

# Inhaltsverzeichnis

<b>1. Parameterabhängige Anfangswertprobleme</b>	<b>7</b>
Differentiation der Lösung des AWP's . . . . .	9
Berechnung der Ableitungen: . . . . .	10
Variationsdifferentialgleichungen für Richtungsableitungen . . . . .	12
Ableitungsmatrizen bei abschnittsweise definierter rechter Seite . . . . .	14
Adjungierte Differentialgleichung . . . . .	15
Satz 1.12 . . . . .	15
Bemerkung 1.13 . . . . .	17
Ableitung von Linearkombinationen von $\frac{\partial y}{\partial y_0}$ . . . . .	17
Anwendung: . . . . .	18
Zusammenfassung: . . . . .	18
Simulation und Optimierungsprobleme bei Differentialgleichungen . . . . .	18
<b>2. Formulierung von Parameterschätzproblemen</b>	<b>21</b>
Problemformulierung . . . . .	22
Prozessmodell . . . . .	22
Modell für Beobachtungen . . . . .	22
Lösungsmethoden . . . . .	27
Parametrisierung der Lösung des AWP durch Single Shooting, Multiple Shooting oder Kollokation . . . . .	27
Löse (2.15) mit verallgemeinertem Gauß-Newton-Verfahren . . . . .	27
<b>3. Shooting-Verfahren und Kollokation</b>	<b>29</b>
3.1 Single-Shooting: Einzelschießverfahren . . . . .	29
Algorithmus 3.2: Verallgemeinertes Gauß-Newton-Verfahren . . . . .	30
Algorithmus 3.3: Single-Shooting Gauß-Newton . . . . .	30
Bemerkung 3.4 . . . . .	31
3.2 Multiple Shooting, die Mehrzielmethode . . . . .	32
Bemerkung 3.5 . . . . .	34
3.3 Kollokation . . . . .	34
Kollokations-Diskretisierung . . . . .	34
Wahl der Polynombasis und der Kollokationspunkte . . . . .	35
Konsistenzfehler: . . . . .	36
Beispiel Impliziter Euler, $k=1$ . . . . .	36
Kollokation für beschränkte Parameterschätzprobleme . . . . .	37
3.4 Ansätze zur Optimierung von DGL-Modellen . . . . .	38

3.5 Relaxierte Formulierung von DAEs . . . . .	39
<b>4. Verallgemeinerte Gauß-Newton-Verfahren</b>	<b>41</b>
Algorithmus 4.1 (Verallgemeinertes Gauß-Newton-Verfahren) . . . . .	41
Bemerkung 4.2 . . . . .	42
Annahmen 4.3: Regularitätsannahmen . . . . .	42
Lemma 4.4 . . . . .	42
Lösung der linearen Ausgleichsprobleme . . . . .	42
1. Unbeschränkter Fall . . . . .	42
Bemerkung 4.10 . . . . .	43
2. Beschränkter Fall . . . . .	43
Lemma 4.7 . . . . .	44
Numerische Lösung . . . . .	46
Lemma 4.11 (Berechnung der adjungierten . . . . .	49
3. Variante: Stoer 1979 . . . . .	49
Anwendung auf die Mehrzielmethode . . . . .	49
Algorithmus 4.12 (Eliminationsalgorithmus) . . . . .	50
Algorithmus 4.13: Lösen . . . . .	51
<b>5. Lokale Konvergenz von Newton-Typ-Verfahren</b>	<b>55</b>
Algorithmus 5.1 (Newton-Typ-Verfahren) . . . . .	55
Bemerkung 5.2 . . . . .	56
Satz 5.3: Lokaler Kontraktionssatz (Bock 1987) . . . . .	56
Korollar 5.4 . . . . .	58
Bemerkung 5.5: Quasi-Newton-Verfahren . . . . .	59
Satz 5.6 (Dennis-Moré . . . . .	59
Varianten von Quasi-Newton-Verfahren . . . . .	60
Bemerkung 5.7 (Bedeutung von $\omega$ ) . . . . .	60
Anwendung auf (verallgemeinerte) Gauß-Newton-Verfahren: . . . . .	60
Bemerkung 5.8: Bedeutung von $\kappa$ . . . . .	61
Bemerkung 5.9 . . . . .	61
Korollar 5.10 . . . . .	61
Newton-Verfahren für die nichtlineare Gleichung $\nabla f(x) = 0$ . . . . .	62
Gauß-Newton-Verfahren für $\min \frac{1}{2} \ F(x)\ _2^2$ . . . . .	62
Newton-Verfahren für $\min \frac{1}{2} \ F(x)\ _2^2$ . . . . .	62
Bemerkung 5.11 . . . . .	62
Satz 5.12: Kleine-Residuen-Probleme . . . . .	63
Satz 5.13 . . . . .	65
Korollar 5.14 . . . . .	65
Fazit . . . . .	65
Statistische Störung des Problems . . . . .	66
Satz 5.15 . . . . .	66
Fazit . . . . .	67

<b>6 Optimalitätsbedingungen für nichtlineare Optimierungsprobleme</b>	<b>69</b>
6.1 Allgemeine Problembeschreibung . . . . .	69
Definition 6.1 . . . . .	69
Definition 6.2 . . . . .	69
Definition 6.3 . . . . .	69
Definition 6.4 . . . . .	70
6.2 Optimalitätsbedingung im eindimensionalen Fall . . . . .	70
6.3 Unbeschränkter Fall . . . . .	70
Satz 6.7 (Heinreichende Bedingung) . . . . .	70
6.4 Gleichungsbeschränkter Fall . . . . .	71
Definition 6.8 . . . . .	71
Definition 6.9: Tangentialebene . . . . .	71
Definition 6.10 . . . . .	71
Satz 6.12 (Notwendige Bedingung erster Ordnung) . . . . .	72
Bemerkung 6.13 . . . . .	73
Satz 6.12 . . . . .	73
Definition 6.14 . . . . .	73
Bemerkung 6.15 . . . . .	73
Satz 6.16 (Notwendige Bedingungen zweiter Ordnung) . . . . .	74
Satz 6.17 (Hinreichende Bedingung) . . . . .	74
Satz 6.18: Stabilität . . . . .	75
6.5 Probleme mit Ungleichungsbeschränkungen . . . . .	76
Definition 6.19 . . . . .	76
Definition 6.20: MFCQ: . . . . .	76
Definition 6.21 (LICQ) . . . . .	77
Satz 6.22 . . . . .	77
Satz 6.23 Notwendige Bedingungen . . . . .	77
Lemma 6.24 . . . . .	78
Satz 6.25 (Hinreichende Bedingung) . . . . .	79
Definition 6.26: Strikte Komplementarität . . . . .	79
Satz 6.27: Stabilität . . . . .	80
<b>SQP-Verfahren</b>	<b>81</b>
Lemma 7.1 . . . . .	81
Lemma 7.2 . . . . .	82
Algorithmus 7.3: SQP-Verfahren für gleichungsbeschränkte Probleme . . . . .	82
Korollar 7.4 . . . . .	83
Bemerkung 7.5 . . . . .	83
Lösung von QPs mit Gleichungsbeschränkungen . . . . .	83
7.3: Quasi-Newton-SQP mit Update . . . . .	84
Wichtigstes Beispiel: BFGS (Broyden, Fletcher, Goldfarb, Shanno) . . . . .	85
Definition 7.5: Sekantenbedingung . . . . .	85
DFP (Davidon, Fletcher, Powell) . . . . .	85
Eigenschaften von BFGS und DFP . . . . .	85

Satz 7.7 . . . . .	86
Resultate zur lokalen Konvergenz von BFGS-Quasi-Newton-SQP-Verfahren . .	86
Satz 7.8: Dennis-Moré im beschränkten Fall . . . . .	86
Bemerkung 7.9 . . . . .	86
Satz 7.10 (Superlineare Konvergenz von BFGS-SQP-Verfahren . . . . .	87
7.4 SQP-Verfahren für NLPs mit Ungleichungsbeschränkungen . . . . .	87
Lemma 7.11 . . . . .	87
Lemma 7.12 . . . . .	88
Folgerung 7.13 . . . . .	88
Lösung von Ungleichungsbeschränkten QP mit einer Active-Set-Methode . . .	88
Active-Set-Strategie . . . . .	89
Algorithmus 7.14: Active-Set-Methode für konvexe QP . . . . .	89
Bemerkung 7.15 . . . . .	91
Satz 7.16 . . . . .	91
Innere-Punkt-Methode . . . . .	91
<b>8. Optimale Steuerung</b>	<b>93</b>
Problemformulierung . . . . .	93
Beispiel: Fahrdynamik . . . . .	93
Allgemeine Problemformulierung . . . . .	94
Modellgleichung . . . . .	94
Randbedingungen . . . . .	95
Steuerungsbeschränkungen . . . . .	95
Zustandsbeschränkungen . . . . .	95
Zielfunktional . . . . .	95
Bemerkung 8.1 . . . . .	96
Bemerkung 8.2 . . . . .	96
Bemerkung 8.3 . . . . .	96
Bemerkung 8.4 . . . . .	96
8.3 Ansätze zur Lösung von Optimalsteuerungsproblemen . . . . .	96
First Optimize then Discretize . . . . .	96
First Discretize then Optimize . . . . .	97
<b>9. Direkter Ansatz der optimalen Steuerung</b>	<b>99</b>
Bemerkung 9.1 . . . . .	99
Bemerkung 9.2 . . . . .	99
Umformulierung der Steuerungsbeschränkungen . . . . .	100
9.2 Direktes Single-Shooting . . . . .	101
9.3 Direktes Multiple Shooting . . . . .	102
Mehrzielmethode / Multiple Shooting . . . . .	102
Kondensieralgorithmus (Plitt 1981, Bock, Plitt 1984 . . . . .	106
Umsortieren der Variablen . . . . .	106
Vorteile von Multiple Shooting gegenüber Single Shooting . . . . .	106

<b>10 Indirekter Ansatz der optimalen Steuerung</b>	<b>107</b>
Optimalsteuerungsproblem . . . . .	107
Definition 10.1 . . . . .	107
Notwendige Optimalitätsbedingung . . . . .	107
Anwendungsbeispiel: Energieoptimaler Raketenwagen . . . . .	108
Prinzip der indirekten Methode . . . . .	110
Bemerkung 10.3: Variation, Beispiel 1 . . . . .	115
Bemerkung 10.4: Maximumprinzip mit Randbedingungen . . . . .	116
Bemerkung 10.5: freie Endzeit $t_{end}$ . . . . .	117
Spezialfälle des Maximumprinzips . . . . .	118
<b>11 Optimale Versuchsplanung</b>	<b>121</b>
11.1 Anwendungsbeispiel Urethan-Reaktion . . . . .	121
Statistische Analyse der Lösung des Parameterschätzproblems . . . . .	123
Modell für die Messungen . . . . .	123
Bemerkung 11.1: Eigenschaften von C . . . . .	126
Lemma 11.2 . . . . .	127
Korollar 11.3 . . . . .	128
Konfidenzgebiete . . . . .	128
Linearisierte Näherung . . . . .	128
Lemma 11.4 . . . . .	129
Korollar 11.5 . . . . .	129
11.3 Das Versuchsplanungs- und Optimierungsproblem . . . . .	129
11.6 . . . . .	130
Bemerkung 11.7 . . . . .	130
Optimale Versuchsplanung . . . . .	130
Versuchsraum . . . . .	130
Varianten . . . . .	132
Feste und variable Experimente . . . . .	132
Mehrfachexperiment-VP-Problem . . . . .	133
Bemerkung 11.8 . . . . .	134
Bemerkung 11.11 . . . . .	134
Robustifizierung der Versuchsplanung gegenüber Parameterunsicherheit . . . . .	135
Lemma 11.9 . . . . .	135
Bemerkung 11.10 . . . . .	136
11.4 Numerische Methoden zur Lösung von Versuchsplanungsproblemen . . . . .	136
Lemma 11.12: Ableitung der inversen Matrix . . . . .	138
5. Ausnutzung der Mehrfachexperimentstruktur . . . . .	139
Anwendungsbeispiel: Ergebnisse . . . . .	139



# 1. Parameterabhängige Anfangswertprobleme

*Parameterabhängige gewöhnliche Differentialgleichung* mit Parametervektor  $p$  und parameterabhängiger Anfangsbedingung:

$$\dot{y}(t) = f(t, y(t), p) \quad (1.1)$$

$$y(t_0) = y_0(p) \quad (1.2)$$

wobei

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

*Variante:* Differentiell-algebraische Gleichungssysteme (DAE):

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

Das DAE benötigt nur Anfangswerte für  $y$ :

$$y(t_0) = y_0(p) \quad (1.4)$$

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

$$g(t_0, y(t_0), z(t_0), p) = 0 \quad (1.5)$$

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

$$\frac{\partial g}{\partial z} \quad \text{hat den Rang } n_z \quad (1.6)$$

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

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

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

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):

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

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

$$\begin{aligned}\dot{y}(t) &= f(t, y, p), & y(t_0) &= y_0(p) \\ \dot{v}(t) &= f(t, v, p_1), & v(t_0) &= y_0(p + \delta p)\end{aligned}$$

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

$$\begin{aligned}\|\delta y_0\| &:= \left\| \frac{\partial y_0}{\partial p}(p) \delta p \right\| \leq \varepsilon_1 \\ \|\delta f\| &:= \left\| \frac{\partial f}{\partial p}(t, y, p) \delta p \right\| \leq \varepsilon_2\end{aligned}$$

*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 AWP

*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:

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

*Definition 1.4:* Ableitungen der Lösung des AWP

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

### Berechnung der Ableitungen:

Integralform des AWP:

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

#### a) Differenziere nach Anfangswerten:

$$\begin{aligned} \frac{\partial}{\partial y_0} y(t) &= \frac{\partial y_0}{\partial y_0} + \int_{t_0}^t \frac{\partial}{\partial y} f(\tau, y(\tau), p) \frac{\partial y}{\partial y_0}(\tau) \, d\tau \\ G(t) &= I + \int_{t_0}^t \frac{\partial}{\partial y} f(\tau, y(\tau), p) G(\tau) \, d\tau \end{aligned} \quad (1.13)$$

Dies ist äquivalent zum Anfangswertproblem

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

zur sogenannten Variationsdifferentialgleichung (VDE) nach Anfangswerten.

**Differenziere nach dem Anfangszeitpunkt:**

$$\begin{aligned}\frac{\partial}{\partial t_0} y(t) &= \frac{\partial y_0}{\partial t_0} + \frac{\partial}{\partial t_0} \int_{t_0}^t f(\tau, y(\tau), p) \, d\tau \\ G_{t_0}(t) &= 0 - f(t_0, y(t_0), p) + \int_{t_0}^t \frac{\partial f}{\partial y} G_{t_0}(\tau) \, d\tau\end{aligned}\quad (1.15)$$

Dies ist äquivalent zum Anfangswertproblem

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

zur VDE nach dem Anfangszeitpunkt.

**c) Differenziere nach den Parametern:**

$$\begin{aligned}\frac{\partial}{\partial p} y(t) &= \frac{\partial y_0}{\partial p} + \int_{t_0}^t \frac{\partial f}{\partial y}(\tau, y(\tau), p) \frac{\partial y}{\partial p}(\tau) + \frac{\partial f}{\partial p}(\tau, y(\tau), p) \, d\tau \\ G_p(t) &= \frac{\partial y_0}{\partial p} + \int_{t_0}^t \frac{\partial f}{\partial y} G_p(\tau) + \frac{\partial f}{\partial p} \, d\tau\end{aligned}\quad (1.17)$$

Dies ist äquivalent zum Anfangswertproblem

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

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  $y: = -Gf(t_0, y_0, p)$  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  $\Delta y_0$  ist gegeben durch

$$\begin{aligned} \frac{\partial y}{\partial y_0}(t; t_0, y_0, p)\Delta y_0 &:= \left. \frac{\partial}{\partial h} y(t; t_0, y_0 + h\Delta y_0, p) \right|_{h=0} \\ &= \lim_{h \rightarrow 0} \frac{y(t; t_0, y_0 + h\Delta y_0, p) - y(t; t_0, y_0, p)}{h} \in \mathbb{R}^{n_y} \end{aligned} \quad (1.19a)$$

Die Richtungsableitung erfüllt

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

Dies ist ein  $(n_y\text{-dimensionales})$  VDE-Anfangswertproblem für jede Richtung  $\Delta y_0$ . hat man mehrere Richtungen, kann man diese spaltenweise in einer Richtungsmatrix  $S$  zusammenfassen:

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

Zur Berechnung aller zugehörigen Richtungsableitungen muss man also lösen:

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

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:

$$\begin{aligned}\frac{\partial}{\partial t}(G_p(t; t_0, y_0, p)\Delta p) &= \frac{\partial f}{\partial y}(t, y, p)G_p(t; t_0, y_0, p)\Delta p + \frac{\partial f}{\partial p}(t, y, p) \\ G_p(t_0; t_0, y_0, p)\Delta p &= \frac{\partial y_0}{\partial p}\Delta p\end{aligned}\quad (1.22)$$

### Eigenschaften der Ableitungsmatrizen

Füge an der Stelle  $t_0 < s < t_{end}$  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-Lindelöf ist die abschnittsweise Lösung die selbe wie für

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

*Definition 1.7: Wronski-Matrix*

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

*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  $t_1, \dots, t_n$  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:*

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

*Satz 1.10*

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

*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

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

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

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

gegeben sei.

*Satz 1.11:*

Dann gilt:

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

*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.

$$\begin{aligned}
 0 &= Q(t_s, y(t_s), p) \\
 \Rightarrow 0 &= \frac{\partial Q}{\partial t} \frac{\partial t}{\partial y} + \frac{\partial Q}{\partial y} \Big|_{y=y_s} \\
 \Rightarrow \frac{\partial t_s}{\partial y_s} &= - \left( \frac{\partial Q}{\partial t} \right)^{-1} \frac{\partial Q}{\partial y_s} \Big|_{t=t_s, y=y_s} \quad (\text{Satz für implizite Funktionen})
 \end{aligned}$$

$$\frac{\partial y}{\partial t_s}(t) = -W(t, t-s) f_2(t_s, y_s, p) \quad (\text{Satz 1.6})$$

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

## Adjungierte Differentialgleichung

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

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

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:

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

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

*Beweis:*

Es gilt:

$$\begin{aligned} \dot{y} - f(t, y, p) &= 0 \\ \Rightarrow \int_{t_0}^T \Lambda^T(\dot{y} - f(t, y, p)) dt &= 0 \end{aligned} \quad (1.32)$$

Leite (1.32) nach  $y_0$  ab:

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

Partielle Integration:

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

Daraus folgt:



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

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

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

### Bemerkung 1.13

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

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

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

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

## Inhaltsverzeichnis

mit dem Anfangswert

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

hat die Lösung

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

$$\text{und } \int_{t_0}^T (u^T \Lambda(t)^T) \frac{\partial f}{\partial p}(t, y, p) dt = u^T \frac{\partial y}{\partial p}(T) \quad (1.36)$$

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:

$$\begin{aligned} \Phi(y(T; t_0, y_0, p)) & \quad \text{„Zielfunktion“} \\ \Phi : \mathbb{R}^{n_y} & \rightarrow \mathbb{R} \end{aligned}$$

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 möglichst 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:



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 Experimentator 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

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

### Modell für Beobachtungen

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

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

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

$$\eta_i = h_i(t_i, y(t_i), \bar{p}) + \varepsilon_i \quad (2.2)$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2) \quad (2.3)$$

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  $\sigma$  :

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

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

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

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:

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

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 Problem nicht.

## Inhaltsverzeichnis

- 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:*

$$\begin{aligned}\eta &= \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_M \end{pmatrix} \\ h = h(y, p) &= \begin{pmatrix} h_1(t_1, y(t_1), p) \\ \vdots \\ h_M(t_M, y(t_M), p) \end{pmatrix} \\ \varepsilon &= \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_M \end{pmatrix} \\ S &= \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_M^2 \end{pmatrix}\end{aligned}$$

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) multinomialverteilte Messfehler ist  $S$  symmetrisch und positiv definit.

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

mit Varianzen  $\sigma_i^2$  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:

$$\begin{aligned} \frac{1}{2}(\eta - h)^T S^{-1}(\eta - h) &= \frac{1}{2}(\eta - h)L^{-T}D^{-\frac{1}{2}}D^{-\frac{1}{2}}L^{-1}(\eta - h) \\ &= \frac{1}{2}\|D^{-\frac{1}{2}}L^{-1}(\eta - h)\|_2^2 \quad (2.9) \end{aligned}$$

$$D^{-\frac{1}{2}}L^{-1}(\eta - h(y, \bar{p})) \sim \mathcal{N}(0, I)$$

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“

$$q = \infty$$

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

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  $r(y(t_0), y(t_1), \dots, y(t_{K-1}), y(t_{end}), p) = 0$

Beispiele und Spezialfälle:

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

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

Bemerkung zu Ungleichbedingungen: 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 Ungleichbedingungen gelöst haben kann folgendes passieren:

- Im Lösungspunkt sind die Ungleichungen inaktiv, d. h. der Lösungspunkt erfüllt die echte Ungleichung  $s(y(t_0), \dots, y(t_k), p) > 0$ . 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

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

oder mit DAE

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

## 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  $F_{1,i}(p) = \sigma_i^{-1}(\eta_i - h_i(t_i, y(t_i; t_0, y_0, p), p))$ ,  $i = 1, \dots, M$  (3.2). Setze die Lösung außerdem an den Randbedingungspunkten in die Randbedingung ein:  $F_2(p) = r(y(t_0; t_0, y_0, p), \dots, y(t_k; t_0, y_0, p), p)$  (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  $J_2$ .

$$\begin{aligned}
 J_1(p) &:= \frac{\partial F_1}{\partial p}(p) \\
 &= \frac{\partial}{\partial p} \left( \frac{\eta_i - h_i(\dots)}{\sigma_i} \right)_{i=1, \dots, M} \\
 &= \left( -\frac{1}{\sigma_i} \left( \frac{\partial h_i}{\partial y}(\dots) \frac{\partial y}{\partial p}(t_i; t_0, y_0, p) + \frac{\partial h_i}{\partial p}(t_i; t_0, y_0, p) \right) \right)_{i=1, \dots, M} \quad (3.6)
 \end{aligned}$$

$$\begin{aligned}
 J_2(p) &= \frac{\partial F_2}{\partial p}(p) \\
 &= \sum_{i=0}^k \frac{\partial r}{\partial y_i} \frac{\partial y}{\partial p}(t_i; t_0, y_0, p) + \frac{\partial r}{\partial p} \quad (3.7)
 \end{aligned}$$

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

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

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  $\delta p^k$  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  $\alpha^k$ , 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  $\delta p^k$  durch Lösen des linearen Ausgleichsproblems (3.9).
  - Berechne eine Schrittweite  $\delta p^k$  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:

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

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

### Bemerkung 3.4

Wenn wir keine Randbedingung  $r(\dots) = 0$  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 Parameter 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 Parameterschätz-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  $t_0 = \tau_0 < \tau_1 < \dots < \tau_m = t_{end}$ . Die  $\tau_i$  nennt man Mehrzielknoten. Löse das AWP nur auf diesen Intervallen  $i = 0, \dots, m-1$ :  $\dot{y} = f(t, y, p)$ ,  $t \in [\tau_i, \tau_{i+1}]$  (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  $s_i$ ,  $i = 0, \dots, m-1$  sind zusätzliche Variablen des Problems. An den Mehrzielknoten  $\tau_i$  formulieren wir zusätzliche Stetigkeitsbedingungen, sog. „Anschlussbedingungen“:

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

Für DAEs ist  $s_i = (s_i^y, s_i^z)$  und im Punkt  $\tau_i$  müssen die Konsistenzbedingungen  $g(\tau_i, s_i^y, s_i^z, p) = 0$ ,  $i = 0, \dots, m-1$  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  $x = (s_0, \dots, s_{m-1}, p)$  (3.16)
- den Zielfunktionstermen

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

mit  $t_j \in [\tau_{i_j}, \tau_{i_j+1}]$ ,  $j = 1, \dots, M$

- den Nebenbedingungen  $F_2(x)$ , die bestehen aus:
  - den Randbedingungen:  $r(y(t_0; \tau_0, s_0, p), \dots, y(t_k; \tau_{i_k}, s_{i_k}, p), p) = 0$  (3.18) mit  $t_j \in [\tau_{i_j}, \tau_{i_j+1}]$ ,  $i = 1, \dots, k$
  - den Anschlussbedingungen  $s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) = 0$ ,  $i = 0, \dots, m-2$  (3.19)
  - und für DAEs den Konsistenzbedingungen  $g(\tau_i, s_i^y, s_i^z, p) = 0$ ,  $i = 0, \dots, m-1$  (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  $(s_0^*, s_1^*, \dots, s_{m-1}^*, p)$  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:

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

mit

$$\begin{aligned} D_1^i &= \frac{\partial F_1}{\partial s_i} & i = 0, \dots, m-1 \\ D_1^p &= \frac{\partial F_1}{\partial p} & \text{Ableitung der ZF-Terme} \\ D_2^o &= \frac{\partial r}{\partial s_i} & i = 0, \dots, m-1 \\ D_2^p &= \frac{\partial r}{\partial p} & \text{Ableitung der Randbedingungen} \end{aligned}$$

Ableitung der Anschlussbedingungen:

$$\begin{aligned} -s_{i+1} + y(\tau_{i+1}; \tau_i, s_i, p) &= 0 & i = 0, \dots, m-2 \\ \text{nach } s_i : \frac{\partial}{\partial s_i} y(\tau_{i+1}; \tau_i, s_i, p) &=: G_i \\ &\text{nach } s_{i+1} : -I \\ \text{nach } p : \frac{\partial}{\partial p} y(\tau_{i+1}; \tau_i, s_i, p) &=: G_i^p & (\text{sonst } 0) \end{aligned}$$

Ableitung der Konsistenzbedingungen:

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

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

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  $t_0 = \tau_0 < \tau_1 < \dots < \tau_{m-1} < \tau_m = t_{end}$  (3.25). Auf jedem dieser Teilintervalle wird die Lösung dargestellt durch

$$y_j^\tau(t; s_j) := \sum_{l=0}^k s_{jl} \chi_l \left( \frac{t - \tau_j}{\tau_{j+1} - \tau_j} \right) \quad (3.2)$$

$$t \in [\tau_j, \tau_{j+1}], \quad j = 0, \dots, m-1$$

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

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

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

$$t_{jl} = \bar{c}_j + \rho_l h_j \quad l = 1, \dots, k, \quad j = 0, \dots, m-1 \quad (3.28)$$

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

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

- Die approximierte Lösung  $y^\tau$  soll an den Gitterpunkten  $\tau_j$  stetig sein:

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

- Anfangsbedingung:

$$y^\tau(\tau_0; s_0) = y_0(p) \quad (3.31)$$

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:

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

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:

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

und die Stetigkeitsbedingung:

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

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

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

$$\text{mit } \Psi_l(\rho) = \int_0^\rho L_l(\bar{\rho}) d\bar{\rho}$$

$$\Psi_l(1) = \int_0^1 L_l(\rho) d\rho$$

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

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:

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

Die Konsistenzordnung  $q$  hängt ab von der Wahl der  $\rho_l$ ,  $l = 1, \dots, k$ . 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:

$$s_{j-10} + h_{j-1} s_{j-1} - s_{j0} = 0$$

## Kollokation für beschränkte Parameterschätzprobleme

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

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

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

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

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 verschiedenen hochdimensionalen Suchräumen berechnet. Dies kann bessere Konvergenzeigenschaften bedeuten.

Verallgemeinerung: Lifting von beliebigen nichtlinearen Problemen in höherdimensionale Räume:

$$x^{16} = 2$$

$$\Leftrightarrow$$

$$x_1^2 - x_2 = 0$$

$$x_2^2 - x_3 = 0$$

$$x_3^2 - x_4 = 0$$

$$x_4^2 - 2 = 0$$

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

### 3.5 Relaxierte Formulierung von DAEs

DAE:

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

Anfangsbedingung:  $y(t_0) = y_0, \quad z(t_0) = z_0$   
 Konsistenzbedingung:  $0 = g(t_0, y_0, z_0, p)$

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:

$$\begin{aligned} \dot{y} &= f(t, y, z, p) \\ 0 &= g(t, y, z, p) - \beta(t)g(t_0, y_0, z_0, p) \end{aligned} \quad (3.40)$$

$$\begin{aligned} y(t_0) &= y_0 \\ z(t_0) &= z_0 \end{aligned}$$

mit  $\beta(t_0) = 1$   
 z. B.  $\beta(t) \equiv 1$   
 oder  $\beta(t) = e^{-\alpha(t-t_0)}$

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  $i = 0, \dots, m-1$ .

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:

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

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

### Algorithmus 4.1 (Verallgemeinertes Gauß-Newton-Verfahren)

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

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

- Löse das lineare Ausgleichsproblem

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

Lösung:  $\Delta x^k$

- 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 Vollschrte  $\alpha^k = 1$  erwartet werden.

## Annahmen 4.3: Regularitätsannahmen

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

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

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.  $\Delta x^*$  ist Lösung von (4.5) genau dann, wenn es das sogenannte Normalgleichungssystem löst:

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

$\Delta x^*$  ist Minimum genau dann wenn

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

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

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

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

#### Bemerkung 4.10

Wenn  $\text{Rg} J_1 < n$ , 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

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

Sei  $\Delta x^*$  Minimum.

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

mit  $\Delta y$  so dass

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

Gelte umgekehrt (4.8) für  $\Delta x^*$  und  $\lambda$ . Dann folgt  $f'(0) = 0$  für alle  $\Delta y \in \ker(J_2)$ . Aus (PD) folgt, dass  $\Delta x^*$  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: Satz

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

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  $J_1 = QR = \overline{Q}\overline{R}$

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

- 2. Variante: Singulärwertzerlegung von  $J_1$ :

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

mit  $U, V$  orthogonal und  $\Sigma$  Diagonalmatrix mit Singulärwerten von  $J_1$  auf der Diagonalen.

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

### Beschränkter Fall

- 1. Variante („Bildraumvariante“):  $\Delta x$  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“):

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

1. Bestimme Lösungsmenge von (4.11) durch LR-Zerlegung von  $P_2 J_2$ :

$$P_2 J_2 = L_2 \begin{pmatrix} R_2 & D_2 \end{pmatrix} = \begin{pmatrix} P_2 J_{21} & P_2 J_{22} \end{pmatrix} \quad (4.12)$$

Mit  $R_2$  reguläre obere Dreiecksmatrix,  $L_2$  normierte untere Dreiecksmatrix,  $P_2$  Permutationsmatrix zur Zeilenpivotierung. Voraussetzung ist (CQ). Spalte  $\Delta x$  auf:

$$\begin{aligned} \Delta x &= \begin{pmatrix} \underbrace{\Delta y}_{m_2} & \underbrace{\Delta z}_{n-m_2} \end{pmatrix} \\ P_2 J_2 \Delta x &= L_2 (R_2 \Delta y + D_2 \Delta z) = -P_2 F_2 \\ \Rightarrow \Delta y &= -R_2^{-1} (D_2 \Delta z + L_2^{-1} P_2 F_2) \quad (4.13) \end{aligned}$$

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

$$\begin{aligned} J_1 &= \begin{pmatrix} \underbrace{J_{11}}_{m_1 \times m_2} & \underbrace{J_{12}}_{m_1 \times (n-m_2)} \end{pmatrix} \\ \text{Einsetzen: } \frac{1}{2} \|F_1 + J_1 \Delta x\|_2^2 &= \frac{1}{2} \|F_1 + J_{11} \Delta y + J_{12} \Delta z\|_2^2 \\ &= \frac{1}{2} \left\| \underbrace{F_1 + J_{11} R_2^{-1} L_2^{-1} P_2 F_2}_{=: b} + \underbrace{(J_{12} - J_{11} R_2^{-1} D_2) \Delta z}_{=: B} \right\|_2^2 \\ &= \frac{1}{2} \|b + B \Delta z\|_2^2 \end{aligned}$$

QR-Zerlegung von  $B$ :

$$\begin{aligned} B &= Q_1 R_1 = (\overline{Q}_1 \quad \hat{Q}_1) \begin{pmatrix} \overline{R}_1 \\ 0 \end{pmatrix} = \overline{Q}_1 \overline{R}_1 \quad (4.15) \\ B^T B \Delta z &= \overline{R}_1^T \overline{Q}_1^T \overline{Q}_1 \overline{R}_1 \Delta z = -B^T b = -\overline{R}_1^T \overline{Q}_1^T b \\ \Delta z &= -\overline{R}_1^{-1} \overline{Q}_1^T b \quad (4.16) \\ \text{Einsetzen: } \Delta y &= -R_2^{-1} (D_2 \Delta z + L_2^{-1} P_1 F_2) \end{aligned}$$

Formale Darstellung:

$$\begin{aligned}
 \begin{pmatrix} P_2 J_2 \\ \dots \\ J_1 \end{pmatrix} &= \begin{pmatrix} P_1 J_{21} & \vdots & P_2 J_{22} \\ \dots & & \dots \\ J_{11} & \vdots & J_{12} \end{pmatrix} \\
 &= \begin{pmatrix} L_2 & \vdots & 0 \\ \dots & & \dots \\ L_1 & \vdots & I \end{pmatrix} \begin{pmatrix} R_2 & \vdots & D_2 \\ \dots & & \dots \\ 0 & \vdots & B \end{pmatrix} \\
 &= \underbrace{\begin{pmatrix} L_2 & \vdots & 0 \\ \dots & & \dots \\ L_1 & \vdots & Q_1 \end{pmatrix}}_{=:T} \begin{pmatrix} R_2 & \vdots & D_2 \\ \dots & & \dots \\ 0 & \vdots & R_1 \end{pmatrix} \quad (4.17) \\
 L_1 &:= J_{11} R_2^{-1} \\
 \Rightarrow J_{12} &= L_1 D_2 + B = L_1 D_2 + Q_1 R_1
 \end{aligned}$$

Rechte Seite entsprechend:

$$- \begin{pmatrix} P_2 F_2 \\ F_1 \end{pmatrix} = T \begin{pmatrix} -L_2^{-1} P_2 F_2 \\ -Q_1^T b \end{pmatrix} \quad (4.18)$$

nachrechnen:

$$\begin{aligned}
 -L_2 L_2^{-1} P_2 F_2 &= -P_2 F_2 \\
 -L_1 L_2^{-1} P_2 F_2 - \underbrace{Q_1 Q_1^T}_{=:I} \underbrace{(F_1 - \overbrace{J_{11} R_2^{-1} L_2^{-1} P_2 F_2}^{=:L_1})}_b &= -F_1
 \end{aligned}$$

Lösen:

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



**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:  $\lambda$  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:

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

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

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

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

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

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

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

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

Die führt auf die kompakte Rekursion:

#### Algorithmus 4.12 (Eliminationsalgorithmus)

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

Für  $i = m-1, \dots, 1$ :

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

Es entstehen folgende Matrix und Vektor:

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

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

**Algorithmus 4.13: Lösen**

- Löse beschränktes Ausgleichsproblem nur in den Variablen  $\Delta s_0$  und  $\Delta p$ :

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

z. B. mit Variante 2.

- Berechne  $\Delta s_1, \dots, \Delta s_{m-1}$  durch:

Für  $i = 1, \dots, m-1$ :

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

**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:

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

Insgesamt:

$$\begin{pmatrix} I & E^1 & 0 & \dots & 0 \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{pmatrix} \dots \begin{pmatrix} I & 0 & \dots & 0 & E^{m-1} \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{pmatrix} (M Z M)$$

$$\begin{aligned}
 &= \begin{pmatrix} I & E^1 & \dots & E^{m-1} \\ & \ddots & & \\ & & \ddots & \\ & & & I \end{pmatrix} (MZM) \\
 &\begin{pmatrix} E^0 & 0 & \dots & \dots & 0 & P^0 \\ G_0 & -I & & & & G_0^p \\ & G_1 & -I & & & G_1^p \\ & & \ddots & \ddots & & \vdots \\ & & & G_{m-2} & -I & G_{m-2}^p \end{pmatrix} \begin{pmatrix} u^0 \\ d_0 \\ d_1 \\ \vdots \\ d_{m-2} \end{pmatrix} \quad (4.24)
 \end{aligned}$$

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

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

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

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

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/Kondensung.



## 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

$F: \mathbb{R}^n \rightarrow \mathbb{R}^m$     $M(x)$ : 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  $\alpha^k$  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.

Sei  $J := \frac{\partial F}{\partial x}$  (5.3) 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  $M(x, \lambda) = \nabla_{x,\lambda}^2 L(x, \lambda)^{-1}$  oder  $M(x, \lambda) \cong \nabla_{x,\lambda}^2 L(x, \lambda)^{-1}$
- 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)

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

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  $x^j \rightarrow x^*$  ( $j \rightarrow \infty$ )
- 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:

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

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:

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

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  $i \rightarrow \infty$ :

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

#### Korollar 5.4

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

Beweis:

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

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

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

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 $\omega$ )

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

- 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  $x^*$

$\omega$  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  $\kappa$** 

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

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

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

$$\|M(y)R(x)\| = \|(M(y) - M(x))R(x)\| \leq \rho L\|y - x\| \\ \kappa = \rho L$$

$\kappa < 1$  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  $\omega$  ein Maß für die Nichtlinearität des Problems und  $\kappa$  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 $\nabla f(x) = 0$

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

$\Delta x^k$  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

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

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  $F_i(x)$ ,  $i = 1, \dots, M$  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),  $\rho$  : Spektralradius.

Beweis:

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

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  $\kappa$ -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:

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

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

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

Spiegel die Messfehler an den geschätzten Modellantworten.

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

$\hat{\eta}$ : 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 gespiegelter Problem:

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

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  $\tau$  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

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

$I + \tau \hat{E}(x^*)$  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

$$\begin{aligned} \min_{x \in \mathbb{R}^n} f(x) \quad & f \in C^2(\mathbb{R}^n, \mathbb{R}), \text{ „Zielfunktion“ (6.1)} \\ g(x) = 0 \quad & g \in C^2(\mathbb{R}^n, \mathbb{R}^{m_1}), \text{ „Gleichungsbedingungen“} \\ h(x) \leq 0 \quad & h \in C^2(\mathbb{R}^n, \mathbb{R}^{m_2}), \text{ „Ungleichungsbedingung“} \end{aligned}$$

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 sein Zielfunktionswert kleiner oder gleich den Zielfunktionswerten aller zulässigen Punkte in einer Umgebung von  $x^*$  ist.

$$g(x^*) = 0 \wedge h(x^*) \leq 0 \wedge \exists \text{ Umgebung } U \text{ von } x^* : \forall x \in U \text{ mit } g(x) = 0, h(x) \leq 0 \text{ ist } f(x^*) \leq f(x) \quad (6.2)$$

### Definition 6.3

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

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

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  $f : (a, b) \subset \mathbb{R} \rightarrow \mathbb{R}$ ,  $f \in C^2$ . 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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f \in C^2$ .  $x^*$  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 (Hinreichende Bedingung)

Sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f \in C^2$ . 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:  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , 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).

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

### Definition 6.8

Ein Punkt  $x^*$  heißt regulär, wenn er die Constraint Qualification (CQ) erfüllt:  $\text{Rg}(g_x(x^*)) = m$  (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 Linearkombination 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  $L: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ ,  $(x, \lambda) \mapsto L(x, \lambda) := f(x) + \lambda^T g(x)$  (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:

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

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

**Satz 6.12 (Notwendige Bedingung erster Ordnung)**

Betrachte das Problem  $\min f(x)$  s. t.  $g(x) = 0$  mit  $f \in C^2(\mathbb{R}^n, \mathbb{R})$ ,  $g \in C^2(\mathbb{R}^n, \mathbb{R}^m)$ ,  $m \leq n$ . 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  $\nabla_\lambda L(x^*, \lambda) = g(x^*) = 0$  (6.13) bzw.  $\nabla L(x, \lambda) = 0$ .

Beweis:

$x^*$  regulär  $\Rightarrow g(x^*) = 0 \Rightarrow$  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.

$x = \varphi(t)$  mit  $\varphi(0) = x^*$ ,  $\varphi(1) = \bar{x} \neq x^*$  zulässiger Punkt  
und  $g(\varphi(t)) \equiv 0$

$$\begin{aligned}\varphi(t) &= \begin{pmatrix} y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ z^* + t(\bar{z} - z^*) \end{pmatrix} \\ 0 &= \left. \frac{\partial}{\partial t} g(\varphi(t)) \right|_{t=0} = g_y(x^*) y'(0) + g_z(x^*) z'(0) \\ \Rightarrow \varphi'(0) &= \begin{pmatrix} y'(0) \\ z'(0) \end{pmatrix} = \begin{pmatrix} -g_y(x^*)^{-1} g_z(x^*) \\ I \end{pmatrix} h\end{aligned}$$

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

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  $\lambda \in \mathbb{R}^m$  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  $\lambda$  positiv semidefinit auf  $T(x^*) = \ker g_x(x^*)$ , d. h.

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

Beweis: Weiter im Beweis von 6.12:

Es gilt:

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

Außerdem gilt:

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

$\varphi'(0)$  ist irgendein Vektor aus  $T(x^*) \Rightarrow \nabla_{xx}^2 L(x^*, \lambda)$  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  $p^T \nabla_{xx}^2 L(x^*, \lambda) p > 0 \forall p \in T(x^*), p \neq 0$ . 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  $x^*$

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

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

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

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

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

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  $\tau$  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:

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

ist regulär weil  $\nabla_x g(x^*, 0)$  vollen Rang hat und  $\nabla_{xx}^2 L(x, \lambda, 0)$  positiv definit auf  $\ker \nabla_x g(x^*, 0)^T$ . Also hängen  $x(\tau), \lambda(\tau)$  stetig differenzierbar von  $\tau$  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

$$\min f(x) \quad f \in C^2(\mathbb{R}^n, \mathbb{R}) \quad (6.18)$$

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

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

Die zulässige Menge von (6.18) ist  $S := \{x : g(x) = 0, h(x) \leq 0\}$ . Sei  $J := \{1, \dots, m_2\}$  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  $\nabla g_i(x)$ ,  $i = 1, \dots, m_1$  linear unabhängig sind.
- ein Vektor  $p \in \mathbb{R}^n$  existiert mit  $\nabla g_i(x)^T p = 0$ ,  $i = 1, \dots, m_1$  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  $\nabla g_i(x)$ ,  $i = 1, \dots, m_1$  und  $\nabla h_i$ ,  $i \in I(x)$  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  $\nabla h_i$ :

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

### 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  $L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x)$ .

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:

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

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 gleichungsbeschränkten Problems. Es folgt aus dem gleichungsbeschränkten Fall die notwendige Bedingung zweiter Ordnung. Komplementarität ist durch die Wahl von  $\mu$  erfüllt:  $\mu^T h(x^*) = 0$ . Zu zeigen bleibt:  $\mu_i \geq 0$ ,  $i \in I(x^*)$ . Siehe folgendes Lemma:

## Lemma 6.24

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

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

$$\begin{pmatrix} \nabla g^T \\ \nabla h_1^T \\ \vdots \\ \nabla \tilde{h}_{i-1}^T \\ \nabla \tilde{h}_i^T \\ \nabla \tilde{h}_{i+1}^T \\ \vdots \\ \nabla \tilde{h}^T \end{pmatrix} (x^*)p = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\varepsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$p$  ist tangential zu allen Gleichungsbedingungen und aktiven Ungleichungsbedingungen außer  $i$ .

$$\begin{aligned} \tilde{h}_i(\varphi(t)) &= \underbrace{\tilde{h}_i(x^*)}_{=0 \text{ aktiv}} + \underbrace{\nabla \tilde{h}_i(x^*)^T p}_{=-\varepsilon \text{ nach Konstr.}} \cdot t + \underbrace{\mathcal{O}(\|p\|^2 t^2)}_{\mathcal{O}(\varepsilon^2)} \\ &\leq 0 \text{ für } \varepsilon > 0 \text{ klein genug} \end{aligned}$$

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

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

## Satz 6.25 (Hinreichende Bedingung)

$(x^*, \lambda, \mu)$  erfülle die notwendige Bedingung erster Ordnung,  $x^*$  sei regulär. Sei  $p^T \nabla_{xx}^2 L(x^*, \lambda, \mu) p > 0 \forall p \in T^+(x^*), p \neq 0$  (6.22) mit  $t^+(x^*) = \{p : \nabla g(x^*)^T p = 0, \nabla h_i(x^*)^T p = 0 \forall i \in I^+(x^*)\}$ ,  $I^+(x^*) = \{i \in I(x^*), \mu_i > 0\}$  (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  $\left. \frac{\partial}{\partial t} f(\varphi(t)) \right|_{t=0} > 0$ , sofern alle  $\mu_i > 0$  sind (analog zu oben). Dazu muss man den Satz für implizite Funktionen anwenden, dafür muss  $\text{Rg}(\nabla g \quad \nabla \tilde{h})^T = m_1 + |I(x^*)|$  in  $x^*$  gelten (Regularität).

## Definition 6.26: Strikte Komplementarität

In einem KKT-Punkt  $(x^*, \lambda, \mu)$ ,  $\mu \geq 0$  gilt strikte 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  $(x^*, \lambda^*, \mu^*)$  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  $\tau$  um 0 ( $\tau = 0$  ist das ursprüngliche Problem). D. h. es existieren Umgebungen  $U$  von  $(x^*, \lambda^*, \mu^*)$  und  $V$  von  $\tau = 0$  und eine stetig differenzierbare Abbildung  $(x, \lambda, \mu): V \rightarrow U$ ,  $\tau \mapsto (x(\tau), \lambda(\tau), \mu(\tau))$  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  $(x, \lambda, \mu)$  in  $\tau = 0$ :

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

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
- $\nabla_{xx}^2 L(x(\tau), \lambda(\tau), \mu(\tau))$  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 Optimalitätsbedingungen 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  $(x, \lambda)$ . 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.

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

## Lemma 7.1

Sei  $\Delta x$  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  $\lambda$

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

Beweis: Lagrangefunktion des QP mit Multiplikator  $u$

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

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

Umgekehrt:

## Lemma 7.2

Falls  $\Delta \lambda$  existiert, so dass  $(\Delta x \quad \Delta \lambda)^T$  die Gleichung (7.6) erfüllt und wenn  $H(x)$  positiv definit auf  $\ker A(x)$ , dann ist  $\Delta x$  Minimum von (7.5).

## Algorithmus 7.3: SQP-Verfahren für gleichungsbeschränkte Probleme

- Startwert  $x^0$ , ggf.  $\lambda^0$ ,  $j := 0$
- Solange ein Abbruchkriterium verletzt ist:
  - Berechne  $\Delta x^j$  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

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

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

## Korollar 7.4

Für die Wahl  $H^j = \nabla_{xx}^2 L(x^j, \lambda^j)$  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  $H^j \cong \nabla_{xx}^2 L(x^j, \lambda^j)$  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

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

mit  $p \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{m \times n}$ ,  $m \leq n$ ,  $\text{Rga} = m$ ,  $H \in \mathbb{R}^{n \times n}$  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  $A^T$ :

$$\begin{aligned}
 A^T &= Q^T \tilde{L}^T = \begin{pmatrix} \underbrace{Q_1^T}_{n \times m} & \underbrace{Q_2^T}_{n \times (m-n)} \end{pmatrix} \begin{pmatrix} L^T \\ 0 \end{pmatrix} = Q_1^T L^T \\
 y &:= Qp = \begin{pmatrix} Q_1 p \\ Q_2 p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
 Ap &= \begin{pmatrix} L & 0 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} p = LQ_1 p \\
 &= Ly_1 = -b \\
 y_1 &= -L^{-1}b
 \end{aligned}$$

$y_2$  zunächst frei.

$$AQ_2^T = LQ_1Q_2^T = 0$$

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  $Q$ :

$$QH Q^T y + QA^T u = -Qg$$

hat zwei Teile:

$$\begin{aligned}
 - A_2 H(Q_1^T y_1 + Q_2^T y_2) + \underbrace{Q_2 A^T}_{=0} u &= -Q_2 g \text{ bzw. } \underbrace{Q_2 H Q_2^T}_{(*)} y_2 = -Q_2 (y - H Q_1^T L^{-1} b) \\
 (*) &: \text{ auf } \ker A \text{ projiziertes } H, \text{ positiv definit, zerlege mit Cholesky } \Rightarrow y_2 \text{ und } p = Q^T y \\
 - Q_1 (Hp + g) + Q_1 A^T u &= 0, Q_1 A^T = L^T \Rightarrow u = -L^{-T} Q_1 (Hp + g).
 \end{aligned}$$

### 7.3: Quasi-Newton-SQP mit Update

Ziel: berechne  $H^k \cong \nabla_{xx}^2 L(x^k, \lambda^k)$  „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)

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

mit

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

### 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  $x^k$

$$\begin{aligned} \nabla_{xx}^2 L(x^{k+1}, \lambda^{k+1})(x^{k+1} - x^k) &= \nabla_x L(x^{k+1}, \lambda^{k+1}) - \nabla_x L(x^k, \lambda^{k+1}) + \mathcal{O}(\|x^{k+1} - x^k\|^2) \\ \nabla_{xx}^2 L(x^{k+1}, \lambda^{k+1}) s^k &= y^k + \mathcal{O}(\|x^{k+1} - x^k\|^2) \end{aligned}$$

### DFP (Davidon, Fletcher, Powell)

$$\begin{aligned} H^{k+1} &= H^k + U_{DFP}(H^k, y^k, s^k) \\ w^k &:= y^k - H^k s^k \\ &= H^k + \frac{w^k (y^k)^T + y^k (w^k)^T}{(y^k)^T s^k} - \frac{(s^k)^T w^k}{(y^k)^T s^k} \frac{y^k (y^k)^T}{(y^k)^T s^k} \quad (7.11) \end{aligned}$$

### Eigenschaften von BFGS und DFP

- BFGS und DFP-Updates erfüllen die Sekantenbedingung.
- Ist  $H^k$  symmetrisch, dann ist auch die BFGS- bzw. DFP-Approximation  $H^{k+1}$  symmetrisch.

- Ist  $(y^k)^T s^k > 0$  (7.12) und  $H^k$  positiv definit, dann ist auch die BFGS- bzw. DFP-Approximation  $H^{k+1}$  positiv definit. Wenn  $(y^k)^T s^k \geq 0$  kann man das Update dämpfen:  $\tilde{H}^{k+1} = (1 - \alpha)H^k + \alpha H^{k+1}$  mit  $\alpha < 1$  geeignet.  $\tilde{H}^{k+1}$  erfüllt dann nicht mehr die Sekantenbedingung!.

## Satz 7.7

Die Inverse des BFGS-Updates ist der DFP-Updates mit  $y^k$  und  $s^k$  vertauscht:

$$\begin{aligned} H^{k+1} &= H^k + U_{BFGS}(H^k, y^k, s^k) \text{ und} \\ B^{k+1} &= B^k + U_{DFP}(B^k, s^k, y^k) \text{ Dann gilt:} \\ \text{Ist } H^k B^k &= I \text{ dann ist auch} \\ H^{k+1} B^{k+1} &= I \end{aligned}$$

In der Praxis kann man die Updates auch direkt auf die Zerlegungen anwenden.

## Resultate zur lokalen Konvergenz von BFGS-Quasi-Newton-SQP-Verfahren

### Satz 7.8: Dennis-Moré im beschränkten Fall

Sei  $x^*$  ein lokales Minimum,  $x^*$  regulär, gelte die hinreichende Bedingung zweiter Ordnung. Seien  $f, g \in C^2$  mit Lipschitz-stetigen zweiten Ableitungen. Im SQP-Verfahren werde die Hessematrix  $\nabla_{xx}^2 L$  durch eine Quasi-Approximation  $H^k$  angenähert. Die Folge  $\{x_k\}$  konvergiere gegen  $x^*$ . Die Konvergenz ist genau dann superlinear, wenn

$$\lim_{k \rightarrow \infty} \frac{\|P^k(H^k - \nabla_{xx}^2 L(x^*))(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} = 0$$

Dabei ist  $P^k$  eine Matrix, die auf dem Nullraum von  $\nabla g^{kT}$  projiziert.

### Bemerkung 7.9

$$P^k = I - \nabla g^k (\nabla g^{kT} \nabla g^k)^{-1} \nabla g^{kT}$$

bzw.  $P^k = Z^k Z^{kT}$ , wenn die Spalten von  $Z^k$  den Nullraum von  $\nabla g^{kT}$  aufspannen, z. B.  $\nabla g^k = (Q_1^T \quad Q_2^T) \begin{pmatrix} L \\ = \end{pmatrix}$ ,  $Z^k = Q_2^T$ .

Nebenrechnung:

$$\begin{aligned}
p &= p_1 + p_2 \\
p_1 &= \nabla g q \\
p_2 &= Q_2^T r \\
p^k p &= Q_2^T Q_2 \\
&= Q_2^T \underbrace{Q_2 \nabla g q}_{=0} + \underbrace{Q_2^T Q_2}_{=I} r \\
&= Q_2^T r = p_2
\end{aligned}$$

### Satz 7.10 (Superlineare Konvergenz von BFGS-SQP-Verfahren)

Gelten die Annahmen wie in Satz 7.8. Seien  $\nabla_{xx}L(x^*)$  und  $H^0$  symmetrisch und positiv definit. Wenn  $\|x^0 - x^*\|$  genügend klein und  $\|\nabla_{xx}^2 L(x^*) - H^0\|$  genügend klein sind erfüllen die BFGS-Approximationen  $H^k$  die Formel (7.12). Die Iterierten konvergieren superlinear gegen  $x^*$ . Dieses Resultat gilt auch für Ungleichungsbeschränkungen.

## 7.4 SQP-Verfahren für NLPs mit Ungleichungsbeschränkungen

$$\min f(x) \text{ s. t. } g(x) = 0 \text{ und } h(x) \leq 0 \quad (7.13)$$

SQP-Verfahren: wie Algorithmus 7.3 aber löse ungleichungsbeschränkte QPs:

$$\begin{aligned}
\min & \frac{1}{2} \Delta x^T H^k \Delta x + \nabla f(x^k)^T \Delta x \quad (7.14) \\
0 &= g(x^k) + \nabla g(x^k)^T \Delta x \\
0 &\geq h(x^k) + \nabla h(x^k)^T \Delta x
\end{aligned}$$

mit Lösung  $\Delta x^k$  und Lagrange-Multiplikatoren  $\mu^k$  und  $v^k \geq 0$ . Iteriere  $x^{k+1}$ ,  $\lambda^{k+1}$  und  $\mu^{k+1} := \mu^k + \alpha^k(v^k - \mu^k)$ . Wenn  $\mu^0 \geq 0$  und  $\alpha \in (0, 1]$  ist  $\mu^k \geq 0$ .

### Lemma 7.11

Sei  $x^*$  Lösung von 7.13. Dann gilt:

$$\begin{aligned}
\min & \frac{1}{2} \Delta x^T H \Delta x + \nabla f(x^*)^T \Delta x \quad (7.15) \\
0 &= g(x^*) + \nabla g(x^*)^T \Delta x \\
0 &\geq h(x^*) + \nabla h(x^*)^T \Delta x
\end{aligned}$$

mit  $H$  positiv definit auf  $T(x^*) = \{p : \nabla g(x^*)^T p = 0, \nabla h_i(x^*)^T p = 0, i \in I(x^*)\}$  hat die Lösung  $\Delta x = 0$  und die Lagrangemultiplikatoren  $u = \lambda^*, v = \mu^*$ . Die aktiven Ungleichungen des QP sind  $I(x^*)$ . Beweis:

$$\begin{aligned} L_{QP} &= \frac{1}{2} \Delta x^T H \Delta x + \nabla f^{*T} \Delta x + u^T (g^* + \nabla g^{*T} \Delta x) + v^T (h^* + \nabla h^{*T} \Delta x) \\ \nabla_{\Delta x} L_{QP} &= H \Delta x + \nabla f^* + \nabla g^* u + \nabla h^* v = 0 \end{aligned}$$

für  $\Delta x = 0, u = \lambda^*, v = \mu^*$  wegen Minimalität von  $x^*$ . Gleichungs- und Ungleichungsbeschränkungen sind erfüllt,  $v \geq 0$ .

## Lemma 7.12

Sei  $\hat{x} \in U_\varepsilon(x^*)$  für geeignetes  $\varepsilon > 0$ . Sei  $H$  positiv definit auf  $T(x^*)$ . Betrachte das QP (7.15) mit  $\hat{x}$  statt  $x^*$ . Für dessen Lösung gilt  $\hat{x} \cong 0, \hat{\mu} \cong \lambda^*, \hat{v} \cong \mu^*$  und die aktiven Ungleichungen sind  $I(x^*)$ . Beweis: Störungssatz 6.27.

## Folgerung 7.13

- Für festes  $H$  werden die aktiven Indizes bestimmt, wenn  $\hat{x} \rightarrow x^*$ .
- Wenn das SQP-Verfahren konvergiert:  $x^k \rightarrow x^*$  und die Folge der  $H^k$  beschränkt bleibt, dann existiert  $k_0$ , so dass  $\forall k \geq k_0$  das QP (7.14) die aktiven Indizes  $I(x^*)$  liefert.

## Lösung von Ungleichungsbeschränkten QP mit einer Active-Set-Methode

QP: Schreibe  $x$  statt  $\Delta x$ , eigentlich Unterproblem des SQP-Verfahrens. Annahme:  $H$  positiv definit.

$$\begin{aligned} \min & \frac{1}{2} x^T H x + g^T x \text{ (quadratische Zielfunktion)} \\ \text{s. t. } & b_i = a_i^T x, i \in E \text{ (Gleichungsbeschränkungen)} \\ & b_i \geq a_i^T x, i \in I \text{ (Ungleichungsbeschränkungen)} \end{aligned}$$

$A(x) := \{i \in E \cup I : a_i^T x = b\}$  Indexmenge alle mit Gleichheit erfüllten Beschränkungen. Wenn man  $A(x^*)$  kennt kann man die Lösung  $x^*$  durch Lösen des Gleichungsbeschränkten QPs



$$\min \frac{1}{2} x^T H x + g^T x \text{ s. t. } b_i = a_i^T x, i \in A(x^*)$$

bestimmen.

## Active-Set-Strategie

hier: primale Variante.

Starte mit einer Schätzung  $W^0$  des active Set, ändere  $W^k$  („Working Set“) sukzessive.  $W^k$  besteht aus allen Gleichungsbeschränkungen und einigen Ungleichungsbeschränkungen, so dass immer die  $a_i^T$ ,  $i \in W^k$  linear unabhängig sind.

### Algorithmus 7.14: Active-Set-Methode für konvexe QP

- Starte mit einem zulässigen Startpunkt  $x^0$ . Sei  $W^0$  eine Teilmenge von  $A(x^*)$  mit linear unabhängigen  $a_i^T$  und sei  $k := 0$ .

Iteriere  $k = 0, 1, 2, \dots$

- Bestimme Schritt  $p^k$  durch Lösung von  $\min \frac{1}{2} p^T H p + c^k p$  so dass  $0 = a_i^T p$ ,  $i \in W^k$  mit  $c^k = H x^k + g$ .  $p^k$  erfüllt  $a_i^T (x^k + \alpha p^k) = a_i^T x^k = b_i \forall i \in W^k$ , d. h. Schritte in Richtung  $p^k$  bleiben zulässig  $\forall i \in W^k \forall \alpha \in \mathbb{R}$ .
- Wenn  $p^k = 0$ : Berechne Lagrangemultiplikatoren  $\lambda^k$  für (7.17). Diese erfüllen

$$\begin{aligned} 0 &= H p^k + x^k + \sum_{i \in W^k} \lambda_i^k a_i \\ \Rightarrow - \sum_{i \in W^k} \lambda_i^k a_i &= c^k = H x^k + g \end{aligned}$$

- Wenn  $\lambda_i^k \geq 0 \forall i \in W^k \cap I$  und  $\lambda_i^k := 0$  für  $i \notin W^k$ , dann sind alle Optimalitätsbedingungen des QP (7.16) erfüllt:

$$\begin{aligned} H x^k + g + \sum_{i \in E \cup I} \lambda_i^k a_i &= 0 \\ a_i^T x^k &= b_i \quad i \in E \\ \lambda_i^k &\geq 0 \quad \forall i \in I \\ a_i^T x^k &= b_i \quad i \in I \end{aligned}$$

$x^* := x^k$  ist Lösung, Stop.

- Wenn  $\lambda_i^k < 0$  für ein  $j \in W^k$ , dann kann man durch Weglassen dieser Nebenbedingung den Wert der Zielfunktion verkleinern. Ändere den Working set  $W^{k+1} := W^k \setminus \{j_0\}$  mit  $j_0$ : ein Index des kleinsten  $\lambda_i^k < 0$ . Dann gilt: Im nächsten Schritt bleibt die Nebenbedingung  $j_0$  zulässig:  $x^{k+1} = x^k$ ,  $p^{k+1}$ ,  $\lambda^{k+1}$  erfüllt

$$\begin{aligned} - \sum_{i \in W^{k+1}} \lambda_i^{k+1} a_i &= H p^{k+1} + c^{k+1} \\ &= H p^{k+1} + H x^k + g \\ \text{und } a_i^T p^{k+1} &= 0 \quad \forall i \in W^{k+1} \end{aligned}$$

$$\text{Ergibt: } - \sum_{i \in W^{k+1}} (\lambda_i^{k+1} - \lambda_i^k) a_i + \lambda_{j_0}^k a_{j_0} = H p^{k+1}$$

$$\begin{aligned} \lambda_{j_0}^k p^{k+1 T} a_{j_0} &= p^{k+1 T} H p^{k+1} \geq 0 \\ \text{und somit } a_{j_0}^T (x^{k+1} + \alpha p^{k+1}) &= a_{j_0}^T x^k + \alpha a_{j_0}^T p^{k+1} \leq b_{j_0} \quad \forall \alpha \geq 0 \end{aligned}$$

- Wenn  $p^k \neq 0$ : Bestimme Schrittweite  $\alpha^k \in [0, 1]$ , so dass für  $x^{k+1} = x^k + \alpha^k p^k$  alle Nebenbedingungen, auch  $i \notin W^k$  erfüllt sind. Falls  $a_i^T p^k \leq 0$  für  $i \notin W^k$  dann ist  $a_i^T (x^k + \alpha p^k) = a_i^T x^k + \alpha a_i^T p^k \leq b_i \quad \forall \alpha \geq 0$ . Wenn  $a_i^T p^k > 0$  muss  $\alpha \leq \frac{b_i - a_i^T x^k}{a_i^T p^k}$

$$\text{erfüllt sein. Wähle } \alpha^k := \min \left\{ 1, \underbrace{\min_{i \notin W^k, a_i^T p^k > 0} \frac{b_i - a_i^T x^k}{a_i^T p^k}}_{(*)} \right\} \quad (*): \text{Nebenbedingungen,}$$

die das erfüllen heißen blockierende Nebenbedingungen.  $\alpha^k$  kann auch 0 sein. Wenn es blockierende Nebenbedingungen gibt, bilde  $W^{k+1} := W^k \cup \{i_0\}$  mit einer blockierenden NB  $i_0$ . Ansonsten setze  $W^{k+1} = W^k$ . Iteriere  $X^{k+1} := x^k + \alpha^k p^k$ . Dabe gilt: Die Zielfunktion nimmt in Richtung  $p^k$  strikt ab: Weil  $p^k$  minimal ist für (7.17) gilt:

$$\frac{1}{2} p^{k T} H p^k + c^{k T} p^k < \frac{1}{2} 0^T H 0 + c^{k T} 0 = 0$$

also

$$\begin{aligned} \frac{1}{2} (x^k + \alpha p^k)^T H (x^k + \alpha p^k) + g^T (x^k + \alpha p^k) \\ &= \frac{1}{2} x^{k T} H x^k + g^T x^k + \underbrace{\alpha x^{k T} p^k}_{<0} + \frac{1}{2} \alpha^2 \underbrace{p^{k T} H p^k}_{>0} \\ &< \frac{1}{2} x^{k T} H x^k + g^T x^k \end{aligned}$$

für  $\alpha$  klein genug. ggf. verkleinere  $\alpha$  durch Linesearch

### **Bemerkung 7.15**

In jeder Iteration bleiben alle Nebenbedingungen erfüllt: Die, die in  $W^k$  bleiben, siehe Schritt 1. Die, die nicht in  $W^k$  waren siehe Schritt 5. Die, die aus  $W^k$  entfernt wurden siehe Schritt 4.

### **Satz 7.16**

Unter bestimmten Voraussetzungen terminiert Algorithmus 7.14 in endlich vielen Iterationen.

Alternative:

### **Innere-Punkt-Methode**

Phase 1: Löse ein Hilfs-QP (LP) zur Bestimmung eines zulässigen  $x^0$ .



## 8. Optimale Steuerung

### Problemformulierung

#### Beispiel: Fahrdynamik

Bringe Fahrzeug (Bahn, Auto) von A nach B. Fahrzeug soll am Anfang und Ende in Ruhe sein. Minimiere dabei die Zeit oder den Energieverbrauch oder maximiere die Reichweite etc.

Modell: Gewöhnliche Differentialgleichung:

$$\ddot{s} = \frac{\partial^2}{\partial t^2} s = \frac{1}{M} f(s, \dot{s}, u)$$

$s: [0, T] \rightarrow \mathbb{R}^n$  „Ort“  
 $v = \dot{s}: [0, T] \rightarrow \mathbb{R}^n$  „Geschwindigkeit“  
 $\ddot{s}: [0, T] \rightarrow \mathbb{R}^n$  „Beschleunigung“  
 $M$  : Masse  
 $f$  : Kräfte, hängen ab von

- Ort  $s$ , z. B. Gefälle
- Geschwindigkeit (z. B. Reibung)
- Steuerung  $u$ , z. B. Gaspedal, Bremse, Lenkrad

Steuerungen sind in Grenzen frei wählbar.

Spezialfall: Raketenwagen

$$\ddot{s} = u$$
$$\underline{u} \leq u(t) \leq \bar{u} \quad \text{Steuerungsbeschränkung}$$

Randbedingungen:

$$\begin{aligned}s(0) &= A \\ s(T) &= B \\ \dot{s}(0) &= v(0) = 0 \\ \dot{s}(T) &= v(T) = 0\end{aligned}$$

Transformation auf System erster Ordnung:

$$\begin{aligned}\dot{v} &= u \\ \dot{s} &= v\end{aligned}$$

Zeitintervall  $[0, T]$ . Die Endzeit  $T$  kann fest oder frei sein.

Zustandsbeschränkung:  $v_{\min} \leq v(t) \leq v_{\max}$ . Erste Frage: Gibt es zulässige Lösungen, also Steuerungen  $u$  und Zustände  $(s, v)$  so dass die Differentialgleichungen, Randbedingungen, Randbedingungen, Steuerungsbeschränkungen und Zustandsbeschränkungen erfüllt sind? Sinnvolles Setting im Beispiel:  $\underline{u} < 0 < \bar{u}$ .  $T$  muss ausreichend groß sein.  $v_{\min} \leq 0$ , zumindest lokal im Ort,  $v_{\max} > 0$ . Wenn  $T$  genügend groß ist gibt es im Allgemeinen mehrere Lösungen.

Wenn  $T$  so klein wie möglich gewählt wird, gibt es in der Regel nur eine zulässige Lösung.

Zweite Frage: Es gebe zulässige Lösungen, suche optimale Lösung.

Mögliche Kriterien:

- zeitoptimal:  $\min T = \int_0^T 1 \, dt$
- energieoptimal:  $\min \frac{1}{2} \int_0^T u(t)^2 \, dt$
- reichweitenoptimal:  $\max s(T)$ , lasse Randbedingungen  $s(t) = B$ ,  $v(T) = 0$  weg, zusätzliche Nebenbedingung:  $\frac{1}{2} \int_0^T u(t)^2 \, dt \leq E_{\max}$ .
- Kombinationen dieser Kriterien:  $\min w_1 T + w_2 \frac{1}{2} \int_0^T u(t)^2 \, dt$ .

## Allgemeine Problemformulierung

### Modellgleichung

$$\dot{y}(t) = f(t, y(t), q, u(t)) \quad (8.1)$$

$y: [t_0, t_{\text{end}}] \rightarrow \mathbb{R}^{n_y}$  : Zustände

$q \in \mathbb{R}^{n_q}$  : Steuergrößen  $u: [t_0, t_{\text{end}}] \rightarrow \mathbb{R}^{n_u}$  Steuerfunktionen, Steuerungen

Bemerkung: Steuerungen und Steuergrößen können vom Experimentator gewählt werden.  $t_0$  und  $t_{\text{end}}$  können fest oder frei sein.

## Randbedingungen

$$y(t_0) = y_0 \quad \text{Anfangsbedingung}$$

$$r(t_0, y(t_0), t_1, y(t_1), \dots, t_{end}, y(t_{end}), q) = 0 \quad \text{oder} \quad \geq 0$$

## Steuerungsbeschränkungen

$$u(t) \in \Omega(t) \subset \mathbb{R}^{n_u}$$

$$\Omega(t) : \text{Steuerungsbereich}$$

$u$  ist messbar.

Beispiele:

- keine Beschränkungen
- $\Omega$  Polyeder,  $\Omega = \{u : a_j^T u \leq b_j, j = 1, \dots, m\}$
- $\Omega$  Quader,  $\Omega = \prod_{i=1}^{n_u} [\underline{u}_i, \bar{u}_i]$
- $\Omega$  diskret,  $\Omega = \{1, 2, 3, 4, R\}$  (Gangschaltung o. ä.).
- Kombinationen davon

## Zustandsbeschränkungen

$$b(t, y(t), q, u(t)) = 0 \quad \text{oder} \quad \geq 0 \quad \forall t \in [t_0, t_{end}]$$

## Zielfunktional

- Mayersche Form:

$$\min_{u, q, y, t_{end}} \Phi(y(t_{end}), t_{end}, q)$$

- Lagrange Form:

$$\min_{u, q, y, t_{end}} \int_{t_0}^{t_{end}} L(t, y(t), q, u(t)) dt$$

- Bolza-Form:

$$\min_{u, q, y, t_{end}} \Phi(y(t_{end}), t_{end}, q) + \int_{t_0}^{t_{end}} L(t, y(t), q, u(t)) dt$$

### Bemerkung 8.1

Statt einer ODE kann auch eine DAE betrachtet werden. Dann sind die Konsistenzbedingungen zu berücksichtigen, siehe oben.

### Bemerkung 8.2

Da die Steuerfunktion im Allgemeinen beliebige messbare Funktionen sein können, ergeben Punktauswertungen von  $u$  keinen Sinn.  $u(t)$  geht deshalb nicht in die Randbedingung und nicht in den Meyer-ZF-Term ein.

### Bemerkung 8.3

Die drei Standardformen des Zielfunktional sind äquivalent und können ineinander übergeführt werden, siehe Übungsaufgabe.

### Bemerkung 8.4

Probleme mit freier Endzeit können in Probleme mit fester Endzeit transformiert werden:  
Definiere normierte Zeit

$$\begin{aligned}\tau &\in [0, 1] \\ t &= t_0 + \tau(t_{\text{end}} - t_0) = \varphi(\tau) \\ dt &= (t_{\text{end}} - t_0) d\tau =: \hat{t} d\tau \\ \text{also } \frac{\partial y}{\partial \tau} &= t_{\text{end}} - t_0 = \hat{t} \\ \text{Dann ist } y(t) &= y(\varphi(\tau)) =: \bar{y}(\tau) \\ u(t) &= u(\varphi(\tau)) =: \bar{u}(\tau) \\ \frac{\partial \bar{y}}{\partial \tau} &= \frac{\partial y}{\partial t} \partial \varphi \partial \tau = f(\varphi(\tau), y(\varphi(\tau)), q, u(\varphi(\tau))) - \hat{t} \\ &=: \bar{f}(\tau, \hat{y}(\tau), \hat{q}, \hat{u}(\tau)) \\ \text{mit } \bar{q} &= (q, \hat{t})\end{aligned}$$

Die Intervalllänge  $\hat{t}$  ist zu einer zusätzlichen Steuergröße geworden. Das so transformierte DGL-System hat das feste Zeitintervall  $\tau \in [0, 1]$ .  $\hat{x}$ ,  $\hat{u}$  setzt man in Zielfunktion, Randbedingungen, Steuerungsbeschränkungen und Zustandsbeschränkungen ein.

## 8.3 Ansätze zur Lösung von Optimalsteuerungsproblemen

### First Optimize then Discretize

Formuliere die Optimalitätsbedingungen des Optimalsteuerungsproblems im Funktionenraum. Hier: Pontryaginsches Maximumsprinzip. Erhalte Randwertproblem. „Rate“



Struktur der optimalen Lösung. Verifiziere Optimalität. Vorteile:

- Rigorose mathematische Untersuchung der optimalen Lösung.

Nachteile:

- Für komplexe Probleme schwierig zu formulieren.
- Struktur der Lösung muss erraten werden.

Siehe Kapitel 10: Indirekter Ansatz.

### **First Discretize then Optimize**

Parametrisiere die Steuerfunktion in einem endlichdimensionalen Unterraum. Erhalte endlichdimensionales, ggf. hochdimensionales aber strukturiertes Optimierungsproblem. Löse diese mit maßgeschneiderten NLP-Methoden.

Nachteile:

- Es werden nur suboptimale Lösungen berechnet.
- Strukturelle Aussagen sind schwer möglich.

Vorteile:

- In der Praxis werden Steuerungen oft genau so realisiert, wie sie hier parametrisiert werden.
- Die Methode kann sehr flexibel und robust
- Zur Simulation der Modellgleichungen, Auswertung der Ableitungen und zur nichtlinearen Optimierung können maßgeschneiderte numerische State-of-the-art-Methoden eingesetzt werden, siehe Kapitel 9: Direkter Ansatz.



## 9. Direkter Ansatz der optimalen Steuerung

Betrachte o. B. d. A. Probleme mit fester Anfangs- und Endzeit und Lagrange-Zielfunktional.

Schwierigkeit:  $u: [t_0, t_{end}] \rightarrow \mathbb{R}^{n_u}$  ist Element eines unendlichdimensionalen Funktionenraumes. Idee: Drücke  $u$  durch endlich viele Variablen  $q_i$  aus, zum Beispiel durch lokale Ansatzfunktionen: Wähle Gitter:  $t_0 = \tau_0 < \tau_1 < \dots < \tau_m = t_{end}$ . Ersetze  $u(t)$  auf  $t \in [\tau_i, \tau_{i+1})$  durch eine lokale Ansatzfunktion  $u(t) = \varphi_i(t, q_i)$ ,  $t \in [\tau_i, \tau_{i+1})$ ,  $i = 0, \dots, m-1$  mit endlich vielen Parametrisierungsvariablen  $q_i \in \mathbb{R}^{n_{q_i}}$ .

Typische Beispiele: Stückweise Polynome, stückweise linear, stückweise linear und stetig.

Stetigkeitsbedingung:  $\varphi_i(t_{i+1}, q_i) = \varphi_{i+1}(t_{i+1}, q_{i+1})$

### Bemerkung 9.1

Stückweise lineare stetige Steuerfunktionen können auch durch Integration von stückweise konstanten Steuerfunktionen gebildet werden. Zusätzliche DGL:  $\dot{y}_{n+1} = u$  hat als Lösung eine stückweise lineare Funktion.

### Bemerkung 9.2

- Verschiedene Komponenten von  $u$  können verschieden behandelt werden.
- Wenn man Ansatzfunktionen mit lokalem Träger (insbesondere keine Spline-Funktionen) verwendet, dann werden die Größen  $q_i$  nur auf dem Intervall  $[\tau_i, \tau_{i+1})$  vorkommen  $\Rightarrow$  Entkopplung  $\Rightarrow$  Blockstruktur der Hessematrix, siehe unten.
- Prinzipiell können im direkten Ansatz aber beliebige Parametrisierungen mit endlich vielen Freiheitsgraden verwendet werden.

Damit haben wir endlich viele Variablen erhalten:

- $q$ : Zeitunabhängige Steuerungen
- $q_0, \dots, q_{m-1}$  Parametrisierungsvariablen auf  $[\tau_i, \tau_{i+1})$

Fasse zusammen:  $\bar{q} := (q, q_0, \dots, q_{m-1}) \in \mathbb{R}^{n_q}$ .

## Umformulierung der Steuerungsbeschränkungen

Beschränkungen an  $u(t)$  gehen über in Beschränkungen an  $q_i$ , z. B.  $u_l \leq u(t) \leq u_u$  geht über in  $q_l \leq q_i \leq u_u$ ,  $i = 0, \dots, m-1$  bei stückweise konstanter Parametrisierung. Bei stückweise linearer Parametrisierung: am Anfang und Ende der Intervalle:  $u_l \leq q_{i0} \leq u_u$ ,  $u_l \leq q_{i0} + (\tau_{i+1} - \tau_i)q_{i1} \leq u_u$ .

Auch die Zustandsbeschränkungen  $b(t, y(t), q, u(t)) \geq 0$  sind unendlichdimensional. Behandlung: Punktweise Auswertung auf einem Gitter  $t_0 \leq t_1 < \dots < t_k \leq t_{end}$ ,  $b(t_j, y(t_j), q, \varphi_{i(j)}(t_j, q_{i(j)})) \geq 0$ ,  $j = 1, \dots, k$ ,  $i(j)$  so, dass  $t_j \in [\tau_{i(j)}, \tau_{i(j)+1})$ . Alternative: Penalty-Term in der Zielfunktion, kontinuierliche Behandlung.

Insgesamt erhalten wir:

$$\begin{aligned} \min_{q, \bar{q}} \sum_{i=0}^{m-1} \int_{\tau_i}^{\tau_{i+1}} L(t, y(t), q, \varphi_i(t, q_i)) dt &= \sum_{i=0}^{m-1} \int_{\tau_i}^{\tau_{i+1}} \bar{L}(t, y(t), \bar{q}) dt \quad (9.3) \\ \text{s. t. } \dot{y} &= f(t, y(t), q, \varphi_i(t, q_i)) \quad t \in [\tau_i, \tau_{i+1}), \quad i = 0, \dots, m-1 \\ &= \bar{f}(t, y(t), \bar{q}) \end{aligned}$$

und Randbedingungen:  $y(t_0) = y_0(q)$

$r(t_0, y(t_0), \dots, t_{end}, y(t_{end}), q) = 0$

Punktweise Bedingungen:  $b(t_j, y(t_j), q, \varphi_{i(j)}(t_j, q_{i(j)})) \quad j = 1, \dots, k$   
 $= \bar{b}(t_j, y(t_j), \bar{q}) \geq 0$

Steuerungsbeschränkungen:  $q_l \leq \bar{q} \leq q_u$

$b_l \leq A\bar{q} \leq b_u$  mit geeigneten  $q_l, q_u, b_l, b_u, A$

zu tun bleibt: Parametrisierung der Zustandsvariablen.

Beispiel: Bimolekulare Katalyse, siehe auch Kapitel 2.



Molzahlen

$$n_1, n_2, n_3, t \in [0, 10]$$

$$k = k_1 + k_2$$

$$= f_1 \exp\left(-\frac{E_1}{RT}\right) + c_{kat} \exp(-\lambda t) f_2 \exp\left(-\frac{E_2}{RT}\right)$$

$$\dot{n}_1 = -V k \frac{n_1}{V} \frac{n_2}{V}, n_1(0) = n_{10}$$

$$\dot{n}_2 = -V k \frac{n_1}{V} \frac{n_2}{V}, n_2(0) = n_{20}$$

$$\dot{n}_3 = V k \frac{n_1}{V} \frac{n_2}{V}, n_3(0) = 0$$

Der Prozessverlauf wird bestimmt durch:

- Parameter  $f_1, E_1, f_2, E_2, \lambda$ , jetzt Konstanten.
- Steuergrößen:

$$q = \begin{cases} V : & \text{Volumen} \\ n_{10} : & \text{Anfangsmolzahl von A} \\ n_{20} : & \text{Anfangsmolzahl von B} \\ c_{kat} : & \text{Katalysatorkonzentration} \end{cases}$$

Mögliche Zielfunktion: Maximiere die Molzahl des Reaktionsproduktes:  $\min -n_3(t_{end})$  (Meyer-Zielfunktion). Parametrisierung der Steuerfunktion: Stückweise linear und stetig auf 3 Intervallen:

$$\begin{aligned} t \in [0, 2] : T(t) &= q_5 + tq_6 \\ t \in [2, 8] : T(t) &= q_7 + (t - 2)q_8 \\ t \in [8, 10] : T(t) &= q_9 + (t - 8)q_{10} \end{aligned}$$

Nebenbedingungen:

$$\begin{aligned} q_6 &= 0 \\ q_7 &= q_5 \\ q_7 + 6q_8 &= q_9 \\ q_{10} &= 0 \end{aligned}$$

Grenzen:

$$\begin{aligned} q_5 &\in [293, 393] \\ q_9 &\in [293, 393] \end{aligned}$$

## 9.2 Direktes Single-Shooting

Wähle Werte für die Anfangswerte  $y(t_0) = s_0$  (oder konstant oder  $q_j$ ) und für die Steuerungen  $\bar{q} = (q, q_0, \dots, q_{m-1})$ . Löse Anfangswertproblem  $\dot{y} = f(t, y(t), q, \varphi_0(t, q_0))$ ,  $t \in [\tau_0, \tau_1)$ ,  $y(t_0) = s_0$ .  $\dot{e} = L(t, y(t), q, \varphi_0(t, q_0))$ ,  $t \in [\tau_0, \tau_1)$ ,  $e(t_0) = 0$  (9.4) und sukzessive  $\dot{y} = f(t, y, q, \varphi_i(t, q_i))$ ,  $\dot{e} = L(t, y, q, \varphi_i(t, q_i))$ ,  $t \in [\tau_i, \tau_{i+1})$ . Liefert Lösung  $y(t, s_0, q, q_0, q_1, \dots, q_{m-1})$  und Zielfunktionswert  $e(t, s_0, q, q_0, q_1, \dots, q_{m-1})$ . Interessant ist nur  $e(t_{end}, s_0, q, q_0, \dots, q_{m-1})$

Bemerkung:  $y(t, s_0, q, q_0, q_{m-1})$  hängt nur von den  $q_i$ ,  $i = 0, \dots, j$  ab für  $t \leq \tau_{j+1}$ . Damit ergibt sich ein endlichdimensionales Optimierungsproblem in den Variablen  $X = (s_0, q, q_0, \dots, q_{m-1})$  (9.3)

$$\begin{aligned} \min_x F_1(x) &= e(t_{end}, x) \\ \text{s. t. } F_2(x) &= r(t_0, y(t_0, x), \dots, t_{end}, y(t_{end}, x), q) =: \bar{r}(x) = 0, \quad j = 1, \dots, k \\ F_3(x) &\geq 0 \quad \text{umfasst alle Ungleichungen: } \bar{b}(t_j, y(t_j, x), x) \geq 0, \quad q_l \leq \bar{q} \leq q_u, \\ b_l &\leq A\bar{q} \leq b_u \end{aligned}$$

Zur Auswertung von Zielfunktion und Nebenbedingungen von (9.6) sind jeweils die Anfangswertprobleme (9.4) zu lösen. Das Optimierungsproblem:

$$\min_x F_1(x) \text{ s. t. } F_2(x) = 0 \text{ s. t. } F_3(x) \geq 0$$

ist ein NLP und kann mit NLP-Methoden gelöst werden, am besten mit SQP-Verfahren, siehe Kapitel 7.

### 9.3 Direktes Multiple Shooting

Problem beim Single-Shooting: Mit schlechten Startwerten starten wir weit weg von der Lösung und können Schwierigkeiten bei der Konvergenz des SQP-Verfahrens bekommen. Wenn wir über die Steuergrößen  $(q, q_0, \dots, q_{m-1})$  keine Informationen haben, bekommen wir ggf. Trajektorien  $y(t)$ , die sehr weit von der optimalen Lösung entfernt sind oder evtl. nicht auf dem ganzen Intervall  $[t_0, t_{end}]$  existieren. Aber: Der ungefähre Verlauf von  $y(t)$  ist oft durchaus bekannt.

### Mehrzielmethode / Multiple Shooting

- Wähle ein Gitter der Mehrzielknoten  $t_0 = \tau_0 < \tau_1 < \dots < \tau_m = t_{end}$  (9.7). Zur Vereinfachung: gleiches Gitter wie für Parametrisierung der Steuerungen, aber das ist nicht notwendig.
- Wähle Werte für die Anfangswerte der Zustände in den Mehrzielknoten:  $y(\tau_i) = s_i$ ,  $i = 0, \dots, m - 1$ . Hier: Möglichkeit, Vorinformationen einzubringen.
- Wähle Werte für die parametrisierten Steuerungen  $q_i$ ,  $i = 0, \dots, m - 1$ . o. B. d. A. diskutieren wir die nicht zeitabhängigen Steuergrößen  $q$  weg. Dafür gibt es zwei Möglichkeiten:
  - Formuliere sie als stückweise konstant und stetig parametrisierte Steuerfunktion.

- Formuliere zusätzliche Zustände.

$$\begin{pmatrix} y_{n_y+1} \\ \vdots \\ y_{n_y+n_q} \end{pmatrix}$$

mit  $\dot{y}_{n_y+j} \equiv 0$ .

Der Integrator sollte diese Zustände nicht mitintegrieren und keine VDE dafür lösen.

- Variablen sind also  $X = (s_0, q_0, s_1, q_1, \dots, s_{m-1}, q_{m-1})$  (9.8)
- Löse Anfangswertproblem

$$\begin{aligned} \dot{y} &= f(t, y, \varphi_i(t, q_i)) = \bar{f}(t, y, q_i) & (9.9) \quad y(\tau_i) &= s_1 \\ \dot{e} &= L(t, y, \varphi_i(t, q_i)) = \bar{L}(t, y, q_i) \\ e(\tau_i) &= 0 \quad i = 0, 1, \dots, m-1 \end{aligned}$$

Es wird eine unstetige Trajektorie, bestehend auf Teiltrajektorien  $y(t, \tau_i, s_i, q_i)$ ,  $t \in [\tau_i, \tau_{i+1})$ .

- Die Gesamttrajektorie muss stetig gemacht werden durch Anschlussbedingungen:

$$h_i(s_i, q_i, s_{i+1}) := y(\tau_{i+1}, \tau_i, s_i, q_i) - s_{i+1} = 0 \quad i = 0, \dots, m-2$$

Anfangsbedingung bei  $t_0 = \tau_0$ :  $y(t_0) = s_0$ . Das sind insgesamt  $((m-1) + 1) \cdot n_y = m \cdot n_y$  Bedingungen für die  $m \cdot n_y$  Zustandsvariablen.

- Zielfunktional:

$$\begin{aligned} \min \sum_{i=0}^{m-1} e(\tau_{i+1}, \tau_i, 0, q_i) & \quad (9.11) \\ &= F_1(X) \end{aligned}$$

- Nebenbedingungen: Setze  $y(t, \tau_i, s_i, q_i)$  ein in Randbedingungen:

$$r(\tau_0, s_0, \tau_1, s_1, \dots, \tau_{m-1}, s_{m-1}, t_{end}, y(t_{end}, \tau_{m-1}, s_{m-1}, q_i)) = 0$$

Besser: separierbare Randbedingungen:

$$\sum_{i=0}^{m-1} r_i(\tau_i, s_i) + r_m(t_{end}, y(t_{end}, \tau_{m-1}, s_{m-1}, q_{m-1})) = 0 \quad (9.12)$$

verwende diese Variante

– In Zustandsbeschränkung:

$$b(\tau_i, s_i, q_i) \geq 0, \quad i = 0, \dots, m-1$$

Zur Vereinfachung: gleiches Gitter (nicht notwendig).

Weitere Nebenbedingungen: Anschlussbedingungen, Steuerungsbeschränkungen

$$\begin{aligned} q_{lo} &\leq q \leq q_{up} \\ v_{lo} &\leq Aq \leq b_{up} \end{aligned}$$

Zur Auswertung von Zielfunktion und Nebenbedingungen werden als Anfangswertprobleme auf den Mehrzielintervallen gelöst:  $t \in [\tau_i, \tau_{i+1})$ . Diese Integrationen sind entkoppelt und können unabhängig voneinander parallel durchgeführt werden, dadurch kann das Problem auf modernen Mehrprozessor-Rechnern schneller als mit Single-Shooting gelöst werden. Zielfunktion und Nebenbedingungen sind separierbar, sie bestehen aus Beiträgen der einzelnen Mehrzielintervalle, die höchstens linear (hier als Summanden) gekoppelt sind.

Zielfunktion:

$$\sum_{i=0}^{m-1} e(\tau_{i+1}, \tau_i, s_i, q_i)$$

Anschlussbedingungen:

$$y(\tau_{i+1}, \tau_i, s_i, q_i) - s_{i+1} = 0$$

Randbedingungen:

$$\sum_{i=0}^{m-1} r_i(\tau_i, s_i) + r_m(t_{end}, y(t_{end}, \tau_{m-1}, s_{m-1}, q_{m-1})) = 0$$

Resultierendes Optimierungsproblem:

$$\begin{aligned} \min_x F_1(x) \quad & (9.13) \\ F_2(x) = 0 \quad & \text{Anschlussbedingungen, Randbedingungen} \\ F_3(x) \geq 0 \quad & \text{Zustandsbeschränkungen, Steuerungsbeschränkungen} \end{aligned}$$

Das Optimierungsproblem (9.13) hat die gleiche Lösung wie das aus der Single-Shooting-parametrisierung hervorgehende (9.6). Einerseits hat (9.13) viel mehr Variablen:  $mn_y + mn_u f$  gegenüber  $n_y + mn_u f$ , andererseits hat es auch viel mehr Struktur:

- 1): Die Jacobimatrix  $A$  der Nebenbedingungen ist dünn besetzt.



- 2): Die Hesse-Matrix der Lagrange-Funktion ist Block-diagonal.

Zu 1):

Anschlussbedingungen:

$$\begin{aligned}\frac{\partial h_i}{\partial x} &= \left(0, \dots, 0, \frac{\partial h_i}{\partial s_i}, \frac{\partial h_i}{\partial q_i}, \frac{\partial h_i}{\partial s_{i+1}}, 0, \dots, 0\right) \\ &= \left(0, \dots, 0, \frac{\partial y}{\partial s_i}(\tau_{i+1}, \tau_i, s_i, q_i), \frac{\partial y}{\partial q_i}(\tau_{i+1}, \tau_i, s_i, q_i), -I, 0, \dots, 0\right) \\ &= (0, \dots, 0, G_i, \Gamma_i, -I, 0, \dots, 0)\end{aligned}$$

Randbedingungen:

$$\frac{\partial r}{\partial x} = \left(\frac{\partial r_0}{\partial s_0}, 0, \frac{\partial r_1}{\partial s_1}, 0, \dots, \frac{\partial r_{m-1}}{\partial s_{m-1}} + \frac{\partial r_m}{\partial y} \frac{\partial y}{\partial s_{m-1}}(t_{end}, \tau_{m-1}, s_{m-1}, q_{m-1}), \frac{\partial r_m}{\partial y} \frac{\partial y}{\partial q_{m-1}}(t_{end}, \tau_{m-1}, s_{m-1}, q_{m-1})\right)$$

$\nabla F_3(x)^T$  : Zustandsbeschränkungen:

$$\frac{\partial}{\partial x} b_i = \left(0, \dots, 0, \frac{\partial b_i}{\partial s_i}, \frac{\partial b_i}{\partial q_i}, 0, \dots, 0\right)$$

Zu 2):

$$\begin{aligned}\mathcal{L}(x, \lambda) &= \sum_{i=0}^{m-1} e(\tau_{i+1}, \tau_i, 0, q_i) + \sum_{i=0}^{m-1} \lambda_{Ai}^T h_i(s_i, q_i, s_{i+1}) + \lambda_R^T \left(\sum_{i=0}^{m-1} r_i + r_m\right) \\ &+ \sum_{i=0}^{m-1} \lambda_{Bi}^T b_i(\tau_i, s_i, q_i) + \lambda_s^T (\text{Steuerbeschr.})\end{aligned}$$

Offenbar:

$$\frac{\partial^2 L}{\partial \begin{pmatrix} s_i \\ q_i \end{pmatrix} \partial \begin{pmatrix} s_j \\ q_j \end{pmatrix}} = 0 \quad \text{für } i \neq j$$

$\Rightarrow$  Blockdiagonalität.

Wir berechnen  $H$  durch Low-Rank-Updates, z. B. BFGS, siehe Kapitel 7. Statt ganz  $H$  aufzudatieren Rang 2 pro Schritt, kann man auch einzelne Blöcke mit BFGS aufdatieren, Rang 2 pro Block, Rang  $2 \cdot m$  insgesamt. Die Hessematrix wird viel besser approximiert, dadurch verbessert sich das Konvergenzverhalten des SQP-Verfahrens drastisch, siehe Satz von Dennis-More.

Bei der QP-Lösung können/müssen die Strukturen in der Mehrzielmatrix ausgewertet werden:

$$\begin{pmatrix} H & A^t \\ A & 0 \end{pmatrix}$$

### Kondensieralgorithmus (Plitt 1981, Bock, Plitt 1984)

wird alle  $\Delta s_i, i = 1, \dots, m-1$  aus QP raus

$$\Delta s_{i+1} = G_i s_i + \Gamma_i \Delta q_i + h_i \quad \mathcal{O}(n_y^2)$$

Neues QP in den Variablen  $\Delta s_0, \Delta q_0, \dots, \Delta q_{m-1}$ :

$$\begin{pmatrix} H_{voll} & B^T \\ B & 0 \end{pmatrix}$$

Matrix hat keine „schöne“ Struktur mehr, Lösen in  $\mathcal{O}(n^3)$

Problem wird reduziert auf Dimension Anzal  $s_0$  + Anzahl  $q_0, \dots, q_{m-1} \Rightarrow$  gleiche Problemgröße wie bei Single Shooting.

### Umsortieren der Variablen

$(s_0, q_0, \lambda_0, s_1, q_1, \lambda_1, \dots) \Rightarrow$  Blockstruktur der Gesamtmatrix, symmetrisch aber nicht positiv definit.  $\Rightarrow$  Benutze iterative oder Sparse-Löser. Aufwand:  $\mathcal{O}(m)$ .

### Vorteile von Multiple Shoting gegenüber Single Shooting

- Es können Informationen über den Verlauf der Trajektorien benutzt werden. Dadurch startet man ggf. näher bei der Lösung und im Einzugsbereich der lokalen Konvergenz des SQP-Verfahrens.
- Durch die kürzeren Integrationsintervalle wird die Nichtlinearität des Problems reduziert (Trompetenabschätzung). Die Existenz der Zustandstrajektorie ist auf dem ganzen Intervall gesichert.
- Die Integration auf den Teilintervallen ist entkoppelt und kann parallel erfolgen, in Realtime somit schneller als Single-Shooting.
- Die Blockbandstruktur der Hessematrix erlaubt Updates von höherem Rang und damit bessere Approximation der Hessematrix der Lagrange-Funktion und schnellere Konvergenz
- das QP hat zwar mehr Variablen und mehr Nebenbedingungen, aber auch Strukturen. Werden diese z. B. durch Kondensierung ausgenutzt ist der Aufwand zum Lösen des QP im wesentlichen auch nicht größer als beim Single-Shooting.

Bemerkung: Auch Kollokation kann für die direkte Methode der Optimalsteuerung eingesetzt werden, analog zu Kapitel 4(2).

# 10 Indirekter Ansatz der optimalen Steuerung

## Optimalsteuerungsproblem

$$\min \int_{t_0}^{t_{end}} L(t, y(t), u(t)) dt + \Phi(t_{end}, y(t_{end})) \quad (10.1)$$

$$\text{s. t. } \dot{y}(t) = f(t, y(t), u(t)) \quad (10.2)$$

$$u \in \Omega \quad (10.3)$$

$$y(t_0) = y_0, r(t_{end}, y(t_{end})) = 0 \quad (10.4)$$

$$b(t, y(t), u(t)) \geq 0 \quad (10.5)$$

Variablen:  $u, y$ , ggf.  $t_{end}, t_0, y_0$ .

Grundannahme:  $L, \Phi, f, r, b$  sind  $C^3$

### Definition 10.1

Eine Steuerung  $u$  heißt zulässig, wenn  $y$  auf  $[t_0, t_{end}]$  existiert, so dass  $(u, y)$  die Nebenbedingungen (10.2) - (10.5) erfüllt. Die Steuerung  $u^*$  heißt optimal, wenn  $u^*$  zulässig ist und  $(u^*, y^*)$  den kleinsten Zielfunktionswert aller zulässigen  $(u, y)$  ergibt.

### Notwendige Optimalitätsbedingung

Satz 10.2: Pontryaginsches Maximumsprinzip (1959).

Sei  $u^*$  eine optimale Steuerung mit zugehörigem  $y^*$ . Definiere:

- Hamiltonfunktion:  $H(t, y, u, \lambda) := -\gamma L(t, y, u) + \lambda^T f(t, y, u) \quad (10.6)$
- Erweiterte Zielfunktion  $\Psi(t_{end}, y(t_{end}), \alpha) = \gamma \Phi(t_{end}, y(t_{end})) + \alpha^T r(t_{end}, y(t_{end})) \quad (10.7)$ .  $\lambda(t) \in \mathbb{R}^{n_y}, \alpha \in \mathbb{R}^{n_r}, \gamma \in \mathbb{R}$ , o. B. d. A.  $\gamma = 1$  falls  $\gamma \neq 0$ .

Dann gilt:  $\exists \gamma \geq 0, \alpha \in \mathbb{R}^{n_y}, \lambda : [t_0, t_{end}] \rightarrow \mathbb{R}^{n_y}$  und  $(\alpha, \lambda(t)) \neq 0$  ( $\alpha, \gamma$  heißen „Lagrange-Parameter“ bzw. adjungierte Variablen), so dass:

- $\lambda(t)$  erfüllen die adjungierten Differentialgleichungen  $\dot{\lambda}(t)^T = \gamma L_y(t, y^*, u^*(t)) - \lambda(t)^T f_y(t, y^*(t), u^*(t)) = -H_y(t, y^*(t), u^*(t), \lambda(t)) \quad (10.8)$ .

- Transversalitätsbedingung:  $\lambda(t_{\text{end}})^T = -\Psi_y(t_{\text{end}}, y^*(t_{\text{end}}), \alpha)$  (10.9). Falls  $t_{\text{end}}$  frei:  $H(t_{\text{end}}, y^*(t_{\text{end}}), u^*(t_{\text{end}}), \lambda(t_{\text{end}})) = \Psi_{t_{\text{end}}}(t_{\text{end}}, y^*(t_{\text{end}}), \alpha)$  (10.10)
- $u^*$  erfüllt das Maximumsprinzip:  $H(t, y^*(t), u^*(t), \lambda(t)) \geq H(t, y^*(t), v(t), \lambda(t))$  fast überall auf  $[t_0, t_{\text{end}}]$  für  $v \in \Omega$  beliebig. D. h.  $u^*$  ist die Lösung von  $\max_{v \in \Omega} H(t, y^*(t), v(t), \lambda(t)) = H(t, y^*(t), u^*(t), \lambda(t))$  fast überall (10.12) bzw.  $u^*(t) = \arg \max_{v \in \Omega} H(t, y^*(t), v(t), \lambda(t))$  fast überall (10.13).

## Anwendungsbeispiel: Energieoptimaler Raketenwagen

$$\begin{aligned}
 \dot{s} &= v & s(0) &= 0 & s(T) &= d \\
 \dot{v} &= u & v(0) &= 0 & v(T) &= 0 & u &\in [-1, 1] \\
 \min \int_0^T \frac{1}{2} u(t)^2 dt \\
 y &:= \begin{pmatrix} s \\ v \end{pmatrix} \\
 y(0) &= y_0 & & & & = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 r(T, y(T)) &= \begin{pmatrix} s(T) - d \\ v(T) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 t_0 &= 0 \\
 t_{\text{end}} &= T \text{ fest}
 \end{aligned}$$

Hamiltonfunktion:

$$H = -\frac{u^2}{2} + \lambda_s v + \lambda_v u = H(v, \lambda_s, \lambda_v, u)$$

adjungierte Differentialgleichung:

$$\begin{aligned}
 \dot{\lambda}_s &= -\frac{\partial H}{\partial s} = 0 \Rightarrow \lambda_s \text{ konstant} \\
 \dot{\lambda}_v &= -\frac{\partial H}{\partial v} = -\lambda_s \Rightarrow \lambda_v \text{ Gerade}
 \end{aligned}$$

Maximumsprinzip

$$H_u = -u + \lambda_v = 0 \Rightarrow u = \lambda v$$

$$u \text{ Gerade: } u = at + b$$

$$\Rightarrow v \text{ Parabel: } v = \frac{a}{2}t^2 + tb + c, \quad c = 0$$

$$\Rightarrow s \text{ kubisches Polynom: } s = \frac{a}{6}t^3 + \frac{b}{2}t^2 + c' \quad c' = 0$$

$T$  einsetzen,  $a, b$  ausrechnen:

$$a = -\frac{12d}{T^3}$$

$$b = 6\frac{d}{T^2}$$

TODO: Bildchen malen mit Geraden (fallend), Parabeln (konkav)

Falls  $b \leq 1$  ist diese Lösung zulässig  $\Leftrightarrow 6d \leq T^2$ . Was passiert, wenn  $T < \sqrt{6d}$ , d. h.  $6d > T^2$ ? Genauer:

$$\max_{u \in [-1, 1]} H(v, \lambda_s, \lambda_v, u) = -\frac{u^2}{2} + \lambda_s v + \lambda_v u$$

Fälle:

- $u = \lambda v$ , falls  $\lambda v \in (-1, 1)$
- $u = -1$ , falls  $\lambda v \leq -1$
- $u = 0$ , falls  $\lambda v \geq 1$

$H$  ist hier konkav in  $u$ , d. h.  $H_{uu} \leq 0$ . Es gilt: Falls  $y, \lambda$  stetig sind,  $H_{uu}$  strikt negativ definit und  $\Omega$  konvex  $\Rightarrow u^*(t)$  stetig. Annahme: Struktur der Lösung: Erst  $u = 1$  von 0 bis  $t_1$ , dann  $u = \lambda v$  von  $t_1$  bis  $t_2$ , dann  $u = -1$  von  $t_2$  bis  $T$ .

Erfüllt Maximumprinzip falls  $b < 1$ , falls  $b > 1$  wegen Konkavität von  $H$ .

Zurück zum Raketenwagen:

$$v(t) = \begin{cases} t & t \in [0, t_1] \\ at^2/2 + bt + c & t \in [t_1, t_2] \\ T - t & t \in [t_2, T] \end{cases} \quad s(t) = \int_0^t v(\tau) d\tau$$

Bestimmung der Variablen:  $a, b, c, t_1, t_2$ . Stetigkeit von  $v$ :  $at_1^2/2 + bt_1 + c = t_1$ ,  $at_2^2/2 + bt_2 + c = T - t_2$ . Randbedingung:  $s(T) = d$ . Stetigkeit von  $u$ :  $at_1 + b = 1$ ,  $at_2 + b = -1$ . 5 (nichtlineare) Bedingungen für 5 Variablen.

Formulierung als Mehrpunkt-Randwertproblem:

$$u(t) = \begin{cases} 1 & 0 \leq t \leq t_1 \\ \lambda v & t_1 \leq t \leq t_2 \\ -1 & t_2 \leq t \leq T \end{cases} = v$$

$$\dot{\lambda}_v = -\lambda_s$$

$$\dot{v} = u$$

$$\dot{\lambda}_s = 0$$

Randbedingungen:

$$s(0) = 0$$

$$s(T) = d$$

$$v(0) = 0$$

$$v(T) = 0$$

Innere-Punkt-Bedingungen:

$$\lambda v(t_s) = 1$$

$$\lambda v(t_2) = -1$$

Das ist ein Randwertproblem mit 4 Differentialgleichungen, 2 zu bestimmenden Größen und 4 Randbedingungen, ist also im Prinzip lösbar, numerisch mit einem Randwertproblemlöser.

## Prinzip der indirekten Methode

- Stelle adjungierte Gleichung auf
- Stelle Transversalitätsbedingungen auf
- Werte Maximumsprinzip aus: vermute Struktur der Steuerung
- Formuliere als Mehrpunkt-Randwertproblem

- Löse durch numerische „Diskretisierung“

(First optimize, then discretize).

Beweis von Satz 10.2 (Maximumprinzip):

Wir beweisen das Maximumprinzip zunächst für das Problem

$$\begin{aligned} \min_{u,y} \int_{t_0}^{t_{end}} L(t, y, u) dt + \Phi(t_{end}, y(t_{end})) &=: J(y, u) \quad (10.14) \\ \dot{y} &= f(t, y, u) \\ y(t_0) &= y_0 \\ t &\in [t_0, t_{end}] \\ u &\in \Omega \end{aligned}$$

d. h. ohne  $r$ ,  $b$  und mit  $t_0, t_{end}$  fest. Grundidee: Konstruktion zulässiger Konkurrenten (Variationen) zur optimalen Lösung  $u^*$  mit Antwort  $y^*$ .

$\bar{u}(t) := u^*(t) + \delta u(t)$  mit zugehöriger Antwort  $\bar{y}(t) := y^*(t) + \delta y(t)$ , die wie folgt konstruiert wird:

- Wähle  $\delta u \Rightarrow \bar{u}$
- Löse DGL  $\Rightarrow \bar{y}$
- $\delta y = \bar{y} - y^*$

$\delta y$  löst die DGL  $\delta \dot{y} = \dot{\bar{y}} - \dot{y}^* = f(t, y^* + \delta y, u^* + \delta u) - f(t, y^*, u^*)$ .

Aus Stabilitätssatz (Satz 1.1) folgt:

$$\|\delta y\| = \mathcal{O}(\|\delta u\|) \text{ Norm im Funktionenraum}$$

Konstruktion der Störung: Beispiel 1:  $\Delta u(t)$  beliebig,  $\delta_\varepsilon u(t) := \varepsilon \Delta u(t)$ ,  $\varepsilon > 0$  klein, so dass

$$\begin{aligned}
 \bar{u} &= u^* + \varepsilon \Delta u \in \Omega \Rightarrow \|\delta_\varepsilon u\| = \mathcal{O}(\varepsilon) \\
 \delta_\varepsilon \dot{y}(t) &= f(t, y^* + \delta_\varepsilon y, u^* + \delta_\varepsilon u) - f(t, y^*, u^*) \\
 &= \frac{\partial f}{\partial y} \delta_\varepsilon y(t) + \frac{\partial f}{\partial u} \delta_\varepsilon u(t) + \mathcal{O}(\delta_\varepsilon^2) \\
 \delta_\varepsilon y(t_0) &= 0 \\
 A(t) &:= \frac{\partial f}{\partial y} \\
 B(t) &:= \frac{\partial f}{\partial u} \\
 \delta_\varepsilon y(t) &= \int_{t_0}^t A(\tau) \delta_\varepsilon y(t\tau) + B(\tau) \delta_\varepsilon u(\tau) \, d\tau + \mathcal{O}(\delta_\varepsilon^2) \\
 \Rightarrow \|\delta_\varepsilon y\| &= \mathcal{O}(\varepsilon)
 \end{aligned}$$

Beispiel:

$$\begin{aligned}
 \bar{u}(t) &= \begin{cases} u^*(t) & \text{falls } t \notin [\hat{t}, \hat{t} + \varepsilon) \\ v \in \Omega \text{ bel.} & t \in [\hat{t}, \hat{t} + \varepsilon) \end{cases} \\
 \|\delta_\varepsilon u\| &= \|\bar{u} - u^*\| = \mathcal{O}(\varepsilon)
 \end{aligned}$$

Im Sinne einer Integralnorm (nicht  $\|\cdot\|_\infty$ ).

Setze Variation in problem (10.14) ein:



$$\begin{aligned}
0 &\leq J(\bar{y}, \bar{u}) - J(y^*, u^*) \\
&= \Phi(t_{end}, \bar{y}(t_{end})) - \Phi(t_{end}, y^*(t_{end})) + \int_{t_0}^{t_{end}} L(t, \bar{y}, \bar{u}) - L(t, y^*, u^*) \, dt \\
&\quad + \int_{t_0}^{t_{end}} \lambda(t)^T \underbrace{(\dot{\bar{y}} - f(t, \bar{y}, \bar{u}) + f(t, y^*, u^*) - \dot{y}^*)}_{=0 \text{ für bel. } \lambda(t) \text{ und bel. } \bar{u} \text{ mit zug. } \bar{y}} \, dt \\
&= \Phi(t_{end}, \bar{y}(t_{end})) - \Phi(t_{end}, y^*(t_{end})) \\
&\quad + \int_{t_0}^{t_{end}} -L(t, y^*, u^*) + \lambda^T f(t, y^*, u^*) + L(t, \bar{y}, \bar{u}) - \lambda^T f(t, \bar{y}, \bar{u}) \, dt \\
&\quad + \int_{t_0}^{t_{end}} \lambda(t)^T \delta \dot{y}(t) \, dt \\
&= \Phi(t_{end}, \bar{y}(t_{end})) - \Phi(t_{end}, y^*(t_{end})) \\
&\quad + \int_{t_0}^{t_{end}} H(t, y^+, u^*, \lambda) - H(t, \bar{y}, \bar{u}, \lambda) \, dt \\
&\quad + [\lambda(t)^T \delta y(t)]_{t_0}^{t_{end}} - \int_{t_0}^{t_{end}} \dot{\lambda}(t)^T \delta y(t) \, dt
\end{aligned}$$

Taylorentwicklung von  $H$ , erst nach  $y$  um  $y^*$ , dann nach  $u$  um  $u^*$ .

$$\begin{aligned}
&= \frac{\partial \Phi}{\partial y}(t_{end}, y^*(t_{end})) \delta y(t_{end}) + \mathcal{O}(\varepsilon^2) \\
&+ \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, \bar{u}, \lambda) - \frac{\partial H}{\partial y}(t, y^*, \bar{u}, \lambda) \delta y(t) + \mathcal{O}(\varepsilon) dt \\
&+ \lambda(t_{end})^T \delta y(t_{end}) - \lambda(t_0)^T \underbrace{\delta y(t_0)}_{=0} - \int_{t_0}^{t_{end}} \lambda(t)^T \delta y(t) dt \\
&= \left( \frac{\partial \Phi}{\partial y}(t_{end}, y^*(t_{end})) + \lambda(t_{end})^T \right) - \int_{t_0}^{t_{end}} \dot{\lambda}(t)^T \delta y(t) dt + \mathcal{O}(\varepsilon^2) \\
&+ \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, \bar{u}, \lambda) - \frac{\partial H}{\partial y}(t, y^*, u^*, \lambda) \delta y(t) - \underbrace{\frac{\partial^2 H}{\partial u \partial y}(t, y^*, u^*, \lambda) \delta u \delta y}_{=\mathcal{O}(\varepsilon^2)} + \mathcal{O}(\varepsilon^2) \\
&= (\lambda(t_{end})^T + \frac{\partial \Psi}{\partial y}(t_{end}, y^*(t_{end})) \delta y(t_{end})) \\
&+ \int_{t_0}^{t_{end}} (-\dot{\lambda}(t)^T - \frac{\partial H}{\partial y}(t, y^*, u^*, \lambda)) \delta y(t) dt \\
&+ \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, \bar{u}, \lambda) dt + \mathcal{O}(\varepsilon^2)
\end{aligned}$$

Wähle  $\lambda(t)$ , das Transversalitätsbedingung und adjungierte Differentialgleichung erfüllt.  
 $\lambda(t)$  ist eindeutig bestimmt.

$$\Rightarrow 0 \leq \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, \bar{u}, \lambda) dt + \mathcal{O}(\varepsilon^2)$$

Für Beispiel 2:

$$0 \leq \int_{\hat{t}}^{\hat{t}+\varepsilon} H(t, y^*, u^*, \lambda) - H(t, y^*, v, \lambda) dt + \mathcal{O}(\varepsilon^2)$$

Lasse  $\varepsilon$  gegen Null gehen:

$$\begin{aligned}
0 &\leq \varepsilon(H(\hat{t}, y^*(\hat{t}), u^*(\hat{t}), \lambda(\hat{t})) - H(\hat{t}, y^*(\hat{t}), v(\hat{t}), \lambda(\hat{t}))) + \mathcal{O}(\varepsilon^2) \\
0 &\leq H(\hat{t}, y^*(\hat{t}), u^*(\hat{t}), \lambda(\hat{t})) - H(\hat{t}, y^*(\hat{t}), v(\hat{t}), \lambda(\hat{t})) + \mathcal{O}(\varepsilon) \\
\Rightarrow H(\hat{t}, y^*(\hat{t}), u^*(\hat{t}), \lambda(\hat{t})) &\geq H(t, y^*(t), v(t), \lambda(t)) \forall v \in \Omega \forall t \in [t_0, t_{end}] \\
&\rightarrow \text{Maximumprinzip}
\end{aligned}$$

### Bemerkung 10.3: Variation, Beispiel 1

$$\begin{aligned}
\delta_\varepsilon u(t) &= \varepsilon \Delta u(t) \\
0 &\leq \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, u^* + \delta_\varepsilon u, \lambda) dt + \mathcal{O}(\varepsilon^2) \\
&= \int_{t_0}^{t_{end}} \frac{\partial H}{\partial u}(t, y^*, u^*, \lambda)(-\varepsilon \Delta u(t)) dt + \mathcal{O}(\varepsilon^2)
\end{aligned}$$

Wähle  $\Delta u = \eta \cdot \frac{\partial H^T}{\partial u}$ ,  $\eta > 0$  hinreichend klein, so dass  $u^* + \delta u \in \Omega$ .

$$\begin{aligned}
\cdots &= \int_{t_0}^{t_{end}} -\varepsilon \eta \frac{\partial H}{\partial u} \partial H^T \partial u dt + \mathcal{O}(\varepsilon^2) \\
\varepsilon \rightarrow 0 &\Rightarrow \int_{t_0}^{t_{end}} \frac{\partial H}{\partial u} \frac{\partial H^T}{\partial u} dt = 0 \\
&\Rightarrow 0 = \frac{\partial H}{\partial u}(t, y^*(t), u^*(t), \lambda(t)) = 0 \text{ fast überall}
\end{aligned}$$

$$\begin{aligned}
H(t, y, u, \lambda) &:= -L(t, y, u) + \lambda^T f(t, y, u) \\
\Psi(t_{end}, y(t_{end}), \alpha) &:= \Phi(t_{end}, y(t_{end})) + \alpha^T r(t_{end}, y(t_{end}))
\end{aligned}$$

Seien  $u^*, y^*$  optimal, dann existieren  $\lambda, \alpha$  so dass:

$$\begin{aligned}\dot{\lambda}(t)^T &= -\frac{\partial H}{\partial y}(t, y^*(t), u^*(t), \lambda(t)) \\ \lambda(t_{end}) &= -\frac{\partial \Psi}{\partial y}(t_{end}, y^*(t_{end}), \alpha) \\ H(t_{end}, y^*(t_{end}), u^*(t_{end}), \lambda(t_{end})) &= \frac{\partial \Psi}{\partial t}(t_{end}, y^*(t_{end}), \alpha)\end{aligned}$$

$u^*$  erfüllt  $H(t, y^*(t), u^*(t), \lambda(t)) \geq H(t, y^*(t), v(t), \lambda(t)) \quad \forall v \in \Omega$ .

**Bemerkung 10.4: Maximumprinzip mit Randbedingungen**

$$\begin{aligned}0 \leq & J(\bar{y}, \bar{u}) - J(y^*, u^*) + \int_{t_0}^{t_{end}} \lambda^T (\dot{\bar{y}} - f(t, \bar{y}, \bar{u}) + \dot{y}^* + f(t, y^*, u^*)) dt \\ & + \alpha^T \underbrace{(r(t_{end}, \bar{y}(t_{end})) - r(t_{end}, y^*(t_{end})))}_{=0}\end{aligned}$$

Dabei sind  $u^*, y^*$  optimal und  $\bar{u} = u^* + \delta u$ ,  $\bar{y} = y^* + \delta y$  sind zulässige Konkurrenten. Taylorentwicklung:

$$\begin{aligned}0 \leq & \left( \frac{\partial \Phi}{\partial y}(t_{end}, y^*(t_{end})) + \lambda(t_{end})^T \right) \delta y(t_{end}) \\ & + \int_{t_0}^{t_{end}} \left( -\dot{\lambda}(t)^T - \frac{\partial H}{\partial y}(t, y^*, u^*, \lambda) \right) \delta y(t) dt \\ & + \int_{t_0}^{t_{end}} H(t, y^*, u^*, \lambda) - H(t, y^*, \bar{u}, \lambda) dt \\ & + \alpha^T \frac{\partial r}{\partial y}(t_{end}, y^*(t_{end})) \delta y(t_{end}) + \text{Terme höherer Ordnung}\end{aligned}$$

Wenn wir  $\alpha, \lambda$  haben, die folgendes erfüllen:

- adjungierte Differentialgleichung
- Transversalitätsbedingung:

$$\lambda(t_{end})^T = \left( -\frac{\partial \Phi}{\partial y}(t_{end}, y(t_{end})) - \alpha^T \frac{\partial r}{\partial y}(t_{end}, y(t_{end})) \right)$$

folgt wie oben das Maximumsprinzip.

Nicht gezeigt: Existenz von  $\alpha$  und  $\lambda$

**Bemerkung 10.5: freie Endzeit  $t_{end}$**

Es gibt eine optimale Endzeit  $t_{end}^*$ , eine optimale Steuerung  $u^*$  und die zugehörige  $y^*$ .  $\delta t_{end} := \overline{t_{end}} - t_{end}^*$  zu  $\bar{u}$  und  $\bar{y}$ . Setze  $t_{end}^*$  und  $\overline{t_{end}}$  analog zu (10.15) an, Taylorentwicklung auch in  $t_{end}$  um  $t_{end}^*$ . Zusätzliche Terme sind dann:

$$\begin{aligned} & \frac{\partial \Phi}{\partial t}(t_{end}^*, y(t_{end}^*)) + \alpha^T \frac{\partial r}{\partial t}(t_{end}^*, y(t_{end}^*)) \delta t_{end} \\ & + \underbrace{\int_{t_{end}^*}^{t_{end}^* + \delta t_{end}} L(t, \bar{y}, \bar{u}) - \lambda(t)^T f(t, \bar{y}, \bar{u}) dt}_{= -H(t_{end}, y(t_{end}))} \\ & + \underbrace{\int_{t_{end}^*}^{t_{end}^* + \delta t_{end}} \lambda(t)^T \dot{\bar{y}}(t) dt}_{= 0 \text{ durch Wahl von } \lambda \equiv 0 \text{ auf } [t_{end}^*, t_{end}^* + \delta t_{end}]} \\ & + \text{Terme höherer Ordnung} + \text{TODO} \end{aligned}$$

Wennes  $\alpha, \lambda$  gibt, die die adjungierte Differentialgleichung, obige Transversalitätsbedingung und  $H(t_{end}^*, y^*(t_{end}^*), u^*(t_{end}^*), \lambda(t_{end}^*)) = \frac{\partial \Phi}{\partial t}(t_{end}^*, y^*(t_{end}^*)) + \alpha^T \frac{\partial r}{\partial t}(t_{end}^*, y^*(t_{end}^*))$  erfüllen, folgt das Maximumprinzip.

Andere Möglichkeit: Transformation auf feste Endzeit.

Beispiel:

Zielfunktional:  $\min t_{end} \ (t_0 = 0)$

- a)  $\Phi = t_{end}, L = 0$
- b)  $\Phi = 0, L = 1$

Randbedingung  $r(y(t_{end}))$  hängt nicht explizit von  $t_{end}$  ab.

Transversalitätsbedingungen:

- a)  $H = \lambda^T f(t, y, u), H(t_{end}) = 1$
- b)  $H = -1 + \lambda^T f(t, y, u), H(t_{end}) = 0$

Transversalitätsbedingung zum Eliminieren von  $\alpha$ :

$$\begin{aligned} \lambda(t_{end}) &= -\frac{\partial \Phi}{\partial y}(t_{end}, y(t_{end}))^T - \frac{\partial r}{\partial y}(t_{end}, y(t_{end}))^T \alpha \\ H(t_{end}) &= \frac{\partial \Phi}{\partial t} + \frac{\partial r^T}{\partial t} \alpha \text{ Wenn } t_{end} \text{ frei} \end{aligned}$$

Wenn  $\frac{\partial r^T}{\partial y}$  Vollrang hat, können wir die  $\alpha$  eliminieren:

$$\frac{\partial r}{\partial y}^T = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

mit  $A_1$  regulär (ggf. nach Zeilenpermutationen)

$$\begin{aligned} \alpha &= A_1^{-1} \begin{pmatrix} -\lambda_I & -\frac{\partial \Phi}{\partial y_I} \end{pmatrix} \\ \lambda_I &= \begin{pmatrix} \lambda_1 \\ \dots \\ \lambda_\nu \end{pmatrix} \\ \lambda_{II} &= \begin{pmatrix} \lambda_{\nu+1} \\ \dots \\ \lambda_n \end{pmatrix} \\ \lambda_{II} &= -\frac{\partial \Phi}{\partial y_{II}}^T - A_2 \alpha, \alpha \text{ einsetzen} \end{aligned}$$

Es fallen die  $\alpha$  und  $\nu$  Bedingungen weg.

### Spezialfälle des Maximumprinzips

$H$  heißt regulär, wenn  $\Omega$  konvex ist und  $-\frac{\partial^2 H}{\partial^2 u}$  positiv definit ist.

$$\forall t, y^*(t), \lambda(t), u \in \Omega$$

Dann ist das Maximierungsproblem im Maximumsprinzip konvex. Es hat eine eindeutige Lösung  $u^*(t)$ .

$$h^*(t) := \max_{u \in \Omega} h(t, u) := \max_{u \in \Omega} H(t, y^*, u(t), \lambda(t))$$

Dann ist  $u^*(t)$  stetig und  $h^*(t)$  stetig.

Beweis: Annahme  $\exists \hat{t}$  mit  $u(\hat{t}_-) \neq u(\hat{t}_+)$ . Sei  $h(\hat{t}, u(\hat{t}_-)) < h(\hat{t}, u(\hat{t}_+))$ ,  $h(t, u) = H(t, y^*, u, \lambda)$  ist stetig in  $u$ .  $\exists \varepsilon : h(t, u(t)) < h(t, u(\hat{t}_+)) \forall t \in (\hat{t} - \varepsilon, \hat{t})$ . Widerspruch zum Maximumprinzip. Auf ganz  $(\hat{t} - \varepsilon, \hat{t})$  ist  $u(t)$  nicht optimal.

Wenn  $\Omega$  nicht konvex ist kann  $u^*$  von einer Ecke in eine andere springen am Zeitpunkt  $\hat{t}$ . Wenn  $h^*(t)$  stetig ist muss gelten:  $h(u(\hat{t}_-), \hat{t}) = h(u(\hat{t}_+), \hat{t})$ . Das legt den Sprungzeitpunkt  $\hat{t}$  fest.

$h(u, t)$  hängt nichtlinear von  $u$  ab. Möglichkeiten:

- a)  $u^* \in \circ\Omega \Rightarrow \frac{\partial h}{\partial u}(u^*(t), t) = 0$  und  $\frac{\partial^2 h}{\partial^2 u}(u^*(t), t) \leq 0 \Rightarrow$  Wir können lokal eine Funktion  $\hat{u}(t, y, \lambda) := \arg \max H(t, y^*, u, \lambda)$  finden.
- b)  $u^* \in \partial\Omega$  Dann gelten die KKT-Bedingungen, falls  $\partial\Omega$  lokal durch Ungleichungen und Gleichungen beschrieben wird.

Beim Übergang von a) nach b) und umgekehrt treten Schaltpunkte auf.

Beispiel: Raketenwagen  $u \in [-1, 1] \Rightarrow$  entweder  $u \in \{-1, 1\}$  oder  $\frac{\partial h}{\partial u} = 0$ .

$H$  ist linear in  $u$ .

$$\begin{aligned} H(t, y, u, \lambda) &= H_0(t, y, \lambda) + H_u(t, y, \lambda)u \\ \dot{y}(t, y, u) &= f_0(t, y) + f_u(t, y)u \\ L &= L_0(t, y) + L_u(t, y)u \\ H_0 &= -L_0 + \lambda^T f_0 \\ H_u &= -L_u + \lambda^T f_u \end{aligned}$$

Sei z. B.  $u \in [-1, 1]$ .

- a)  $H_u > 0 \Rightarrow u^* = +1$
- b)  $H_u < 0 \Rightarrow u^* = -1$

Fall 1:

$$u^*(t) = \begin{cases} +1 & t < \hat{t} \\ -1 & t > \hat{t} \end{cases}$$

Im Punkt  $\hat{t}$  gilt ( $H$  stetig in  $u$ ):

$$\begin{aligned} H(\hat{t}_-) &= H_0 + H_u \cdot (+1) = H_0 + H_u \cdot (-1) = H(\hat{t}_+) \\ \Rightarrow H_u &= 0 = -L_u + \lambda^T f_u \end{aligned}$$

Heißt Schaltfunktion. Die Nullstelle legt den Schaltpunkt in  $\hat{t}$  fest.

Fall 2:

b)  $H_u = 0$  gilt auf einem ganzen Intervall  $[t_1, t_2]$

$$\frac{\partial H}{\partial u}(t, y, (t), \lambda(t)) \equiv 0 \text{ auf } [t_1, t_2]$$

$u^*$  kann nicht durch das Maximumprinzip bestimmt werden. Dann gilt („Indexreduktion“):

- $\left(\frac{\partial}{\partial t}\right)^T H_u(t, y(t), \lambda(t)) \equiv 0 \forall i$  auf  $[t_1, t_2]$
- Dies hängt, wenn überhaupt, nur für  $i = 2j$  von  $u$  ab. Das kleinste solche  $j$  heißt Ordnung.

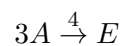
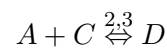
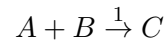
Sei z. B.  $j = 1$ : Singulärer Bogen der Ordnung 1.  $u$  wird festgelegt durch  $\ddot{U}_u \equiv 0$ .  $H_u(t_1) = 0 = \dot{H}_u(t_1)$  oder  $H_u(t_1) = 0 = H_u(t_2) = 0$  legen  $t_1$  und  $t_2$  fest.





# 11 Optimale Versuchsplanung

## 11.1 Anwendungsbeispiel Urethan-Reaktion



- A: Isocyanat
- B: Butanol
- C: Urethan
- D: Allophanat
- E: Isozyanurat
- L: Dimethylsulfonid

TODO: Bild und so

DAE-Modell:  $n_1, \dots, n_6$ : Molzahlen von  $A, \dots, L$

$$\dot{n}_3 = V(r_1 - r_2 + r_3)$$

$$n_3(t_0) = 0$$

$$\dot{n}_4 = V(r_2 - r_3)$$

$$n_4(t_0) = 0$$

$$\dot{n}_5 = Vr_4$$

$$n_5(t_0) = 0$$

$$0 = n_1 + n_3 + 2n_4 + 3n_5 - n_{1,0} - n_{1,0,Z_1} feed_1(t) \quad Z_1: \text{Zulauf 1}$$

$$0 = n_2 + n_3 + n_4 - n_{2,0} - n_{2,0,Z_2} feed_2(t)$$

$$0 = n_6 - n_{6,0} - n_{6,0,Z_1} feed_1(t) - n_{6,0,Z_2} feed_2(t)$$

Algebraische Gleichungen ergeben sich aus Stöchioetrie.

$$\begin{aligned}
 \dot{T} &= dT(t) \\
 T(t_0) &= T_0 \\
 \dot{feed}_1 &= dfeed_1(t) \\
 feed_1(t_0) &= 0 \\
 \dot{feed}_2 &= dfeed_2(t) \\
 feed_2(t_0) &= 0
 \end{aligned}$$

Reaktionsgeschwindigkeiten:

$$\begin{aligned}
 r_1 &= k_1 \frac{n_1}{V} \frac{n_2}{V} \\
 r_2 &= k_2 \frac{n_1}{V} \frac{n_3}{V} \\
 r_3 &= k_3 \frac{n_4}{V} \\
 r_4 &= k_4 \left( \frac{n_1}{V} \right)^2
 \end{aligned}$$

Geschwindigkeitskoeffizienten:

$$\begin{aligned}
 k_i &= k_{ref} \exp \left( -\frac{E_{ai}}{R} \left( \frac{1}{T(t)} - \frac{1}{T_{refi}} \right) \right) \quad i = 1, 2, 4 \\
 \frac{k_2}{k_3} &= K_{c,2} \exp \left( -\frac{\Delta H_2}{R} \left( \frac{1}{T(t)} - \frac{1}{T_{g2}} \right) \right)
 \end{aligned}$$

Volumen:

$$V = \sum_{i=1}^6 \frac{M_i n_i}{\rho_i}$$

Das Modell enthält:

- 9 Zustände:  $n_1, \dots, n_6, T, feed_1, feed_2$
- 8 unbekannte Parameter:  $k_{ref1}, E_{a1}, k_{ref2}, E_{a2}, k_{ref4}, E_{a4}, K_{c2}, \Delta H_2$
- 3 Steuerfunktionen  $u(t)$ :  $dT(t), dfeed_1(t), dfeed_2(t)$
- 8 Steuergrößen:  $n_{1,0}, n_{2,0}, n_{6,0}, n_{1,0,Z_1}, n_{6,0,Z_1}, n_{2,0,Z_2}, n_{6,0,Z_2}$
- Restliche Größen sind Konstanten:  $R, T_{ref}, M_i, \rho_i$

Ziele:

- Schätzungen der Modellparameter auf experimentellen Daten.
- Design der Experimente, so dass die geschätzten Parameter möglichst kleine statistische Unsicherheiten haben, also optimale Versuchsplanung.

## Statistische Analyse der Lösung des Parameterschätzproblems

Modellgleichung (o. B. d. A. ODE-RWP).

$$\begin{aligned}\dot{y} &= f(t, y(t), p, q, u(t)) \quad (11.1) \\ y(t_0) &= y_0(p, q)\end{aligned}$$

- $p \in \mathbb{R}^{n_p}$ : Parameter, werden durch die Natur bestimmt und sind erstmal unbekannt bzw. müssen geschätzt werden.
- $q \in \mathbb{R}^{n_q}$ : Steuergrößen und  $u: [t_0, t_{end}] \rightarrow \mathbb{R}^{n_u}$ : Steuerfunktionen werden durch den Experimentator festgelegt.
- $r(t_1, y(t_1), \dots, t_k, y(t_k), p, q) = 0$ : Randbedingungen (11.2)

## Modell für die Messungen

$$\begin{aligned}\eta_i &= h_i(t_i, y(t_i), p, q) + \varepsilon_i, \quad i = 1, \dots, M \quad (11.3) \\ \varepsilon_i &\sim \mathcal{N}\left(0, \frac{\sigma_i^2}{w_i}\right), \text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j\end{aligned}$$

$w_i \in \{0, 1\}$ : Wird Messung  $i$  tatsächlich durchgeführt? Für gegebene Daten: alle  $w_i = 1$ . Für zu planende Experimente: Die Wahl von  $w$  entscheidet über die Durchführung der Messung  $i$ .

- $w_i = 1$ :  $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$
- $w_i = 0$ : Die Messung  $i$  hat unendliche Varianz.
- $w_i \in [0, 1]$ , z. B.  $w_i = \frac{1}{2}$ . „Halbe Messung“ mit doppelter Varianz ( $\rightarrow$  sinnvolle Relaxierung).

Parameterschätzproblem:

$$\begin{aligned}
 & \min_{p,y} \frac{1}{2} \sum_{i=1}^M \frac{w_i}{\sigma_i^2} (\eta_i - h_i(t_i, y(t_i), p, q))^2 \\
 & \text{s. t. } \dot{y} = f(t, y, p, q, u) \\
 & y(t_0) = y_0(p, q) \\
 & r(t_1, y(t_1), \dots, t_k, y(t_k)) = 0 \\
 & \text{mit } \xi := (q, u, w) \text{ fest}
 \end{aligned}$$

Nach Parametrisierung der Dynamik, siehe Kapitel 3:

$$\begin{aligned}
 & v = (p, s) \\
 & \min_v \frac{1}{2} \|F_1(v, \xi)\|_2^2 \quad (11.5) \\
 & \text{s. t. } 0 = F_2(v, \xi) \\
 & F_1(v, \xi) = \left( \frac{\sqrt{w_i}}{\sigma_i} (j_i(t_i, y(t_i, v, \xi), p, q)) \right)_{i=1, \dots, M}
 \end{aligned}$$

$F_2(v, \xi)$  enthält die Randbedingungen und die zusätzlichen Gleichungsbeschränkungen aus der Parametrisierung.

Jacobi-Matrix:

$$\begin{aligned}
 \partial_{1ij} &= \frac{\partial F_{1i}}{\partial v_j} = -\frac{\sqrt{w_i}}{\sigma_i} \left( \frac{\partial h_i}{\partial y} \frac{\partial y(t_i)}{\partial v_j} + \frac{\partial h_i}{\partial v_j} \right) \\
 \partial_{2ij} &= \frac{\partial F_{2i}}{\partial v_j} = \frac{\partial F_{2i}}{\partial y} \frac{\partial y}{\partial v_j} + \frac{\partial F_{2i}}{\partial v_j}
 \end{aligned}$$

$\frac{\partial y}{\partial v}$  sind Lösungen von Variationsdifferentialgleichungen nach Parametern und Anfangswerten.

Verallgemeinerte Inverse:

Lineares Problem:

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

Lösungsoperator:

$$J^\dagger = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix}$$

unbeschränkter Spezialfall:  $J^\dagger$   $= (J_1^T J_1)^{-1} J_1^T$

$$\Delta v = -J^\dagger F, \quad F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}$$

Im Beispiel: Messwerte gegeben für Massenprozent von  $A, C, D, E$  mit verschiedenen Varianzen an 5 Messzeitpunkten.

TODO: Bildchen

Ergebnis: Gesamtresiduum wird reduziert von  $3 \cdot 10^6$  auf  $4 \cdot 10^1$ , die Kurven werden gut gefittet  $\rightarrow$  „Guter Fit“

Aber: Ist das auch eine gute Schätzung der Parameter?

Es gilt: die Messdaten sind Zufallsgrößen. Der Lösungsalgorithmus des Parameterschätzproblems bildet die Messdaten auf den Parameterschätzer  $\hat{v}_{ab}$ . Also sind auch die geschätzten Parameter Zufallsgrößen.

TODO: mehr Bildchen wagen

Wir betrachten die Fehlerfortpflanzung für das linearisierte Parameterschätzproblem im Lösungspunkt  $\hat{v}$ . Dann gilt:

$$\hat{v} \sim \mathcal{N}(v, C)$$

(ist die Kovarianzmatrix des Parameterschätzers).

Erwartungswert  $E$ :

$$\begin{aligned}
 E(\hat{v}) &= E(v_{wahr} + \Delta v) = E(v_{wahr}) + E(-J^\dagger F) \\
 &= E(v_{wahr}) - J^\dagger E(F) = E(v_{wahr}) - J^\dagger E \begin{pmatrix} S^{-\frac{1}{2}}(\eta - h) \\ F_2 \end{pmatrix} \\
 &= E(v_{wahr}) - J^\dagger 0 = v_{wahr}
 \end{aligned}$$

$$\text{mit } S = \text{diag} \left( \frac{\sigma_i^2}{w_i}, i = 1, \dots, M \right)$$

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

$$h = \begin{pmatrix} h_1 \\ \dots \\ h_M \end{pmatrix}$$

$$\begin{aligned}
 C &= E(\hat{v}\hat{v}^T) = E((v_{wahr} + \Delta v)(v_{wahr} + \Delta v)^T) = E(\Delta v \Delta v^T) \\
 &= E(J^\dagger F F^T (J^\dagger)^T) \\
 &= J^\dagger E \begin{pmatrix} F_1 F_1^T & F_1 F_2^T \\ F_2 F_1^T & F_2 F_2^T \end{pmatrix} (J^\dagger)^T \\
 &= J^\dagger \begin{pmatrix} E(F_1 F_1^T) & 0 \\ 0 & 0 \end{pmatrix} (J^\dagger)^T
 \end{aligned}$$

$$\begin{aligned}
 \text{mit } E(F_1 F_1^T) &= E(S^{-\frac{1}{2}}(\eta - h)(\eta - h)^T S^{-\frac{1}{2}}) \\
 &= S^{-\frac{1}{2}} E(\varepsilon \varepsilon^T) S^{-\frac{1}{2}} \\
 &= S^{-\frac{1}{2}} S S^{-\frac{1}{2}} = I
 \end{aligned}$$

Also:  $\hat{v}$  ist multinomialverteilt mit Erwartungswert  $v_{wahr}$  und Kovarianzmatrix  $C = J^\dagger \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} (J^\dagger)^T$ .

### Bemerkung 11.1: Eigenschaften von C

- a)  $C$  existiert und ist eindeutig, wenn CQ und PD erfüllt sind.
- b)  $C$  ist positiv semidefinit
- c)  $C$  hängt von Ableitungen des Parameterschätzproblems ab.
- d) Wenn das Modell nichtlinear in  $v$  ist hängt  $C$  von den Parametern  $\hat{v}$  ab.
- e)  $C$  hängt von den Versuchsplanungsgrößen  $\xi = (q, u, v)$  ab.
- f)  $C$  hängt nicht von den Messwerten ab.

Unbeschränkter Fall:

$$J^\dagger = (J_1^T J_1)^{-1} J_1^T$$

$$C = (J_1^T J_1)^{-1} J_1^T J_1 (J_1^T J_1)^{-1} = (J_1^T J_1)^{-1}$$

Beschränkter Fall:

$$J^\dagger = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \begin{pmatrix} J_1^T & 0 \\ 0 & I \end{pmatrix}$$

$$C = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T J_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \begin{pmatrix} I \\ 0 \end{pmatrix} \quad (11.7)$$

### Lemma 11.2

Sei

$$\begin{pmatrix} X & Y^T \\ Y & Z \end{pmatrix} := \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} =: M^{-1}$$

Dann ist die Kovarianzmatrix  $C = X$  und erfüllt

$$\begin{aligned} I &= J_1^T J_1 C + J_2^T Y \\ 0 &= J_2 C \end{aligned}$$

Beweis: Multipliziere Matrix mit  $M^{-1}$ :

$$\begin{aligned} I &= J_1^T J_1 X + J_2^T Y \\ 0 &= J_1^T J_1 Y + J_2^T Z \\ 0 &= J_2 X \\ I &= J_2 Y^T \end{aligned}$$

Setze in (11.7) ein:

$$\begin{aligned}
 C &= \begin{pmatrix} I & 0 \end{pmatrix} M^{-1} \begin{pmatrix} J_1^T J_1 & 0 \\ 0 & 0 \end{pmatrix} M^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} \\
 &= \begin{pmatrix} X & Y^T \end{pmatrix} \begin{pmatrix} J_1^T J_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \\
 &= X J_1^T J_1 X \\
 &= X \underbrace{(J - J_2^T Y)}_{= J_1^T J_1 X} \\
 &= X - \underbrace{(J_2 X)^T Y}_{=0} \\
 &= X
 \end{aligned}$$

### Korollar 11.3

$$C = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 J_2^T & \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix}$$

## Konfidenzgebiete

Das  $(100 \cdot \alpha)\%$ -Konvergenzgebiet ist das Gebiet, in dem die wahren aber unbekannten Parameter  $v_{wahr}$  mit einer Sicherheit von  $\alpha \in (0, 1]$  liegen. Als Näherung betrachten wir das Konfidenzgebiet im Lösungspunkt  $\hat{v}$  des Gauß-Newton-Verfahrens:

$$\begin{aligned}
 G(\alpha, \hat{v}) &= \{v \in \mathbb{R}^n : F_2(x) = 0 \wedge \|F_1(v)\|_2^2 - \|F_1(\hat{v})\|_2^2 \leq \gamma(\alpha)^2\} \\
 \text{mit } \gamma(\alpha)^2 &= \mathcal{X}_{n-n_2}^2(a - \alpha)
 \end{aligned}$$

(Quantil der  $\mathcal{X}^2$ -Verteilung zum Wert  $\alpha$  mit  $n - n_2$  Freiheitsgraden,  $v \in \mathbb{R}^n$ ,  $F_1(v) \in \mathbb{R}^{n_1}$ ,  $F_2(v) \in \mathbb{R}^{n_2}$ ).

### Linearisierte Näherung

$$\begin{aligned}
 G_L(\alpha) &= \{v \in \mathbb{R}^n : F_2(\hat{v}) + J_2(\hat{v})(v - \hat{v}) = 0, \\
 &\quad \|F_1(\hat{v}) + J_1(\hat{v})(v - \hat{v})\|_2^2 - \|F_1(\hat{v})\|_2^2 \leq \gamma(\alpha)^2\} \quad (11.10)
 \end{aligned}$$



**Lemma 11.4**

Das linearisierte Konfidenzgebiet lässt sich in folgender Form schreiben:

$$G_L(\alpha, \hat{v}) = \overline{G}_L(\alpha, \hat{v}) := \left\{ v \in \mathbb{R}^n, v - \hat{v} = J^\dagger(\hat{v}) \begin{pmatrix} \delta w \\ 0 \end{pmatrix}, \delta w \in \mathbb{R}^{n_1}, \|\delta w\| \leq \gamma(\alpha) \right\}$$

Beweis: siehe z. B. Skript zu Numerik 3.

**Korollar 11.5**

Sei  $\theta_i = \gamma(\alpha) \sqrt{c_{ii}(\hat{v})}$ ,  $i = 1, \dots, n$ . Dann gilt:  $G_L(\alpha, \hat{v})$  wird durch einen Quader mit den Seitenlängen  $2\theta_i$  eingeschlossen:

$$G_L(\alpha, \hat{v}) \subseteq [\hat{v}_1 - \theta_1, \hat{v}_1 + \theta_1] \times \dots \times [\hat{v}_n - \theta_n, \hat{v}_n + \theta_n] \quad (11.12)$$

Beweis:

$$\begin{aligned} |v_1 - \hat{v}_1| &\leq \left\| \left( J^\dagger(\hat{v}) \begin{pmatrix} I \\ 0 \end{pmatrix} \right)^T e_i \right\| \|\delta w\|_2 \\ &= \left( e_i^T J^\dagger(\hat{v}) \begin{pmatrix} I \\ 0 \end{pmatrix} (I \quad 0) J^\dagger(\hat{v})^T e_i \right)^{\frac{1}{2}} \|\delta w\|_2 \\ &\leq \sqrt{c_{ii}(\hat{v})} \gamma(\alpha) \end{aligned}$$

**11.3 Das Versuchsplanungs- und Optimierungsproblem**

Wie kann die Güte des parameterschätzproblems verbessert werden?

Unbeschränkter Fall:  $C = (J_1^T J_1)^{-1}$  Konfidenzintervall  $\sqrt{c_{ii}}$ . Mehr Messungen durchführen:

$$C_+ = \left( \begin{pmatrix} J_1 \\ \dots \\ J_1 \end{pmatrix}^T \begin{pmatrix} J_1 \\ \dots \\ J_1 \end{pmatrix} \right)^{-1} = (m J_1^T J_1)^{-1} = \frac{1}{m} (J_1^T J_1)^{-1}$$

Die Konfidenzintervalle werden nur um den Faktor  $m^{-\frac{1}{2}}$  kleiner:  $\sqrt{c_{+ii}} = \sqrt{m^{-1} c_{ii}} = m^{-\frac{1}{2}} \sqrt{c_{ii}}$ .

Stattdessen: Entweder Experimente, die die statistische Unsicherheit der Parameterschätzung minimieren.

Versuchsplanungsgrößen:  $\xi = (q, u, w)$ :

- Steuergrößen  $q \in \mathbb{R}^{n_q}$
- Steuerfunktionen  $u: [t_0, t_{end}] \rightarrow \mathbb{R}^{n_u}$
- Auswahl der Messungen  $w \in \{0, 1\}^M$

$$w_i = \begin{cases} 1 & \text{Messung wird durchgeführt} \\ 0 & \text{Messung wird nicht durchgeführt} \end{cases}$$

Zielfunktion: Informationsfunktion auf der Kovarianzmatrix  $C$  ( $C$  ist positiv semidefinit, symmetrisch).

Üblich sind:

- $\Phi(C) = \sqrt[n]{\det C}$  D-Kriterium
- $\Phi(C) = \frac{1}{n} \text{Spur} C$  A-Kriterium
- $\Phi(C) = \|C\|_2 = \lambda_{\max}(C)$  E-Kriterium
- $\Phi(C) = \max_{i=1, \dots, n} \sqrt{c_{ii}}$  U-Kriterium

## 11.6

Statt auf  $C$  kann man die Kriterien auch auf  $K^T C K$ , die Projektion auf den  $n_k$ -dimensionalen Teilraum,  $K \in \mathbb{R}^{n \times n_k}$  vollrangig, anwenden. Insbesondere für das D-Kriterium, wenn  $C$  nicht positiv definit ist.

## Bemerkung 11.7

Im unbeschränkten Fall korrespondiert das D-Kriterium zum Volumen des Konfidenz-ellipsoids, das A-Kriterium zur Länge der durchschnittlichen Halbachsen, das E-Kriterium zur längsten Halbachsenlänge und das M-Kriterium zum größten Konfidenzintervall.

## Optimale Versuchsplanung

Entwirf Experimente, die die statistische Unsicherheit der Parameterschätzung minimieren.

Versuchsplanungsgrößen:  $\xi = (q, u, w)$

## Versuchsraum

Nebenbedingungen

- Steuerbeschränkungen:  $q_i \in [L_{q_i}, U_{q_i}]$ ,  $u_i(t) \in [L_{u_i}, U_{u_i}]$ ,  $L_c \leq c(q) \leq U_c$
- Zustandsbeschränkungen:  $L_b \leq b(t, y(t), p, q, u(t)) \leq U_b \quad \forall t \in [t_0, t_{end}]$ . Spezialfall:  $L_{y_i} \leq y_i(t) \leq U_{y_i}$ .

- Auswahl von Messungen: Sei  $I_j \subseteq \{1, \dots, M\}$  eine Teilmenge aller möglichen Messungen, z. B. alle Messungen an einem Zeitpunkt, alle Messungen mit einem bestimmten Verfahren.  $L_{I_j} \leq \sum_{i \in I_j} w_i \leq U_{I_j}$ .

- Kostenbeschränkung (lineare Kosten):

$$\sum_{i=1}^M c_i^w w_i + \sum_{i=1}^{n_q} c_i^q + \sum_{i=1}^{n_u} \int_{t_0}^{t_{end}} c_i^u(t) u(t) dt \leq C_{max}$$

- $c_i^w$ : Kosten pro Messung
- $c_i^q$ : Kosten pro Einheit Steuergröße
- $c_i^u$ : Kosten für Steuerfunktion

Insgesamt:

- Ungleichungsbeschränkungen:  $L \leq \Psi(t, y(t), p, q, u(t), w) \leq U$  (11.16)
- Gleichungsbeschränkungen:  $\mathcal{X}(t, y(t), p, q, u(t), w) = 0$  (11.17)
- Ganzzahligkeitsbedingungen:  $w_i \in \{0, 1\}$ ,  $i = 1, \dots, M$ .
- Variationsdifferentialgleichung für  $\frac{\partial y}{\partial v} =: G$

Versuchsplanungsproblem (11.19)

$$\begin{aligned} \min_{q, u, w} & \Phi \left( (I \quad 0) \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} I \\ = \end{pmatrix} \right) \\ \text{s. t. } & J_1 = \left( -\frac{\sqrt{w_i}}{\sigma_i} \left( \frac{\partial h_i}{\partial y} G(t_i) + \frac{\partial h_i}{\partial v} \right) \right)_{i=1, \dots, M} \\ & J_2 = \frac{\partial F_2}{\partial y} G + \frac{\partial F_2}{\partial v} \\ & \dot{y} = f(t, y, p, q, u), \quad y(t_0) = y_0 \\ & r(t_1, y(t_1), \dots, t_k, y(t_k), p, q) = 0 \\ & \dot{G} = \frac{\partial f}{\partial y} G + \frac{\partial f}{\partial v}, \quad G(t_0) = \frac{\partial y_0}{\partial v}(p, q) \\ & L \leq \Psi(t, y(t), p, q, u(t), w) \leq U \\ & 0 = \mathcal{X}(t, y(t), p, q, u(t), q) \\ & w_i \in \{0, 1\} \quad i = 1, \dots, M \end{aligned}$$

Eigenschaften:

- Optimalsteuerungsproblem
- mit ODE und VDE-Nebenbedingung
- Nicht-Standard-Zielfunktion: hochgradig nicht-separiert
- Ganzzahligkeitsbedingungen

## Varianten

- Mehrfachexperimente: Ein Experiment besteht aus einer Durchführung des Prozesses für gegebene Versuchsplanungsgrößen  $\xi = (q, u, w)$ . Oft benötigt man aber Daten aus mehreren Experimenten, um alle Parameter zu schätzen. Wir erlauben, dass die Experimente durch verschiedene Modelle beschrieben werden, diese müssen nur die Parameter gemeinsam haben.  $j = 1, \dots, N_{ex}$ . Pro Experiment:
  - Modellgleichungen:  $\dot{y}^j = f(t, y^j, p, q^j, w^j)$ ,  $y^j(t_0^j) = y_0^j(p, q^j)$ ,  $r^j(t_1^j, y^j(t_1^j), \dots, t_{kj}^j, y^j(t_{kj}^j), p$ ,  $0$  mit  $t \in [t_0^j, t_{end}^j]$
  - Messungen:  $\eta_i^j = h_i^j(t_i^j, y^j(t_i^j), p, q_i^j) + \varepsilon_i^j$ ,  $i = 1, \dots, M^j$ ,  $\varepsilon_i^j \sim \mathcal{N}\left(0, \frac{\sigma_i^{j2}}{w_i^j}\right)$ ,  $\text{cov}(\varepsilon_i^j, \varepsilon_k^l) = 0$ ,  $i \neq k$  oder  $j \neq l$
  -

Führt auf Mehrfachexperiment-Parameterschätz-Problem:

$$\begin{aligned} \min_{p, s^1, \dots, s^{N_{ex}}} \quad & \sum_{j=1}^{N_{ex}} \|F_1^j(p, s^j)\|_2^2 \quad (11.20) \\ \text{s. t. } \quad & F_2^j(p, s^j) = 0 \quad j = 1, \dots, N_{ex} \end{aligned}$$

Struktur.  $F^h$  hängt nur von  $s^j$  ab. Jacobi-Matrizen  $J_i$ ,  $i = 1, 2$ :

$$J_i = \frac{\partial F_i}{\partial(p, s)} = \begin{pmatrix} J_{ip}^1 & J_{is^1}^1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ J_{ip}^{N_{ex}} & 0 & 0 & \dots & J_{is^{N_{ex}}}^{N_{ex}} \end{pmatrix}$$

## Feste und variable Experimente

$$N_{ex} = N_{fix} + N_{var}$$

- $N_{fix}$ : Bereits durchgeführte Experimente, für die schon Messdaten vorliegen und deren Sensitivitäten in der Versuchsplanung als Vorinformationen berücksichtigt werden sollen.

- $N_{var}$ : neu zu planende Experimente, deren  $\xi$ -Variablen in der Versuchsplanung optimal bestimmt werden sollen.

$$\begin{aligned}
 J_i &= \begin{pmatrix} J_{ip}^{fix} & J_{is^{fix}}^{fix} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ J_{ip}^{var} & 0 & 0 & \cdots & J_{is^{var}}^{var} \end{pmatrix} \quad i = 1, 2 \\
 J_1^T J_1 &= \begin{pmatrix} H_{pp} & H_{s^1 p}^T & \cdots & H_{s^{N_{ex} p}}^T \\ H_{s^1 p} & H_{s^1 s^1} & & 0 \\ \vdots & & \ddots & \vdots \\ H_{s^{N_{ex} p}} & 0 & \cdots & H_{s^{N_{ex} s^{N_{ex}}}} \end{pmatrix} \\
 H_{pp} &= \sum_{j=1}^{N_{ex}} J_{1p}^j J_{1p}^j{}^T \\
 &= \underbrace{\sum_{j=1}^{N_{fix}} J_{1p}^j J_{1p}^j{}^T}_{=: H_{pp}^{fix} \text{ fest}} + \underbrace{\sum_{j=N_{fix}+1}^{N_{fix}+N_{var}} J_{1p}^j J_{1p}^j{}^T}_{\text{variabel}}
 \end{aligned}$$

### Mehrfachexperiment-VP-Problem

- Für die alten Experimente  $j = 1, \dots, N_{fix}$  sind die  $q^j, u^j, w^j$  Konstanten. Es müssen keine Steuer-, Zustands-, Ganzzahligkeits- oder Kostenconstraints berücksichtigt werden
- Für die neuen Experimente  $j = N_{fix} + 1, \dots, N_{fix} + N_{neu}$  sind die  $q^j, u^j, w^j$  Variablen und die Nebenbedingungen müssen berücksichtigt werden.

$$\begin{aligned}
 \min \Phi &\left( (I \quad 0) \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} \right) \\
 \text{s. t. } &J_1, J_2 \text{ wie oben}
 \end{aligned}$$

ODE-RWP für Experimente  $q, \dots, N_{ex}$ . VDE für Experimente  $1, \dots, N_{ex}$ , Ungleichungs- Gleichungs- und Ganzzahligkeitsbedingungen für Experimente  $N_{fix} + 1, \dots, N_{fix} + N_{var}$  (11.25)

### Bemerkung 11.8

Das Mehrfachexperiment-VP-Problem (11.25) minimiert die statistische Unsicherheit einer PS aus allen (alten und neuen) Experimenten. Dabei werden die neuen Experimente so ausgelegt, dass sie gegenüber den alten einen maximalen Zuwachs an statistischer Güte ergeben.

Sequentielle Vorgehensweise:

Wenn die Modellantworten  $h_i(t_i, y(t_i, p), p, q)$  nichtlinear in  $p$  sind, d. h.  $J_1 = J_1(p)$  hängt auch  $C$ , das VP-Problem und der optimale Versuchsplan von den aktuellen Werten von  $p$  ab.

Wir führen sukzessive und abwechselnd VP und PS durch:

- Startparametersatz
- Versuchsplanung
- Vorschlag für neue Experimente
- Versuchsdurchführung
- Messdaten
- Parameterschätzung
- Verbesserter Parametersatz
- Versuchsplanung
- ...

### Bemerkung 11.11

- in der Versuchsplanung werden Mehrfach-Experimente geplant unter Berücksichtigung der alten Experimente aus der vorherigen Iterationen und mit Optimierung ggf. neuer Experimente
- Die Parameterschätzung wird mit allen Daten aus allen bisherigen Iterationen durchgeführt.
- Die Versuchsplanung kann robustifiziert werden (s. u.). Der Robustifizierungsanteil wird automatisch zurückgefahren.
- Wenn nötig können in den Iterationen auch Modellmodifikationen vorgenommen werden.
- Man stoppt, wenn die Parameter hinreichend gut bestimmt sind, oder wenn das experimentelle Budget erschöpft ist.

## Robustifizierung der Versuchsplanung gegenüber Parameterunsicherheit

Wenn  $h$  nichtlinear in  $p$  ist, ist  $J = J(p)$  eine Funktion der Parameter und damit auch  $C = C(p)$ . Für schlecht bekannte Parameter können also auch schlechte Experimente berechnet werden. Unsicherheit der Parameter: (a-priori-Information)  $p \sim \mathcal{N}(p_0, \Sigma)$ ,  $\Sigma$  positiv definit mit Konfidenzgebiet  $\{p : \|p - p_0\|_{2, \Sigma^{-1}} \leq \gamma\}$  mit der Norm  $\|\cdot\|_{2, \Sigma^{-1}}$  aus dem Skalarprodukt  $\langle x, y \rangle = x^T \Sigma^{-1} y$ . Worst-Case-Design:

$$\min_{\xi \in \Omega} \max_{\|p - p_0\|_{2, \Sigma^{-1}} \leq \gamma} \Phi(C(\xi, p))$$

Das ist ein semi-infinites Optimierungsproblem. Die Lösung erfordert die Berechnung von globalen Optima von endlichdimensionalen Teilproblemen, das ist NP-schwer.

Vereinfachte Problemstellung: Taylorentwicklung nach  $p$  um  $p_0$ :

$$\min_{\xi \in \Omega} \max_{\|p - p_0\|_{2, \Sigma^{-1}} \leq \gamma} \Phi(C(\xi, p_0)) + \frac{\partial}{\partial p} \Phi(C(\xi, p_0))(p - p_0) \quad (11.27)$$

Das innere Problem kann jetzt explizit gelöst werden:

### Lemma 11.9

$$\begin{aligned} \max_{\|p - p_0\|_{2, \Sigma^{-1}} \leq \gamma} \Phi(C(\xi, p_0)) + \frac{\partial}{\partial p} \Phi(C(\xi, p_0))(p - p_0) \\ = \Phi(C(\xi, p_0)) + \gamma \left\| \frac{\partial}{\partial p} \Phi(C(\xi, p_0)) \right\|_{2, \Sigma} \end{aligned} \quad (11.28)$$

Beweis:

Sei  $\Delta p := p - p_0$ ,  $a := \frac{\partial}{\partial p} \Phi(C(\xi, p_0))^T$ . Das Minimum wird am Rand angenommen, betrachte

$$\min_{\|\Delta p\|_{2, \Sigma^{-1}}^2 = \gamma^2} a^T \Delta p$$

Mit der Lagrangefunktion

$$\begin{aligned}
 L(\Delta p, \lambda) &= a^T \Delta p + \lambda(\Delta p^T \Sigma^{-1} \Delta p - \gamma^2) \\
 \nabla_p L(\Delta p^*, \lambda) &= a + 2\lambda \Sigma^{-1} \Delta p^* = 0 \Rightarrow \Delta p^* = \frac{-\Sigma a}{2\lambda} \\
 \Delta_\lambda L(\Delta p^*, \lambda) &= \Delta p^{*T} \Sigma^{-1} \Delta p^* \gamma^2 = 0 \\
 &= \frac{a^T \Sigma a}{(2\lambda)^2} - \gamma^2 \\
 \Rightarrow \lambda &= \pm \frac{\|A\|_{2,\Sigma}}{2\gamma}
 \end{aligned}$$

Positive Lösung: Minimum. Maximum:

$$\Delta p^* = \gamma \frac{\Sigma a}{\|a\|_{2,\Sigma}}$$

Zielfunktionswert:

$$a^T \Delta p^* = \gamma \frac{a^T \Sigma a}{\|a\|_{2,\Sigma}} = \gamma \|a\|_{2,\Sigma}$$

Robustifiziertes Versuchsplanungsproblem:

$$\min_{\xi \in \Omega} \underbrace{\Phi(C(\xi, p_0))}_{\text{nicht robustes Gütekriterium}} + \gamma \underbrace{\left\| \frac{\partial}{\partial p} \Phi(C(\xi, p_0)) \right\|_{2,\Sigma}}_{\text{Sensitivität des Gütekriteriums nach den Parametern}}$$

Gewichtung durch das Konfidenzlevel der Parameterunsicherheit.

### Bemerkung 11.10

- Gegenüber der nichtrobusten Formulierung kommt eine zusätzliche Ableitung nach  $p$  hinzu.
- Parameterabhängige Nebenbedingungen müssen ebenfalls robustifiziert werden.

## 11.4 Numerische Methoden zur Lösung von Versuchsplanungsproblemen

- Direkter Ansatz der optimalen Steuerung
- Diskretisierung der Zustandsbeschränkungen
- Parametrisierung der Dynamik durch Single Shooting oder mit einer geeigneten Umformulierung durch Multiple Shooting:

$$\min \Phi \left( (I \quad 0) \begin{pmatrix} H_m J_{2m}^T \\ J_{2m} \quad 0 \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} \right) \text{ s. t. } H_j = H_{j-1} + \sum_{i:t_j \in [\tau_i, \tau_{i+1})} J_{1j}^T J_{j1} = H_{j-1} + \sum_{i:t_j \in [\tau_i, \tau_{i+1})}$$



+ Multiple-Shooting-Parametrisierung von ODE und VDE. Dann ist die ZF vom Mayer-Typ, das Problem kann auf Teilintervallen separiert ausgewertet werden. Aber: viele zusätzliche Variablen  $H_0, J_{2j}$ . Strukturen müssen ausgenutzt werden  $\Rightarrow$  Aktuelle Doktorarbeit Dennis Janka.

- Behandlung der Ganzzahligkeit:

0-1-Variablen.  $w_i \in \{0, 1\}$ ,  $i = 1, \dots, M$ . Relaxierung.  $w_i \in [0, 1]$ ,  $i = 1, \dots, M$   $\Rightarrow$  Lösungen liefern untere Schranken für die ganzzahlige Optimallösung. Obere Schranken erhält man durch Heuristiken, z. B. Runden der  $w_i$  auf zulässige ganze Werte, Beispiel: Sum-Up-Rounding. TODO: Bildchen. Addiere Gewichte solange auf, bis die Summe  $\geq 1$  wird und platziere dann eine Messung, fahre mit dem Rest und den weiteren Gewichten fort.

Beobachtung bei numerischen Ergebnissen: Die Lösungen der relaxierten Probleme sind schon vollständig oder beinahe 0-1. Erklärungsansätze:

- Exakte lineare Designs (Pukelsheim, Rieder)
- Bang-Bang-Prinzip der optimalen Steuerung (Sager 2009)

- Lösung des endlichdimensionalen NLP:

Benutze SQP-Verfahren, siehe Kapitel 7. Benötigte Ableitungen:

- Hessematrix der Lagrangefunktion durch Niedrigrang-Aufdatierungen, z. B. BFGS.
- Jacobi-Matrix der Nebenbedingungen
- Gradient der Zielfunktion

$$\begin{aligned}\Phi(C) &= \Phi \left( (I \ 0) M^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} \right) \\ M &= \begin{pmatrix} J_1^T J_1 & J_2^T \\ J_2 & 0 \end{pmatrix} \\ J_1 &= \left( -\frac{\sqrt{w_i}}{\sigma_i} \frac{\partial h_i}{\partial y} \frac{\partial y}{\partial v}(t_i) + \frac{\partial h_i}{\partial v} \right)_{i=1, \dots, M} \\ J_2 &= \frac{\partial F_2}{\partial y} \frac{\partial y}{\partial v} + \frac{\partial f_2}{\partial v}\end{aligned}$$

Gradient von  $\Phi$  nach  $\xi$ :

$$\frac{\partial \Phi}{\partial \xi} = \frac{\partial \Phi}{\partial C} \frac{\partial C}{\partial M} \frac{\partial M}{\partial J} \frac{\partial J}{\partial \xi} \quad (11.30)$$

$\frac{\partial \Phi}{\partial C}$ : Ableitung von Funktionen auf Matrizen

$$\begin{aligned} \frac{\partial \text{Sp} C}{\partial C} \Delta C &:= \lim_{h \rightarrow 0} \frac{\text{Sp}(C + h \Delta C) - \text{Sp} C}{h} \quad (11.31) \\ &= \text{Sp} \Delta C \end{aligned}$$

$$\frac{\partial \det C}{\partial C} \Delta C = \sum_{i,k} (\det C) (C^{-1})_{ik} \Delta C_{ik} \quad (11.32)$$

$$\frac{\partial \lambda_{\max}(C)}{\partial C} \Delta C = z^T \Delta C z$$

Wenn  $z$  auf 1 normierter EV zum einfachen EW  $\lambda_{\max}(C)$  ist.

$$\frac{\partial C}{\partial M} \Delta M = \begin{pmatrix} I & 0 \end{pmatrix} \frac{\partial M^{-1}}{\partial M} \Delta M \begin{pmatrix} I \\ 0 \end{pmatrix}$$

### Lemma 11.12: Ableitung der inversen Matrix

Sei  $A \in \mathbb{R}^{n \times n}$  invertierbar und  $\Delta A \in \mathbb{R}^{n \times n}$ . Dann ist

$$\frac{\partial A^{-1}}{\partial A} \Delta A = -A^{-1} \Delta A A^{-1} \quad (11.34)$$

Beweis: Sei  $X = A^{-1}$ . Es gilt:  $XA = I$ . abgeleitet:

$$\begin{aligned} \frac{\partial X}{\partial A} \Delta A \cdot A + X \frac{\partial A}{\partial A} \Delta A &= 0 \\ \Rightarrow \frac{\partial X}{\partial A} \Delta A &= -X \Delta A A^{-1} = -A^{-1} \Delta A A^{-1} \\ \frac{\partial C}{\partial M} \Delta M &= - \begin{pmatrix} I & 0 \end{pmatrix} M^{-1} \Delta M M^{-1} \begin{pmatrix} I \\ 0 \end{pmatrix} \\ \frac{\partial M}{\partial J} \Delta J &= \begin{pmatrix} J_1^T \Delta J_1 + \Delta J_1^T J_1 & \Delta J_2^T \\ \Delta J_2^T & 0 \end{pmatrix} \frac{\partial J_1}{\partial \xi} \end{aligned}$$

ableiten nach  $w, q$  benötigt insbesondere  $\frac{\partial}{\partial q} \frac{\partial y}{\partial p} =: GAQ$  Lösung einer VDE zweiter Ordnung:

$$\begin{aligned} \dot{y} &= f(t, y, p, q) \\ y(t_0) &= y_0(p, q) \quad (11.35) \end{aligned}$$

Ableiten nach  $p$ :  $G := \frac{\partial y_0}{\partial p}$   
 Ableiten nach  $q$ :

$$GAQ := \frac{\partial^2 y}{\partial q \partial p}$$

$$GAQ(t_0) = \frac{\partial^2 y_0}{\partial q \partial p}$$

$$G\dot{A}Q = \frac{\partial f}{\partial y} GAQ + \left( \frac{\partial^2 f}{\partial y \partial p} GQ + \frac{\partial^2 f}{\partial q \partial y} \right) G + \left( \frac{\partial^2 f}{\partial y \partial p} GQ + \frac{\partial^2 f}{\partial y \partial p} \right)$$

Ableitung nach  $q$ :

$$GQ = \frac{\partial y}{\partial q}$$

$$\dot{G}q = \frac{\partial f}{\partial y} GQ + \frac{\partial f}{\partial q}$$

$$GQ(t_0) = \frac{\partial y_0}{\partial q}$$

Müssen als System für  $y, G, GQ, GAQ$  gelöst werden.  
 TODO: Fehlender Text

## 5. Ausnutzung der Mehrfachexperimentstruktur

### Anwendungsbeispiel: Ergebnisse

Urethan-Reaktion, siehe 11.1

a) Intuitive Versuchsplanung (Chemiker, BASF): Batch-Versuche, isotherm, „Trennung der Reaktion“, 15 Experimente, 90 Messungen

Konfidenzintervalle nach Parameterschätzung:

- $k_{ref1} \pm 0.93\%$
- $E_{a1} \pm 0.88\%$
- $k_{ref2} \pm 0.72\%$
- $E_{a2} \pm 0.26\%$
- $k_{ref4} \pm 0.01\%$
- $DH_2 \pm 32.96\%$
- $KL_2 \pm 22.4\%$

## *Inhaltsverzeichnis*

b) Optimale Versuchsplanung mit VPLAN:

Beschränkungen an die Steuergrößen, maximal 16 Messungen pro Experiment. Tag-Nacht-Schichtrhythmus 80h. Nachts keine Messungen und keine Änderungen der Steuerungen.

→ 2 Experimente sequentiell  $\rho$  alle Parameter mit Standardabweichung  $< 1\%$