rho$$

$$\text{und } L_l(\rho) = \prod_{i \neq l} \frac{\rho - \rho_i}{\rho_l - \rho_i}$$

sind

Man kann also in jedem Intervall $[\tau_j, \tau_{j+1})$ $y_j^\tau(\tau_{j+1}, s_j)$ als Ergebnis eines Schrittes eines impliziten Runge-Kutta-Verfahrens, gestartet bei $y_j^\tau(\tau_j, s_j)$ auffassen: „Kollokationsschema“

Konsistenzfehler:

$$y(\tau_{j+1}) = y(\tau_j) + h_j \sum_{l=1}^{k} f(t_j + \rho_l h_j, y(\tau_j + \rho_l h_j), p) \int_0^1 L_l(\rho)\,\mathrm{d}\rho + h_j \mathcal{O}(h_j^q) \tag{3.36}$$

Die Konsistenzordnung q hängt ab von der Wahl der ρ_l, $l = 1, \cdots, k$. Für Gauß-Punkte (Nullstellen der Legendre-Polynome) erhält man die maximal mögliche Ordnung $q = 2(k-1) + 2 = 2k$. Für die Lobatto-Punkte erhält man die Ordnung $q = 2k - 2$, „Radau-Kollokations-Schema“.

Beispiel Impliziter Euler, k=1

$$y_j^\tau(t; s_j) = s_{j0} + h_j s_{j1}$$

Kollokationsbedingung:

$$s_{j1} = f(\tau_{j+1} + h_j, s_{j0} + h_j s_{j1}, p)$$

Stetigkeitsbedingung:

$$s_{j-1\,0} + h_{j-1} s_{j-1} - s_{j0} = 0$$

Kollokation für beschränkte Parameterschätzprobleme

Approximiere die Lösung der ODE durch eine Kollokationsdiskretisierung $y_j^\tau(t; s_j)$, $j = 0, \cdots, m-1$ Setze diese an den Messzeitpunkten in die Least-Squares-Terme und an den Ranbedingungs-Punkten in die Randbedingung ein. Variablen sind:

$$x = (s_0, s_1, \cdots, s_{m-1}, p) \in \underbrace{\mathbb{R}^{(k+1)n_y} \times \cdots \times \mathbb{R}^{(k+1)n_y}}_{m} \times \mathbb{R}^{n_p}$$

F_1 besteht auf den Least-Squares-Termen, F_2 besteht au den Randbedingungen, den Kollokationsbedingungen (3.29), den Stetigkeitsbedingungen (3.30) und den Anfangsbedingungen (3.31). Es ergibt sich wieder ein beschränktes nichtlineares Ausgleichsproblem:

$$\min_x \frac{1}{2}\|F_1(x)\|_2^2 \quad s.\,t.\ F_2(x) = 0 \quad (3.37)$$

Zur Lösung mit dem verallgemeinerten Gauß-Newton-Verfahren benötigen wir wieder

$$J_1 = \frac{\partial F_1}{\partial x}$$
$$J_2 = \frac{\partial F_2}{\partial y}$$

Bemerkung 3.6: Eigenschaften von $\begin{pmatrix}\partial_1 \\ \partial_2\end{pmatrix}$

- Es müssen keine Variations-DGL gelöst werden. Die Ableitungen nach s und p ergeben sich allein aus den Ableitungen der Kollokationspolynome und Ableitungen von f, Messfunktionen und Randbedingungen.
- Das System ist sehr groß, aber auch sehr dünn besetzt.

Vorteile von Kollokation

- Es werden keine Integratoren und Variations-DGL-Löser benötigt.
- Es können Standard-Sparse-Löser eingesetzt werden
- Für lineare DGL, lineare Messfunktion und lineare Randbedingungen ist das Ausgleichs-Problem linear.
- Simulationsproblem und Optimierungsproblem werden in einer einzigen Schleife gelöst: All-At-Once

Nachteile von Kollokation

- Die DGL wird nicht adaptiv diskretisiert
- Das Problem ist sehr hochdimensional

3.4 Ansätze zur Optimierung von DGL-Modellen

Wir haben bisher kennengelernt:

- Single-Shooting („Black-Box-Ansatz“)
- Multiple Shooting
- Kollokation

Das Optimierungsproblem muss iterativ gelöst werden.
Die konzeptionellen Unterschiede sind:

- Beim Single-Shooting wird in jeder Optimierungsiteration die DGL des AWP komplett gelöst. Das ist der sogenannte sequentielle Ansatz.
- Bei Multiple Shooting wird die Differentialgleichung in jeder Optimierungsiteration nur auf Teilintervallen gelöst, das AWP wird relaxiert, d. h. wir lassen während der Iterationen unstetige Lösungen zu. Erst im Lösungspunkt wird das AWP gelöst, simultan mit Nebenbedingungen und der Optimalität. Das ist ein Beispiel für einen simultanen Ansatz.
- Bei Kollokation wird die DGL vollständig diskretisiert und gemeinsam mit dem Optimierungsproblem gelöst. Das nennt man den All-At-Once-Ansatz.

Unsere Optimierungsverfahren lassen unzulässige Iterierte zu. Die Lösungen dieser Formulierungen sind gleich, da die Formulierungen mathematisch äquivalent sind. Die Iterationen der Verfahren können sich aber stark unterscheiden. Es werden während der Optimierung Schritte in verschieden hochdimensionalen Suchräumen berechnet. Dies kann bessere Konvergenzeigenschaften bedeuten.

Verallgemeinerung: Lifting von beliebigen nichtlinearen Problemen in höherdimensionale Räume:

$$
\begin{aligned}
x^{16} &= 2 \\
&\Leftrightarrow \\
x_1^2 - x_2 &= 0 \\
x_2^2 - x_3 &= 0 \\
x_3^2 - x_4 &= 0 \\
x_4^2 - 2 &= 0
\end{aligned}
$$

siehe Albersmeyer, Diehl: The lifted Newton Method and its Apllication in Optimization

3.5 Relaxierte Formulierung von DAEs

DAE:

$$\begin{aligned} \dot{y} &= f(t, y, z, p) \\ 0 &= g(t, y, z, p) \qquad \text{Annahme: Index 1} \\ \text{Anfangsbedingung: } y(t_0) &= y_0, \quad z(t_0) = z_0 \\ \text{Konsistenzbedingung: } 0 &= g(t_0, y_0, z_0, p) \end{aligned}$$

Für numerische Lösungsverfahren werden konsistente Anfangswerte benötigt. Dafür gibt es zwei Ansätze:

- Berechnung konsistenter Anfangswerte durch Lösen der Konsistenzbedingung mit z. B. einem Newton-Verfahren. Kann Schwierigkeiten bereiten, weil das Newton-Verfahren zunächst nur lokal konvergiert und globalisiert werden muss, z. B. mit Homotopie-Ansätzen. Das muss vor jeder Integration der DAE erfolgen.
- Integration mit „inkonsistenten“ Anfangswerten, relaxierte Formulierung:

$$\begin{aligned} \dot{y} &= f(t, y, z, p) \qquad (3.40) \\ 0 &= g(t, y, z, p) - \beta(t) g(t_0, y_0, z_0, p) \\ y(t_0) &= y_0 \\ z(t_0) &= z_0 \\ \text{mit } \beta(t_0) &= 1 \\ \text{z. B. } \beta(t) &\equiv 1 \\ \text{oder } \beta(t) &= e^{-\alpha(t-t_0)} \end{aligned}$$

Das modifizierte Problem (3.40) ist automatisch konsistent und kann numerisch integriert werden. Interpretation: Die Nebenbedingungen geben eine Mannigfaltigkeit vor, auf der die Lösung liegt. In der relaxierten Formulierung ist diese Mannigfaltigkeit verschoben. Die Konsistenzbedingung $g(t_0, y_0, z_0, p) = 0$ (3.41) löst man im übergeordneten Optimierungsproblem als Nebenbedingung mit.

Bei Single-Shooting wird die Konsistenzbedingung im Anfangszeitpunkt erfüllt, dafür wird das Gleichungssystem gelöst.

Bei Multiple-Shooting wird für jedes Mehrzielintervall eine relaxierte Konsistenzbedingung eingesetzt. In jedem Mehrzielknoten: $g(\tau_i, s_i^y, s_i^z, p) = 0$ für $i = 0, \cdots, m-1$.

Bei Kollokation wird kein Integrator verwendet. Die algebraischen Nebenbedingungen werden zusätzlich in den Kollokationspunkten gefordert. (in RK-Basis-Darstellung):

$$\begin{aligned} y_j^\tau(t_{jl}; s_j) &= s_{jl}^y \\ &= f(t_{jl}, y_j^\tau(t_{jl}; s), s_{jl}^z, p) \\ 0 &= g(t_{jl}, y_j^\tau(t_{jl}; s_j), s_{jl}^z, p) \\ \text{mit } s_{jl}^y &\cong \dot{y}(t_{jl}), \quad s_{jl}^z \cong z(t_{jl}) \end{aligned}$$

Stetigkeitsbedingung nur für s^y:

$$y_{j-1}^\tau(\tau_j; s_{j-1}) - s_{j_0}^y = 0$$

Die Bestimmung konsistenter Werte für die algebraischen Gleichungen wird von der Simulationsaufgabe in die Ebene des Optimierungsproblems verlagert.

4. Verallgemeinerte Gauß-Newton-Verfahren

Problem:

$$\begin{aligned} \min \frac{1}{2}\|F_1(x)\|_2^2 \text{ s. t. } F_2(x) = 0 \quad (4.1) \\ \text{mit } F_1 \colon D \subset \mathbb{R}^n \to \mathbb{R}^{m_1}, \quad F_1 \in C^2(D) \\ F_2 \colon D \subset \mathbb{R}^n \to R^{m_2}, \quad F_2 \in C^2(D) \\ n \geq m_2 \\ m_1 \geq n - m_2 \\ \text{Jacobi-Matrizen: } J_1 \colon = \frac{\partial F_1}{\partial x}, \quad J_2 \colon = \frac{\partial F_2}{\partial x} \end{aligned}$$

- Unbeschränkter Fall: $\min \frac{1}{2}\|F(X)\|_2^2$
- Spezialfall: nichtlineare Gleichung: $F(x) = 0$

Algorithmus 4.1 (Verallgemeinertes Gauß-Newton-Verfahren)

- Startpunkt x^0, $k \colon = 0$
- Solange ein Abbruchkriterium verletzt ist (z. B. $\|\Delta x\| > \varepsilon$):
  - Berechne

$$\begin{aligned} F_1^k &\colon = F_1(x^k) \\ F_2^k &\colon = F_2(x^k) \\ J_1^k &\colon = J_1(x^k) \\ J_2^k &\colon = J_2(x^k) \end{aligned}$$

  - Löse das lineare Ausgleichsproblem

$$\min \frac{1}{2}\|F_1^k + J_1^k \Delta x\|_2^2 \text{ s. t. } F_2^k + J_2^k \Delta x = 0 (4.2)$$

    Lösung: Δx^k

- Bestimme eine Schrittweite $\alpha^k \in (0, 1]$
- Iteriere $x^{k+1} := x^k + \alpha^k \Delta x^k$

„Newton“: löse iterativ, linearisiere in jeder Iteration.
„Gauß“: linearisiere innerhalb der Norm, löse das lineare Ausgleichsproblem.

Bemerkung 4.2

Zur Globalisierung der Konvergenz kann z. B. Linesearch verwendet werden. Es ergeben sich gedämpfte Schritte $\alpha^k < 1$. In der Nähe der lösung können Vollschritte $\alpha^k = 1$ erwartet werden.

Annahmen 4.3: Regularitätsannahmen

- (CQ) „Constraint Qualification“: $\mathrm{Rg} J_2 = m_2$ (J_2 hat vollen Rang, die Nebenbedingungen sind widerspruchsfrei und nicht redundant).
- (PD) „Positive Definiteness“:

$$\mathrm{Rg} \begin{pmatrix} J_1 \\ J_2 \end{pmatrix} = n \quad (4.4)$$

  bedeutet: die nicht durch die Nebenbedingungen festgelegten Variablen können aus den experimentellen Daten eindeutig geschätzt werden.

Lemma 4.4

Gelten (PD) und (CQ), dann gilt:

- Die Matrix $J_1^T J_1$ ist positiv definit auf $\ker J_2$
- Die Matrix

$$\begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}$$

  ist regulär.

Beweis: Übungsaufgabe.

Lösung der linearen Ausgleichsprobleme

1. Unbeschränkter Fall

$$\min \frac{1}{2} \|F_1 + J_1 \Delta x\|_2^2 \quad (4.5)$$

Es gelte (PD), d. h. J_2 habe vollen Rang. Δx^* ist Lösung von (4.5) genau dann, wenn es das sogenannte Normalgleichungssystem löst:

$$J_1^T J_1 \Delta x^* + J_1^T F_1 = 0$$

Δx^* ist Minimum genau dann wenn

$$\begin{aligned}
F_1 + J_1\Delta x^* &\perp \{J_1 \Delta x\} \\
\Leftrightarrow (\Delta x^* J_1^T)(J_1\Delta x^* + F_1) &= 0 \, \forall \Delta x \\
\Leftrightarrow \Delta x^*(J_1^T J_1 \Delta x^* + J_1^T F_1) &= 0 \\
\Leftrightarrow J_1^T J_1 \Delta x^* + J_1^T F_1 &= 0 \\
\Leftrightarrow \Delta x^* &= -(J_1^T J_1)^{-1} J_1^T F_1 \\
&= -J^\dagger F_1 \qquad (4.6)
\end{aligned}$$

$J^\dagger$ heißt Moore-Penrose-Pseudoinverse und erfüllt die vier Moore-Penrose-Axiome:

- $(J^\dagger J)^T = J^\dagger J$
- $(JJ^\dagger)^T = JJ^\dagger$
- $JJ^\dagger J = J$
- $J^\dagger JJ^\dagger = J^\dagger$

Umgekehrt: $J^\dagger$ ist durch die vier Axiome eindeutig bestimmt. Die Lösung der Normalengleichung ist eindeutig, wenn (PD) erfüllt ist.

Bemerkung 4.10

Wenn $\mathrm{Rg} J_1 < n$, also (PD) nicht erfüllt ist, dann ist die Lösung von $\min \frac{1}{2}\|F_1 + J_1\Delta x\|_2^2$ nicht eindeutig. Dann ist

$$\Delta x = -J^\dagger F_1$$

Die Lösung kleinster euklidischer Norm. $J^\dagger$ erfülle dabei die vier Moore-Penrose-Axiome.
Beweis: Übungsaufgabe

2. Beschränkter Fall

Lemma 4.5

Gelten (CQ) und (PD) dann gilt
$\Delta x^* \in \mathbb{R}^n$ ist genau dann Lösung von $\min \frac{1}{2}\|F_1 + F_1\Delta x\|_2^2$, s. t. $F_2 + J_2\Delta x = 0$ (4.6), wenn

$$\Delta x^* = -\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \tag{4.7}$$

Sei Δx^* Minimum.

$$f(t) \colon = \frac{1}{2}\|F_1 + J_1(\Delta x^* + t\Delta y)\|_2^2$$

mit Δy so dass

$$\begin{aligned}
F_2 + J_2(\Delta x^* + t\Delta y) &= 0 \\
\Rightarrow J_2 \Delta y &= 0 \\
\text{d. h. } \Delta y &\in \ker(J_2) \\
0 = f'(0) &= \Delta y^T (J_1^T J_1 \Delta x^* + J_1^T F_1) \\
\Rightarrow J_1^T J_1 \Delta x^* + J_1^T F_1 &\in (\ker J_2)^\perp = \text{Bild}\left(J_2^T\right) \\
\Rightarrow \exists! \lambda \text{ mit } J_1^T J_1 \Delta x^* + J_1^T F_1 + J_2^T \lambda &= 0 \\
\text{außerdem } J_2 \Delta x^* + F_2 &= 0 \quad \text{(KKT-Bedinungen)} \\
\Rightarrow \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \begin{pmatrix} \Delta x^* \\ \lambda \end{pmatrix} &= -\begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \qquad (4.8) \\
\Rightarrow \Delta x^* &= -\underbrace{\begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix}}_{=: J^+ \quad (4.9)} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \\
\Delta x^* &= -J^+ F
\end{aligned}$$

Gelte umgekehrt (4.8) für Δx^* und λ. Dann folgt $f'(0) = 0$ für alle $\Delta y \in \ker(J_2)$. Aus (PD) folgt, dass Δx^* ein Minimum ist.

Definition 4.6: Verallgemeinerte Inverse

Der Lösungsoperator J^+ von (4.8) heißt verallgemeinerte Inverse von

$$J = \begin{pmatrix} J_1 \\ J_2 \end{pmatrix}$$

Lemma 4.7

J^+ erfüllt das Moore-Penrose-Axiom $J^+ J J^+ = J^+$.

Beweis: Setz

$$F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} J_1 J^+ y \\ J_2 J^+ y \end{pmatrix}$$

in (4.6) ein:

$$\min \frac{1}{2} \|F_1 + J_1 \Delta x\|_2^2 \text{ s. t. } F_2 + J_2 \Delta x = 0$$

Erste Lösung:

$$\begin{aligned} \Delta x_1 &= -J^+ F \\ &= -J^+ (J J^+ y) \end{aligned}$$

Zweite Lösung:

$$\begin{aligned} \Delta x_2 &= -J^+ y \\ \text{NR: } \underbrace{J_1 J^+ y}_{=F_1} - J_1 J^+ y &= 0 \\ \underbrace{J_2 J^+ y}_{=F_2} - J_2 J^+ y &= 0 \end{aligned}$$

Wegen der Eindeutigkeit der Lösung ist $\Delta x_1 = \Delta x_2$. $-J^+ J J^+ y = -J^+ y$ für beliebige y, $J^+ J J^+ = J^+$.

Lemma 4.8

J^+ Erfüllt die Moore-Penrose-Axiome 1,2 und 4. Beweis: Übungsaufgabe.

Bemerkung 4.9

Die Moore-Penrose-Pseudoinverse für den beschränkten Fall würde das folgende Problem lösen:

$$\min_{\Delta x} \frac{1}{2} \left\| \begin{pmatrix} J_1 \\ J_2 \end{pmatrix} \Delta x + \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \right\|_2^2$$

Das ist nicht äquivalent zu

$$\begin{aligned} &\min_{\Delta x} \frac{1}{2} \|F_1 + J_1 \Delta x\|_2^2 \\ &\text{s. t. } F_2 + J_2 \Delta x = 0 \end{aligned}$$

Numerische Lösung

Man stellt nicht J^+ oder $J^\dagger$ auf, sondern zerlegt J_1 und J_2 geeignet.

1. Unbeschränkter Fall

$$J_1^T J_1 \Delta x = -J_1^T F_1$$

- 1. Variante: QR-Zerlegung von $J_1 = QR = \overline{QR}$

$$\begin{aligned} J_1^T J_1 \Delta x &= \overline{R}^T \underbrace{\overline{Q}^T \overline{Q}}_{=I} \overline{R} \Delta x \\ &= -\overline{R}^T \overline{Q}^T F_1 \\ \Delta x &= -\overline{R}^{-1} \overline{Q}^T F_1 \\ J^\dagger &= \overline{R}^{-1} \overline{Q}^T \end{aligned}$$

- 2. Variante: Singulärwertzerlegung von J_1:

$$J_1 = U \Sigma V^T$$

mit U, V orthogonal und Σ Diagonalmatrix mit Singulärwerten von J_1 auf der Diagonalen.

$$\Delta x = -V \begin{pmatrix} \sigma_1^{-1} & & \\ & \ddots & 0 \\ & & \sigma_n^{-1} \end{pmatrix} U^T F_1$$

$$J^\dagger = V \begin{pmatrix} \sigma_1^{-1} & & \\ & \ddots & 0 \\ & & \sigma_n^{-1} \end{pmatrix} U^T$$

Beschränkter Fall

- 1. Variante („Bildraumvariante“): Δx ist Lösung von

$$\begin{pmatrix} J_q^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda \end{pmatrix} = - \begin{pmatrix} J_1^T F_1 \\ F_2 \end{pmatrix}$$

Diese Matrix ist symmetrisch aber indefinit, also wende Gaußsche LR-Zerlegung an

- 2. Variante („Nullraumvariante“):

$$\min_{\Delta x} \frac{1}{2}\|F_1 + J_1\Delta x\|_2^2 \text{ (4.10) s. t. } F_2 + J_2\Delta x = 0 \text{ (4.11)}$$

1. Bestimme Lösungsmenge von (4.11) durch LR-Zerlegung von P_2J_2:

$$P_2J_2 = L_2 \begin{pmatrix} R_2 & D_2 \end{pmatrix} = \begin{pmatrix} P_2J_{21} & P_2J_{22} \end{pmatrix} \quad (4.12)$$

Mit R_2 reguläre obere Dreiecksmatrix, L_2 normierte untere Dreiecksmatrix, P_2 Permutationsmatrix zur Zeilenpivotierung. Vorraussetzung ist (CQ). Spalte Δx auf:

$$\begin{aligned} \Delta x &= \begin{pmatrix} \underbrace{\Delta y}_{m_2} & \underbrace{\Delta z}_{n-m_2} \end{pmatrix} \\ P_2J_2\Delta x &= L_2(R_2\Delta y + D_2\Delta z) = -P_2F_2 \\ \Rightarrow \Delta y &= -R_2^{-1}(D_2\Delta z + L_2^{-1}P_2F_2) \quad (4.13) \end{aligned}$$

2. Minimiere (4.10) auf der Lösungsmenge von (4.11). Spalte J_1 analog auf.

$$\begin{aligned} J_1 &= \begin{pmatrix} \underbrace{J_{11}}_{m_1\times m_2} & \underbrace{J_{12}}_{m_1\times(n-m_2)} \end{pmatrix} \\ \text{Einsetzen: } \frac{1}{2}\|F_1 + J_1\Delta x\|_2^2 &= \frac{1}{2}\|F_1 + J_{11}\Delta y + J_{12}\Delta z\|_2^2 \\ &= \frac{1}{2}\|\underbrace{F_1 + J_{11}R_2^{-1}L_2^{-1}P_2F_2}_{=:b} + \underbrace{(J_{12} - J_{11}R_2^{-1}D_2)}_{=:\,B}\Delta z\|_2^2 \\ &= \frac{1}{2}\|b + B\Delta z\|_2^2 \end{aligned}$$

QR-Zerlegung von B:

$$\begin{aligned} B = Q_1R_1 = \begin{pmatrix} \overline{Q}_1 & \hat{Q}_1 \end{pmatrix} \begin{pmatrix} \overline{R}_1 \\ 0 \end{pmatrix} = \overline{Q}_1\overline{R}_1 \quad (4.15) \\ B^TB\Delta z = \overline{R}_1^T\overline{Q}_1^T\overline{Q}_1\overline{R}_1\Delta z = -B^Tb = -\overline{R}_1^T\overline{Q}_1^Tb \\ \Delta z = -\overline{R}_1^{-1}\overline{Q}_1^Tb \quad (4.16) \\ \text{Einsetzen: } \Delta y = -R_2^{-1}(D_2\Delta z + L_2^{-1}P_1F_2 \end{aligned}$$

Formale Darstellung:

$$\begin{aligned}
\begin{pmatrix} P_2J_2 \\ \cdots \\ J_1 \end{pmatrix} &= \begin{pmatrix} P_1J_{21} & \vdots & P_2J_{22} \\ \cdots & & \cdots \\ J_11 & \vdots & J_{12} \end{pmatrix} \\
&= \begin{pmatrix} L_2 & \vdots & 0 \\ \cdots & & \cdots \\ L_1 & \vdots & I \end{pmatrix} \begin{pmatrix} R_2 & \vdots & D_2 \\ \cdots & & \cdots \\ 0 & \vdots & B \end{pmatrix} \\
&= \underbrace{\begin{pmatrix} L_2 & \vdots & 0 \\ \cdots & & \cdots \\ L_1 & \vdots & Q_1 \end{pmatrix}}_{=:T} \begin{pmatrix} R_2 & \vdots & D_2 \\ \cdots & & \cdots \\ 0 & \vdots & R_1 \end{pmatrix} \qquad (4.17) \\
L_1 &:= J_{11}R_2^{-1} \\
\Rightarrow J_{12} &= L_1D_2 + B = L_1D_2 + Q_1R_1
\end{aligned}$$

Rechte Seite entsprechend:

$$-\begin{pmatrix} P_2F_2 \\ F_1 \end{pmatrix} = T \begin{pmatrix} -L_2^{-1}P_2F_2 \\ -Q_1^Tb \end{pmatrix} \qquad (4.18)$$

nachrechnen:

$$-L_2L_2^{-1}P_2F_2 = -P_2F_2$$

$$-L_1L_2^{-1}P_2F_2 - \underbrace{Q_1Q_1^T}_{=I}(\underbrace{F_1 - \overbrace{J_{11}R_2^{-1}}^{=L_1}L_2^{-1}P_2F_2}_{b}) = -F_1$$

Lösen:

$$\begin{aligned}
R_2\Delta y + D_2\Delta z &= -L_2^{-1}P_2F_2 \\
\overline{R}_1\Delta z &= -\overline{Q}_1^Tb \\
\begin{pmatrix} R_2 & D_2 \\ 0 & \hat{R}_1 \end{pmatrix} \begin{pmatrix} \Delta y \\ \Delta z \end{pmatrix} &= -T^{-1} \begin{pmatrix} P_2F_2 \\ F_1 \end{pmatrix} \\
&= -T^{-1} \begin{pmatrix} 0 & P_2 \\ I & 0 \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \\
\Delta x &= -\underbrace{\left[\begin{pmatrix} R_2 & D_2 \\ 0 & \overline{R}_1 \end{pmatrix}^{-1} \vdots 0 \right] T1^{-1} \begin{pmatrix} 0 & P_2 \\ I & 0 \end{pmatrix}}_{=J^+} \underbrace{\begin{pmatrix} F_1 \\ F_2 \end{pmatrix}}_{F} \qquad (4.19) \\
&= -J^+F
\end{aligned}$$

Lemma 4.11 (Berechnung der adjungierten

$$\begin{aligned}\lambda &= -P_2 L_2^{-T} R_2^{-T} J_{11}^T (J_1 \Delta x + F_1) \\ &= -P_2 L_2^{-T} L_1^T (J_1 \Delta x + F_1) \quad (4.20)\end{aligned}$$

Beweis: λ erfüllt eindeutig die Gleichungen

$$\begin{aligned}J_1^T J_1 \Delta x + J_2^T \lambda &= -J_1^T F_1 \quad (\lambda \in \mathbb{R}^{m_2} \\ \text{bzw. } J_1^T (J_1 \Delta x + F_1) &= -J_2^T \lambda\end{aligned}$$

Es reicht, m_2 dieser Gleichungen zu betrachten:

$$\begin{aligned}J_{11}^T (J_1 \Delta x + F_1) = -J_{21}^T \lambda &= -R_2^T L_2^T P_2^{-1} \lambda \\ \Rightarrow \lambda &= -P_2 L_2^{-T} \underbrace{R_2^{-T} J_{11}^T}_{=L_1} (J_1 \Delta x + F_1)\end{aligned}$$

3. Variante: Stoer 1979

QR-Zerlegung von J_2^T: $J_2 = L_2 Q_2$. Transformation von rechts:

$$Q_1 \begin{pmatrix} J_2 \\ \cdots \\ J_1 \end{pmatrix} Q_2^T = TODO!!!!!$$

Rechte Seite transformieren mit Q_1, $\Delta\tilde{y}$, $\Delta\tilde{z}$ ausrechnen:

$$\Delta x = Q_2^T \begin{pmatrix} \Delta\tilde{y} \\ \Delta\tilde{z} \end{pmatrix}$$

Anwendung auf die Mehrzielmethode

TODO!!!!! Große Matrix
mit $\Delta x = (\Delta)$, $G, G_i^p D_1^{\dot{c}}, D_1^p, D_2^i, D_2^p$ wie in (3.23) und

$$r_1 = (\sigma_j^{-1}(\eta_j - h_j(t_j, y(t_j; \tau_{ij}, s_{ij}, p), p))), \quad j = 1, \cdots, M, \quad t_j \in [\tau_{ij}, \tau_{ij+1}].r_2 \quad = r(y(t_o; \tau_0, s_0, p), \cdots, y(t_K, \tau$$

mit $t_j \in [\tau_{ij}, \tau_{ij+1}], \quad j = 1, \cdots, K$

$$d_i = -s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) \quad i = 0, \cdots, m-2$$

Lösen:

- Wähle $-I$ unten rechts als Blockpivot-Element. Multiplizieren unterste Blockzeile mit D_i^{m-1}, $i = 1, 2$ und addiere auf die i-te Zeile $i = 1, 2$.
- Es entsteht

$$D_i^0 D_i^1, \cdots, (\underbrace{D_i^{m-2} + D_1^{m-1} \cdot G_{m-2}}_{(*)})(\underbrace{D_1^{m-1} + D_i^{m-1} \cdot (-I)}_{=0})(D_i^p + D_i^{m-1} \cdot G_{m-2}^p)$$

- Eliminiere $(*)$ ebenso mit der zweituntersten Blockzeile usw.

Die führt auf die kompakte Rekursion:

Algorithmus 4.12 (Eliminationsalgorithmus)

$$u^{m-1} := \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$$
$$E^{m-1} := \begin{pmatrix} D_1^{m-1} \\ D_2^{m-1} \end{pmatrix}$$
$$P^{m-1} := \begin{pmatrix} D_1^p \\ D_2^p \end{pmatrix}$$

Für $i = m-1, \cdots, 1$:

- $u^{i-1} := u^i + E^i d_{i-1}$
- $p^{i-1} := p^i + E^i G_{i-1}^p$
- $E^{i-1} := D^{i-1} + E^i G_{i-1}$

Es entstehen folgende Matrix und Vektor:

$$\begin{pmatrix} E^0 & 0 & \cdots & \cdots & 0 & P^0 \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix} \begin{pmatrix} u^0 \\ d_0 \\ d_1 \\ \vdots \\ d_{m-2} \end{pmatrix} \tag{4.22}$$

$$E^0 = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$$
$$P^0 = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$$
$$u^0 = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

Algorithmus 4.13: Lösen

- Löse beschränktes Ausgleichsproblen nur in den Variablen Δs_0 und Δp:

$$\min_{\Delta s_0, \Delta p} \frac{1}{2}\|E_1\Delta s_0 + P_1\Delta p + u_1\|_2^2 \text{ s. t. } E_2\Delta s_0 + P_2\Delta p = 0$$

  z. B. mit Variante 2.

- Berechne $\Delta s_1, \cdots, \Delta s_{m-1}$ durch:
  Für $i = 1, \cdots, m-1$:

$$\Delta s_i = G_{i-1}\Delta s_{i-1} + G_{i-1}^p \Delta p + d_{i-1} \quad (4.23)$$

Korollar:

Das Lösen der linearen beschränkten Probleme erfordert für Multiple Shooting im wesentlichen den selben Aufwand wie für Single Shooting.

Bemerkung 4.14:

- Es reicht, u, P und E abzuspeichern, da im Normalfall u^i, P^i und E^i nicht im Speicher gehalten werden müssen.

- Diese blockweise Gauß-Elimination kann als Dreieckszerlegung aufgefasst werden:

$$\begin{pmatrix} I & 0 & \cdots & 0 & E^{m-1} \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{pmatrix} \begin{pmatrix} D^0 & D^1 & \cdots & \cdots & D^{m-1} & D^p \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix}$$

$$= \begin{pmatrix} D^0 & D^1 & \cdots & E^{m-2} & 0 & P^{m-2} \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix}$$

  Insgesamt:

$$\begin{pmatrix} I & E^1 & 0 & \cdots & 0 \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{pmatrix} \cdots \begin{pmatrix} I & 0 & \cdots & 0 & E^{m-1} \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{pmatrix} (MZM)$$

$$= \begin{pmatrix} I & E^1 & \cdots & E^{m-1} \\ & \ddots & & \\ & & \ddots & \\ & & & I \end{pmatrix} (MZM)$$

$$\begin{pmatrix} E^0 & 0 & \cdots & \cdots & 0 & P^0 \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix} \begin{pmatrix} u^0 \\ d_0 \\ d_1 \\ \vdots \\ d_{m-2} \end{pmatrix} \tag{4.24}$$

Variante: Man kann auch Blockspalten-Elimination durch Transformationen von rechts durchführen.

$$\begin{pmatrix} E^0 & 0 & \cdots & \cdots & 0 & P^0 \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix}$$

$$= \begin{pmatrix} (D^0 + D^1 G_0) & D^1 & \cdots & \cdots & D^{m-1} & (D^p + D^1 D_0^p) \\ 0 & -I & 0 & \cdots & \cdots & 0 \\ (G_1 G_0) & G_1 & -I & 0 & \cdots & (G_1^p + G_1 G_0^p) \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix} \begin{pmatrix} I & & & & & \\ -G_0 & I & & & & -G_0^p \\ & & I & & & \\ & & & \ddots & & \\ & & & & I & \\ & & & & & I \end{pmatrix}$$

Transformierte Variablen:

$$\begin{pmatrix} \Delta\tilde{s} \\ \Delta\tilde{p} \end{pmatrix} = \begin{pmatrix} I & & & & & \\ -G_0 & I & & & & -G_0^p \\ & & I & & & \\ & & & \ddots & & \\ & & & & I & \\ & & & & & I \end{pmatrix} \begin{pmatrix} \Delta s \\ \Delta p \end{pmatrix}$$

$\Delta\tilde{s}_1 := d_0$ kann gesetzt werden

Streiche zweite Blockzeile und zweite Blockspalte und Variable $\Delta\tilde{s}_1$. Nach dem Streichen hat das Restsystem die gleiche Struktur wie das Ausgangssystem. Fahre mit nächster Blockspalte fort usw. Das führt auf ein kondensiertes System:

$$\begin{pmatrix} E^{m-1} & P^{m-1} \end{pmatrix}$$

Aus diesem können $\Delta\tilde{s}$, $\Delta\tilde{p}$ berechnet werden durch Lösen des linearen Ausgleichsproblems $\min \frac{1}{2}\|E_q^{m-1}\Delta\tilde{s}_0 + P^{m-1}\Delta\tilde{p}\|_2^2$,

Rücktransformation der Variablen liefert $\Delta s_0, \cdots, \Delta s_{m-1}, \Delta p$.

Diese Elimination kann parallelisiert werden:

- Führe den Schritt für jede zweite Spalte aus.
- Streiche Zeilen, Spalten, Variablen; Matrix hat wieder Ausgangsstruktur.
- Algorithmus kann $\frac{n}{2}$ Prozessoren beschäftigen
- Die Berechnung der G_i-Matrizen durch Lösen des Systems ODE/VDE kann auch im Rahmen dieser Parallelisierung erfolgen.
- Dadurch wird die „Wall-Time"-Berechnungszeit für Multiple Shooting sogar kürzer als für Single Shooting.
- Literatur: Gallitendörfer: Parallele Algorithmen für Optimierungsprobleme, Dissertation 1997
- Es gibt eine dritte Variante: Sukzessive QR-Elimination von rechts, ebenfalls parallelisierbar.

Bemerkung 4.15

Die strukturausnutzenden Zerlegungen für die „Mehrzielmatrizen" nennen Kondensierung/Kondensing.

5. Lokale Konvergenz von Newton-Typ-Verfahren

Newton-Typ-Verfahren:

- Löse iterativ mit einem guten Startwert $x_0 \in \mathbb{R}^n$
- Löse in jeder Iteration ein linearisiertes Problem
- Wende eine Globalisierungsstrategie an

$$F\colon \mathbb{R}^n \to \mathbb{R}^m \quad M(x)\colon \text{ Lösungsoperator des linearen Problems}$$

Algorithmus 5.1 (Newton-Typ-Verfahren)

- Startwert x^0, $k := 0$
- Solange ein geeignetes Abbruchkriterium verletzt ist:
  - Berechne $\Delta x^k := -M(x^k)F(x^k)$ (5.1)
  - Berechne α^k auf einer Globalisierungsstrategie
  - Iteriere $x^{k+1} := x^k + \alpha^k \Delta x^k$ (5.2)

Kann angewendet werden:

- Zur Bestimmung von Nullstellen von $F(\cdot)$, $m = n \Rightarrow$ Newton-Verfahren oder Quasi-Newton-Verfahren
- Insbesondere zur Bestimmung von Nullstellen von $\Delta L(x, \lambda) = 0$ (notwendige Optimalitätsbedingung) $\Rightarrow$ SQP-Verfahren.
- Zur Bestimmung von Lösungen unbeschränkter nichtlinearer Ausgleichsprobleme $\min \frac{1}{2}\|F(x)\|_2^2$, $m \geq n$
- Zur Bestimmung von Lösungen beschränkter nichtlinearer Ausgleichsprobleme $\min \frac{1}{2}\|F_1(x)\|$ s. t. $F_2(x) = 0$, $m = m_1 + m_2 \geq n$, $m_2 \leq n \Rightarrow$ Verallgemeinertes Gauß-Newton-Verfahren.

$$\text{Sei } J := \frac{\partial F}{\partial x} \quad (5.3) \text{ Jacobimatrix}$$

Bemerkung 5.2

- Bei Newton-Verfahren ist $M(x) = J(x)^{-1}$ die Inverse von J.
- bei Quasi-Newton-Verfahren ist $M(x) \cong J(x)^{-1}$
- Bei SQP-Verfahren ist $M(x,\lambda) = \nabla^2_{x,\lambda} L(x,\lambda)^{-1}$ oder $M(x,\lambda) \cong \nabla^2_{x,\lambda} L(x,\lambda)^{-1}$
- Bei Gauß-Newton-Verfahren ist $M(x) = (J(x)^T J(x))^{-1} J(x)^T$ die Moore-Penrose-Pseudoinverse.
- Bei Verallgemeinerten Gauß-Newton-Verfahren ist

$$M(x) = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1(x)^T J_1(x) & J_2(x)^T \\ J_2(x) & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1(x)^T 0 & \\ 0 & I \end{pmatrix}$$

  die verallgemeinerte Inverse von J.

Satz 5.3: Lokaler Kontraktionssatz (Bock 1987)

$$\text{Sei } F\colon D \subset \mathbb{R}^n \to \mathbb{R}^m, \quad F \in \mathbb{C}^1(D, \mathbb{R}^m), \quad J := \frac{\partial F}{\partial x}$$

Für alle $x, y \in D$ mit $y - x = -M(x)F(x)$ und $\theta \in [0,1]$ gelte:

- Es existiert ein $\omega < \infty$ so dass $\|M(y)(J(x+\theta(y-x)) - J(x))(y-x)\| \leq \omega\theta\|x-y\|^2$ (5.4)
- Es existiert ein $\kappa(x) \leq \kappa < 1$, so dass $\|M(y)R(x)\| \leq \kappa(x)\|y-x\|$ (5.5) für das Residuum $R(x) := F(x) - J(x)M(x)F(x)$
- Sei $x_0 \in D$ gegeben mit $\Delta x^j := -M(x^j)F(x^j)$, $\delta_j := \kappa + \frac{\omega}{2}\|\Delta x^j\|$, $\delta_0 = \kappa + \frac{\omega}{2}\|\Delta x^0\| < 1$ (5.6)
- $$D^0 := \left\{ z : \|z - z_0\| \leq \frac{\|\Delta x^0\|}{1 - \delta 0} \right\} \subset D$$

Dann gilt:

- a) Die Iterierten $x^{j+1} = y^j + \Delta x^j$ sind wohldefiniert und bleiben in D^0.
- b) Es existiert ein $x^* \in D^0$ so dass $x^j \to x^*$ $(j \to \infty)$
- c)
$$\|x^{j+k} - x^*\| \leq \frac{\Delta x^0}{1 - s_j} s_j^k \text{ „hoch k“ (a-priori-Abschätzung)}$$

- d)

$$\|\Delta x^{j+1}\| \leq s_j \|\Delta x^j\| = \kappa \|\Delta x^j\| + \frac{\omega}{2}\|\Delta x^j\|^2 \quad (5.7)$$

Beweis:

$$\begin{aligned}
\|\Delta x^{j+1}\| &= \|M(x^{j+1})F(x^{j+1})\| =: \|M^{j+1}F^{j+1}\| \\
&= \|M^{j+1}(F^{j+1} - F^j - J^j\Delta x^j) + M^{j+1}R^j\| \\
&\leq \|M^{j+1}\left(\int_0^1 J(x^j + t\Delta x^j)\Delta x^j \,\mathrm{d}t - \int_0^1 J^j \Delta x^0 \,\mathrm{d}t\right)\| + \|M^{j+1}R^j\| \\
&\leq \int_0^1 \|M^{j+1}(J(x^j + t\Delta x^j) - J^j)\Delta x^j\| \,\mathrm{d}t + \|M^{j+1}R^j\| + \kappa\|\Delta x^j\| \\
&\leq \int_0^1 \omega t \|\Delta x^j\|^2 \,\mathrm{d}t + \kappa\|\Delta x^j\| \\
&= \frac{\omega}{2}\|\Delta x^j\|^2 + \kappa\|\Delta x^j\| \\
&= \delta_j \|\Delta x^j\| \Rightarrow d)
\end{aligned}$$

Zeige: $(\delta_j)_{j\in\mathbb{N}}$, $(\|\Delta x^j\|)_{j\in\mathbb{N}}$ sind monoton fallend:
Induktion:

$$\begin{aligned}
\delta_j - \delta_{j+1} &= \frac{\omega}{2}(\|\Delta x^j\| - \|\Delta x^{j+1}\|) \\
&\geq \frac{\omega}{2}(\|\Delta x^j\| - \delta_j\|\Delta x^j\|) \\
&\geq \frac{\omega}{2}\|\Delta x^j\|(1 - \delta_j) > 0 \\
\|\Delta x^{k+k}\| &\leq \delta_{j+k}\|\Delta x^{j+k-1}\| \\
&\leq \delta_{j+k-1}\cdots\delta_j\|\Delta x_j\| \\
&\leq \delta_j^k\|\Delta x^j\| \\
\|x^{j+2} - x^0\| &\leq \|\Delta x^j + 1\| + \cdots + \|\Delta x^0\| \\
&\leq (\delta_0^{j+1} + \cdots + \delta s_0^1)\|\Delta x^0\| \\
&\leq \frac{1}{1-\delta_0}\|\Delta x^0\| \Rightarrow a)
\end{aligned}$$

$(x^j)_{j\in\mathbb{N}}$ ist Cauchy-Folge:

$$
\begin{aligned}
\|x^{i+j+1} - x^i\| &\leq \sum_{k=0}^{j} \|\Delta x^{i+k} \\
&\leq \sum_{k=0}^{j} \delta_j^k \|\Delta x^i\| \\
&\leq \sum_{k=0}^{i} \delta_0^{i+k} \|\Delta x^0\| \\
&\leq \delta_0^i \left(\sum_{k=0}^{\infty} \delta_0^k \right) \|\Delta x^0\| \\
&= \delta_0^i \frac{\|\Delta x^0\|}{1-\delta_0} \to 0 (i \to \infty \\
&\Rightarrow x^j \to x^* \\
D \text{ kompakt } &\Rightarrow x^* \in D^0 \Rightarrow b)
\end{aligned}
$$

Beweis:

$$
\begin{aligned}
\|x^j + k + i - x^{j+k}\| &\leq \frac{\|\Delta x^{j+k}\|}{1-\delta_{j+k}} \\
&\leq \delta_j^k \frac{\|\Delta x^j\|}{1-\delta_j}
\end{aligned}
$$

$\forall i \geq 0$ also auch für $i \to \infty$:

$$
\|x^* - x^{j+k}\| \leq \delta_j^k \frac{\|\Delta x^j\|}{1-\delta_j} \Rightarrow c)
$$

Korollar 5.4

Für $F(x) = 0$, $F\colon \mathbb{R}^n \to \mathbb{R}^m$ konvergiert das Newton-Verfahren mit $M(x) = J(x)^{-1}$ lokal quadratisch.

Beweis:

$$
\begin{aligned}
\|M^{j+1}R^j\| &= \|(J^{j+1})^{-1}(I - J^j(J^j)^{-1})F^j\| = 0 \Rightarrow \kappa = 0 \\
\|\Delta x^{j+1}\| &\leq \frac{\omega}{2}\|\Delta x^j\|^2 \\
\|x^{j+1} - x^*\| &= \|x^j - x^* + \Delta x^j\| = \|M^j(J^j(x^j - x^*) - (F^j - F^*))\| \quad \text{(*): Im Lösungspunkt} \\
&= \left\| M^j \int_0^1 (J^j - J(x^j + t(x^j - x^*)))(x^j - x^*)\,\mathrm{d}t \right\| \quad \text{(HDI)} \\
&\leq \underbrace{\|M^j J^*\|}_{\leq \Gamma} \int_0^1 \|M^*(J^j - J(x^j + t(x^j - x^*)))(x^j - x^*)\|\,\mathrm{d}t \\
&\leq \Gamma \int_0^1 \omega t \|x^j - x^*\|^2\,\mathrm{d}t \\
&= \Gamma\frac{\omega}{2}\|x^j - x^*\|^2
\end{aligned}
$$

Bemerkung 5.5: Quasi-Newton-Verfahren

Für näherungsweise Newton-Verfahren („Quasi-Newton-Verfahren“) ist $x^{j+1} = x^j - M(x^j)F(x^j)$ mit $M(x^j) \cong J(x^j)^{-1}$ $\kappa > 0$, zur Konvergenz muss $\kappa < 1$ sein:

$$
\begin{aligned}
\|M^{j+1}R^j\| &= \|M^{j+1}((M^j)^{-1} - J^j)M^jF^j\| \\
&= \|M^{j+1}((M^j)^{-1} - J^j)(x^{j+1} - x^j)\| \\
&\leq \kappa\| < x^{j+1}x^j\|
\end{aligned}
$$

Notwendig für Konvergenz ist also

$$\|M^j\| \leq \gamma \text{ und } \|(M^j)^{-1} - J^j\| \text{ klein}$$

Satz 5.6 (Dennis-Moré

Sei $F\colon \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar. Betrachte die Iteration $x^{j+1} = x^j + \Delta x^j$ und sei Δx^j gegeben durch $\Delta x^j = -M(x^j)F(x^j)$. Wir nehmen an, dass die Folge der x^j gegen einen Punkt x^j mit $F(x^j) = 0$ konvergiert mit $J(x^*)$ regulär. Dann konvergiert $(x^j)_{j\in\mathbb{N}}$ Q-superlinear gegen x^*, d. h. $\lim_{j\to\infty} \frac{\|x^{j+1}-x^*\|}{\|x^j-x^*\|} = 0$ genau dann wenn

$$\lim_{j\to\infty} \frac{\|(M(x^j)^{-1} - J(x^*))\Delta x^j\|}{\|\Delta x^j\|} = 0 \quad (5.8)$$

Beweis: siehe Vorlesung Algorithmische Optimierung 1.

Varianten von Quasi-Newton-Verfahren

$$\Delta x^j = -M(x^j)F(x^j), \quad M(x^j) \cong J(x^j)^{-1}$$

- Berechne $J(x^j)$ durch Differenzenquotienten
- Halte $M(x^j)$ fest
  - für alle Iterationen: $M(x^j) = J(x_0)^{-1}$
  - für einige Iterationen: $M(x^j) = J(x^{\bar{j}})^{-1}$ solange $\frac{\Delta x^{j+1}}{\|\Delta x^j\|} \leq \delta$ z. B. $\delta = \frac{1}{4}$, danach neues $\bar{j} = j$
- Nähere $J(x^j)$ bzw. $M(x^j)$ durch Update-Formeln aus $J(x^{j-1})$ bzw. $M(x^{j-1})$ an, siehe unten.

Bemerkung 5.7 (Bedeutung von ω)

$$\|M(y)(J(x + t(y - x)) - J(x))(x - y)\| \leq t\omega\|y - x\|^2$$

- Wenn $M(y)$ in einer Umgebung von x^* beschränkt ist: $\|M(y)\| \leq \gamma$, $\gamma < \infty$ und
- J eine Lippschitz-Bedingung erfüllt: $\|(J(x+t(x-y))-J(x))(y-x)\| \leq \beta t\|y-x\|^2$, $\beta < \infty$, dann ist $\omega = \gamma\beta < \infty$ in einer Umgebung von x^*

ω kann sehr groß werden wenn

- $\|M\|$ sehr groß ist, d. h. J ist fast singulär.
- die erste Ableitung von J bzw. die zweite Ableitung von F groß ist, d. h. das Problem ist sehr nichtlinear.

Anwendung auf (verallgemeinerte) Gauß-Newton-Verfahren:

$$M(x) = J^+(x) = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix}$$

Es gelte (CQ) und (PD). $\|M\|$ ist groß, wenn

$$\begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}$$

fast singulär ist.
Residuum:

$$\begin{aligned} R(x) &= F(x) - J(x)M(x)F(x) \\ &= F(x) + J(x)\Delta x \end{aligned}$$

Im Lösungspunkt: $\Delta x^* = 0$: $R(x^*) = F(x^*)$,

$$\|R(x^*)\| = \|F_1(x^*)\|, \quad \text{da } F_2(x^*) = 0$$

Bemerkung 5.8: Bedeutung von κ

Wenn $M(x)$ stetig differenzierbar ist gilt in D^0:

$$\begin{aligned} \|M(y) - M(x)\| &\leq L\|y - x\| \text{ und} \\ \|R(x)\| = \|F(x) - J(x)M(x)F(x)\| &\leq \rho \end{aligned}$$

Außerdem ist $M(x)R(x) = (\underbrace{M(x) - M(x)J(x)M(x)}_{=0 \text{ (4. Moore-Penrose-Axiom)}})F(x) = 0$

$$\begin{aligned} \|M(y)R(x)\| = \|(M(y) - M(x))R(x)\| &\leq \rho L\|y - x\| \\ \kappa &= \rho L \end{aligned}$$

$\kappa < 1$ falls

- das Residuum R klein ist.
- M eine Lippschitz-Bedingung mit kleinem L erfüllt, d. h. falls die erste Ableitung von M klein ist.

Bemerkung 5.9

Seien $\|M\|$ und $\|M^*\|$ beschränkt, dann ist ω ein Maß für die Nichtlinearität des Problems und κ ein Maß für die Inkompatibilität zwischen Modell und Daten. Probleme mit $\kappa < 1$ heißen Kleine-Residuen-Probleme, Probleme mit $\kappa > 1$ heißen Große-Residuen-Probleme. Anwendung des Kontraktionssatzes:

Korollar 5.10

Wenn der Startwert x^0 nahe der Lösung x^* gewählt wird, d. h. wenn $\kappa + \frac{\omega}{2}\|\Delta x^0\| < 1$ dann konvergiert das (verallgemeinerte) Gauß-Newton-Verfahren und die Konvergenz ist linear:

$$\|\Delta x^{k+1}\| \leq \kappa\|\Delta x^k\| + \frac{\omega}{2}\|\Delta x^k\|^2$$

Notwendig dafür ist $\kappa < 1$. Wenn $\kappa > 1$ ist konvergiert das Verallgemeinerte Gauß-Newton-Verfahren nicht.

Warum sollte man nichtlineare Ausgleichsprobleme nicht mit den Newton-(SQP)-Verfahren lösen?

Betrachte den unbeschränkten Fall: $\min \frac{1}{2}\|F(x)\|_2^2 = \frac{1}{2}F(x)^T F(x) =: f(x)$. Optimalitätsbedingung:

$$\nabla f(x) = J^T(x)F(x) =: g(x) = 0$$

$$\text{Hessematrix: } H_f(x) = \nabla^2 f(x) = \underbrace{J^T(x)J(x)}_{=:B(x)} + \underbrace{\sum_{i=1}^{n} F_i(x)\frac{\partial J_i}{\partial x}(x)}_{=:E(x)} =: H_f(x)$$

Newton-Typ-Verfahren: $x^{k+1} = x^k + \alpha^k \Delta x^k$. Schreibweisen: $F := F(x^k)$, $J := J(x^k)$, $f := f(x^k)$, $g := g(x^k)$, $H := H(x^k)$, $B := B(x^k)$, $E := E(x^k)$

Newton-Verfahren für die nichtlineare Gleichung $\nabla f(x) = 0$

Δx^k löst $\nabla f + \nabla^2 f \Delta x = 0$

$$\Rightarrow \Delta x^k = -(\nabla^2 f)^{-1}\nabla f = -(B+E)^{-1}J^T F = -Mg \quad (M = (B+E)^{-1}$$

Daraus folgt lokal quadratische Konvergenz.

Gauß-Newton-Verfahren für $\min \frac{1}{2}\|F(x)\|_2^2$

Δx^k löst $\min \frac{1}{2}\|F + J\Delta x\|_2^2 = \frac{1}{2}F^T f + F^T J\Delta x + \frac{1}{2}\Delta x^T J J^T \Delta x$ bzw. $J^T J\Delta x + J^T f = 0$, siehe auch (4.6). $\Rightarrow \Delta x^k = -(J^T J)^{-1}J^T F = -Mg$ (5.10) mit $M = (J^T J)^{-1} = B^{-1}$ (5.11)

Newton-Verfahren für $\min \frac{1}{2}\|F(x)\|_2^2$

$$\begin{aligned} \nabla f + \nabla^2 f \Delta x &= 0 \\ \Delta x^k &= -(\nabla^2 f)^{-1}\nabla f \\ &= -(B+E)^{-1}J^T F \\ M &= (B+E)^{-1} \end{aligned} \tag{5.13}$$

Bemerkung 5.11

Die Hessematrix

$$\begin{aligned} H(x) &= \nabla^2 f(x) \\ &= J^T(x)J(x) + \sum_{i=1}^{n} F_j \frac{\partial J_i}{\partial x}(x) \end{aligned} \tag{5.9}$$

ist die Summe aus

- dem Gauß-Newton-Anteil $B(x) = J^T(x)J(x)$ (deterministischer Anteil) und

- dem Anteil $E(x) = \sum_{i=1}^{n} F_j \frac{\partial J_i}{\partial x}(x)$, der von den zweiten Ableitungen $\frac{\partial J_i}{\partial x}(x)$ und vom Residuum $F_i(x)$, $i = 1, \cdots, M$ d. h. von den zufallsbehafteten Messfehlern abhängt. $E(x)$ ist der Zufallsanteil der Hessematrix.

Satz 5.12: Kleine-Residuen-Probleme

Äquivalent zu $\kappa < 1$ ist $\rho(B(x^*)1{-}1E(x^*)) < 1$ (5.14), ρ : Spektralradius.

Beweis:

$$
\begin{aligned}
M &= J^\dagger \\
&= (J^T j)^{-1} J \\
R &= F - JJ^\dagger F \\
y - x &= -J^\dagger(x)F(x) \\
J^\dagger(y)R(x) &= \underbrace{J^\dagger(x)R(x)}_{=0} + \left(\frac{\partial J^\dagger}{\partial y}(y)\Big|_{y=x} (y-x) \right) R(x) + \mathcal{O}(\|y-x\|^2) \\
&= ((J(x)^T J(x))^{-1} \left(\frac{\partial J(y)^T}{\partial y} y \Big|_{y=x} (y-x) \right) R(x) + \left(\frac{\partial J(y)^T J(y)}{\partial y} y \Big|_{y=x} (y-x) \right) \underbrace{J^T(x)R(x)}_{=0} \\
&\quad + \mathcal{O}(\|y-x\|^2) \\
&= B(x)^{-1} \left(\frac{\partial J(y)^T}{\partial y}\Big|_{y=x} (y-x) \right) F(x) - B(x)^{-1} \left(\frac{\partial J(y)^T}{\partial y}\Big|_{y=x} (y-x) \right) J(x) \underbrace{J^\dagger(x)F(x)}_{=-(y-x)} \\
&\quad + \mathcal{O}(\|y-x\|^2) \\
&= B(x)^{-1}E(x)(y-x) + \mathcal{O}(\|y-x\|^2)
\end{aligned}
$$

Sei $\rho(B(x^*)1{-}1E(x^*)) =: \kappa_1 < 1$. Wähle eine Umgebung von x^* und eine Norm, so dass $\|B(x)^{-1}E(x)\| \leq \kappa_2 < 1$ für alle x aus dieser Umgebung. Verkleinere die Umgebung evtl., so dass

$$\mathcal{O}(\|y-x\|^2) \leq \frac{1-\kappa_2}{2}\|y-x\|$$

für alle x, y aus dieser Umgebung. Dann ist

$$\begin{aligned}\|J^\dagger(y)R(x)\| &= \|B(x)^{-1}E(x)(y-x) + \mathcal{O}(\|y-x\|^2)\| \\ &\leq \kappa_2\|y-x\| + \frac{1-\kappa_2}{2}\|y-x\| \\ &= \frac{1+\kappa_2}{2}\|y-x\| \\ &=: \kappa\|y-x\| \\ \kappa :&= \frac{1+\kappa}{2} < 1\end{aligned}$$

Sei umgekehrt die κ-Bedingung (5.5) erfüllt, d. h. $\exists\kappa < 1$ so dass

$$\|J^\dagger(y)R(x)\| = \|B(x)^{-1}E(x)(y-x) + \mathcal{O}(\|y-x\|^2) \leq \kappa\|y-x\|$$

Dann ist

$$\|B(x)^{-1}E(x)(y-x)\| - \mathcal{O}(\|y-x\|^2) \leq \kappa\|y-x\|$$

Mache die Umgebung so klein, dass

$$\mathcal{O}(\|y-x\|^2) \leq \frac{1+\kappa}{2}\|y-x\|$$

Dann ist

$$\begin{aligned}\|B(x)^{-1}E(x)(y-x)\| &\leq \frac{1+\kappa}{2}\|y-x\| \\ &=: \kappa_1\|y-x\| \\ \kappa_1 :&= \frac{1+\kappa}{2} < 1\end{aligned}$$

Daher ist

$$\|B(x)^{-1}E(x)\| \leq \kappa_1 < 1$$

nach Definition der Norm für Matrizen. Dann ist auch

$$\rho(B(x)^{-1}E(x)) < 1$$

Anders ausgedrückt: Bei Kleine-Residuen-Problemen wird der Zufallsanteil E der Hesse-Matrix H relativ beschränkt durch den deterministischen Anteil $B(x)$.

Satz 5.13

Sei x^* ein statischer Punkt des Gauß-Newton-Verfahrens, d. h. $J(x^*)^T F(x^*) = 0$ und $\rho(B(x^*)^{-1}E(x^*)) < 1$. Dann ist $H(x^*)$ positiv definit.

Beweis: $J(x^*)$ hat vollen Rang $\Rightarrow$ $B(x^*) = J(x^*)^T J(x^*)$ ist positiv definit. Dann existiert $B(x^*)^{\frac{1}{2}}$. Es gilt:

$$\begin{aligned} H(x^*) &= B(x^*) + E(x^*) \\ &= B^{\frac{1}{2}}(x^*) \left(I + B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{-\frac{1}{2}} \right) B(x^*)^{\frac{1}{2}} \text{ positiv definit} \\ &\Leftrightarrow I + B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{\frac{1}{2}} \text{ positiv definit} \end{aligned}$$

Die Umformung

$$B(x^*)^{-1}E(x^*) = B(x^*)^{-\frac{1}{2}} \left(B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{-\frac{1}{2}} \right) B(x^*)^{\frac{1}{2}}$$

ist eine Ähnlichkeitstransformation, daher hat $B(x^*)^{-1}E(x^*)$ die selben Eigenwerte wie

$$B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{-\frac{1}{2}}$$

Also ist

$$\rho(B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{-\frac{1}{2}}) < 1$$

und die Eigenwerte von

$$I + B(x^*)^{-\frac{1}{2}} E(x^*) B(x^*)^{-\frac{1}{2}}$$

liegen in $(0, 2)$, also ist $H(x^*)$ positiv definit.

Korollar 5.14

In Satz 5.14 ist x^* nicht nur stationärer Punkt sondern striktes lokales Minimum und stabil gegen Störungen.

Beweis: hinreichende Bedinung zweiter ordnung für lokale Minima für unbeschränkte Optimierungsprobleme, siehe unten.

Fazit

Für Kleine-Residuen-Probleme

- konvergiert das Newton-Verfahren mit $M = (B + E)^{-1}$ gegen ein stabiles lokales Minimum

- konvergiert das Gauß-Newton-Verfahren

Für Große-Residuen-Problem ($\kappa > 1$ bzw. $\rho(B(x^*)^{-1}E(x^*)) > 1$) konvergiert das Gauß-Newton-Verfahren nicht. Das Newton-Verfahren konvergiert lokal gegen ein Minimum x^*. Was passiert in diesem Punkt?

Statistische Störung des Problems

Sei $F(x) = (\eta_i - h_i(x))_{i=1,\cdots,M}$, (o. B. d. A. $\sigma_i \equiv 1$), η : Messdaten
wobei mit den wahren werten $\overline{x} : \eta_j - h_i(\overline{x}) \sim \mathcal{N}(0,1)$ und unabhängig

$$F(x^*) = \eta - h(x^*), \quad \eta = h(x^*) + F(x^*)$$

Spiegel die Messfehler an den geschätzten Modellantworten.

$$\hat{\eta} := h(x^*) - F(x^*)$$
$$F(x^*) = h(x^*) - \hat{\eta}$$

$\hat{\eta}$: gestörte Messdaten.
Dann gilt:

$$\|\hat{\eta} - h(x^*)\| = \|F(x^*)\| = \|\eta + h(x^*)\|$$

Betrachte eine Homotopie:

$$\tilde{F}(x,\tau) := F(x) + (\tau - 1)F(x^*), \quad \tau \in [-1,1]$$

zwischen Originalproblem und gespiegeltem Problem:

$$\tau = +1: \quad \tilde{F}(x,+1) = F(x) = \eta - h(x)$$
$$\tau = -1: \quad \tilde{F}(x,-1) = F(x) - 2F(x^*) = \eta - h(x) - \eta + h(x^*) - h(x^*) + \hat{\eta}$$
$$= \hat{\eta} - h(x)$$

Betrachte die Probleme:

$$\min_x \frac{1}{2}\|\tilde{F}(x,\tau)\|_2^2, \quad \tau \in [-1,1] \quad (5.13)$$

Satz 5.15

Sei x^* ein Minimum von

$$\min \frac{1}{2}\|F(x)\|_2^2$$

mit $\tilde{\kappa} := \rho(B(x^*)^{-1}E(x^*)) > 1$. Dann gilt:

- a) Für alle τ ist x^* ein stationärer Punkt von $\min \frac{1}{2}\|\tilde{F}(x,\tau)\|_2^2$
- b) Die Hessematrix in x^* ist $\tilde{H}(x^*,\tau) = B(x^*) + \tau E(x^*)$
- c) $\tilde{H}(x^*,\tau)$ ist für alle $\tau < -\frac{1}{\kappa} \geq -1$ nicht positiv definit

Beweis: Es gilt

$$\begin{aligned}
J(x^*)^T F(x^*) &= 0 \\
\tilde{J}(x,\tau) &= J(x) \\
\tilde{J}(x^*),\tau)^T \tilde{F}(x^*,\tau) &= J(x^*)^T(F(x^*) + (\tau-1)F(x^*)) = 0 \Rightarrow \text{ a)} \\
\tilde{H}(x,\tau) &= \tilde{J}(x,\tau)^T \tilde{J}(x,\tau) + \tilde{F}(x,\tau)\frac{\partial \tilde{J}}{\partial x}(x,\tau) \\
&= B(x) + (F(x) + (\tau-1)F(x^*))\frac{\partial J}{\partial x}(x) \\
\tilde{H}(x^*,\tau) &= B(x^*) + \tau F(x^*)\frac{\partial J}{\partial x}(x^*) \\
&= B(x^*) + \tau E(x^*) \Rightarrow \text{ b)} \\
&\tilde{H}(x^*,\tau) \text{ positiv definit} \\
&\Leftrightarrow \underbrace{I + B(x^*)^{-\frac{1}{2}}\tau E(x^*)B(x^*)^{-\frac{1}{2}}}_{=:I+\tau\hat{E}(x^*)} \text{ positiv definit} \\
\rho\left(B(x^*)^{-1}E(x^*)\right) &= \rho\left(B(x^*)^{-\frac{1}{2}}E(x^*)B(x^*)^{-\frac{1}{2}}\right) \\
&= \rho\left(\hat{E}(x^*)\right) \\
&= \tilde{\kappa} > 1
\end{aligned}$$

$I + \tau\hat{E}(x^*)$ hat einen Eigenwert $1 + \tau\tilde{\kappa}$, dieser ist < 0 falls $\tau < -\frac{1}{\tilde{\kappa}} \geq -1$, dann ist $\tilde{H}(x^*,\tau)$ nicht positiv definit $\Rightarrow$ c)

Fazit

Für Große-Residuen-Probleme ist das Minimum von $\min \frac{1}{2}\|F(x)\|_2^2$ nicht stabil gegen Störungen. x^* kann ein Sattelpunkt oder ein Maximum werden. Das Minimum springt von x^* weg. Große-Residuen-Minima sind nicht statistisch stabil. Sie sind Minima, aber keine Parameterschätzer. Das Gauß-Newton-Verfahren konvergiert nicht gegen sie.

6 Optimalitätsbedingungen für nichtlineare Optimierungsprobleme

6.1 Allgemeine Problembeschreibung

$$\min_{x \in \mathbb{R}^n} f(x) \quad f \in C^2(\mathbb{R}^n, \mathbb{R}),\ \text{„Zielfunktion“} \tag{6.1}$$

$$g(x) = 0 \quad g \in C^2(\mathbb{R}^n, \mathbb{R}^{m_1}),\ \text{„Gleichungsbedingungen“}$$

$$h(x) \leq 0 \quad h \in C^2(\mathbb{R}^n, \mathbb{R}^{m_2}),\ \text{„Ungleichungsbedingung“}$$

Wir betrachten endlichdimensionale, kontinuierliche, glatte Probleme.
Wir unterscheiden

- unbeschränkte
- gleichungsbeschränkte
- Gleichungs- und Ungleichungs-beschränkte Probleme

Definition 6.1

Ein Punkt $c \in \mathbb{R}^n$ heißt zulässiger Punkt des Problems (6.1), wenn er alle Nebenbedingungen erfüllt: $g(x) = 0$, $h(x \leq 0)$.

Definition 6.2

Ein Punkt $x^* \in \mathbb{R}^n$ heißt lokales Minimum von (6.1) wenn er zulässig ist und wenn ein Zielfunktionswert kleiner oder gleich den Zielfunktionswerten aller zulässigen Punkte in einer Umgebung von x^* ist.

x^* lokales Minimum $\Leftrightarrow g(x^*) = 0 \wedge h(x^*) \leq 0 \wedge \exists$ Umgebung U von $x^* : \forall x \in U$ mit $g(x) = 0, h(x) \leq 0$ ist $f(x^*) \leq$

Definition 6.3

Ein Punkt heißt globales Minimum, wenn er zulässig ist und sein Zielfunktionswert kleiner gleich den Zielfunktionswerten aller anderen zulässigen Punkte.

x^* globales Minimum $\Leftrightarrow g(x^*) = 0 \wedge h(x^*) \leq 0 : \forall x$ mit $g(x) = 0, h(x) \leq 0$ ist $f(x^*) \leq f(x)$ (6.3)

Die Bestimmung von globalen Optima heißt globale Optimierung. Wir beschränken uns in dieser Vorlesung auf die Berechnung lokaler Optima.

Definition 6.4

Ein Minimum x^* heißt strikt, wenn

$$f(x^*) < f(x) \forall x \in U, x \neq x^*, x \text{ zulässig (lokal)}$$
$$f(x^*) < f(x) \forall x \text{ zulässig (global)}$$

6.2 Optimalitätsbedingung im eindimensionalen Fall

Lemma 6.5: Sei $f : (a, b) \subset \mathbb{R} \to \mathbb{R}$, $f \in C^2$. Dann gilt:

- Wenn $x^* \in (a, b)$ ein lokales Minimum von f ist, ist $f'(x) = 0$ (notwendige Bedingung erster Ordnung) und $f''(x^*) \geq 0$ (notwendige Bedingung 2ter Ordnung)
- Wenn $f'(x^*) = 0$ und $f''(x^*) > 0$, dann ist x^* ein striktes lokales Minimum von f (hinreichende Bedingung 2ter Ordnung).

Beweis: Analysis 1

6.3 Unbeschränkter Fall

Satz 6.6 (Notwendige Bedingungen erster und zweiter Ordnung)
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$. x^* sei ein lokales Minimum von f.

- a) $\nabla f(x^*) = 0$ (6.4)
- b) $\nabla^2 f(x^*) = 0$ (6.5)

Beweis: Zurückführung auf den eindimensionalen Fall entlang beliebiger Kurven von Konkurrenten.
Sei $p \in \mathbb{R}^n$ beliebig. Die Funktion $\tilde{f}(t) := f(x^* + tp)$, $t \in \mathbb{R}$ hat in x^* bzw. $t = 0$ ein lokales Minimum. Also gilt nach Lemma 6.5 (eindimensionaler Fall):

- a) $0 = \tilde{f}'(0) = \frac{\partial}{\partial t} f(x^* + tp)|_{t=0} = f'(x^*)p = \nabla f(x^*)^T p \Rightarrow \nabla f(x^*) = 0$, da p beliebig.
- b) $0 \leq \tilde{f}''(0) = \frac{\partial^2}{\partial t^2} f(x^* + tp)|_{t=0} = p^T \nabla^2 f(x^*) p \Rightarrow \nabla^2 f(x^*)$ ist positiv semidefinit.

Satz 6.7 (Heinreichende Bedingung)

Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$. Sei $x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv definit (6.6). Dann gilt: x^* ist ein striktes lokales Minimum von f.

Beweis: Es existiert eine Umgebung U von x^* so dass die Hesse-Matrix $\nabla^2 f$ positiv definit für alle x in der Umgebung U, da die Eigenwerte stetig von den Einträgen abhängen und die Einträge stetig von x abhängen. Entwicklung in eine Taylorreihe um x^*:

$$f(x) = f(x^*) + \nabla f(x^*)^T(x - x^*) + \underbrace{\frac{1}{2}(x - x^*)^T \nabla^2 f(\tilde{x})(x - x^*)}_{>0} \text{ mit } \tilde{x} \in U$$

$i \Rightarrow f(x) > f(x^*)$

6.4 Gleichungsbeschränkter Fall

Notation: $f : \mathbb{R}^n \to \mathbb{R}$, Ableitung $\frac{\partial f}{\partial x} f(x) = f'(x) = f_x(x) \in \mathbb{R}^n$ Zeilenvektor. Gradient: $\nabla f(x) = f_x(x)^T \in \mathbb{R}^n$ Spaltenvektor. Hessematrix $\nabla^2 f(x) = f_{xx}(x) = \frac{\partial}{\partial x} \nabla f(x) = \nabla \frac{\partial}{\partial x} f(x) \in \mathbb{R}^{n \times n}$ (symmetrische Matrix nach Satz von Schwarz).

$g : \mathbb{R}^n \to \mathbb{R}^m$, $n \leq m$, $\frac{\partial g}{\partial x}(x) = g_x(x) \in \mathbb{R}^{m \times n}$, $\nabla g(x) = g_x(x)^T \in \mathbb{R}^{n \times m}$. Menge aller zulässigen Punkte: $S := \{x : g(x) = 0\}$

Definition 6.8

Ein Punkt x^* heißt regulär, wenn er die Constraint Qualification (CQ) erfüllt: $\mathrm{Rg}(g_x(x^*)) = m$ (6.7).

Definition 6.9: Tangentialebene

Die Menge $T(x^*) = \{p : g_x(x^*)p = 0\}$ (6.8) heißt Tangentialebene an S in $x^* \in S$.

Laufe entlang der zulässigen Menge $S = \{x : g(x) = 0\}$. Entweder wir schneiden die Höhenlinien von f, z. B. in $\hat{x}$. Dann erhöht oder erniedrigt ein kleiner Schritt den Zielfunktionswert, $\hat{x}$ ist also kein Optimum. Oder wir berühren eine Höhenlinie von f tangential, hier in x^*. Dann kann x^* ein Optimum sein.

Im lokalen Minimum gilt: Die Tangentialebene von S und die Höhenlinien von f sind parallel, also sind die Normalenvektoren parallel:

$$-\nabla f(x^*) = \lambda \nabla g(x^*), \quad \lambda \in \mathbb{R}$$

Verallgemeinerung für m Nebenbedingungen:

$-\nabla f(x^*)$ ist eine Linearbkombination der Gradienten der Nebenbedingungen:

$$\exists \lambda \in \mathbb{R}^m : -\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) = \nabla g(x^*)\lambda \quad (6.9)$$

Definition 6.10

Die Funktion $L : \mathbb{R}^n \times \mathbb{R}^m : (x, \lambda) \mapsto L(x, \lambda) := f(x) + \lambda^T g(x)$ (6.10) heißt Lagrangefunktion (Lagrangian) des beschränkten Optimierungsproblems $\min f(x)$ s. t. $g(x) = 0$.

Bemerkung 6.11

Gradient der Lagrangsfunktion:

$$\begin{aligned}\nabla_x L(x,\lambda) &= \nabla f(x) + \nabla g(x)\lambda \quad (6.11)\\ \nabla_\lambda L(x,\lambda) &= g(x)\end{aligned}$$

Hessematrix der Lagrangefunktion:

$$\nabla^2 L(x,\lambda) = \begin{pmatrix}\nabla^2_{xx} L(x,\lambda) & \nabla g(x)\\ \nabla g(x)^T & 0\end{pmatrix} \quad (6.12)$$

$$\text{mit } \nabla x x^2 L(x,\lambda) = \nabla^2_{xx} f(x) + \underbrace{\nabla^2_{xx} g(x)\lambda}_{\in \mathbb{R}^{n\times n}}$$

Satz 6.12 (Notwendige Bedingung erster Ordnung

Betrachte das Problem $\min f(x)$ s. t. $g(x) = 0$ mit $f \in C^2(\mathbb{R}^n, \mathbb{R})$, $g \in C^2(\mathbb{R}^m, \mathbb{R}^n)$, $m \leq n$. Sei x^* ein lokales Minimum, x^* regulär. Dann existiert $\lambda \in \mathbb{R}^m$, so dass $\nabla_x L(x^*, \lambda) = \nabla f(x^*) + \nabla g(x^*)\lambda = 0$ und $\nabla_\lambda L(x^*, \lambda) = g(x^*))9$ (6.13) bzw. $\nabla L(x.\lambda) = 0$.

Beweis:

x^* regulär $\Rightarrow g_x(x^*) = 0 \Rightarrow$ man kann x zerlegen in $x = (y, z)$ so dass $g_y(x^*) \in \mathbb{R}^{m\times m}$ invertierbar. Führe die Fragestellung zurück auf den eindimensionalen Fall entlang Kurven von zulässigen Konkurrenzpunkten.

$$\begin{aligned}x = \varphi(t) \text{ mit } \varphi(0) = x^*, \varphi(1) = \overline{x} \neq x^* \text{ zulässiger Punkt}\\ \text{und } g(\varphi(t)) \equiv 0\\ \varphi(t) = \begin{pmatrix}y(t)\\ z(t)\end{pmatrix} = \begin{pmatrix}y(t)\\ z^* + t(\overline{z} - z^*)\end{pmatrix}\\ 0 = \frac{\partial}{\partial t} g(\varphi(t))\Big|_{t=0} = g_y(x^*)y'(0) + g_z(x^*)z'(x^*)\\ \Rightarrow \varphi'(0) = \begin{pmatrix}y'(o)\\ z'(0)\end{pmatrix} = \begin{pmatrix}-g_y(x^*)^{-1}g_z(x^*)\\ I\end{pmatrix} h\end{aligned}$$

$f(\varphi(t))$ ist minimal in $t = 0$. Daher ist $0 = \frac{\partial}{\partial t} f(\varphi(t))|_{t=0} = f_x(x^*)\varphi'(0) = (-f_y(x^*)g(x^*)^{-1}g_z(x^*) + f_z(x^*))h$.

Definiere $\lambda^T := -f_y(x^*)g_y(x^*)^{-1}$. Dann ist $f_y(x^*)+\lambda^T g_y(x^*) = 0$. $f_z(x^*)+\lambda^T g_z(x^*) = 0$ also ist $f_x(x^*) + \lambda^T g_x(x^*) = 0$ d. h. $\nabla L(x, \lambda) = 0$.

$\nabla_\lambda L(x^*, \lambda) = g(x^*) = 0$, da x^* zulässig.

Bemerkung 6.13

Die Vektoren aus der Tangentialebene $T(x^*) = \ker g_x(x^*)$ haben die Gestalt:

$$p = \begin{pmatrix} p_y \\ p_z \end{pmatrix} = \begin{pmatrix} -g_y(x^*)^{-1} g_z(x^*) \\ I \end{pmatrix} p_z \text{ denn}$$
$$0 = g_x(x^*)p = g_y(x^*)p_y + g_z(x^*)p_z$$

Alle zulässigen Kurven laufen also in der Tangentialebene in x^* ein:

$$\varphi(o) = x^*$$
$$\varphi'(0) \in T(x^*)$$

Die Spalten von

$$\begin{pmatrix} -g_y(x^*)^{-1} g_z(x^*) \\ I \end{pmatrix}$$

bilden eine Basis von $T(x^*)$.

Satz 6.12

x^* optimal $\Rightarrow \nabla L(x^*, \lambda) = 0$ für ein $\lambda \in \mathbb{R}^m$.

$$0 = \nabla_x L(x^*, \lambda) = \nabla f(x^*) + \nabla g(x^*)\lambda$$
$$0 = \nabla_\lambda L(x^*, \lambda) = g(x^*)$$

Definition 6.14

Die notwendigen Bedingungen (6.13) heißen KKT-Bedingungen. Reguläre Punkte $x^* \in \mathbb{R}^n$, für die es $\lambda \in \mathbb{R}^n$ gibt, so dass $\nabla_{x,\lambda} L(x^*, \lambda) = 0$ heißen KKT-Punkte bzw. stationäre Punkte.

Bemerkung 6.15

Stationäre Punkte können Minima aber auch Maxima oder Sattelpunkte sein.

Satz 6.16 (Notwendige Bedingungen zweiter Ordnung

Seien die Voraussetzungen wie in Satz 6.12. Sei x^* ein lokales Minimum, x^* regulär. Dann ist die Hessematrix von L nach x und x positid semidefinit auf $T(x^*) = \ker g_x(x^*)$, d. h.

$$p^T \nabla^2_{xx} L(x^*, \lambda) p \geq 0 \quad \forall p \in T(x^*)$$

Beweis: Weiter im Beweis von 6.12:
Es gilt:

$$\begin{aligned} 0 \leq \frac{\partial^2}{\partial t^2} f(\varphi(t))\Big|_{t=0} &= \frac{\partial}{\partial t} f_x(\varphi(t))\varphi'(t)\Big|_{t=0} \\ &= \underbrace{\varphi'(t)^T f_x x(\varphi(t))\varphi'(t)}_{\text{Krümmung von } f}\Big|_{t=0} + \underbrace{f_x(\varphi(t))\varphi'(t)}_{\text{Krümmung von } \varphi}\Big|_{t=0} \end{aligned}$$

Außerdem gilt:

$$\begin{aligned} &\frac{\partial^2}{\partial t^2} \lambda^T g(\varphi(t)) = 0 \text{ da } g(\varphi(t)) \equiv 0 \\ &\frac{\partial^2}{\partial t^2} \lambda^T g(\varphi(t)) = \varphi'(t)^T (\lambda^T g_{xx}(\varphi(t)))\varphi'(t) + \lambda^T g_x(\varphi(t))\varphi''(t) \\ &\qquad \text{Also } 0 \leq \varphi'(0)^T (f_{xx}(x^*)^T + \lambda^T g_{xx}(x^*))\varphi'(0) + \underbrace{(f_{xx}(x^*) + \lambda^T g_x(x^*))}_{L_x(x^*,\lambda)=0} \varphi''(t) \\ &\qquad\qquad = \varphi'(0)^T \nabla^2_{xx} L(x^*, \lambda)\varphi'(0) \end{aligned}$$

$\varphi'(0)$ ist irgendein Vektor aus $T(x^*) \Rightarrow \nabla^2_{xx} L(x^*, \lambda)$ ist positiv semidefinit auf $T(x^*)$.

Satz 6.17 (Hinreichende Bedingung)

Seien die Voraussetzungen wie in Satz 6.12, aber $f, g \in C^3$. Wenn gilt: $\exists \lambda \in \mathbb{R}^n$, so dass $\nabla \lambda L(x^*, \lambda) = 0$ und $p^T \nabla^2_{xx} L(x^*, \lambda) p > 0 \forall p \in T(x^*), p \neq 0$. Für einen regulären Punkt x^*, dann ist x^* striktes lokales Minimum.

Beweis: Erfülle x^* die hinreichende bedingung. Sei $\hat{x} \neq x^*$ zulässig und hinreichend nahe bei x^*. Betrachte die Kurve $\varphi(t)$ mit $\varphi(0) = x^*$, $\varphi(1) = \hat{x}$,

$$\varphi(t) := \begin{pmatrix} \gamma(z(t)) \\ x^* + t(\hat{z} - z^*) \end{pmatrix}$$

so dass $g(\varphi(t)) = 0$ in einer kleinen Umgebung von x^*

$$
\begin{aligned}
h &:= \hat{z} - z^*, \quad \|h\| \le \varepsilon_z \text{ Dann gilt} \\
\varphi'(t) &= \begin{pmatrix} \gamma_z(z(t)) \\ I \end{pmatrix} h \\
\|\varphi'(t)\| &\le c_1 \|h\| \\
\varphi'(0) &= \begin{pmatrix} \gamma_z(z^*)h \\ h \end{pmatrix} \in T(x^*) \le c_1 \varepsilon_z \\
\varphi''(t) &= \begin{pmatrix} (\gamma_{zz}h)h \\ 0 \end{pmatrix} \\
\|\varphi''(t)\| &\le c_2 \|h\|^2 \le c_2 \varepsilon_z^2 \\
\varphi'''(t) &= \begin{pmatrix} ((\gamma_{zzz}h)h)h \\ 0 \end{pmatrix} \\
\|\varphi''(t)\| &\le c_3 \|h\|^2 \le c_3 \varepsilon_3^3
\end{aligned}
$$

Taylorentwicklung von $L(\varphi(t), \lambda)$ um $t = 0$:

$$
\begin{aligned}
f(\hat{x}) - f(x^*) &= L(\hat{x}, \lambda) - L(x^*, \lambda) = L(\varphi(t), \lambda) - L(\varphi(0), \lambda) \\
&= \underbrace{\nabla_x L(\varphi(0), \lambda)^T}_{=0} + \frac{1}{2} \underbrace{\varphi'(0)^T \nabla^2_{xx} L(\varphi(0), \lambda) \varphi'(0)}_{=(*)} + \frac{1}{2} \underbrace{\nabla_x L(\varphi(0), \lambda)^T}_{=0} \varphi''(0) \\
&\quad + \frac{1}{6} \left[\underbrace{\varphi'(\tilde{t})^T (\nabla^3_{xxx} L(\varphi(\tilde{t}), \lambda)^T \varphi'(\tilde{t})) \varphi'(\tilde{t})}_{\|\cdot\| \le b\|h\|^2} + 3 \underbrace{\varphi'(\tilde{t})^T \nabla^2_{xx} L(\varphi(\tilde{t}), \lambda) \varphi''(\tilde{t})}_{\|\cdot\| \le b\|h\|^3} + \underbrace{\nabla_x L(\varphi(\tilde{t}), \lambda)^T \varphi'''(\tilde{t})}_{\|\cdot\| \le b\|h\|^3} \right] \\
(*) &= h^T \begin{pmatrix} \gamma_z(z^*) \\ I \end{pmatrix}^T \nabla^2_{xx} L(x^*, \lambda) \begin{pmatrix} \gamma_z(z^*) \\ I \end{pmatrix} h = h^T H h > 0
\end{aligned}
$$

H ist positiv definit und hat kleinsten Eigenwert $\lambda_{min} > 0$. Also ist $(*) \le \lambda_{\min} \|h\|^2$,

$$
f(\hat{x}) - f(x^*) \ge \frac{1}{2} \lambda_{min} \|h\|^2 - \frac{5}{6} b \|h\|^3
$$

Für $\|h\| > 0$ genügend klein ist dies > 0, also $f(\hat{x}) > f(x^*)$. Damit haben wir gezeigt, dass x^* ein striktes lokales Minimum ist.

Satz 6.18: Stabilität

Betrachte das gestörte Problem

$$\min_x f(x,\tau) \quad (6.16)$$
$$\text{s. t.} g(x,\tau) = 0$$

Gelten in x^*, $\tau = 0$ (ungestörter Lösungspunkt): Regularität, notwendige Bedingung erster Ordnung, hinreichende Bedingung zweiter Ordnung. Dann hängt die Lösung von (6.16) in einer Umgebung von x^*, $\tau = 0$ stetig differenzierbar von τ ab. Die $x(\tau)$ sind alle strikte lokale Minima.

Beweis: Wende den Satz für implizite Funktionen auf das System $\nabla_{x,\lambda} L(x,\lambda,\tau) =: F(x,\lambda,\tau) = 0$ an:

$$\frac{\partial}{\partial(x,\lambda)} F(x,\lambda,\tau)\Big|_{t=0} = \begin{pmatrix} \nabla^2_{xx} L(x^*,\lambda,0) & \nabla_x g(x^*,0) \\ \nabla_x g(x^*,0)^T & 0 \end{pmatrix}$$

ist regulär weil $\nabla_x g(x^*,0)$ vollen Rang hat und $\nabla^2_{xx} L(x,\lambda,0)$ positiv definit auf $\ker \nabla_x g(x^*,0)^T$. Also hängen $x(\tau), \lambda(\tau)$ stetig differenzierbar von τ ab und sind stationäre Punkte in einer Umgebung von $\tau = 0$. Die hinreichende Bedingung zweiter Ordnung gilt in einer Umgebung von $\tau = 0$, also sind die $x(\tau)$ strikte lokale Minima. Die Lösung ist also stabil gegenüber kleiner Störungen der Zielfunktion und der Nebenbedingungen.

6.5 Probleme mit Ungleichungsbeschränkungen

Definition 6.19

Betrachte

$$\min f(x) f \in C^2(\mathbb{R}^n, \mathbb{R}) \quad (6.18)$$
$$g(x) = 0 g \in C^2(\mathbb{R}^n, \mathbb{R}^{m_1})$$
$$h(x) \leq 0 h \in C^2(\mathbb{R}^n, \mathbb{R}^{m_2})$$

Die zulässige Menge von (6.18) ist $S := \{x : g(x) = 0, h(x) \leq 0\}$. Sei $J := \{1, \cdots, m_2\}$ die Indexmenge der Ungleichungen. Sei $x \in S$. Die Indexmenge der aktiven Ungleichungen ist $I(x) = \{i \in J : h_i(x) = 0\}$. Die Indexmenge der inaktiven Ungleichungen ist $I^\perp(x) = \{i \in J : h_i(x) < 0$. x^* ist regulär, wenn er die MFCQ oder die LICQ erfüllt.

Definition 6.20: MFCQ:

Sei $x \in S$ zulässig und $I(x)$ die Indexmenge der in x aktiven Ungleichungen. x erfüllt die Mangasarian-Fromowitz-Constraint-Qualifications (MFCQ), wenn:

- die Gradienten $\nabla g_i(x =), i = 1, \cdots, m_1$ linear unabhängig sind.
- ein Vektor $p \in \mathbb{R}^n$ existiert mit $\nabla g_i(x)^T p = 0$, $i = 1, \cdots, m_1$ und $\nabla h_i(x)^T p < 0$, $i \in I(x)$.

Definition 6.21 (LICQ)

Sei $x \in S$ Indexmenge der in x aktiven Ungleichungen. x erfüllt die Linear Independence Constraint Qualifications, wenn die Gradienten $\nabla g_i(x)$, $i = 1, \cdots, m_1$ und ∇h_i, $i \in I(x)$ linear unabhängig sind, also die Menge $\{\nabla g_i\} \cup \{\nabla h_i\}$ linear unabhängig ist.

Satz 6.22

LICQ $\Rightarrow$ MFCQ. Die Umkehrung gilt nicht. Beweis: Übungsaufgabe

Im Minimum gilt: $-\nabla f$ ist Linearkombination mit positiven Koeffizienten (konische Kombination) der ∇h_i:

$$\nabla f + \nabla h \mu = 0 \quad \mu \geq 0$$

Satz 6.23 Notwendige Bedingungen

- Notwendige Bedingung erster Ordnung: Sei x^* ein lokales Minimum von (6.18), x^* regulär. Dann existiert $\lambda \in \mathbb{R}^{m_1}$, $\mu \in \mathbb{R}^{m_2}$, $\mu_i \geq 0$ so dass $\nabla f(x^*) + \nabla g(x^*)\lambda + \nabla h(x^*)\mu = 0$, d. h. $\nabla_x L(x^*, \lambda, \mu) = 0$ für die Lagrangefunktion $L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \lambda^T h(x)$.

  Außerdem gilt Komplementarität:

  $$\sum_{i=1}^{m_2} \mu_i h_i(x^*) = \mu^T h(x^*) = 0$$

  gleichbedeutend:

  $$h_i(x^*) < 0 \Rightarrow \mu_i = 0, \quad \mu_i > 0 \Rightarrow h_i(x^*) = 0$$

  Für Punkte x, die die Komplementaritäts-Bedingung erfüllen ist

  $$I^+(x) = \{i \in I(x), \mu_i > 0\}$$

  die Indexmenge der strikt aktiven Ungleichungen.

- Notwendige Bedingung zweiter Ordnung. Sei x^* ein lokales Minimum, x^* regulär. Seien $\tilde{h}$ die Komponenten von h, die in x^* aktiv sind und $T(x^*) := \{p : \nabla g(x^*)^T p = 0, \nabla \tilde{h}(x^*)^T p = 0\}$ die Tangentialebene an S in x^*. Dann gilt:

  $$p^T \nabla^2_{xx} L(x^*, \lambda, p) p \geq 0 \quad \forall p \in T(x^*) (6.21)$$

Beweis: Schränke ein auf Umgebung U von x^*, so dass $h_i(x) < 0 \forall x \in U \cap S \forall i \in I^{\perp}(x^*)$. Betrachte zulässige Konkurrenten $x \in U \cap S$, lasse nur solche zu, die auch $h_i(x) = 0$ erfüllen $\forall i \in I(x^*)$. Wegen Regularität: $|I(x^*)| + m_1 \leq n$

$$\tilde{S} := \{x : \tilde{h}(x) = 0\} x \in U \cap S \cap \tilde{S}$$

x^* ist ein lokales Minimum des gleichungsbeschränkten Problems bzgl. $S \cap \tilde{S}$. Setze $\mu_i = 0$, $i \in I^{\perp}(x^*)$. Aus dem Gleichungsbeschränkten Fall folgt dann (6.19). $T(x^*)$ ist Tangentialebene des gleichungbeschränkten Problems. Es folgt aus dem gleischungsbeschränkten Fall die notwendige Bedingung zweiter Ordnung. Komplementarität ist durch die Wahl von μ erfüllt: $\mu^T h(x^*) = 0$. Zu zeigen bleibt: $\mu_i \geq 0$, $i \in I(x^*)$. Siehe folgendes Lemma:

Lemma 6.24

Sei (x^*, λ, μ) KKT-Punkt, x^* regulär. Sei $\mu_i < 0$ für ein $i \in I(x^*)$. Dann existiert $\varepsilon > 0$ und eine Kurve $\varphi : [0, \varepsilon] \to S$, so dass $\varphi(0) = x^*$ und $f(\varphi(t)) < f(\varphi(0))$ für alle $t \in [0, \varepsilon]$.

Beweis: betrachte Kurve φ mit $\dot{\varphi}(0) = p$ mit

$$\begin{pmatrix} \nabla g^T \\ \nabla \tilde{h}_1^T \\ \vdots \\ \nabla \tilde{h}_{i-1}^T \\ \nabla \tilde{h}_i^T \\ \nabla \tilde{h}_{i+1} \\ \vdots \\ \nabla \tilde{h}^T \end{pmatrix} (x^*)p = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\varepsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

p ist tengential zu allen Gleichungsbedingungen und aktiven Ungleichungsbedingungen außer i.

$$\tilde{h}_i(\varphi(t)) = \underbrace{\tilde{h}_i(x^*)}_{=0 \text{ aktiv}} + \underbrace{\nabla \tilde{h}_i(x^*)^T p}_{=-\varepsilon \text{ nach Konstr.}} \cdot t + \underbrace{\mathcal{O}(\|p\|^2) t^2}_{\mathcal{O}(\varepsilon^2)}$$

≤ 0 für $\varepsilon > 0$ klein genug

Aus Regularität und Satz für implizite Funktionen folgt: Es gibt eine Kurve φ mit $\dot{\varphi}(0) = p$ und

$$\begin{aligned}
\varphi(t) &\in \{x : g(x) = 0, \tilde{h}_j(x) = 0, j \neq i\} \\
\frac{\partial d}{\partial t} f(\varphi(t))\Big|_{t=0} &= \nabla f(x^*)^T p \\
&= -(\lambda^T \nabla g(x^*)^T + \mu^T \nabla h(x^*)^T) p \\
&= -(-\mu_i \varepsilon) \\
&= \mu_i \varepsilon < 0
\end{aligned}$$

Satz 6.25 (Hinreichende Bedingung)

(x^*, λ, μ) erfülle die notwendige Bedingung erster Ordnung, x^* sei regulär. Sei $p^T \nabla^2_{xx} L(x^*, \lambda, \mu) p > 0 \forall p \in T^+(x^*), p \neq 0$ (6.22) mit $t^+(x^*) = \{p : \nabla g(x^*) p = 0, \nabla h_i(x^*)^T p = 0 \forall i \in I^+(x^*)\}$, $I^+(x^*) = \{i \in I(x^*), \mu_i > 0\}$ (strikt aktive Ungleichung). Dann ist x^* striktes lokales Minimum von (6.18). Es ist sogar auch striktes lokales Minimum von

$$\min f(x) \text{ s. t. } g(x) = 0, h_i(x) = 0, i \in I^+(x^*)$$

d. h. $h_i(x) = 0$ mit $\mu_i = 0$ und $h_i(x) < 0$ weggelassen.
Beweis:
Aus hinreichender Bedingung im gleichungsbeschränkten Fall folgt:

- x^* ist striktes lokales Minimum von (6.23).
- x^* bleibt striktes lokales Minimum von (6.23) wenn man weitere Bedingungen, für die x^* zulässig ist, hinzufügt; $h_i(x) \leq 0$, $i \in I^+(x^*)$. Für die weiteren zulässigen Punkte x (in einer Umgebung von x^* mit $h_i(x) < 0$, $i \in I^+(x^*)$ kann man eine Kurve $\varphi(t)$ konstruieren, so dass $\frac{\partial}{\partial t} f(\varphi(t))\big|_{t=0}$, sofern alle $\mu_i > 0$ sind (analog zu oben). Dazu muss man den Satz für implizite Funktionen anwenden, dafür muss $\text{Rg}\begin{pmatrix} \nabla g & \nabla \tilde{h} \end{pmatrix}^T = m_1 + |I(x^*)|$ in x^* gelten (Regularität).

Definition 6.26: Strikte Komplementarität

In einem KKT-Punkt (x^*, λ, μ), $\mu \geq 0$ gilt strike Komplementarität, wenn:

- $h_i(x^*) = 0 \Rightarrow \mu_i > 0$
- bzw. $\mu_i = 0 \Rightarrow h_i(x^*) < 0$
- bzw. $I^+(x^*) = I(x^*)$
- bzw. $h_i(x^*) = 0$ und $\mu_i = 0$ nicht gleichzeitig.

Bemerkung: Wenn strikte Komplementarität gilt ist $T^+(x^*) = T(x^*)$. In der hinreichenden Bedingung ist dann diese Lücke zu notwendigen Bedingung geschlossen.

Satz 6.27: Stabilität

Sei (x^*, λ^*, μ^*) ein KKT-Punkt, gelte strikte komplementarität, sei x^* regulär und die hinreichende Bedingung zweiter Ordnung sei erfüllt. Dann ist das Minimierungsproblem

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

stabil gegen Störungen von τ um 0 ($\tau = 0$ ist das ursprüngliche Problem). D. h. es existieren Umgebungen U von (x^*, λ^*, μ^*) und V von $\tau = 0$ und eine stetig differenzierbare Abbildung $(x, \lambda, \mu) \colon V \to U$, $\tau \mapsto (x(\tau), \lambda(\tau), \mu(\tau))$ mit $x(0) = x^*$, $\lambda(0) = \lambda^*$, $\mu(0) = \mu^*$ und $x(\tau)$ ist striktes lokales Minimum .

Beweis: Betrachte

$$\begin{aligned} \nabla f(x, \tau) + \nabla g(x, \tau)\lambda + \nabla \tilde{h}(x, \tau)\mu &= 0 \\ g(x, \tau) &= 0 \\ \tilde{h}(x, \tau) &= 0 \end{aligned}$$

Jacobi-Matrix bezüglich (x, λ, μ) in $\tau = 0$:

$$\begin{pmatrix} \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) & \nabla g(x^*) & \nabla \tilde{h}(x^*) \\ \nabla g(x^*)^T & 0 & 0 \\ \nabla \tilde{h}(x^*)^T & 0 & 0 \end{pmatrix}$$

ist regulär.

Mache V möglicherweise kleiner, so dass

- $\mu_i(\tau) > 0$ ($\mu_i(0) > 0$ wegen strikter Komplementarität)
- $h_i(x(\tau)) < 0$ für $i \notin I(x^*)$. Setze deren $\mu_i(\tau) = 0$
- $\begin{pmatrix} \nabla g^T \\ \nabla \tilde{h}^T \end{pmatrix} (x(\tau))$ vollranging
- $\nabla^2_{xx} L(x(\tau), \lambda(\tau), \mu(\tau))$ positiv definit auf $T(x(\tau)) \Rightarrow$ hinreichende Bedingung für $x(\tau), \lambda(\tau), \mu(\tau)$, $\tau \in V$

Bemerkung: Wir haben in den Beweisen LICQ als Regularitätsbedingung benutzt. MFCQ kann auch verwendet werden, Beweise siehe z. B. Geiger Kanzov: Nichtlineare Optimierung, Springer.

SQP-Verfahren

Zunächst: gleichungsbeschränkter Fall

$$\min f(x) \text{ s. t. } g(x) = 0 \quad (7.1)$$

Die notwendigen Optimalitätsbedinungen erster Ordnung lauten:

$$\begin{aligned} \nabla_x L(x,\lambda) &= \nabla f(x) + \nabla g(x)\lambda = 0 \quad (7.2) \\ \lambda_\lambda L(x,\lambda) &= g(x) = 0 \\ \text{bzw. } \nabla L(x,\lambda) &= 0 \end{aligned}$$

Das ist ein nichtlineares Gleichungssystem in (x, λ). Dieses wollen wir mit dem Newton-Verfahren lösen. Ein Schritt $(\Delta x, \Delta\lambda)$ erfüllt dabei das LGS

$$\nabla L(x,\lambda) + \nabla^2 L(x,\lambda) \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = 0$$

(Iterationsindex k weggelassen) bzw.

$$\begin{pmatrix} \nabla f(x) + \nabla g(x)\lambda \\ g(x) \end{pmatrix} + \begin{pmatrix} \nabla^2_{xx} L(x,\lambda) & \nabla g(x) \\ \nabla g(x)^T & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Lemma 7.1

Sei Δx die Lösung des QP

$$\min \frac{1}{2} \Delta x^T H(x) \Delta x + \nabla f(x)^T \Delta x \text{ s. t. } 0 = A(x)\Delta x + g(x) \quad (7.5)$$

mit Matrixfunktionen A, H.
Dann existiert ein $\Delta\lambda$, so dass für beliebige λ

$$\begin{pmatrix} \nabla f(x) + A(x)^T\lambda \\ g(x) \end{pmatrix} + \begin{pmatrix} H(x) & A(x)^T \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = 0 \quad (7.6)$$

Beweis: Lagrangefunktion des QP mit Multiplikator u

$$
\begin{aligned}
L(\Delta x, u) &= \frac{1}{2}\Delta x^T H \Delta x + \nabla f^T \Delta x + u^T(A\Delta x + g) \\
\nabla_{\Delta x} L(\Delta x, u) &= H\Delta x + \nabla f + A^T u - A^T\lambda + A^T\lambda \qquad = 0 \\
\nabla_u L(\Delta x, u) &= A\Delta x + g \\
\begin{pmatrix} \nabla f + A^T\lambda \\ g \end{pmatrix} + \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ u - \lambda \end{pmatrix} &= 0
\end{aligned}
$$

mit $\Delta\lambda = u - \lambda$.
Umgekehrt:

Lemma 7.2

Falls $\Delta\lambda$ existiert, so dass $\begin{pmatrix}\Delta x & \Delta\Lambda\end{pmatrix}^T$ die Gleichung (7.6) erfüllt und wenn $H(x)$ positiv definit auf $\ker A(x)$, dann ist Δx Minimum von (7.5).

Algorithmus 7.3: SQP-Verfahren für gleichungsbeschränkte Probleme

- Startwert x^0, ggf. λ^0, $j := 0$
- Solange ein Abbruchkriterium verletzt ist:
  - Berechne Δx^j als Lösung des folgenden QPs und ggf. die Lagrangemultiplikatoren u^j:

$$\min_{\Delta x} \frac{1}{2}\Delta x^T H^j \Delta x + \nabla f(x^j)^T \Delta x \text{ s. t. } 0 = A^j \Delta x + g(x^j)$$

    mit

$$H^j \cong \nabla^2_{xx} L(x^j, \lambda^j), A^j \cong \nabla g(x^j)^T$$

  - Iteriere:

$$
\begin{aligned}
x^{j+1} &:= x^+ + \alpha^j \Delta x^j \\
\lambda^{j+1} &:= \lambda^k + \alpha^j (\underbrace{u^j - \lambda^j}_{\Delta\lambda^j})
\end{aligned}
$$

    mit $\alpha^j \in (0, 1]$ aus einer Globalisierungsstrategie (z. B. Linesearch).

Bemerkung: λ^j wird benötigt zur Berechnung von H^j (siehe unten) und zur Ungleichungsbehandlung. Wenn $\alpha^j = 1$ hängt λ^{j+1} nicht von λ^j ab.

Korollar 7.4

Für die Wahl $H^j = \nabla^2_{xx} L(x^j, \lambda^j)$ und $A^j = \nabla g(x^j)^T$ ist das SQP-Verfahren ein Newton-Verfahren für die KKT-Bedingung des gleichungsbeschränkten NLP. Nach Korollar 5.4 konvergiert es lokal quadratisch, wenn $\alpha^j \equiv 1 \forall j \geq \bar{j}$.

Bemerkung 7.5

Für $H^j \cong \nabla^2_{xx} L(x^j, \lambda^j)$ und $A^j = \nabla g(x^j)^T$ ist das SQP-Verfahren ein Quasi-Newton-Verfahren, genannt Partial-Quasi-Newton-SQP-Verfahren. Wenn außerdem $A^j \cong \nabla g(x^j)^T$ nennt man das Verfahren Total-Quasi-Newton-SQP-Verfahren.

Die Konvergenzrate hängt von der Wahl der Approximationen ab.

Lösung von QPs mit Gleichungsbeschränkungen

$$\min_p \frac{1}{2} p^T H p + g^T p \text{ s. t. } Ap + b = 0 \quad (7.7)$$

mit $p \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $m \leq n$, Rg$a = m$, $H \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit auf $\ker A$, $g \in \mathbb{R}^n$, $b \in \mathbb{R}^m$.

- zulässige Menge ist konvex
- Zielfunktion ist konvex auf der zulässigen Menge

$\Rightarrow$ es existiert genau ein Minimum.
Äquivalent:

$$\exists u : \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ u \end{pmatrix} = - \begin{pmatrix} g \\ b \end{pmatrix}$$

Variante 1: Bildraummethode: Numerisch instabil, im Allgemeinen teurer als Variante 2, nur geeignet wenn H insgesamt positiv definit.

- Multipliziere die erste Blockzeile mit H^{-1} und A und ziehe von der zweiten Blockzeile ab:
$$\begin{pmatrix} H & A^T \\ 0 & -AH^{-1}A^T \end{pmatrix} \begin{pmatrix} p \\ u \end{pmatrix} = - \begin{pmatrix} g \\ b - AH^{-1}g \end{pmatrix}$$
erfordert eine Cholesky-Zerlegung von H und eine Matrixmultiplikation.
- Zerlege $-AH^{-1}A^T$ (sog. Schur-Komplement), berechne u
- Aus erster Blockzeile:
$$p = H^{-1}(-g + A^T u)$$

Variante 2: Nullraummethode: stabil, deutlich schneller als Variante 1, wenn $m > \frac{n}{2}$.

- QR-Zerlegung von A^T:

$$A^T = Q^T \tilde{L}^T = \begin{pmatrix} \underbrace{Q_1^T}_{n\times m} & \underbrace{Q_2^T}_{n\times(m-n)} \end{pmatrix} \begin{pmatrix} L^T \\ 0 \end{pmatrix} = Q_1^T L^T$$

$$y := Qp = \begin{pmatrix} Q_1 p \\ Q_2 p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$Ap = \begin{pmatrix} L & 0 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} p = LQ_1 p$$

$$= Ly_1 = -b$$

$$y_1 = -L^{-1}b$$

  y_2 zunächst frei.

$$AQ_2^T = LQ_1 Q_2^T = 0$$

  Spalten von Q_2^T bilden eine Basis von $\ker A$.

$$p = Q^T y = Q_1 y_2 + Q_2^T y_2$$

- Einsetzen in die erste Blockzeile und Multiplikation mit Q:

$$QHQ^T y + QA^T u = -Qg$$

  hat zwei Teile:

  - $A_2 H(Q_1^T y_1 + Q_2^T y_2) + \underbrace{Q_2 A^T}_{=0} u = -Q_2 g$ bzw. $\underbrace{Q_2 H Q_2^T}_{(*)} y_2 = -Q_2(y - HQ_1^T L^{-1} b)$

    (*): auf $\ker A$ projeziertes H, positiv definit, zerlege mit Cholesky $\Rightarrow y_2$ und $p = Q^T y$

  - $Q_1(Hp + g) + Q_1 A^T u = 0$, $Q_1 A^T = L^T \Rightarrow u = -L^{-T} Q_1(Hp + g)$.

7.3: Quasi-Newton-SQP mit Update

Ziel: berechne $H^k \cong \nabla^2_{xx} L(x^k, \lambda^k)$ „gut und billig“.

Niedrigrang-Aufdatierungen:

$$H^{k+1} = H^k + \overbrace{\underbrace{ab^T}_{\text{hat Rang 1}} + cd^T}^{\text{hat Rang 2}}$$

mit $a, b, c, d \in \mathbb{R}^n \setminus \{0\}$

Wichtigstes Beispiel: BFGS (Broyden, Fletcher, Goldfarb, Shanno)

$$H^{k+1} = H^k + \frac{y^k {y^k}^T}{{y^k}^T y^k} - \frac{(H^k s^k)(H^k s^k)^T}{{s^k}^T H^k s^k} \quad (7.9)$$

mit

$$y^k = \nabla_x L(x^{k+1}, \lambda^{k+1}) - \nabla_x L(x^k, \lambda^{k+1})$$
$$s^k = x^{k+1} - x^k = \alpha^k \Delta x^k$$

Definition 7.5: Sekantenbedingung

H^{k+1} erfüllt die Sekantenbedingung, wenn

$$H^{k+1} s^k = y^k \quad (7.10)$$

Das bedeutet, dass H^{k+1} in erster Ordnung korrekt ist. Taylorentwicklung um x^{k+1} angewendet bei x^k

$$\nabla^2_{xx} L(x^{k+1}, \lambda^{k+1})(x^{k+1} - x^k) = \nabla_x L(x^{k+1}, \lambda^{k+1}) - \nabla_x L(x^k, \lambda^{k+1}) + \mathcal{O}(\|x^{k+1} - x^k\|^2)$$
$$\nabla^2_{xx} L(x^{k+1}, \lambda^{k+1}) s^k = y^k + \mathcal{O}(\|x^{k+1} - x^k\|^2)$$