Analysis und Numerik für Algebro-Differentialgleichungen[1]

Roswitha März Renate Winkler
Humboldt-Universität Berlin
Institut für Mathematik

23. Mai 2001

[1]nach einer Vorlesung von Prof.R. März im WS 97/98 und SS 00

Inhaltsverzeichnis

Kapitel 0

Einführung

Wir betrachten quasilineare Algebro-Differentialgleichungen (ADGln) der Form

$$A(x(t),t)x'(t) + b(x(t),t) = 0$$

wobei $A(x(t),t)$ singulär ist, so dass die ADGl nicht nach $x'(t)$ aufgelöst werden kann.

Spezialfälle:

- lineare ADGln $\qquad A(t)x'(t) + B(t)x(t) = q(t)$
- lineare ADGln mit konst. Koeff. $\qquad A\,x'(t) + B\,x(t) = q(t)$
- semiexplizite ADGln z.B. als Grenzwert singulär gestörter DGln
$$\begin{aligned} u'(t) &= \varphi(u(t),v(t),t) \\ 0 &= \psi(u(t),v(t),t) \end{aligned}$$
- Steuerprobleme z. B. mit $u \in \mathbb{R}^m$, einer Steuerfunktion $v \in \mathbb{R}$ und dem Ziel $\quad u_m(t) = g(t)$
$$\begin{aligned} u'(t) &= \varphi(u(t),v(t),t) \\ 0 &= u_m(t) - g(t) \end{aligned}$$

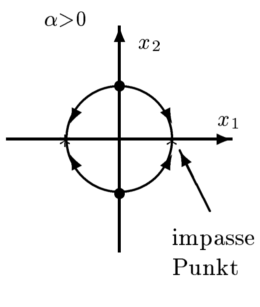
Herkunft z. B.

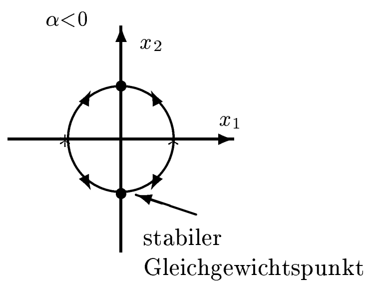
- Netzwerkmodellierung
- mechanische Systeme (Euler-Lagrange-Gleichungen)
- chemische Kinetik (Grenzfall singulär gestörter Differentialgleichungen)

Kleine Beispiele:

Beispiel 1:

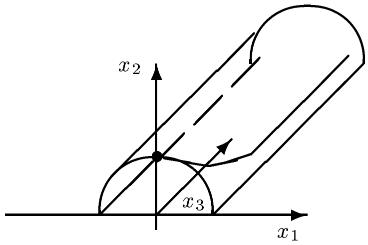
$$\begin{aligned} x_1' &= \alpha x_1 \\ 0 &= x_1^2 + x_2^2 - 1 \end{aligned}$$





Beispiel 2:

$$\begin{aligned} x_1' &= \alpha x_1 \\ x_2' &= x_3/x_2 \\ 0 &= x_1^2 + x_2^2 - 1 \end{aligned}$$



mit $x_2 > 0$

Wenn wir annehmen, dass $x(t)$ eine Lösung ist, folgt

$$\begin{aligned} & 0 &\equiv\ & x_1(t)^2 + x_2(t)^2 - 1 \\ \Rightarrow\ & 0 &=\ & 2x_1(t)x_1'(t) + 2x_2(t)x_2'(t) \\ & &=\ & 2x_1 \cdot \alpha x_1 + 2x_2 \cdot x_3/x_2 \\ \Rightarrow\ & x_3 &=\ & -\alpha x_1^2 \end{aligned}$$

Es gibt also die versteckte Nebenbedingung $x_3 = -\alpha x_1^2$ und jede Lösung von Beispiel 2 ist Lösung von

$$\begin{aligned} x_1' &= \alpha x_1 \\ 0 &= x_3 + \alpha x_1^2 \\ 0 &= x_1^2 + x_2^2 - 1 \end{aligned}$$

Beispiel 3:

$$\begin{aligned} x_1' &= (1 + x_3)x_1 \\ x_2' &= -x_3 x_2 \\ 0 &= x_1 x_2 \end{aligned}$$

Wenn wir annehmen, dass $x(t)$ eine Lösung ist, folgt

$$\begin{aligned} & 0 &\equiv\ & x_1(t)x_2(t) \\ \Rightarrow\ & 0 &=\ & x_1'x_2 + x_1x_2' \\ & &=\ & (1 + x_3)x_1x_2 - x_1x_3x_2 \\ & &=\ & x_1x_2 \end{aligned}$$

Hier ergibt sich keine neue Bedingung. Für $x_1x_2 = 0$ ist x_3 durch die Gleichungen nicht bestimmt.

Beispiel 4:

$$\begin{aligned} x_1' + x_2' &= |t| + 1 \\ (x_1' + x_2')x_1 &= t^2 + 3|t| + 1 \end{aligned}$$

Umstellen nach x_1 liefert $x_1(t) = \frac{t^2+3|t|+1}{|t|+1}$.

Damit ist x_1 in $t = 0$ nicht differenzierbar.

Schreibt man$(x_1 + x_2)'$ statt $x_1' + x_2'$ erhält man dennoch die wohldefinierte Lösung

$$(x_1 + x_2)(t) = (x_1 + x_2)(t_0) + \int_{t_0}^{t} (|s| + 1)ds \quad , \quad x_2 = (x_1 + x_2) - x_1 \quad .$$

Beispiel 5:

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{A} x' + \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}}_{B} x = q \quad \hat{=} \quad \begin{array}{ccccc} x_1' & + & x_1 & = & q_1 \\ & & x_1 & = & q_2 \end{array}$$

Hier ist x_1 überbestimmt und x_2 nicht bestimmt. Solcherart Aufgaben sollte man in der Modellierung vermeiden. Wir beobachten, dass die Matrix

$$\lambda A + B = \begin{pmatrix} \lambda + 1 & 0 \\ 1 & 0 \end{pmatrix} \qquad \text{für alle} \lambda \in \mathbb{C} \text{ singulär ist.}$$

Beispiel 6:

$$\underbrace{\begin{pmatrix} 0 & 1 & \\ & 0 & 1 \\ & 0 & 0 \end{pmatrix}}_{A} x' + \underbrace{I}_{B} x = q \quad \hat{=} \quad \begin{array}{ccccc} x_2' & + & x_1 & = & q_1 \\ x_3' & + & x_2 & = & q_2 \\ & & x_3 & = & q_3 \end{array}$$

Hier ergibt sich als Lösung $\quad x_3 = q_3, \quad x_2 = q_2 - q_3', \quad x_1 = q_1 - (q_2 - q_3')'.$
Um diese Lösung in Abhängigkeit von der rechten Seite q zu bestimmen, müssen Teile der rechten Seite sogar zweimal differenziert werden. Für keine Komponente können Anfangswerte frei vorgegeben werden.
Hier ist jedoch die Matrix

$$\lambda A + B = \begin{pmatrix} 1 & \lambda & 0 \\ 0 & 1 & \lambda \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{für alle} \lambda \in \mathbb{C} \text{ regulär .}$$

Die Art, wie man Anfangswerte vorgeben kann, hängt vom Gleichungstyp ab.

Wir unternehmen nun einen ersten Versuch, ein Integrationsverfahren für ADGln zu finden, und betrachten das explizite Eulerverfahren angewendet auf $Ax' + Bx = q$:

$$\begin{array}{rll} & A \cdot \frac{x_\ell - x_{\ell-1}}{h} + Bx_{\ell-1} & = q_{\ell-1} \\ \rightsquigarrow & \frac{1}{h} A \cdot x_\ell & = (\frac{1}{h} A - B) x_{\ell-1} + q_{\ell-1} \end{array}$$

Dies ist <u>nicht</u> nach x_ℓ auflösbar. Das explizite Eulerverfahren ist daher <u>nicht</u> durchführbar.

Dagegen gibt es gute Erfahrungen mit dem implizitem Eulerverfahren und der BDF (Gear 71) aus extensiven Anwendungen für die Chip-Entwicklung in der Elektronik-Industrie.
Das implizite Eulerverfahren führt auf:

$$\begin{array}{rll} & A \cdot \frac{x_\ell - x_{\ell-1}}{h} + Bx_\ell & = q_\ell \\ \rightsquigarrow & (\frac{1}{h} A + B) x_\ell & = \frac{1}{h} A x_{\ell-1} + q_\ell \end{array}$$

Die Matrix $(\frac{1}{h} A + B)$ sollte für hinreichend kleine h regulär sein, damit das Verfahren durchführbar ist.

Bemerkung:

Selbst bei regulären impliziten Differentialgleichungen ($A(x,t)$ regulär) gibt es gute Gründe, die Matrix $A(x,t)$ nicht zu invertieren:

- in Anwendungen hat diese Matrix oft eine spezielle, schwach besetzte Struktur, die beim Berechnen der Inversen nicht gewahrt würde,
- das Lösen eines Gleichungssystems verlangt nur eine LU-Zerlegung und eine Rückelimination, was weniger aufwändig als die Berechnung der Inversen ist,
- für die Integration steifer Differentialgleichungen sind sowieso implizite Verfahren angezeigt.

Beispiel 7:

$$\underbrace{\begin{pmatrix} \delta-1 & \delta t \\ 0 & 0 \end{pmatrix}}_{A(t)} x'(t) + \mu \underbrace{\begin{pmatrix} \delta-1 & \delta t \\ \delta-1 & \delta t-1 \end{pmatrix}}_{B(t)} x(t) = 0, \quad \delta \neq 1, \quad \mu \neq 0$$

$$\hat{=} \quad \begin{array}{rcrcl} (\delta-1)x_1' + \delta t x_2' & + & \mu(\delta-1)x_1 + \mu\delta t x_2 & = & 0 \\ & & \mu(\delta-1)x_1 + \mu(\delta t-1)x_2 & = & 0 \end{array}$$

Aus der zweiten Gleichung folgt

$$x_1(t) = \frac{1-\delta t}{\delta-1} x_2(t) \quad ,$$

eingesetzt in die erste Gleichung ergibt sich folgende explizite Differentialgleichung in x_2 :

$$x_2'(t) = (\delta-\mu)x_2(t) \quad .$$

Erfreulich wäre es, wenn die Ergebnisse des impliziten Eulerverfahrens angewandt auf die ADGl mit denen bei Anwendung des impliziten Eulerverfahrens auf die herausgearbeitete Differentialgleichung in x_2 und der Zuweisung für x_1 übereinstimmen würden. Doch wie sieht es in diesem Beispiel aus?
Das implizite Eulerverfahren mit konstanter Schrittweite h angewandt auf die (lineare, zeitabhängige, homogene) ADGl führt zu

$$\begin{array}{rcl} A(t_\ell) \cdot \dfrac{x_\ell - x_{\ell-1}}{h} + B(t_\ell)x_\ell & = & 0 \qquad \text{bzw.} \\ (\dfrac{1}{h}A(t_\ell) + B(t_\ell))x_\ell & = & \dfrac{1}{h}A(t_\ell)x_{\ell-1} \quad . \end{array}$$

Hier ist

$$\frac{1}{h}A(t_\ell) + B(t_\ell) = \begin{pmatrix} \mu(\delta-1) + \frac{1}{h}(\delta-1) & \frac{1}{h}\delta t_\ell + \mu\delta t_\ell \\ \mu(\delta-1) & \mu(\delta t_\ell - 1) \end{pmatrix}$$

und nach weiterer Rechnung erhält man als Ergebnis :

$$x_{\ell,1} = \frac{1-\delta t_\ell}{\delta-1} x_{\ell,2} \quad , \quad x_{\ell,2} = \frac{1+h\delta}{1+h\mu} x_{\ell-1,2} \quad .$$

Nun können wir die asymptotischen Eigenschaften der exakten und der numerischen Approximation vergleichen. Aus der hergeleiteten Differentialgleichung für x_2 folgt, dass die exakte Lösung für Parameterkonstellationen mit $\delta - \mu < 0$ monoton fällt und für $t \to \infty$ gegen 0 konvergiert. Da das implizite Eulerverfahren (bei Anwendung auf explizite gewöhnliche Differentialgleichungen) A-stabil ist, würde das implizite Eulerverfahren bei Anwendung auf die hergeleitete explizite Differentialgleichung für x_2 unabhängig von der Größe der Schrittweite Approximationen mit derselben qualitativen Eigenschaft liefern. Doch wir erhalten Ergebnisse, bei denen nur gilt

$$x_{\ell,2} \to 0 \quad , \text{ falls } \quad |\frac{1+h\delta}{1+h\mu}| < 1 \quad .$$

Anliegen dieser Vorlesung sind:

- Formulierung von Anfangswertaufgaben, so dass sie eindeutig lösbar sind.
- Übertragung von Eigenschaften des Flusses, die wir von den regulären Differentialgleichungen kennen, wie etwa die Stabilität im Sinne von Lyapunov.
- Analyse numerischer Integrationsverfahren.
  Gilt alles das, was wir aus der Theorie der Integrationsverfahren für reguläre Differentialgleichungen kennen, auch bei ADGln? (einerseits Genauigkeit der Integration, andererseits Verhalten auf langen Intervallen)

Kapitel 1

Grundlagen aus der linearen Algebra

Ein wichtiges Hilfsmittel für die Analyse von ADGln werden Projektoren sein. Wir stellen daher Definitionen und wichtige Eigenschaften hier vor:

- Eine Matrix $Q \in L(\mathbb{R}^m)$ heißt Projektor oder Projektormatrix falls gilt $Q^2 = Q$.
- Ein Projektor $Q \in L(\mathbb{R}^m)$ zerlegt den Raum $\mathbb{R}^m$ in die direkte Summe $\mathbb{R}^m = \operatorname{im} Q \oplus \ker Q$.
- Für eine gegebene Zerlegung des Raumes $\mathbb{R}^m$ als direkte Summe zweier Unterräume $\mathbb{R}^m = N \oplus S$ existiert genau ein Projektor Q mit $\operatorname{im} Q = N$ und $\ker Q = S$. Wir sagen dann, Q sei der Projektor auf N längs S.
- Falls Q und $\bar{Q}$ zwei Projektoren auf den selben Unterraum N sind (d.h. $\operatorname{im} Q = \operatorname{im} \bar{Q} = N$) , so gilt $Q\bar{Q} = \bar{Q}$ und $\bar{Q}Q = Q$.

  Für die komplementären Projektoren $P := I - Q$ und $\bar{P} := I - \bar{Q}$ längs N (d.h. $\ker P = \ker \bar{P} = N$) gilt $P\bar{P} = P$ und $\bar{P}P = \bar{P}$.
- Ist $A \in L(\mathbb{R}^m)$ und $N := \ker A$ so gilt für jeden Projektor Q auf N, dass $AQ = 0$ und $AP = A$.
- Das Spektrum eines Projektors Q besteht nur aus den Eigenwerten 0 mit dem invarianten Unterraum $\ker Q$ und 1 mit dem invarianten Unterraum $\operatorname{im} Q$.
- Die Projektion erfolgt orthogonal gdw. der Projektor eine symmetrische Matrix ist. (Das ist nicht zu verwechseln mit orthogonalen Matrizen, die immer regulär sind und deren Eigenwerte alle den Betrag 1 haben.)

Nun wenden wir uns geordneten Paaren von Matrizen A, B zu. Zu einem geordneten Matrixpaar $\{A, B\}$ heißt die Menge der Matrizen $\{\lambda A + B : \lambda \in \mathbb{C}\}$ Matrixbüschel (im Englischen: matrix pencil) .

Definition: *Das geordnete Matrixpaar (Matrixbüschel) $\{A, B\}$ heißt regulär, falls $\det(\lambda A + B) \not\equiv 0$. Die Menge $\sigma(A, B) := \{\lambda \in \mathbb{C} : det(\lambda A + B) = 0\}$ heißt (finites) Spektrum des Matrixpaares bzw. Matrixbüschels.*

$\{A, B\}$ regulär ist hinreichend dafür, dass $(\frac{1}{h}A + B)$ für hinreichend kleine h regulär ist, da, falls $\{A, B\}$ regulär ist, das Polynom $\det(\lambda A + B)$ höchstens m Nullstellen hat.

Jedes reguläre Matrixpaar $\{A, B\}$ kann in (Weierstrass-)Kronecker-Normalform überführt werden, d. h. es existieren reguläre Matrizen E, F mit

$$\bar{A} := EAF = \begin{pmatrix} I & \\ & J \end{pmatrix}, \quad \bar{B} := EBF = \begin{pmatrix} W & \\ & I \end{pmatrix},$$

wobei J eine nil-potente Jordan-Block-Matrix (zum Eigenwert 0)ist.
Es sei $\mu \in I\!N$ mit $J^\mu = 0,\ J^{\mu-1} \neq 0$ (d.h. $\mu = \operatorname{ind}(J)$).

Der Beweis für reguläre Büschel stammt von Weierstrass, Kronecker hat dies auf möglicherweise singuläre Büschel ausgedehnt. Den Beweis findet man z.B. im Anhang des Buches von März/Griepentrog(86) oder Gantmacher(66).

Definition: *Für jedes reguläre Matrixpaar $\{A, B\}$ heißt $\mu = \operatorname{ind}\{A, B\} := \operatorname{ind}(J)$ Index des Paares $\{A, B\}$. Für lineare ADGln mit konstanten Koeffizienten A, B heißt $\mu = \operatorname{ind}\{A, B\}$ auch Kronecker-index der ADGl.*

Betrachten wir wieder die lineare ADGl mit konstanten Koeffizienten

$$Ax' + Bx = q \quad .$$

Falls das Matrixpaar $\{A, B\}$ singulär ist, hat die homogene Anfangswertaufgabe $Ax' + Bx = 0 \quad , \quad x(0) = 0$ ein Kontinuum von Lösungen. Wir werden daher im folgenden nur solche Aufgaben untersuchen, bei denen $\{A, B\}$ regulär ist. Dann ist $Ax' + Bx = q$ äquivalent zu

$$\underbrace{EAF}_{\begin{pmatrix} I & \\ & J \end{pmatrix}} \underbrace{F^{-1}x'}_{\bar{x}'} + \underbrace{EBF}_{\begin{pmatrix} W & \\ & I \end{pmatrix}} \underbrace{F^{-1}x}_{\bar{x}} = \underbrace{Eq}_{\bar{q}}$$

Zerlegen wir nun

$$F^{-1}x = \bar{x} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} \quad \text{und} \quad Eq = \bar{q} = \begin{pmatrix} \bar{r} \\ \bar{s} \end{pmatrix} \quad ,$$

so können wir die äquivalente ADGl schreiben als

$$\begin{array}{ccccc} \bar{u}' & + & W\bar{u} & = & \bar{r} \quad , \\ J\bar{v}' & + & \bar{v} & = & \bar{s} \quad . \end{array}$$

Im ersten Teil erkennen wir eine explizite gewöhnliche Differentialgleichung in $\bar{u}$. Die Lösbarkeitseigenschaften des zweiten Teils hängen vom Index der ADGl ab:

- Für $\mu = 1$ ist $J = 0$ und wir erhalten als Lösung $\bar{v} = \bar{s}$.
- Für $\mu = 2$ gibt es in J mindestens ein Jordankästchen der Länge 2. Es ist $J^2 = 0$, aber $J \neq 0$. Nach Multiplikation mit J erhalten wir

$$J^2\bar{v}' + J\bar{v} = J\bar{s} \quad \text{und damit} \quad J\bar{v} = J\bar{s} \quad \text{und nach Differentiation}$$

$$(J\bar{v})' = (J\bar{s})' \quad \text{und schließlich} \quad \bar{v} = \bar{s} - (J\bar{s})' \quad .$$

- Für allgemeinere Fälle verweisen wir z.B. auf März(92).

Untersuchen wir speziell die homogene ADGl $Ax' + Bx = 0$ so ergibt sich unabhängig vom Index für die Lösung $\bar{v} = 0$ und insgesamt

$$\begin{array}{rcl} \bar{u}' + W\bar{u} &=& 0 \quad , \\ \bar{v} &=& 0 \quad . \end{array}$$

Dann ist $x = F\bar{x} = F \begin{pmatrix} \bar{u} \\ 0 \end{pmatrix} \in \operatorname{im} F \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} F^{-1}$ (dies ist der finite Eigenraum) .

Andererseits müssen Lösungen der homogenen ADGl auch mindestens die (offensichtlichen) Nebenbedingungen erfüllen. Für allgemeine Matrizen A lassen sich diese als $Bx(\cdot) \in \operatorname{im} A$ formulieren. Bezeichnen wir den Teilraum von $\mathbb{R}^m$, wo dies erfüllt ist, mit

$$S := \{z \in \mathbb{R}^m : Bz \in \operatorname{im} A\} \quad ,$$

so folgt als Bedingung an die Lösung, dass gilt $x(\cdot) \in S$.

Nur für $\operatorname{ind}\{A, B\} = 1$ stimmt der Raum S mit dem finiten Eigenraum überein. Im allgemeinen ist der finite Eigenraum nur ein Teilraum von S .

Ein Analogon der Weierstrass-Kronecker-Normalform gibt es leider im allgemeinen nicht einmal für lineare zeitabhängige ADGln, und noch weniger für nichtlineare ADGln. Dennoch bleibt unser Ziel, Lösungsanteile analog $\bar{u}$, $\bar{v}$ und, noch detaillierter, diejenigen Anteile von $\bar{v}$, die erst nach einem oder mehreren Differentiationsschritten bestimmt werden können, herauszukristallisieren. Wir nutzen dazu relevante Unterräume, die auch für allgemeinere ADGln sinnvoll definiert werden können.

Lemma 1.1 *Seien* $A, B \in L(\mathbb{R}^m), N := \ker A$; $S := \{z \in \mathbb{R}^m : Bz \in$ *im* $A\}$.
Dann sind folgende Aussagen äquivalent:

(i) $N \cap S = \{0\}$

(ii) $N \oplus S = \mathbb{R}^m$

(iii) $A + BQ$ *ist regulär für jeden Projektor* $Q \in L(\mathbb{R}^m)$ *auf* N,
(d. h. $Q^2 = Q$, *im* $Q = N$*).*

(iv) $\{A, B\}$ *regulär mit* $\operatorname{ind}\{A, B\} = 1$.

(v) $A + WB$ *ist regulär für jeden Projektor* W *längs* $\operatorname{im} A$.

Beweis: Wir zeigen: (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii). Aus (ii) folgt offensichtlich (i) , so dass dann die Äquivalenz der ersten drei Aussagen folgt. Weiter wird gezeigt (i) $\Leftrightarrow$ (iv). Den Beweis der Äquivalenz (i) $\Leftrightarrow$ (v) lassen wir als Übungsaufgabe.

(i) $\Rightarrow$ (iii) : Sei $N \cap S = \{0\}$. Wir zeigen , dass dann $A + BQ$ injektiv ist.
$(A + BQ)z = 0 \;\Rightarrow\; Az + BQz = 0 \;\Rightarrow\; Qz \in S$ nach Definition von S .
Weiter gilt aber $Qz \in N$ nach Definition von Q. $\Rightarrow Qz \in N \cap S \Rightarrow Qz = 0$.
$\Rightarrow Az = 0 \Rightarrow z \in \ker A = N \Rightarrow z = Qz = 0$.

(iii) $\Rightarrow$ (ii) : Sei $A+BQ$ regulär. Wir betrachten

$$Q_S := Q(A+BQ)^{-1}B$$

und zeigen, dass dies ein Projektor auf N längs S ist.

a) Q_S ist ein Projektor, denn

$$Q_S^2 \;=\; Q(A+BQ)^{-1} \underbrace{BQ}_{=(A+BQ)Q} (A+BQ)^{-1}B \;=\; Q_S\,.$$

b) Es ist $\operatorname{im} Q_S = N$, denn

b1) nach Konstruktion ist $\operatorname{im} Q_S \subseteq \operatorname{im} Q = N$.

b2) es ist $N \subseteq \operatorname{im} Q_S$, denn für $z \in N$ gilt

$$z = Qz = Q^2 z = Q(A+BQ)^{-1}(A+BQ)Qz = Q(A+BQ)^{-1}BQz = Q_S z \quad .$$

c) Es ist $\ker Q_S = S$, denn

c1) Es ist $S \subseteq \ker Q_S$, denn denn für $z \in S$ gilt $Bz \in \operatorname{im} A$,
d.h. $\exists\, w : Bz = Aw$ und es ist $z \in \ker Q_S$,da

$$Q_S z \;=\; Q(A+BQ)^{-1}Bz \;=\; Q(A+BQ)^{-1} \underbrace{A}_{(A+BQ)P} w \;=\; QPw \;=\; 0\,.$$

c2) Es ist $\ker Q_S \subseteq S$, denn für $z \in \ker Q_S$ ist

$$0 \;=\; Q_S z \;=\; Q(A+BQ)^{-1}Bz \;=\; (I-P)(A+BQ)^{-1}Bz$$

$$\Rightarrow \quad (A+BQ)^{-1}Bz \;=\; P(A+BQ)^{-1}Bz$$

$$\Rightarrow \quad Bz \;=\; \underbrace{(A+BQ)P}_{=A}(A+BQ)^{-1}Bz \in \operatorname{im} A$$

(i) $\Leftrightarrow$ (iv) : Zunächst zeigen wir, dass aus (i) die Regularität des Matrixpaares folgt. Im Extremfall $A=0$ gilt einerseits $det(\lambda A + B) = detB \not\equiv 0$ gdw. B regulär , und andererseits $N \cap S = \mathbb{R}^m \cap \ker B = \ker B = \{0\}$ gdw. B regulär, so dass die Behauptung folgt.

Sei nun o.B.d.A. $\operatorname{rank} A = r > 0$. Dann existieren reguläre Matrizen V, W , so dass gilt $VAW = \begin{pmatrix} I_r & \\ & 0 \end{pmatrix}$ (womit man die AGDl auf semiexplizite Form transformieren kann).

Mit $\quad VBW =: C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ folgt dann einerseits

$$\begin{aligned}
N \cap S &= \{z \in \mathbb{R}^m : Az = 0\,,\; Bz \in \operatorname{im} A\} \\
&= \{Wy \in \mathbb{R}^m : VAWy = 0\,,\; VBWy \in \operatorname{im} VAW\} \\
&= \{Wy \in \mathbb{R}^m : \begin{pmatrix} I_r & \\ & 0 \end{pmatrix} y = 0\,,\; Cy \in \operatorname{im} \begin{pmatrix} I_r & \\ & 0 \end{pmatrix}\} \\
&= \{Wy \in \mathbb{R}^m : y_1 = 0\,,\; C_{21}y_1 + C_{22}y_2 = 0\} \\
&= \{0\} \quad \Leftrightarrow \quad C_{22} \text{ regulär,}
\end{aligned}$$

d.h. es gilt (i) gdw C_{22} regulär, und andererseits

$$
\begin{aligned}
&\det(\lambda A + B) \not\equiv 0 \\
\Leftrightarrow \quad & \det V(\lambda A + B)W \;=\; \det(\begin{pmatrix} \lambda I_r & \\ & 0 \end{pmatrix} + C) \;=\; \det \begin{pmatrix} \lambda I_r + C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \\
&= \lambda^r \det \begin{pmatrix} I_r & C_{11} \\ 0 & C_{22} \end{pmatrix} + p_1(\lambda) \;=\; \lambda^r \det C_{22} + p_1(\lambda) \;\not\equiv 0,
\end{aligned}
$$

wobei p_1 ein Polynom mit Grad kleiner r ist.
Aus (i) folgt nach obigem die Regularität von C_{22}, was garantiert, dass der Koeffizient vor λ^r nicht verschwindet. Damit ist gezeigt, dass aus (i) die Regularität des Matrixpaares $\{A\,,\,B\}$ folgt. .
Weiter betrachten wir nun der Einfachheit halber den Fall, dass $\{A, B\}$ in Kronecker-Normalform vorliegen, d.h. $A = \begin{pmatrix} I & \\ & J \end{pmatrix}$, $B = \begin{pmatrix} W & \\ & I \end{pmatrix}$. Dann berechnet man

$$
\begin{aligned}
N \;&=\; \{z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 = 0\,,\, Jz_2 = 0\} \\
S \;&=\; \{z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_2 \in \operatorname{im} J\} \\
N \cap S \;&=\; \{z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 = 0\,,\, z_2 \in \ker J \cap \operatorname{im} J\} \;=\; \{0\}^{m_1} \times \ker J \cap \operatorname{im} J \\
&=\; 0 \quad \Leftrightarrow \quad \operatorname{ind} J = 1
\end{aligned}
$$

□

Corollar 1.2 *Seien A, B, N, S wie in Lemma 1.1, $N \cap S = \{0\}$, und sei Q ein beliebiger Projektor auf N.*
Dann ist $Q_S := Q(A + BQ)^{-1}B$ der Projektor auf N längs S.

Q_S heißt auch <u>kanonischer Projektor auf N</u> .

Im weiteren werden wir zeitabhängige Unterräume $N(t)$ zulassen.
Definition: *Ein zeitabhängiger Unterraum $N(t) \subseteq \mathbb{R}^m$ heißt <u>glatt</u> ($N \in C^1$)*
für $t \in \mathcal{I}$, ($\mathcal{I} \subseteq \mathbb{R}$ ein Intervall), falls er durch eine Basis von C^1-Funktionen aufgespannt wird, d.h. falls

$$N(t) \;=\; \operatorname{span}\{n_1(t), \ldots, n_d(t)\}\,,\; t \in \mathcal{I},\; \textit{mit } n_i \in C^1(\mathcal{I}, \mathbb{R}^m)\,,\; i = 1, \ldots, d\,.$$

Mit Hilfe einer Basis von $N(t)$ kann man den Orthoprojektor auf $N(t)$ beschreiben:

Wir bilden $F(\cdot) = [n_1(\cdot), \ldots, n_d(\cdot)] \in L(\mathbb{R}^d, \mathbb{R}^m)$
Dann ist $Q^\perp(t) := F(t) \left\{ F(t)^T F(t) \right\}^{-1} F(t)^T$ der Orthoprojektor auf $N(t)$.

(Es gilt nach Konstruktion erstens, dass $Q^\perp(t)^2 = Q^\perp(t)$, womit folgt, dass $Q^\perp(t)$ ein Projektor ist, und zweitens, dass $Q^\perp(t)$ symmetrisch ist, woraus wegen $(\ker Q^\perp)^\perp = \mathrm{im}\, Q^{\perp T} = \mathrm{im}\, Q^\perp$ folgt, dass die Projektion orthogonal ist.)

Falls $N(\cdot) \in C^1$, ist offensichtlich auch $Q^\perp(\cdot) \in C^1$.

Umgekehrt kann man eine glatte Basis auf $N(\cdot)$ finden, falls es einen glatten Projektor $Q(\cdot) \in C^1$ auf $N(\cdot)$ gibt:

Denn sei $t_* \in \mathcal{I}$, $N(t_*) = \mathrm{span}\, \{n_1^*, \ldots, n_{m-r}^*\}$.
Nun betrachten wir die Anfangswertaufgaben (AWAn):

$$\left\{ \begin{array}{lcll} n'(t) & = & Q'(t)\; n(t), & t \in \mathcal{I} \\ n(t_*) & = & n_j^* & j = 1, \ldots, m-r \end{array} \right.$$

und zeigen: Die Lösungen $n_j(\cdot)$ dieser AWAn, bilden eine Basis von $N(\cdot)$.
Sind die Anfangswerte linear unabhängig, so sind die Lösungen linear unabhängig für alle t. Weiter gilt $n_j(t) \in N(t)$ für alle t, denn :

$$\begin{array}{rcl} n_j' & = & Q' n_j \\ P n_j' & = & P Q' n_j \quad \text{mit} \quad P := I - Q \\ (P n_j)' - P' n_j & = & -P' Q n_j \\ (P n_j)' & = & P' P n_j \end{array}$$

Damit genügt $(I - Q) n_j$ einer linearen Differentialgleichung mit dem Anfangswert $(I - Q(t_*)) n_j(t_*) = 0$, so dass folgt $(I - Q(t)) n_j(t) = 0$, d.h.$n_j(t) \in N(t)$ für alle t.

Kapitel 2

Allgemeine Annahmen

Wir betrachten (quasilineare) Algebro-Differentialgleichungen der Form

$$\boxed{f(x'(t), x(t), t)\ (=\ A((x(t), t)x'(t) + b(x(t), t)\ )\ =\ 0 \qquad (2.1)}$$

mit $f : \mathbb{R}^m \times D_f \times I_f \to \mathbb{R}^m$, $\quad D_f \times I_f \subseteq \mathbb{R}^m \times \mathbb{R}$ offen
f, f'_y, f'_x (für $f = f(y, x, t)$) stetig , $\quad G_f := \mathbb{R}^m \times D_f \times I_f$

Wir nennen $f'_y(y, x, t)$ ($= A(x, t)$) die <u>führende Jacobimatrix</u> und machen folgende grundlegende Voraussetzung:

- $\ker f'_y(y, x, t)$ ($= \ker A(x, t)$) hänge nicht von y, x ab, d. h.

  $$N(t) := \ker f'_y(y, x, t)\ (= \ker A(x, t)\ ).$$

- $N(\cdot)$ sei glatt, d. h. es existiere eine Basis
  $n_1(\cdot), \ldots, n_{m-r}(\cdot) \in C^1(I_f; \mathbb{R}^m)$ mit $N(t) =$ span $\{n_1(t), \ldots, n_{m-r}(t)\}$
  für $t \in I_f$.
  Dies bedeutet zugleich, dass $\dim N(t) \equiv m - r$ und rank $f'_y(y, x, t) \equiv r$.

Im folgenden bezeichne <u>$Q(\cdot)$</u> immer einen beliebigen glatten <u>Projektor auf $N(\cdot)$</u>
und <u>$P(\cdot) = I - Q(\cdot)$</u> den <u>zu Q komplementären Projektor</u> .

Lemma 2.1 *Es gilt* $f(y, x, t) = f(P(t)y, x, t) \quad \forall (y, x, t) \in G_f$.

<u>Beweis:</u>

$$f(y, x, t) - f(P(t)y, x, t) = \int_0^1 \underbrace{f'_y(sy + (1-s)P(t)y, x, t)(I - P(t))y}_{=0}\ ds\ =\ 0$$

□

(2.1) ist daher äquivalent zu $f(P(t)x'(t), x(t), t) = 0$. Dies ist unabhängig von der Wahl des Projektors P (P ein beliebiger glatter Projektor längs N) .

Für $x(\cdot) \in C^1$ gilt aber $P(t)x'(t) = (Px)'(t) - P'(t)x(t)$.

Für den Ausdruck auf der rechten Seite benöten wir weniger Glattheit von $x(.)$, es genügt $(Px)(\cdot) \in C^1$. Wir präzisieren (2.1) daher zu

$$\boxed{f((Px)'(t) - P'(t)x(t),\, x(t), t) = 0.} \tag{2.2}$$

Im Fall linearer zeitabhängiger ADGln entsteht so

$$A((Px)' - P'x) + Bx \;=\; A(Px)' + \underbrace{(B - AP')}_{:= B_0} x \;=\; q$$

Der Raum, in dem wir nach Lösungen suchen, ist

$$\underline{C_N^1 := \left\{ x(\cdot) \in C \;:\; (Px)(\cdot) \in C^1 \right\}}$$

Auf kompakten Intervallen ist C_N^1 mit $\|x\|_{C_N^1} := \|x\|_\infty + \|(Px)'\|_\infty$ ein Banach-Raum. Wir bemerken noch einmal, dass sowohl die Darstellung (2.2) als auch der Lösungsraum C_N^1 unabhängig vom Projektor sind, d. h. falls Q, $\hat{Q}$ zwei glatte Projektoren auf N sind, so gilt

$$f((Px)' - P'x, x, t) = f((\hat{P}x)' - \hat{P}'x, x, t)$$

und da $\bar{P}x = \bar{P}Px$, folgt aus $Px \in C^1$ auch $\bar{P}x \in C^1$.

<u>Beispiel:</u> Bei semiexpliziten ADGln ist $A = \begin{pmatrix} I & \\ & 0 \end{pmatrix}$, $N = \{0\}^{m_1} \times \mathbb{R}^{m_2}$.

Der Orthoprojektor längs N ist $P^\perp = \begin{pmatrix} I & \\ & 0 \end{pmatrix}$,

beliebige Projektoren längs N haben die Form $P(t) = \begin{pmatrix} I & 0 \\ P_{21}(t) & 0 \end{pmatrix}$.

Der Lösungsraum ist $C_N^1 = \{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : x_1 \in C^1 , x_2 \in C \}$.

<u>*Bemerkung:*</u>

Die gemachten Voraussetzungen schließen aus, dass der Nullraum der führenden Jacobimatrix von der Lösung selbst abhängt. Für diesen Fall gibt es die Möglichkeit, die ADGl wie folgt umzuformatieren:

Falls ker $f_y'(y, x, t)$ doch von (y, x) abhängig ist, gehen wir über zum erweiterten System

$$\left\{ \begin{array}{lcl} x'(t) - y(t) & = & 0 \\ f(y(t), x(t), t) & = & 0 \end{array} \right. \tag{2.3}$$

(2.3) entspricht der ADGl $\hat{f}(\hat{x}'(t), \hat{x}(t), t) = 0$ mit

$$\hat{f}\,(\hat{y}, \hat{x}, t) = \begin{pmatrix} x_p - y \\ f(y, x, t) \end{pmatrix}, \text{ wobei } \hat{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \hat{y} = \begin{pmatrix} x_p \\ y_p \end{pmatrix}.$$

Dann ist

$$\hat{f}'_{\hat{y}} = \begin{pmatrix} I & \\ & 0 \end{pmatrix}, \hat{f}'_{\hat{x}} = \begin{pmatrix} 0 & -I \\ f'_x & f'_y \end{pmatrix}, \quad C^1_{\hat{N}} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in C : x \in C^1 \right\}.$$

Wir werden später sehen, dass diese Umformung nicht ratsam ist, wenn man sie nicht vornehmen muss. Ein sofort sichtbarer Nachteil ist, dass dann ganz x und nicht nur Px in C^1 sein muss.

Für nichtlineare Aufgaben $f(x'(t), x(t), t) = 0$ betrachten wir nun punktweise die Matrizen, die bei Linearisierung der nichtlinearen ADGl in diesem Punkt entstehen, und die dazugehörigen Unterräume. Wir verwenden im weiteren folgende Bezeichnungen:

Matrizen/ Unterräume/ Mannigfaltigkeiten

Sei $(y, x, t) \in G_f$. Dann bezeichne

$$\begin{array}{lcl}
A(y,x,t) & := & f'_y(y,x,t) \\
N(t) & := & \ker A(y,x,t) \\
Q(t) & & \text{ein glatter Projektor auf } N(t) \\
P(t) & := & I - Q(t) \\
 & & \\
B(y,x,t) & := & f'_x(y,x,t) \\
B_0(y,x,t) & := & B(y,x,t) - A(y,x,t)P'(t) \\
S(y,x,t) & := & \{z \in \mathbb{R}^m : B(y,x,t)z \in \operatorname{im} A(y,x,t)\} \\
 & & \quad \text{die Menge derjenigen } z \text{, die die homogenen} \\
 & & \quad\quad \text{linearisierten Nebenbedingungen erfüllen} \\
 & & \\
\mathcal{M}_0(t) & := & \{x \in D_f : \exists w \in \mathbb{R}^m : f(w,x,t) = 0\} \\
 & & \quad \text{“constrained manifold”,} \\
 & & \quad \text{die Menge derjenigen } x \text{, die die Nebenbedingungen erfüllen} \\
 & & \quad = \{x \in D_f : b(x,t) \in \operatorname{im} A(x,t)\} \text{ für quasilineare ADGln} \\
 & & \\
G_1(y,x,t) & := & A(y,x,t) + B(y,x,t)Q(t) \\
A_1(y,x,t) & := & A(y,x,t) + B_0(y,x,t)Q(t) \\
 & & \\
N_1(y,x,t) & := & \ker A_1(y,x,t) \\
S_1(y,x,t) & := & \{z \in \mathbb{R}^m : B(y,x,t) \cdot P(t) \cdot z \in \operatorname{im} A_1(y,x,t)\} \\
 & & \\
Q_1(y,x,t) & & \text{Projektor auf } N_1(y,x,t) \\
G_2(y,x,t) & = & A_1(y,x,t) + B(y,x,t)P(t)Q_1(y,x,t)
\end{array}$$

Es gilt folgende Beziehung zwischen A_1 und G_1:

$$A_1 = A + [B - AP']Q = G_1 - G_1PP'Q = G_1(I - PP'Q) .$$

Dabei ist $(I - PP'Q)$ regulär, da $(I - PP'Q)^{-1} = (I + PP'Q)$.

Folglich gilt: $\ker A_1 = (I + PP'Q) \ker G_1$ und $\operatorname{im} A_1 = \operatorname{im} G_1$.

Folgerungen aus Lemma 1.1:

$$\begin{aligned}
S(y,x,t)\cap N(t)=\{0\} \quad &\Leftrightarrow \quad G_1(y,x,t) \text{ regulär}\\
&\Leftrightarrow \quad A_1(y,x,t) \text{ regulär}\\
&\Leftrightarrow \quad \text{Büschel } \{A(y,x,t),B(y,x,t)\} \text{ regulär mit index 1}\\
&\Leftrightarrow \quad \text{Büschel } \{A(y,x,t),B_0(y,x,t)\} \text{ regulär mit index 1}
\end{aligned}$$

$$\begin{aligned}
S_1(y,x,t)\cap N_1(y,x,t)=\{0\} \quad &\Leftrightarrow \quad G_2(y,x,t) \text{ regulär}\\
&\Leftrightarrow \quad \text{Büschel } \{A(\ ),B_0(\ )\} \text{ regulär mit index 2}\\
&\qquad\ \ \text{(das Büschel } \{A(\ ),B(\ )\} \text{ kann singulär sein.)}
\end{aligned}$$

Definition: *Sei $G \subseteq G_f$ offen.*

(i) Die ADGl (2.1) hat den Traktabilitäts(oder Behandelbarkeits-)index 1 auf G
$\Leftrightarrow N(t)\cap S(y,x,t)=\{0\}$ *für alle* $(y,x,t)\in G$

(ii) Die ADGl (2.1) hat den Traktabilitäts(oder Behandelbarkeits-)index 2 auf G
$\Leftrightarrow$ $\dim(N(t)\cap S(y,x,t))$ *ist konstant* >0 *auf G und*
$N_1(y,x,t)\cap S_1(y,x,t)=\{0\}$ *für alle* $(y,x,t)\in G$

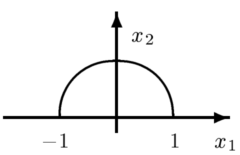
Bemerkung:

$N(t)\cap S(y,x,t)$ hat konstante Dimension genau dann, wenn $A_1(y,x,t)$ bzw. $G_1(y,x,t)$ konstanten Rang haben.
Für den Fall linearer ADGln mit konstanten Koeffizienten
$$f(x',x,t)=Ax'+Bx-q(t)$$
stimmt der Traktabilitätsindex mit dem Kroneckerindex überein.

Beispiele:

Beispiel 0: Explizite DGln $x'=g(x,t)$ aufgefasst als spezielle ADGln ergeben $f(y,x,t)=y-g(x,t)$, und $f'_y=I$, $f'_x(y,x,t)=g'_x(x,t)$.

Beispiel 1: $m=2$, $\quad \left\{\begin{array}{l} x_1'-\alpha x_1=0 \\ x_1^2+x_2^2=1 \end{array}\right. , \quad D_f=\{x\in I\!R^2 : x_2>0\},\ I_f=I\!R$



$$A=f'_y=\begin{pmatrix}1 & \\ & 0\end{pmatrix} \quad , \quad B(x)=f'_x=\begin{pmatrix}-\alpha & 0\\ 2x_1 & 2x_2\end{pmatrix}$$

$$
\begin{aligned}
N(t) &= \{z \in I\!R^2 : z_1 = 0\} =: N \quad , \\
S(y,x,t) &= \left\{ z \in I\!R^2 : \begin{pmatrix} -\alpha & 0 \\ 2x_1 & 2x_2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in \text{ im } \begin{pmatrix} 1 & \\ & 0 \end{pmatrix} \right\} \\
&= \{z \in I\!R^2 : 2x_1 z_1 + 2x_2 z_2 = 0\} = S(0,x,0) =: S(x) \\
N \cap S(x) &= \{0\} \quad \forall x \in D_f \, .
\end{aligned}
$$

Betrachten wir nun $\mathcal{M}_0 = \{x \in D_f : x_1^2 + x_2^2 = 1\}$, die Mannigfaltigkeit, die durch die Nebenbedingungen bestimmt wird, und zu $\mathcal{M}_0$ den Tangentialraum in $x \in \mathcal{M}_0$:

$$
T_x\mathcal{M}_0 = \left\{ z \in I\!R^2 : 2x_1 z_1 + 2x_2 z_2 = 0 \right\} = S(x)
$$

$T_x\mathcal{M}_0$ ist nur für $x \in \mathcal{M}_0$ definiert, $S(x)$ ist für alle $x \in D_f$ definiert. Mit Projektoren $P = A = \begin{pmatrix} 1 & \\ & 0 \end{pmatrix}$ und $\quad Q = (I-P) = \begin{pmatrix} 0 & \\ & 1 \end{pmatrix}$ gilt

$$
PT_x\mathcal{M}_0 = P \cdot \left\{ z = \begin{pmatrix} z_1 \\ -\frac{x_1}{x_2} z_1 \end{pmatrix}, z_1 \in I\!R \right\} = \left\{ \begin{pmatrix} z_1 \\ 0 \end{pmatrix}, z_1 \in I\!R \right\} = \text{ im } P.
$$

<u>Beispiel 2:</u> $\quad m = 3\,, \quad \begin{aligned} x_1' - \alpha x_1 &= 0, & D_f &= \{x \in I\!R^3 : x_2 > 0\} \\ x_2' - \tfrac{x_3}{x_2} &= 0, \ , & & \\ x_1^2 + x_2^2 &= 1, & I_f &= I\!R \end{aligned}$

$$
A = f_y' = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 0 \end{pmatrix}, \quad B(x) = f_x' = \begin{pmatrix} -\alpha & 0 & 0 \\ 0 & \frac{x_3}{x_2^2} & -\frac{1}{x_2} \\ 2x_1 & 2x_2 & 0 \end{pmatrix},
$$

$$
N(t) = \{z \in I\!R^3 : z_1 = 0, \ z_2 = 0\} = \text{ span } \{e_3\} =: N \quad , \quad Q = \begin{pmatrix} 0 & & \\ & 0 & \\ & & 1 \end{pmatrix}
$$

$$
S(y,x,t) = \{z \in I\!R^3 : 2x_1 z_1 + 2x_2 z_2 = 0\} = S(0,x,0) =: S(x)
$$

$$
N \cap S(x) = \{z \in I\!R^3 : z_1 = 0, z_2 = 0\} = \text{ span } \{e_3\}
$$

$$
A_1 = G_1 = f_y' + f_x' \cdot Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{1}{x_2} \\ 0 & 0 & 0 \end{pmatrix}, \qquad \text{rank } A_1 = 2 < 3 = m
$$

$$
\begin{aligned}
N_1(y,x,t) &= \ker A_1(y,x,t) = \left\{ z \in I\!R^3, \ z_1 = 0, \ z_2 - \tfrac{1}{x_2} \cdot z_3 = 0 \right\} =: N_1(x) \\
S_1(y,x,t) &= \{z \in I\!R^3 : f_x'(y,x,t) \cdot P(t) z \in \text{ im } A_1(y,x,t)\} \\
&= \{z \in I\!R^3 : 2x_1 z_1 + 2x_2 z_2 = 0\} =: S_1(x)
\end{aligned}
$$

$$
\begin{aligned}
\leadsto N_1(y,x,t) \cap S_1(y,x,t) &= N_1(x) \cap S_1(x) \\
&= \left\{ z \in I\!R^3 : z_1 = 0, \ z_2 = \tfrac{1}{x_2} z_3, \ 2x_1 z_1 + 2x_2 z_2 = 0 \right\} \\
&= \{0\}
\end{aligned}
$$

Sei $\mathcal{M}_1$ die Mannigfaltigkeit, auf der sowohl die Nebenbedingung $x_1^2 + x_2^2 = 1$ als auch die versteckte Nebenbedingung $x_3 + \alpha x_1^2 = 0$ erfüllt sind.

$$\begin{aligned}
\mathcal{M}_1 &= \{x \in I\!R^3 : x_1^2 + x_2^2 - 1 = 0, \quad x_3 + \alpha x_1^2 = 0\} \\
T_x\mathcal{M}_1 &= \{z \in I\!R^3 : 2x_1 z_1 + 2x_2 z_2 = 0, \;\; z_3 + 2\alpha x_1 z_1 = 0\} \;=\; \mathrm{span}\left\{ \begin{pmatrix} 1 \\ -\frac{x_1}{x_2} \\ -2\alpha x_1 \end{pmatrix} \right\} \\
PT_x\mathcal{M}_1 &= \mathrm{span}\left\{ \begin{pmatrix} 1 \\ -\frac{x_1}{x_2} \\ 0 \end{pmatrix} \right\} \subset \ \mathrm{im}\, P = \mathrm{span}\,\{e_1, e_2\} \\
S_1(x) &= \mathrm{span}\left\{ \begin{pmatrix} 1 \\ -\frac{x_1}{x_2} \\ 0 \end{pmatrix}, e_3 \right\} \;=\; PS_1(x) \cup N \\
PS_1(x) &= PT_x\mathcal{M}_1
\end{aligned}$$

Kapitel 3

Analysis für ADGln vom Index 1

3.1 Lineare ADGln vom Index 1

Wir betrachten die zeitabhängige lineare ADGL

$$A(t)x'(t) + B(t)x(t) - q(t) = 0 \quad , \quad t \in I_f \tag{3.1}$$

Es gelten die allgemeinen Annahmen aus dem vorangegangenen Kapitel (das bedeutet in diesem Fall, dass die Koeffizienten stetig und der Nullraum von A glatt sind) und die Index-1-Bedingung (d.h. $N(t) \cap S(t) = \{0\}$).
$x(\cdot) \in C_N^1$ sei Lösung.

Bei semi-expliziten ADGln

$$\begin{array}{rcl} x_1' + B_{11}x_1 + B_{12}x_2 & = & q_1 \\ B_{21}x_1 + B_{22}x_2 & = & q_2 \end{array}$$

gelangt man zu einer Lösungsdarstellung, indem man die Nebenbedingungen (im nichtlinearen Fall nach dem impliziten Funktiontheorem) nach x_2 in Abhängigkeit von x_1 auflöst und dies in die obere Blockzeile einsetzt.
Im Ergebnis erhält man eine reguläre gewöhnliche Differentialgleichung in den "differentiellen" Lösungskomponenten x_1 und eine Zuweisung für die "algebraischen" Lösungskomponenten x_2.
Sind die betrachteten ADGln nicht mehr semi-explizit, lässt sich die Lösung i.allg. nicht komponentenweise in die relevanten Anteile aufspalten. Wir werden stattdessen erfolgreich mit Anteilen in relevanten, evtl. zeitabhängigen Unterräumen arbeiten.
Zu der Lösung $x(\cdot)$ betrachten wir :

$$w(t) := P(t)x'(t) + Q(t)x(t) \quad , \quad u(t) := P(t)x(t) \quad .$$

Mit diesen Bezeichnungen entspricht (3.1)

$$A(t)w(t) + B(t)(u(t) + Q(t)w(t)) - q(t) = 0 \quad .$$

Unter der Index-1-Voraussetzung lässt sich die Gleichung

$$A(t)w + B(t)(u + Q(t)w) - q(t) = 0 \quad , \quad u, w \in \mathbb{R}^m \tag{3.2}$$

nach w in Abhängigkeit von (u,t) auflösen , denn es ist

$$Aw + B(u + Qw) - q \;=\; \underbrace{(A + BQ)}_{=G_1} w + Bu - q \;=\; 0 \quad ,$$

mit einer nach Index-1-Voraussetzung regulären Matrix G_1 . Wir erhalten

$$w \;= \omega(u,t) \;:=\; -G_1^{-1}(t)[B(t)u - q(t)]$$

Daraus ergibt sich eine Zuweisung für Qx :

$$Qx \;=\; Qw \;=\; Q\omega(u,t) \;=\; -QG_1^{-1}(Bu - q) \;=\; -Q_S u + QG_1^{-1}q$$

und mit

$$(Px)' - P'x \;=\; Px' \;=\; Pw \;=\; P\omega(u,t) \;=\; -PG_1^{-1}(Bu - q)$$

eine Differentialgleichung in $u = Px$:

$$\boxed{u' - P'(u + Q\omega(u,t)) \;=\; P\omega(u,t) \qquad (3.3)}$$

d.h.

$$u' - P'(u - Q_S u + QG_1^{-1}q) \;=\; -PG_1^{-1}(Bu - q) \quad ,$$

d.h.

$$u' - P'P_S u + PG_1^{-1}Bu \;=\; (P'Q + P)G_1^{-1}q \quad .$$

(3.3) ist eine explizite gewöhnliche Differentialgleichung in u. Ausgehend von einer linearen ADGl mit stetigen Koeffizienten entsteht eine lineare gewöhnliche Differentialgleichung mit stetigen Koeffizienten. Für beliebig vorgegebene Anfangswerte $u(t_0) = u_0, \quad u_0 \in I\!R^m$ hat (3.3) eine eindeutige Lösung $u(\cdot\,; t_0, u_0)$. Damit $u(\cdot\,; t_0, u_0)$ als 'P-Komponente' einer Lösung $x(\cdot)$ der ADGl denkbar ist, muss gelten $u(t\,; t_0, u_0) \in \operatorname{im} P(t)$. Wir zeigen, dass das genau dann der Fall ist, wenn der Anfangswert u_0 aus $\operatorname{im} P(t_0)$ gewählt wird:

> Wir nutzen die Äquivalenz $\quad u(t) \in \operatorname{im} P(t) \;\Leftrightarrow\; \mu(t) := Q(t)u(t) = 0$. Damit gilt $u(t_0) \in \operatorname{im} P(t_0) \Leftrightarrow u(t) \in \operatorname{im} P(t) \;\; \forall t \quad$ gdw. $\quad \mu(t_0) = 0 \Leftrightarrow \mu(t) = 0 \;\; \forall t$.
> Sei also $u(\cdot)$ Lösung von (3.3). Wir multiplizieren (3.3) mit $Q(\cdot)$ und erhalten
>
> $$\underbrace{Qu'}_{=(Qu)'-Q'u} \;-\; \underbrace{QP'}_{=-Q'P} u - Q \underbrace{P'Q}_{-PQ'} \omega(u,\cdot) \;=\; QP\omega(u,\cdot)$$
>
> $$\Rightarrow \quad (Qu)' \;\; \underbrace{-Q'u + Q'Pu}_{=-Q'(I-P)u=-Q'Qu} \;\;=\; 0$$
>
> $$\Rightarrow \quad \mu' - Q'\mu \;=\; 0 \quad .$$
>
> Damit ist $\mu(\cdot)$ Lösung einer linearen homogenen Differentialgleichung und es gilt die behauptete Eigenschaft $\mu(t_0) = 0 \Leftrightarrow \mu(t) = 0 \;\; \forall t$.

Wir sagen, die reguläre Differentialgleichung (3.3) hat $\operatorname{im} P(\cdot)$ als <u>invarianten Unterraum</u>.

Wir nennen (3.3) eine inhärente reguläre Differentialgleichung zu (3.1).

Diese inhärente reguläre Differentialgleichung hängt von der speziellen Wahl des Projektors $P(t)$ längs $N(t)$ ab, d.h. von der Wahl von $\operatorname{im} P(t)$ als Komplement zu $N(t)$ in $\mathbb{R}^m = N(t) \oplus \operatorname{im} P(t)$.
Dies erlaubt es solche Projektoren $P(t)$ bzw. ergänzende Unterräume $\operatorname{im} P(t)$ zu wählen, die für weitere Überlegungen günstig erscheinen. Ist $N(t)$ konstant, so lässt sich ein konstanter Projektor wählen.
Ist nun $u(\cdot)$ eine Lösung von (3.3) mit Anfangswert $u(t_0) = u_0 \in \operatorname{im} P(t_0)$, so ist

$$x(\cdot) := u(\cdot) + Q(\cdot)\omega(u(\cdot), \cdot)$$

nach Konstruktion in C_N^1 und Lösung der gegebenen ADGl .
Es gilt:

Satz 3.1 *Die ADGl (3.1) sei index-1-behandelbar (tractable) und die rechte Seite sei stetig ($q(\cdot) \in C(I_f, \mathbb{R}^m)$). Dann gilt:*

(i) Jede Anfangswertaufgabe

$$\begin{aligned} A(t)x'(t) + B(t)x(t) &= q(t)\ ,\ t \in I_f\ , \\ P(t_0)(x(t_0) - x^0) &= 0\ ,\ t_0 \in I_f\ ,\ x^0 \in \mathbb{R}^m \end{aligned} \tag{3.4}$$

ist in $C_N^1(I_f, \mathbb{R}^m)$ eindeutig lösbar.

(ii) Durch (t_0, x_0), $t_0 \in I_f$, $x_0 \in S(t_0)$ verläuft genau eine Lösung der zugehörigen homogenen ADGl ((3.1) mit $q = 0$).

(iii) Für die Lösungen $x(\cdot\,; t_0, x_0)$ der Anfangswertaufgaben aus (i) gibt es auf jedem kompakten Teilintervall von I_f eine Konstante K, so dass die folgende Abschätzung gilt:

$$\|x\|_{C_N^1} := \|x\|_\infty + \|(Px)'\|_\infty \le K(|P(t_0)x^0| + \|q\|_\infty)$$

Beweis:**(i)** folgt nach obiger Lösungsdarstellung mit $u_0 := P(t_0)x^0$ als Anfangswert für die inhärente reguläre Differentialgleichung (3.3).
zu (ii) : . Für die Lösung $x(\cdot\,; t_0, x_0)$der Anfangswertaufgabe aus (i) mit $q = 0$ gilt

$$x = u + Q\omega(u, \cdot) = u - \underbrace{QG_1^{-1}B}_{=Q_S} u = (I - Q_S)u = P_S u\,.$$

Für $x^0 \in S(t_0)$ ist $P_S(t_0)x^0 = x^0$. Mit $u_0 := P(t_0)x^0$ ergibt sich dann für die Lösung $x(t_0) = P_S(t_0)u_0 = P_S(t_0)P(t_0)x^0 = P_S(t_0)x^0 = x^0$.
zu (iii) : Die Lösung u der inhärenten regulären linearen Differentialgleichung

$$u' - P'P_S u + PG_1^{-1}Bu \ = \ (P'Q + P)G_1^{-1}q \quad .$$

hängt auf jedem kompakten Intervall stetig (d.h. beschränkt) von der rechten Seite ab und diese lässt sich durch eine Konstante mal $\|q\|_\infty$ abschätzen. Aus der Lösungsdarstellung

$$x = u + Q\omega(u, \cdot) = u - Q_S u + QG_1^{-1}q = P_S u + QG_1^{-1}q$$

folgt damit die behauptete Abschätzung. □

Bemerkung:

1.) Die zu (3.1) gehörenden Anfangswertaufgaben lassen sich als lineare Operatorgleichungen in geeigneten Banachräumen schreiben. Im Index-1-Fall gehören zu (3.1),(3.4) lineare Operatoren

$$\begin{array}{rcll} \mathfrak{A} & : & C_N^1 \to C & \text{mit} \quad \mathfrak{A}x := Ax' + Bx \\ \mathfrak{B} & : & C_N^1 \to \operatorname{im} P(t_0) \subseteq I\!R^m & \text{mit} \quad \mathfrak{B}x := P(t_0)x(t_0) \quad \text{und} \\ \mathfrak{L} := (\mathfrak{A}, \mathfrak{B}) & : & C_N^1 \to C \times \operatorname{im} P(t_0) & \end{array}$$

mit denen sich die Anfangswertaufgabe (3.1),(3.4) in der Form

$$\mathfrak{L}x := (\mathfrak{A}, \mathfrak{B})x \;=\; (q, P(t_0)x^0) \in C \times \operatorname{im} P(t_0)$$

schreiben lässt. $\mathfrak{L}$ ist auf kompakten Teilintervallen ein linearer beschränkter Operator, $\mathfrak{A}$ und $\mathfrak{B}$ sind surjektiv, $\mathfrak{L}$ ist bijektiv. Da lineare bijektive Abbildungen beschränkte Inverse haben, ist $\mathfrak{L}^{-1}$ beschränkt. Dies sagt auch Teil (iii) der Behauptung von Satz 3.1 .
(3.1),(3.4) ist eine korrekt gestellte Aufgabe im Sinne von Hadamard.

2.) Die Anfangsbedingungen könnte man anstelle von (3.4) auch anders äquivalent formulieren, z.B. als

$$x(t_0) - x^0 \in N(t_0) \qquad \text{oder} \qquad A(t_0)(x(t_0) - x^0) = 0 \qquad \text{oder}$$

$$C_0(x(t_0) - x^0) = 0 \;\text{ mit }\; \ker C_0 = \ker A(t_0) = N(t_0) \quad .$$

3.) Im allgemeinen gilt **nicht** , dass $x(t_0; t_0, x^0) = x^0$.
Wir hatten bereits $x(t_0) = P_S(t_0)x^0$ für die Lösung der homogenen ADGl (d.h. $q = 0$). Insbesondere gilt $x(t_0) \in S(t_0)$, d.h. $x(t_0)$ genügt den Nebenbedingungen der homogenen ADGl. Für allgemeine rechte Seiten $q \in C$ ergibt sich

$$x(t_0) = P_S(t_0)x^0 + (QG_1^{-1}q)(t_0) \quad .$$

Speziell für semiexplizite lineare ADGln

$$\begin{pmatrix} I & \\ & 0 \end{pmatrix} x' + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} x = q$$

bedeutet dies, dass $x(t_0) = \begin{pmatrix} x_1^0 \\ B_{22}^{-1}(q_2 - B_{21}x_1^0) \end{pmatrix}$.

4.) Ähnlich wie bei regulären linearen Differentialgleichungen lässt sich die allgemeine Lösung der homogenen ADGl gut mit Hilfe einer Fundamentalmatrix darstellen. Zu beachten ist hierbei, dass bei ADGln mit Index 1 nur $\operatorname{rank} A$ $(= \operatorname{rank} P)$ linear unabhängige Anfangsbedingungen gestellt werden können, bzw. dass es nur $\operatorname{rank} A$ linear unabhängige Lösungen der homogenen ADGl gibt.
Wir formulieren für die Fundamentalmatrix m Anfangsbedingungen analog

(3.4) und verstehen unter der Fundamentalmatrix $X(\cdot)$ der ADGl die Lösung von

$$\begin{aligned} A(t)X'(t) + B(t)X(t) &= 0 \in \mathbb{R}^{m\times m}, \quad t \in [t_0, T] \quad , \\ P(t_0)(X(t_0) - I) &= 0 \in \mathbb{R}^{m\times m} \quad . \end{aligned}$$

Aus der Darstellung der Lösungen der homogenen ADGl als $x(t) = P_S(t)u(t)$, wobei $u(\cdot)$ die Lösung der zugehörigen inhärenten regulären Differentialgleichung mit dem Anfangswert $u(t_0) = P(t_0)x^0$ ist, ergibt sich auch die Darstellung der Fundamentalmatrix als

$$X(t) = P_S(t)U(t)P(t_0), \quad t \in [t_0, T] \quad ,$$

wobei $U(\cdot)$ die Fundamentalmatrix der inhärenten regulären Differentialgleichung mit $U(t_0) = I$ ist.

$X(t)$ ist für alle $t \in [t_0, T]$ singulär , es gilt $\operatorname{rank} X(t) = \operatorname{rank} A(t)$ und $\ker X(t) = \ker P(t_0) = N(t_0)$ für alle $t \in [t_0, T]$.
Mit Hilfe dieser Fundamentalmatrix können wir nun die Lösung der homogenen ADGl (3.1) (mit $q = 0$) zu der Anfangsbedingung (3.4) darstellen als

$$x(t) = X(t)x^0 \quad , t \in [t_0, T] \quad .$$

Lösungen der inhomogenen Aufgabe ergeben sich zu

$$x(t) = X(t)x^0 + x_q(t) \quad ,$$

wobei $x_q(\cdot)$ die spezielle Lösung der inhomogenen Aufgabe mit dem Anfangswert $P(t_0)x_q(t_0) = 0$ bezeichnet und sich darstellen lässt als

$$x_q(t) = P_S(t)U(t)\int_{t_0}^{t} U(s)^{-1}((P'Q + P)G_1^{-1}q)(s)ds + (QG_1^{-1}q)(t) \quad .$$

3.2 Nichtlineare ADGLn mit Index 1

Wir betrachten nun ADGln der Form

$$f(x'(t), x(t), t) = 0 \tag{3.5}$$

mit der Index-1-Voraussetzung und den Annahmen aus Kapitel 2. In Analogie zu Abschnitt 3.1 untersuchen wir nun die nichtlineare Gleichung

$$F(w, u, t) \;:=\; f(w, u + Q(t)w, t) \;=\; 0$$

Die Jacobimatrix $F'_w(w, u, t) \;=\; f'_y(\ldots) + f'_x(\ldots)Q(t) \;=\; G(w, u + Q(t)w, t)$ ist nach der Index-1-Voraussetzung regulär. Weiter ist F ebenso wie die Ableitungen $F'_w = f'_y + f'_x Q$ und $F'_u \;=\; f'_x$ stetig. Damit können wir das Implizite Funktionentheorem anwenden und auf die Existenz einer Auflösungsfunktion $w \;=\; \omega(u, t)$ in der Umgebung eines Lösungspunktes $(w_0, u_0, t_0) \;\in\; F^{-1}(0)$ schließen. (Im vorangegangenen Abschnitt konnten wir auf einen solchen Bezugspunkt verzichten, da dort lineare Gleichungen zu lösen waren.)
Es gilt:

Lemma 3.2 *Sei* $(y_0, x_0, t_0) \in I\!R^m \times D_f \times I_f$, $f(y_0, x_0, t_0) \;=\; 0$ *und es gelte* $N(t_0) \cap S(y_0, x_0, t_0) \;=\; \{0\}$.
Dann ist die Gleichung $f(w, u + Q(t)w, t) \;=\; 0$ *lokal auflösbar zu* $w \;=\; \omega(u, t)$ *mit* $\omega(u_0, t_0) \;=\; w_0$ *, wobei* $w_0 \;:=\; P(t_0)y_0 + Q(t_0)x_0$, $u_0 \;:=\; P(t_0)x_0$.
Die Auflösungsfunktion ω *ist stetig, es existiert die Ableitung* ω'_u *und ist stetig, und es gilt*

$$\omega'_u(u, t) \;=\; -G_1(Pw, u + Qw, t)^{-1} f'_x(Pw, u + Qw, t) \quad .$$

Beweis : Wir setzen

$$w_0 \;:=\; P(t_0)y_0 + Q(t_0)x_0 \quad , \quad u_0 \;:=\; P(t_0)x_0 \quad .$$

Dann ist $F(w_0, u_0, t_0) \;=\; 0$ und die Behauptung folgt mit dem Impliziten Funktionentheorem. □

Bemerkung:

1.) ω ist fortsetzbar, bis der Rand des Definitionsbereiches $D_f \times I_f$ erreicht wird, bzw. bis zum Rand des Gebietes, wo die Index-1-Bedingung erfüllt ist.

2.) Ist $x(\cdot) \;\in\; C^1_N$ Lösung von (3.5) und gilt die Index-1-Bedingung längs dieser Lösung (d.h. für alle $(y, x, t) \;=\; ((Px)'(t) - P'(t)x(t), x(t), t)$), so erhalten wir analog wie in Abschnitt 3.1 die Lösungsdarstellung

$$x(t) := u(t) + Q(t)\omega(u(t), t) \tag{3.6}$$

mit

$$u'(t) \;=\; P'(t)u(t) + (P + P'Q)(t)\omega(u(t), t)) \quad , \tag{3.7}$$

$$u(t_0) \;=\; P(t_0)x^0 \quad . \tag{3.8}$$

Die inhärente reguläre Differentialgleichung (3.7) in $u(\cdot)$ ergibt sich dabei aus

$$\begin{aligned} P(t)\omega(u(t),t) = P(t)x'(t) &= (Px)'(t) - P'(t)x(t) \\ &= u'(t) - P'(t)(u(t) + Q(t)\omega(u(t),t)) \quad . \end{aligned}$$

Sie "lebt" im $I\!R^m$ und hat $\operatorname{im} P(\cdot)$ als invarianten Unterraum,
d.h. falls für die Lösung von (3.7) gilt $u(\bar{t}) = P(\bar{t})u(\bar{t})$ für ein $\bar{t}$, so gilt $u(t) = P(t)u(t)$ für alle t.

Beispiel 1: Wir betrachten die semi-explizite ADGl $\begin{aligned} x_1' &= \alpha x_1 \\ 0 &= x_1^2 + x_2^2 - 1 \end{aligned}$ für $x_2 > 0$.

Hier ist $f(y,x,t) = \begin{pmatrix} y_1 - \alpha x_1 \\ x_1^2 + x_2^2 - 1 \end{pmatrix}$. Mit den (konstanten) Ortho-Projektoren
$P = \begin{pmatrix} 1 & \\ & 0 \end{pmatrix}$ und $Q = \begin{pmatrix} 0 & \\ & 1 \end{pmatrix}$ ist $u + Qw = \begin{pmatrix} u_1 \\ u_2 + w_2 \end{pmatrix}$
und die nach w in Abhängigkeit von (u,t) aufzulösende Gleichung lautet
$f(w, u + Qw, t) = \begin{pmatrix} w_1 - \alpha u_1 \\ u_1^2 + (u_2 + w_2)^2 - 1 \end{pmatrix} = 0$.

Als Lösung erhalten wir $w = \omega(u,t) = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \alpha u_1 \\ -u_2 + \sqrt{1 - u_1^2} \end{pmatrix}$.

Daraus folgt die Lösungsdarstellung $x = u + Qw = \begin{pmatrix} u_1 \\ u_2 + w_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ \sqrt{1 - u_1^2} \end{pmatrix}$,
wobei u der zugehörigen inhärenten regulären Differentialgleichung genügt.
Diese vereinfacht sich, da der Projektor P konstant ist ($\Rightarrow P' = 0$), zu

$$u' = P\omega(u,t) = \begin{pmatrix} 1 & \\ & 0 \end{pmatrix} \begin{pmatrix} \alpha u_1 \\ \sqrt{1 - u_1^2} \end{pmatrix} = \begin{pmatrix} \alpha u_1 \\ 0 \end{pmatrix} .$$

Man sieht, dass $\operatorname{im} P = I\!R \times \{0\}$ ein invarianter Unterraum dieser regulären Differentialgleichung ist.

In Beispiel 1, wo die Dinge so einfach liegen, könnte man auch $u_1' = \alpha u_1$ als Differentialgleichung in $u_1 = \begin{pmatrix} 1 & 0 \end{pmatrix} Pu$ als kompakte Form der inhärenten regulären Differentialgleichung betrachten. Ein solches Vorgehen bringt jedoch Nachteile mit sich, wenn $\operatorname{im} P(\cdot)$ nicht konstant ist, da dann auch Ableitungen einer Projektion von $\operatorname{im} P(\cdot)$ auf den Raum niedrigerer Dimension $I\!R^{dim\,im\,P}$ eine Rolle spielen. Die entstehenden Differentialgleichungen hängen zusätzlich von der Wahl dieser Projektion ab.
Unter Verwendung der inhärenten regulären Differentialgleichung (3.7) lässt sich der folgende Satz beweisen:

Satz 3.3 *Sei* $(y_0, x_0, t_0) \in I\!R^m \times D_f \times I_f$, *mit* $f(y_0, x_0, t_0) = 0$ *(damit gilt insbesondere* $x_0 \in \mathcal{M}_0(t_0)$ *) und es gelte* $N(t_0) \cap S(y_0, x_0, t_0) = \{0\}$ *. Dann gilt:*

(i) Die Anfangswertaufgabe $f(x'(t), x(t), t) = 0, \quad x(t_0) = x_0$ *ist lokal eindeutig lösbar.*

(ii) *Die Anfangswertaufgaben* $f(x'(t), x(t), t) = 0$, $P(t_0)x(t_0) = P(t_0)x^0$ *mit Anfangswerten* $x^0 \in \mathbb{R}^m$, $|P(t_0)(x^0 - x_0)| \le \tau$ *für hinreichend kleines* $\tau > 0$ *sind lokal eindeutig lösbar mit Lösungen* $x(\cdot, x^0) \in C_N^1$.

(iii) *Die Lösungen* $x(\cdot, x^0)$ *sind stetig differenzierbar bzgl.* x^0 .
Die partielle Ableitung $X(\cdot) := \frac{\partial}{\partial x^0} x(\cdot, x^0)$ *gehört zu* C_N^1 *mit Werten in* $L(\mathbb{R}^m)$ *und genügt der in der Lösung linearisierten homogenen ADGl, der "Gleichung der 1. Variation":*

$$f_y'(x'(t,x^0), x(t,x^0), t)X'(t) + f_x'(\ldots)X(t) = 0 \quad , \quad P(t_0)(X(t_0) - I) = 0 \quad .$$

<u>Beweis :</u> **zu (i):** (i) folgt aus (ii) mit $x^0 = x_0$, denn mit $u_0 := P(t_0)x_0$ erhalten wir die Lösungsdarstellung

$$x(t_0) = u_0 + Q(t_0)\omega(u_0, t_0) = u_0 + Q(t_0)w_0 = P(t_0)x_0 + Q(t_0)x_0 = x_0 \quad .$$

zu (ii): Für hinreichend kleines τ liegt (u_0, t_0) mit $u_0 := P(t_0)x^0$ im Definitionsbereich der impliziert definierten Funktion $\omega(u, t)$. (Wir bemerken, dass die Voraussetzung, dass eine (lokal eindeutige) Auflösungsfunktion existieren muss, bereits bei impliziten regulären Differentialgleichungen nötig ist.)
Dann betrachten wir die Anfangswertaufgabe für die zugehörige inhärente reguläre Differentialgleichung (3.7). Analog wie in Abschnitt 3.2 folgt, dass die Differentialgleichung in (3.7) $\operatorname{im} P(\cdot)$ als invarianten Unterraum hat. Wir bezeichnen die Lösung von (3.7) mit $u(\cdot, u_0)$ und erhalten für die Lösungen der Anfangswertaufgaben aus (ii) die Darstellung

$$x(t) = u(t, P(t_0)x^0) + Q(t)\omega(u(t, P(t_0)x^0), t) \quad .$$

zu (iii): Aus der Lösungsdarstellung folgt, dass

$$\frac{\partial}{\partial x^0}x = \frac{\partial}{\partial x^0}u + Q\omega_u'\frac{\partial}{\partial x^0}u = (I - \underbrace{QG_1^{-1}f_x'P}_{=Q_S})\frac{\partial}{\partial x^0}u = P_S\frac{\partial}{\partial x^0}u = P_S\frac{\partial}{\partial u^0}uP(t_0).$$

Damit ist $PX = \frac{\partial}{\partial x^0}u$ stetig differenzierbar bzgl. t und aus der Identität

$$f(x'(t,x^0), x(t,x^0), t) = 0 \quad \forall t \quad \forall x^0$$

folgt nach Durchdifferenzieren nach x^0 und nach Vertauschen der Ableitungen nach t und x^0 die behauptete Gleichung der ersten Variation. □

In Satz 3.3 wird von einem Punkt $(y_0, x_0, t_0) \in \mathbb{R}^m \times D_f \times I_f$ mit $f(y_0, x_0, t_0) = 0$ ausgegangen. Was bedeutet diese Voraussetzung?

Für quasilineare ADGln $A(x,t)x' + b(x,t) = 0$ gibt es zu jedem $t \in I_f$ und jedem x aus der durch die Nebenbedingungen bestimmten Menge

$$\mathcal{M}_0(t) = \{x \in D_f : b(x,t) \in \operatorname{im} A(x,t)\}$$

Werte $y \in \mathbb{R}^m$ mit $A(x,t)y + b(x,t) = 0$. Im Index-1-Fall folgt aus der Herleitung obigen Satzes

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^m : Q(t)x = Q(t)\omega(P(t)x, t))\} \quad .$$

Durch die Bedingung $A(x,t)y + b(x,t) = 0$ ist $y \in \mathbb{R}^m$ nicht eindeutig bestimmt, insbesondere ist der Anteil $Q(t)y$ unbestimmt. Eine Möglichkeit , ein solches y zu wählen, ist $y := \hat{y}(x,t) := -A(x,t)^+ b(x,t)$, wobei A^+ die Moore-Penrose-Inverse zu A bezeichne. Allgemein ist jeder Wert $y \in \hat{y}(x,t) + \operatorname{im} Q(t)$ möglich.
Nun weisen wir noch auf den Zusammenhang zwischen dem begleitenden Unterraum $S(y,x,t)$ und dem Tangentialraum $T_x\mathcal{M}_0(t)$ hin:
Mit Hilfe von A^+ lässt sich der Ortho-Projektor auf $\operatorname{im} A$ schreiben als AA^+. Damit ist $I - AA^+$ der Ortho-Projektor längs $\operatorname{im} A$ und

$$
\begin{aligned}
\mathcal{M}_0(t) &= \{x \in D_f : ((I - AA^+)b)(x,t) = 0\},\\
T_x\mathcal{M}_0(t) &= \{z \in \mathbb{R}^m : ((I - AA^+)b)'_x(x,t)z = 0\},\\
S(y,x,t) &= \{z \in \mathbb{R}^m : (b'_x(x,t) + [A(x,t)y]'_x)z \in \operatorname{im} A(x,t)\}.
\end{aligned}
$$

Für $y \in \hat{y}(x,t) + \operatorname{im} Q(t) = -A(x,t)^+ b(x,t) + \operatorname{im} Q(t)$ gilt $S(y,x,t) = T_x\mathcal{M}_0(t)$.

Definition: *$(x_0,t_0) \in D_f \times I_f$ heißt* <u>*konsistenter Anfangswert*</u> *zu (3.5), falls es eine Lösung $x \in C^1_N(I,\mathbb{R}^m)$, mit $t_0 \in I$ und $x(t_0) = x_0$ gibt.*
$(y_0,x_0,t_0) \in \mathbb{R}^m \times D_f \times I_f$ heißt <u>*konsistente Initialisierung*</u> *zu (3.5), falls (x_0,t_0) konsistenter Anfangswert ist und $f(y_0,x_0,t_0) = 0$.*

<u>*Bemerkung:*</u>

> Für eine konsistente Initialisierung kann $Q(t_0)y_0$ beliebige Werte haben. Es gilt $P(t_0)y_0 = (Px)'(t_0) - P'(t_0)x(t_0)$.

Als Zusammenfassung der bisherigen Überlegungen zu ADGln vom Index 1 ergibt sich:

Satz 3.4 *Konsistente Anfangswerte für ADGLn vom Index 1 sind genau diejenigen (x_0,t_0) mit $t_0 \in I_f$, $\quad x_0 \in \mathcal{M}_0(t_0)$.*

Nun befassen wir uns mit der Frage, wie **konsistente Anfangswerte** bestimmt werden können. Aufgrund des im allgemeinen nichtlinearen Charakters der Nebenbedingungen können wir nur lokale Lösbarkeitseigenschaften erwarten.

Einerseits bietet es sich an, die Gleichung

$$f(w, P(t)u + Q(t)w, t) = 0$$

zu benutzen, deren lokal eindeutige Auflösbarkeit nach $w = \omega(u,t)$ in Lemma 3.2 gezeigt wurde. Falls es einen konsistenten Anfangswert (x_0,t_0) mit $P(t_0)x_0 = P(t_0)x^0$ gibt, so ergibt sich x_0 mit Hilfe der Lösungsdarstellung $x_0 := P(t_0)x^0 + Q(t_0)w_0$, wenn w_0 Lösung von

$$f(w_0, P(t_0)x^0 + Q(t_0)w_0, t_0) = 0$$

ist. Mit $y_0 := P(t_0)w_0$ oder auch $y_0 := w_0$ ergibt sich eine konsistente Initialisierung.

Eine weitere Möglichkeit, konsistente Anfangswerte bzw. konsistente Initialisierungen zu berechnen bietet der folgende Satz:

Satz 3.5 *Die ADGl (3.5) habe den Index 1 auf $\mathbb{R}^m \times D_f \times I_f$. Es seien $t_0 \in I_f$ und $x^0 \in \mathbb{R}^m$ gegebene Werte. Falls das (im allgemeinen nichtlineare) Gleichungssystem*

$$\left\{ \begin{array}{rcl} f(y_0, x_0, t_0) & = & 0 \\ Q(t_0)y_0 + P(t_0)(x_0 - x^0) & = & 0 \end{array} \right. \tag{3.9}$$

bezüglich x_0, y_0 lösbar ist, so ist die Lösung lokal eindeutig und liefert eine konsistente Initialisierung. Die Jacobimatrix (die Funktionalmatrix) des Systems ist regulär.

Beweis : Zunächst sieht man, dass für Lösungen von (3.9) gilt, dass $x_0 \in \mathcal{M}_0(t_0)$ und (y_0, x_0, t_0) konsistente Initialisierung ist. Es bleibt die Regularität der Jacobimatrix zu zeigen. Die Jacobimatrix zu (3.9) ist

$$\mathcal{F} = \begin{pmatrix} f'_y & f'_x \\ Q & P \end{pmatrix} =: \begin{pmatrix} A & B \\ Q & P \end{pmatrix} \quad .$$

Wir zeigen, dass diese injektiv ist. Sei also

$$\mathcal{F} \begin{pmatrix} \tilde{y} \\ \tilde{x} \end{pmatrix} = 0\,, \text{ d.h. } \begin{array}{ccccc} A\tilde{y} & + & B\tilde{x} & = & 0 \\ Q\tilde{y} & + & P\tilde{x} & = & 0 \end{array}$$

Aus der zweiten Zeile folgt sofort , dass sowohl $Q\tilde{y} = 0$ als auch $P\tilde{x} = 0$ gelten.
Dann bleibt in der ersten Zeile $AP\tilde{y} + BQ\tilde{x} = (A + BQ)(P\tilde{y} + Q\tilde{x}) = 0$.
Wegen der Index-1-Voraussetzung ist die Matrix $(A + BQ)$ regulär und es folgt $P\tilde{y} + Q\tilde{x} = 0$ und daraus wieder sowohl $P\tilde{y} = 0$ als auch $Q\tilde{x} = 0$.
Insgesamt muss dann $\tilde{y} = \tilde{x} = 0$ gelten , womit die Injektivität der Jacobimatrix gezeigt ist. □

Bemerkung:

Für nichtlineare Beispiele muss (3.9) nicht für beliebige vorgegebene Werte $x^0 \in \mathbb{R}^m$ lösbar sein. Beispielsweise gibt es im Beispiel 1, wo Lösungen auf dem Einheitskreis im $\mathbb{R}^2$ liegen müssen, nur dann eine Lösung von (3.9), wenn für die erste Komponente x_1^0 des vorgegebenen Wertes x^0 gilt $|x_1^0| < 1$.
In Anwendungen ist die Bestimmung konsistenter Anfangswerte oft ein schwieriges Problem, oft schwieriger als die nachfolgende Integration der Anfangswertaufgaben, die mit Standardintegrationsverfahren (vgl. das folgende Kapitel) oft erfolgreich gemeistert werden kann.
In speziellen Anwendungen helfen zuweilen spezielle Ideen. In der Schaltungssimulation kann man Arbeitspunkte, d.h. Gleichgewichtslösungen ($x' = 0$) der ADGl als Ausgangspunkt verwenden. Damit ist das Problem dort auf das Finden von derartigen Arbeitspunkten zurückgeführt. Dort braucht man die Regularität der Matrix b'_x. Dies ist im Sprachgebrauch der Schaltungssimulation die "DC-Lösbarkeit".

Beispiel 8: Das folgende Beispiel ist dem Buch von Ascher/Petzold entnommen: Wir betrachten die ADGl

$$\begin{array}{rcl} x'_1 - x_3 & = & q_1 \\ x_2(1 - x_2) & = & q_2 \\ x_1 x_2 + x_3(1 - x_2) & = & q_3 \end{array} \quad .$$

Hier ist $A = \begin{pmatrix} 1 & & \\ & 0 & \\ & & 0 \end{pmatrix}$ und $B(x) = b'_x(x) = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1-2x_2 & 0 \\ x_2 & x_1 - x_3 & 1-x_2 \end{pmatrix}$.
Weiter ist

$$\begin{aligned} N &= \{z \in I\!R^3 : z_1 = 0\} = \operatorname{span}\{e_2, e_3\} \\ S(x) &= \{z \in I\!R^3 : (1-2x_2)z_2 = 0\,,\ x_2 z_1 + (x_1 - x_3)z_2 + (1-x_2)z_3 = 0\} \\ N \cap S(x) &= \{z \in I\!R^3 : z_1 = 0\,,\ (1-2x_2)z_2 = 0\,,\ (x_1-x_3)z_2 + (1-x_2)z_3 = 0\} \\ &= \{0\} \Longleftrightarrow \begin{pmatrix} 1-2x_2 & 0 \\ x_1 - x_3 & 1-x_2 \end{pmatrix} \text{regulär} \\ & \Longleftrightarrow 1 - 2x_2 \neq 0\,,\ 1 - x_2 \neq 0 \\ & \Longleftrightarrow x_2 \neq \frac{1}{2}\,,\ x_2 \neq 1 \quad . \end{aligned}$$

Mit $x_2 = \frac{1}{2}$ und $x_2 = 1$ gibt es zwei parrallele Ebenen im $I\!R^3$ auf denen die Index-1-Bedingung nicht erfüllt ist.
Weiter ist

$$\mathcal{M}_0(t) = \{x \in I\!R^3 : x_2 - x_2^2 = q_2(t)\,,\ x_3(1-x_2) + x_1 x_2 = q_3(t)\} \quad ,$$

insbesondere gilt

$$\mathcal{M}_0(t) \subseteq \{x \in I\!R^3 : x_2 = 1/2 \pm \sqrt{1/4 - q_2(t)}\,\} \quad .$$

Die Lösbarkeitseigenschaften hängen von der speziellen Wahl von $q_2(t)$ ab. Falls etwa $q_2(t) > \frac{1}{4}$ ist, folgt $\mathcal{M}_0(t) = \emptyset$.
Für $q_2(t) \equiv \frac{1}{4}$ liegt ganz $\mathcal{M}_0(t)$ in der Ebene $x_2 = \frac{1}{2}$ auf der die Index-1-Bedingung nicht erfüllt ist.
Für $q_2(t) < \frac{1}{4}$ zerfällt $\mathcal{M}_0(t)$ in zwei disjunkte Teilgebiete $\mathcal{M}_0(t) = \mathcal{M}_0^+(t) \cup \mathcal{M}_0^-(t)$ oberhalb und unterhalb der Ebene $x_2 = \frac{1}{2}$. Die Lösungen in $\mathcal{M}_0^-$ erfüllen immer die index-1-Bedingung. Für die Lösungen in $\mathcal{M}_0^+$ ist dies weiter abhängig von $q_2(t)$
Etwa für $q_2 \equiv 0$ ergibt sich

$$\begin{aligned} \mathcal{M}_0^+(t) &= \{: x_2 = 1\,,\ x_1 = q_3(t)\} \\ \mathcal{M}_0^-(t) &= \{: x_2 = 0\,,\ x_1 = q_3(t)\} \quad . \end{aligned}$$

Hier liegt $\mathcal{M}_0^+(t)$ ganz in der Ebene $x_2 = 1$ auf der die Index-1-Bedingung nicht erfüllt ist.

Kapitel 4

Integrationsverfahren für ADGln vom Index 1

Wir betrachten wieder Algebro-Differentialgleichungen der Form

$$\boxed{f(x'(t), x(t), t) \;=\; A(x(t), t)x'(t) + b(x(t), t) \;=\; 0 \qquad (4.1)}$$

unter den Voraussetzungen von Kapitel 3 .
$x_* \in C^1_N(\mathcal{I}, \mathbb{R}^m)$ sei eine zu approximierende Lösung von (4.1) und (4.1) habe den Index 1 auf einer Umgebung dieser Lösung, d.h. auf einer Umgebung der Menge $\{ \; ((Px_*)'(t) - P'(t)x_*(t) \; , \; x_*(t) \; , \; t) \; , \; t \in \mathcal{I} \; \}$. Wir wollen nach Integrationsverfahren suchen, die direkt auf die ADGl in der Form (4.1) angewendet werden können.
Theoretisch könnte man auch vor Anwendung numerischer Verfahren eine Entkopplung wie in Kapitel 3 vorgestellt vornehmen und dann erst numerische Verfahren auf die dort hergeleitete reguläre inhärente Differentialgleichung anwenden. Wir fassen noch einmal die Gründe zusammen, dies **nicht** zu tun:
Praktisch steht diese Entkopplung **nicht** zur Verfügung. Die rechte Seite der regulären inhärenten Differentialgleichung müsste für alle Werte von u, t , die während der Integration benutzt werden, erst numerisch beschafft werden. Dies ist aufwändig und zerstört oft vorhandene spezielle Strukturen. Es ist so uneffektiv, dass es nicht einmal dann günstig ist, falls die Matrix A konstant ist.

Bei Integrationsverfahren zur approximativen Lösung der ADGl (4.1) arbeiten wir auf einem Gitter

$$t_0 < t_1 < \cdots < t_\ell < \cdots \qquad \text{mit Schrittweiten} \qquad h_\ell = t_\ell - t_{\ell-1}$$

um Näherungen x_ℓ für die Lösungswerte $x_*(t_\ell)$ zu berechnen. Wir haben bereits im einführenden Kapitel gesehen, dass explizite Verfahrensansätze für ADGln nicht geeignet sind, da die entstehenden Gleichungen nicht nach den zu bestimmenden Näherungswerten x_ℓ auflösbar sind. Als gut geeignete Verfahrensansätze für die Integration von Anfangswertaufgaben von ADGln haben sich speziell die rückwärtigen Differentiationsformeln (backward differentiation formula, BDF) von Gear und gewisse implizite Runge-Kutta (IRK) -Verfahren erwiesen:

4.1 BDF-Ansatz

Die BDF-Formeln entstehen , indem die in die ADGl eingehenden Ableitungen der Lösung approximativ durch rückwärtige Differenzenquotienten ersetzt werden:

Für eine gegebene hinreichend glatte Funktion $y(\cdot)$ bezeichne $p_y(\cdot)$ das Interpolationspolynom , das y in den Stützstellen $t_{\ell-k}, \ldots, t_{\ell-1}, t_\ell$ interpoliert, d.h. es sei

$$p_y(t) \;=\; \sum_{j=0}^{k} \prod_{\substack{i=0 \\ i \neq j}}^{k} \frac{t - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}}\, y(t_{\ell-j}) \quad .$$

Dann heißt $p_y'(t_\ell) \;=\; \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{j,\ell} y(t_{\ell-j})$ rückwärtiger Differenzenquotient zu y an der Stelle t_ℓ .

Für die Herleitung der k-schrittigen BDF-Formel wird mit dem Polynom p gearbeitet, dass die zurückliegenden k Näherungswerte $x_{\ell-k}, \ldots, x_{\ell-1}$ **und** den noch zu bestimmenden Wert x_ℓ interpoliert. x_ℓ wird dann aus der (BDF-)Gleichung

$$f(p'(t_\ell), x_\ell, t_\ell) \;=\; f(\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j}, x_\ell, t_\ell) = 0 \quad , \quad \ell \geq k \tag{4.2}$$

bestimmt. Im allgemeinen hängen die Koeffizienten $\alpha_{\ell,j}$, $j = 0, \ldots, k$, von den Schrittweitenverhältnissen ab. Werden konstante Schrittweiten verwendet, ergeben sich konstante Koeffizienten $\alpha_j \;=\; \alpha_{\ell,j}$, $j = 0, \ldots, k$, $l \geq k$.

Für den Spezialfall $f(y, x, t) \;=\; y - g(x, t)$ ergibt sich das bekannte BDF-Verfahren für explizite gewöhnliche Differentialgleichungen.

Für $k \;=\; 1$ ergibt sich das implizite Euler-Verfahren. Die Formel für die zwei-schrittige BDF mit konstanter Schrittweite lautet:

$$\mathrm{BDF}_2 \,:\, f(\frac{1}{h}(\frac{3}{2}x_\ell - 2x_{\ell-1} + \frac{1}{2}x_{\ell-2}, x_\ell, t_\ell) = 0 \quad , \quad \ell \geq k \quad .$$

Als erstes untersuchen wir die Durchführbarkeit der BDF (4.2). Wir fragen, ob (4.2) lokal eindeutig nach x_ℓ auflösbar ist. Dazu betrachten wir die entsprechende Jacobi-Matrix in x_ℓ. Diese hat die Form

$$\frac{\alpha_{\ell,0}}{h_\ell} f_y'(\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j}, x_\ell, t_\ell) + f_x'(\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j}, x_\ell, t_\ell) \;=:\; \frac{\alpha_{\ell,0}}{h_\ell} A_\ell + B_\ell \quad .$$

Wir untersuchen dazu allgemein die Regularität von Matrizen der Form $\frac{1}{h}A + B$. Es gilt:

Lemma 4.1 *Für ein Matrizenpaar* $\{A\,, B\}$ $A\,, B \in L(\mathbb{R}^m)$ *gelte* $N \cap S = \{0\}$.
Dann ist für hinreichend kleine h *die Matrix* $\frac{1}{h}A + B$ *regulär und ihre Kondition verhält sich wie* $\frac{1}{h}$, *d.h.* $\mathrm{cond}(\frac{1}{\mathrm{h}}\mathrm{A} + \mathrm{B}) \;=\; \mathrm{O}(\frac{1}{\mathrm{h}})$.

Beweis: Die Regularität der Matrix $\frac{1}{h}A + B$ für hinreichend kleine h folgt aus der Regularität des Büschels $\{A\,, B\}$, da $\lambda A + B$ nur für endlich viele Werte von λ singulär wird. Um eine Aussage über die Kondition zu treffen, bemühen wir uns um eine Darstellung für die Inverse $(\frac{1}{h}A + B)^{-1}$:
Es sei Q_S der Projektor auf N längs S und $G_S := A + BQ_S$.

Dann gilt nach Corallar 1.2 $Q_S = Q_S G_S^{-1} B$.
Skalieren der Koeffizienten $A\,,B$ mit G_S^{-1} führt dann zu

$$\begin{array}{lll} G_S^{-1}A = P_S & \text{da} & G_S P_S = A P_S = A \quad , \\ G_S^{-1}B = G_S^{-1}BP_S + Q_S & \text{da} & G_S^{-1}B = G_S^{-1}BP_S + G_S^{-1}BQ_S \\ & & \text{und } G_S Q_S = BQ_S \quad . \end{array}$$

Nun sei

$$z = (\frac{1}{h}A + B)^{-1} w \quad \text{d.h.} \quad (\frac{1}{h}A + B) z = w \quad \text{.Dies gilt}$$

$$\Longleftrightarrow (\frac{1}{h}P_S + G_S^{-1}BP_S + Q_S) z = G_S^{-1} w$$

$$\Longleftrightarrow \left\{ \begin{array}{rcl} (\frac{1}{h}P_S + P_S G_S^{-1} B P_S) z & = & P_S G_S^{-1} w \\ (\underbrace{Q_S G_S^{-1} B}_{=Q_S} P_S + Q_S) z & = & Q_S G_S^{-1} w \end{array} \right.$$

$$\Longleftrightarrow \left\{ \begin{array}{rcl} P_S z & = & (I + hP_S G_S^{-1} B)^{-1} h P_S G_S^{-1} w \\ Q_S z & = & Q_S G_S^{-1} w \end{array} \right.$$

$$\Longleftrightarrow z = P_S z + Q_S z = [(I + hP_S G_S^{-1} B)^{-1} h P_S + Q_S] G_S^{-1} w \quad .$$

Damit gilt die Darstellung

$$(\frac{1}{h}A + B)^{-1} = [(I + hP_S G_S^{-1} B)^{-1} h P_S + Q_S] G_S^{-1} \quad .$$

Wegen $\|\frac{1}{h}A + B\| = O(\frac{1}{h})$ und , wie nun ersichtlich, $\|(\frac{1}{h}A + B)^{-1}\| = O(1)$ folgt die Behauptung. □

<u>*Bemerkung:*</u>

1.) Die Einschränkung der Invertierbarkeit nur für hinreichend kleine h ergibt sich aus dem differentiellen Anteil.

2.) Bei regulären Differentialgleichungen hätte man A regulär, $P = I$, $Q = 0\,, G = A$ und $(\frac{1}{h}A + B)^{-1} = (I + hA^{-1}B)^{-1} h A^{-1} = O(h)$, so dass trotz $\|\frac{1}{h}A + B\| = O(\frac{1}{h})$ gilt $\mathrm{cond}(\frac{1}{\mathrm{h}}\mathrm{A} + \mathrm{B}) = \mathrm{O}(1)$.

3.) Eine geeignete Skalierung der zu lösenden linearen Gleichungssysteme innerhalb etwa des Newton-Verfahrens zur Lösung der nichtlinearen Gleichungen in x_ℓ ist bei ADGln besonders wichtig.

4.) $(\frac{1}{h}A + B)^{-1}\frac{1}{h}A$ ist eine Approximation des kanonischen Projektors P_S, denn

$$\begin{array}{rcl} (\frac{1}{h}A + B)^{-1}\frac{1}{h}A & = & [(I + hP_S G_S^{-1} B)^{-1} h P_S + Q_S] \underbrace{G_S^{-1} A}_{P_S} \frac{1}{h} \\ & = & (I + hP_S G_S^{-1} B)^{-1} P_S = (I + o(h)) P_S \quad . \end{array}$$

Nach Lemma 4.1 wissen wir also, dass die entstehenden nichtlinearen Gleichungen in x_ℓ für hinreichend kleine Schrittweiten h regulär lösbar sind. Dies ist neben ausreichender Glattheit auch die Vorauassetzung, die die lokale Konvergenz des Newton-Verfahrens oder Newton-ähnlicher Verfahren garantiert.

Als nächstes fragen wir, ob die BDF bei Anwendung auf index-1-ADGln gleich gute Eigenschaften hat wie bei der Anwendung auf explizite gewöhnliche Differentialgleichungen. Dazu interessieren uns sowohl Verfahrenseigenschaften, die das Verhalten auf einem kompakten Intervall $[t_0, T]$ für $h \to 0$ beschreiben (Konsistenz, Konvergenz, und Stabilität) als auch die Frage, ob abklingende Lösungen von Testgleichungen für **beliebige** Schrittweiten wenigstens abklingend wiedergespiegelt werden. Wie bei der Behandlung expliziter gewöhnlicher Differentialgleichungen bedeutet hierbei

- **Konsistenz**, dass die exakte Lösung eingesetzt in die Näherungsgleichungen zu Fehlern (den lokalen Diskretisierungsfehlern τ_ℓ) führt, die gleichmäßig (d.h. im maximum über alle ℓ) beliebig klein werden, wenn nur die Schrittweiten klein genug gewählt werden.
  Wir definieren

$$\tau_\ell := f(\frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell,j}x_*(t_{\ell-j}), x_*(t_\ell), t_\ell) \quad \ell \geq k \quad .$$

- **Konvergenz**, dass die Differenz zwischen exakter Lösung an den Gitterpunkten und der Näherungslösung (d.h. der globale Fehler $\varepsilon_\ell := x_*(t_\ell) - x_\ell$) gleichmäßig (d.h. im maximum über alle ℓ) beliebig klein wird, wenn die Schrittweiten klein genug und die k Startwerte genau genug gewählt werden, und

- **Stabilität**, dass der Einfluss von Störungen in den Näherungsgleichungen auf deren Lösungen unabhängig davon, wie klein die Schrittweite gewählt wird (und wie viele Schritte demzufolge nötig sind), in der Größenordnung der Störungen bleibt, d.h. abgeschätzt werden kann durch eine von der Schrittweite unabhängige Konstante (die Stabilitätskonstante S) multipliziert mit der Norm der Störungen:

$$\max_{\ell=k,\ldots,N}|x_\ell - \tilde{x}_\ell| \leq S\{\max_{\ell=k,\ldots,N}|\delta_\ell| + \max_{\ell<k}|x_\ell - \tilde{x}_\ell|\} \quad ,$$

  wenn $f(\frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell,j}\tilde{x}_{\ell-j}, \tilde{x}_\ell, t_\ell) = \delta_\ell \quad, \ell = k, \ldots, N$ für hinreichend feine Gitter $t_0 < t_1 < \ldots < t_N = T$.

- Aus der Stabilitätsungleichung folgt die beim numerischen Rechnen wichtige Abschätzung

$$\max_{\ell=k\ldots N}|\tilde{x}_\ell - x_*(t_\ell)| \leq S\{\max_{\ell=k\ldots N}|\delta_\ell - \tau_\ell| + \max_{\ell<k}|\tilde{x}_\ell - x_*(t_\ell)|\} \quad ,$$

  aus der als Spezialfall für $\delta_\ell = 0$ mit der Konsistenz auch die Konvergenz folgt.

Wir wissen, dass wir unter den index-1-Voraussetzungen die ADGl (4.1) in eine nichtlineare Gleichung $f(w, u + Qw, t) = 0$ zur Bestimmung von $w = \omega(u, t)$ und die inhärente reguläre Differentialgleichung (3.7) in $u = Px$ entkoppeln können. Liefert die BDF (4.2) angewendet auf die ADGl (4.1) dieselben Ergebnisse wie bei Anwendung auf die inhärente reguläre Differentialgleichung (3.7) und anschließendem Einsetzen

in die Lösungsdarstellung (3.6), so können alle oben gestellten Fragen positiv beantwortet werden.
Um dies zu entscheiden, entkoppeln wir die BDF-Formel zur Berechnung der Näherungswerte x_ℓ in analoger Art und Weise wie zuvor das kontinuierliche Problem. Der Einfachheit halber betrachten wir konstante Schrittweiten h.

Sei $Q(\cdot)$ ein glatter Projektor auf $N(\cdot)$ und $P(\cdot) = I - Q(\cdot)$.
Bezeichne $Q_\ell := Q(t_\ell)$, $P_\ell := P(t_\ell)$ und $u_\ell := P_\ell x_\ell$.

Mit $w_\ell := P_\ell \frac{1}{h} \sum \alpha_j x_{\ell-j} + Q_\ell x_\ell$ kann die BDF-Formel aufgelöst werden zu

$$f(\underbrace{P_\ell \frac{1}{h} \sum \alpha_j x_{\ell-j}}_{:=P_\ell w_\ell}, \underbrace{x_\ell}_{u_\ell + Q_\ell w_\ell}, t_\ell) = 0 \Longleftrightarrow P_\ell \frac{1}{h} \sum \alpha_j x_{\ell-j} + Q_\ell x_\ell = \omega(P_\ell x_\ell, t_\ell) \quad .$$

Analog zur Lösungsdarstellung für das kontnuierliche Problem erhalten wir

$$x_\ell = u_\ell + Q_\ell \omega(u_\ell, t_\ell) \in \mathcal{M}_0(t_\ell) \quad .$$

Für u_ℓ erhält man als Approximation der inhärenten regulären Differentialgleichung die Rekursion

$$P_\ell \frac{1}{h} \sum \alpha_j x_{\ell-j} = P_\ell \omega(u_\ell, t_\ell) \quad \text{,d.h.}$$

$$\tfrac{1}{h} \sum \alpha_j u_{\ell-j} + \underset{\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast}{\sum \alpha_j \tfrac{1}{h}(P_\ell - P_{\ell-j})(u_{\ell-j} + Q_{\ell-j}\omega(u_{\ell-j}, t_{\ell-j}))} = P_\ell \omega(u_\ell, t_\ell) \quad .$$

Beim Vergleich mit der inhärenten regulären Differentialgleichung

$$u'(t) - \underset{\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast\ast}{P'(t)(u(t) + Q(t)\omega(u(t), t))} = P(t)\omega(u(t), t)$$

sieht man, dass falls P konstant ist ($P' = 0$), in beiden Ausdrücken die mit $_{\ast\ast\ast}$ gekennzeichneten Terme wegfallen und die BDF (4.2) angewendet auf die ADGL (4.1) tatsächlich dieselben Ergebnisse liefert wie bei Anwendung auf die (nur theoretisch verfügbare) inhärente reguläre Differentialgleichung (3.7). Wir sagen, die Diskretisierung mittels der BDF ist dann mit der Entkopplung vertauschbar. Es entsteht ein kommutatives Diagramm:

$$\boxed{\begin{array}{ccc}
f(x'(t), x(t), t) = 0 & \xrightarrow{\text{BDF}} & f(\frac{1}{h} \sum \alpha_j x_{\ell-j}, x_\ell, t_\ell) = 0 \\
\downarrow \text{Entkopplung} & & \downarrow \text{Entkopplung} \\
\begin{aligned} x(t) &= u(t) + Q\omega(u(t), t) \\ u'(t) &= P\omega(u(t), t) \end{aligned} & \xrightarrow{\text{BDF}} & \begin{aligned} x_\ell &= u_\ell + Q\omega(u_\ell, t_\ell) \\ \frac{1}{h} \sum \alpha_j u_{\ell-j} &= P\omega(u_\ell, t_\ell) \end{aligned}
\end{array}}$$

Wir fassen die obigen Überlegungen zu folgendem Satz zusammen:

Satz 4.2 *Die BDF angewendet auf die index-1-ADGl (4.1) erzeugt Werte $x_\ell \in \mathcal{M}_0(t_\ell)$. Falls $N(\cdot)$ konstant ist ($P' = 0$), so sind die BDF-Diskretisierung und die Entkopplung vertauschbar.*

Damit haben wir im Fall, dass N konstant ist, die Grundlage, um Eigenschaften der BDF, die bei Anwendung des Verfahrens auf explizite gewöhnliche Differentialgleichungen bekannt sind, zu übertragen.

Wir wollen im folgenden die Konsistenz und Stabilität der BDF für index-1-ADGln genauer untersuchen. Wir beginnen mit der Konsistenz, die **für hinreichend glatte Lösungen** aus den Eigenschaften des rückwärtigen Differenzenquotienten folgt.

Lemma 4.3 *Sei $y(\cdot) \in C^{k+1}$ und $\frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{j,\ell} y(t_{\ell-j}) = p'_y(t_\ell)$ rückwärtiger Differenzenquotient zu y an der Stelle t_ℓ. Dabei bezeichne $p_y(\cdot)$ das Interpolationspolynom, das y in den Stützstellen $t_{\ell-k}, \ldots, t_{\ell-1}, t_\ell$ interpoliert. Dann gilt*

$$\frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{j,\ell} y(t_{\ell-j}) - y'(t_\ell) \;=\; h_\ell^k \frac{c_{\ell,k}}{(k+1)!} y^{(k+1)}(t_\ell) + O(h_\ell^{k+1}) \quad .$$

<u>Beweis:</u> Wir verwenden die Darstellung des Restgliedes bei der Polynominterpolation. Es ist:

$$\begin{aligned}
y(t) &= p_y(t) + R(t) \\
&= p_y(t) + \prod_{j=0}^{k} (t - t_{\ell-j}) y[t_{\ell-k}, \ldots, t_\ell, t] \quad , \\
y'(t) &= p'_y(t) + \sum_{j=0}^{k} \prod_{i \neq j} (t - t_{\ell-i}) y[t_{\ell-k}, \ldots, t_\ell, t] + \prod_{j=0}^{k} (t - t_{\ell-j}) (y[t_{\ell-k}, \ldots, t_\ell, t])' \quad , \\
y'(t_\ell) &= p'_y(t_\ell) + \sum_{j=0}^{k} \prod_{i \neq j} (t_\ell - t_{\ell-i}) y[t_{\ell-k}, \ldots, t_\ell, t_\ell] \\
&= p'_y(t_\ell) + \underbrace{\prod_{i \neq 0} (t_\ell - t_{\ell-i})}_{= c_{\ell,k} h_\ell^k} \quad \underbrace{y[t_{\ell-k}, \ldots, t_\ell, t_\ell]}_{= \frac{1}{(k+1)!} y^{(k+1)}(\theta_\ell) \text{ mit } \theta_\ell \in [t_{\ell-k}, t_\ell]} \quad .
\end{aligned}$$

Dabei bezeichnet $y[\ldots]$ die rekursiv definierten dividierten Differenzen zu y. Mit $y^{(k+1)}(\theta_\ell) = y^{(k+1)}(t_\ell) + O(h_\ell)$ folgt die Behauptung. □

Falls $x_* \in C^{(k+1)}$, erhalten wir deshalb für den lokalen Fehler

$$\begin{aligned}
\tau_\ell &= f(\frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{j,\ell} x_*(t_{\ell-j}), x_*(t_\ell), t_\ell) \\
&= f(p'_{x_*}(t_\ell), x_*(t_\ell), t_\ell) \\
&= A(x_*(t_\ell), t_\ell) p'_{x_*}(t_\ell) + b(x_*(t_\ell), t_\ell) \\
&= A(x_*(t_\ell), t_\ell) p'_{x_*}(t_\ell) - A(x_*(t_\ell), t_\ell) x'_*(t_\ell) \\
&= A(x_*(t_\ell), t_\ell)(p'_{x_*}(t_\ell) - x'_*(t_\ell)) \\
&= h_\ell^k \frac{c_{\ell,k}}{(k+1)!} A(x_*(t_\ell), t_\ell) x_*^{(k+1)}(\theta_\ell) \;=\; O(h_\ell^k) \quad .
\end{aligned}$$

Ist der Nullraum N (und damit ein möglicher Projektor P) konstant, ergibt sich die Konsistenzordnung k auch, falls nur $Px_* \in C^{(k+1)}$ vorausgesetzt wird. Denn dann gilt $Pp_{x_*}(t) = p_{Px_*}(t)$ und wir können schließen:

$$\begin{aligned}\tau_\ell &= A(x_*(t_\ell),t_\ell)P(p'_{x_*}(t_\ell) - x'_*(t_\ell)) \\ &= h_\ell^k \frac{c_{\ell,k}}{(k+1)!} A(x_*(t_\ell),t_\ell)(Px_*)^{(k+1)}(\tilde{\theta}_\ell) \;=\; O(h_\ell^k) \quad .\end{aligned}$$

Ist weder der Nullraum konstant noch Qx_* hinreichend glatt, geht die Ordnung k ($\tau_\ell = O(h_\ell^k)$) verloren. Wir erhalten dann nur noch

$$\begin{aligned}\tau_\ell &= A(x_*(t_\ell),t_\ell)(\frac{1}{h_\ell}\sum_{j=0}^k \alpha_{\ell,j}x_*(t_{\ell-j}) - x'_*(t_\ell)) \\ &= A(x_*(t_\ell),t_\ell)(\frac{1}{h_\ell}P_\ell\sum_{j=0}^k \alpha_{\ell,j}x_*(t_{\ell-j}) - [(Px_*)'(t_\ell) - P'(t_\ell)x_*(t_\ell)]) \\ &= A(x_*(t_\ell),t_\ell)([\frac{1}{h_\ell}\sum_{j=0}^k \alpha_{\ell,j}P_{\ell-j}x_*(t_{\ell-j}) - (Px_*)'(t_\ell)] \\ &\qquad\qquad - [\frac{1}{h_\ell}\sum_{j=0}^k \alpha_{\ell,j}(P_{\ell-j} - P_\ell)x_*(t_{\ell-j}) \; - P'(t_\ell)x_*(t_\ell)]) \\ &= A(x_*(t_\ell),t_\ell)(h_\ell^k\frac{c_{\ell,k}}{(k+1)!}(Px_*)^{(k+1)}(\tilde{\theta}_\ell) + o(1)) \;=\; o(1) \quad .\end{aligned}$$

Als nächstes wenden wir uns der Frage der Stabilität zu. Auch falls der Nullraum N konstant ist, erfüllen die störungsbehafteten Werte $\tilde{x}_\ell$ im allgemeinen nicht mehr exakt die Nebenbedingungen der ADGl, sie liegen nicht mehr in $\mathcal{M}_0$. Wir können daher auch in diesem Fall die Stabilitätseigenschaften der BDF nicht unmittelbar mittels Satz 4.2 aus denen bei Anwendung auf explixite gewöhnliche Differentialgleichungen folgern. Wir werden jedoch ebenfalls eine Entkopplung , wie sie Satz 4.2 zugrunde liegt, verwenden, um eine Stabilitätsungleichung für die BDF bei Anwendung auf index-1-ADGln selbst bei zeitabhängigem Nullraum $N(\cdot)$ zu beweisen. Es gilt:

Satz 4.4 *Für hinreichend feine äquidistante Gitter*

$$t_0 < t_1 < \ldots < t_n = T, \quad t_\ell = t_0 + \ell h = t_0 + \ell\frac{T-t_0}{n}$$

(mit einer von der Feinheit des Gitters abhängigen Anzahl der Stützstellen n) gilt für die k-schrittige BDF (4.2) für $k \leq 6$ bei Anwendung auf eine index-1-ADGl (4.1) die Stabilitätsabschätzung

$$\max_{\ell=k,\ldots,n}|x_*(t_\ell) - \tilde{x}_\ell| \leq S\{\max_{\ell=0,\ldots,k-1}|x_*(t_\ell) - \tilde{x}_\ell| + \max_{\ell=k,\ldots,n}|\tau_\ell - \delta_\ell|\}$$

mit einer vom Gitter unabhängigen Konstanten S .

Beweis: Im Beweis wird zunächst eine Entkopplung der betrachteten Fehlerterme und anschließend eine Analyse der Rekursion in den “P”-Anteilen analog zu der bei expliziten gewöhnlichen Differentialgleichungen verwendeten Beweistechnik vorgenommen.

Der Einfachheit halber führen wir den Beweis hier nur für lineare zeitabhängige ADGln. Für $x_*(t_\ell)$ und $\tilde{x}_\ell$ gelten dann folgende Rekursionen:

$$\begin{array}{rrrl} A(t_\ell)\frac{1}{h}\sum_{j=0}^k x_*(t_{\ell-j}) & +B(t_\ell)x_*(t_\ell) & -q(t_\ell) & =\tau_\ell \\ A(t_\ell)\frac{1}{h}\sum_{j=0}^k \tilde{x}_{\ell-j} & +B(t_\ell)\tilde{x}_\ell & -q(t_\ell) & =\delta_\ell \end{array} \qquad \ell = k,\dots,n \quad .$$

Mit den Bezeichnungen

$$\varepsilon_\ell = x_*(t_\ell) - \tilde{x}_\ell \quad , \quad A_\ell = A(t_\ell) \quad , \quad B_\ell = B(t_\ell) \quad , \quad \text{usw.}$$

erhalten wir die folgende Fehlerrekursion:

$$A_\ell \frac{1}{h}\sum_{j=0}^k \varepsilon_{\ell-j} + B_\ell\varepsilon_\ell = \tau_\ell - \delta_\ell \quad , \quad l = k,\dots,n \quad .$$

Zu zeigen ist eine Abschätzung

$$\max_{l=k,\dots,n} |\varepsilon_\ell| \leq S\{\max_{\ell=0,\dots,k-1} |\varepsilon_\ell| + \max_{l=k,\dots,n} |\tau_\ell - \delta_\ell|\}$$

Wir skalieren die Fehlerrekursion zunächst je mit $G_{1\ell}^{-1}$ und nutzen dabei die Beziehungen $G_1^{-1}A = P$ und $G_1^{-1}B = G_1^{-1}BP + Q$:

$$P_\ell \frac{1}{h}\sum_{j=0}^k \varepsilon_{\ell-j} + (G_{1\ell}^{-1}B_\ell P_\ell + Q_\ell)\varepsilon_\ell = G_{1\ell}^{-1}(\tau_\ell - \delta_\ell) \quad , \quad \ell = k,\dots,n \quad .$$

Weiter skalieren wir einerseits mit P_ℓ und andererseits mit Q_ℓ :

$$\begin{array}{rrl} P_\ell\frac{1}{h}\sum_{j=0}^k \varepsilon_{\ell-j} & +P_\ell G_{1\ell}^{-1}B_\ell P_\ell\varepsilon_\ell & = P_\ell G_{1\ell}^{-1}(\tau_\ell - \delta_\ell) \\ & Q_\ell G_{1\ell}^{-1}B_\ell P_\ell\varepsilon_\ell + Q_\ell\varepsilon_\ell & = Q_\ell G_{1\ell}^{-1}(\tau_\ell - \delta_\ell) \end{array} \quad , \quad \ell = k,\dots,n \quad .$$

Wir führen Bezeichnungen für die P- und Q-Anteile der Fehler ein. Sei

$$\mu_\ell := P_\ell\varepsilon_\ell \quad \text{und} \quad \nu_\ell := Q_\ell\varepsilon_\ell$$

Aus der zweiten Zeile erhält man ν_ℓ in Abhängigkeit von μ_ℓ:

$$\nu_\ell = -\underbrace{Q_\ell G_{1,\ell}^{-1}B_\ell}_{Q_{S\ell}}\mu_\ell + Q_\ell G_{1\ell}^{-1}(\underbrace{\tau_\ell}_{\in \operatorname{im} A_\ell} - \delta_\ell) = -Q_{S\ell}\mu_\ell - Q_\ell G_{1\ell}^{-1}\delta_\ell \quad .$$

An dieser Stelle weisen wir darauf hin, dass in den ableitungsfreien Anteilen der ADGL kein (lokaler) Diskretisierungsfehler, wohl aber beim praktischen Rechnen Rundungsfehler oder Fehler durch das nur näherungsweise Lösen der Gleichungen auftreten. Das wird beschrieben durch $\tau_\ell \in \operatorname{im} A_\ell$ oder $Q_\ell G_{1,\ell}^{-1}\tau_\ell = 0$. Wir erhalten eine Darstellung des Fehlers in der Form

$$\boxed{\varepsilon_\ell = \mu_\ell + \nu_\ell = \mu_\ell - Q_{S\ell}\mu_\ell - Q_\ell G_{1\ell}^{-1}\delta_\ell = P_{S\ell}\mu_\ell - Q_\ell G_{1\ell}^{-1}\delta_\ell \quad , \quad \ell = k,\dots,n \quad (*)}$$

Um auch für die k Startwerte eine analoge Darstellung benutzen zu können , nehmen wir an, dass die Werte $\tilde{x}_0,\dots,\tilde{x}_{k-1}$ mit geeigneten Werten $\tilde{y}_0,\dots,\tilde{y}_{k-1}$ den Beziehungen

$$A_\ell\tilde{y}_\ell + B_\ell\tilde{x}_\ell - q_\ell \; = \; \delta_\ell \quad , \quad \ell = 0,\dots,k-1$$

genügen und setzen $\tau_0 = \ldots = \tau_{k-1} = 0$. Für μ_ℓ ergibt sich die folgende Rekursion:

$$\begin{aligned}\frac{1}{h}\sum_{j=0}^{k}\alpha_j\mu_{\ell-j} \quad &+ \quad \sum_{j=1}^{k}\alpha_j\frac{1}{h}(P_\ell - P_{\ell-j})\{P_{S\ell-j}\mu_{\ell-j} - Q_{\ell-j}G_{1\ell-j}^{-1}\delta_{\ell-j}\} \\ &+ \quad P_\ell G_{1\ell}^{-1}B_\ell\mu_\ell \quad = \quad P_\ell G_{1\ell}^{-1}(\tau_\ell - \delta_\ell) \quad , \quad \ell = k, \ldots, n \quad .\end{aligned}$$

Wir skalieren mit $\frac{h}{\alpha_0}$ und fassen die Terme in μ_ℓ zusammen:

$$\underbrace{(I + \frac{h}{\alpha_0}P_\ell G_{1\ell}^{-1}B_\ell)}_{=:H_\ell}\mu_\ell + \sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\mu_{\ell-j} + \sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\underbrace{(P_\ell - P_{\ell-j})}_{=0(h)}\{\ldots\} \;=\; \frac{h}{\alpha_0}P_\ell G_{1\ell}^{-1}(\tau_\ell - \delta_\ell)$$

Für hinreichend kleine Schrittweiten h ist die Matrix H_ℓ reguär und ihre Inverse gleichmäßig für alle hinreichend feinen Gitter beschränkt. Man kann diese Aussage auch ohne Einschränkung an die Schrittweiten treffen, wenn die finiten Eigenwerte der Matrixbüschel $\{A_\ell, B_\ell\}$ alle negative Realteile haben.
Wir multiplizieren mit der Inversen $H_\ell^{-1} = I - \frac{h}{\alpha_0}H_\ell^{-1}P_\ell G_{1\ell}^{-1}B_\ell = I + O(h)$, schreiben die Terme $-QG_1^{-1}\delta$ hier wieder als $QG_1^{-1}(\tau - \delta)$ und erhalten eine Rekursion der Form

$$\mu_\ell + \sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\mu_{\ell-j} \quad = \quad h\sum_{j=1}^{k}\mathcal{B}_{\ell j}\mu_{\ell-j} + h\sum_{j=0}^{k}\mathcal{C}_{\ell j}(\tau_{\ell-j} - \delta_{\ell-j}) \quad , \quad l = k, \ldots, n \quad .$$

Diese k-schrittige Rekursion in $\mu_\ell \in \mathbb{R}^m$ wandeln wir um in eine einschrittige Rekursion in $\Upsilon_\ell \in \mathbb{R}^{km}$. Unter Hinzunahme der trivialen Idenditäten

$$\begin{aligned}\mu_{\ell-1} &= \mu_{\ell-1} \\ &\vdots \\ \mu_{\ell-k+1} &= \mu_{\ell-k+1}\end{aligned}$$

ergibt sich

$$\begin{pmatrix}\mu_\ell \\ \mu_{\ell-1} \\ \vdots \\ \mu_{\ell-k+1}\end{pmatrix} = \underbrace{\begin{pmatrix}-\frac{\alpha_1}{\alpha_0}I & \cdots & \cdots & -\frac{\alpha_k}{\alpha_0}I \\ I & 0 & & \\ & \ddots & \ddots & \\ & & I & 0\end{pmatrix}}_{=:\mathfrak{A}}\begin{pmatrix}\mu_{\ell-1} \\ \mu_{\ell-2} \\ \vdots \\ \mu_{\ell-k}\end{pmatrix} +$$

$$+h\begin{pmatrix}\mathcal{B}_{\ell 1} & \cdots & \mathcal{B}_{\ell k} \\ 0 & \cdots & 0 \\ \vdots & & \\ 0 & \cdots & 0\end{pmatrix}\begin{pmatrix}\mu_{\ell-1} \\ \mu_{\ell-2} \\ \vdots \\ \mu_{\ell-k}\end{pmatrix} + h\begin{pmatrix}\sum_{j=0}^{k}\mathcal{C}_{\ell j}(\tau_{\ell-j} - \delta_{\ell-j}) \\ 0 \\ \vdots \\ 0\end{pmatrix}$$

oder in kompakter Form

$$\boxed{\Upsilon_\ell \quad = \quad \mathfrak{A}\Upsilon_{\ell-1} + h\mathcal{B}_\ell\Upsilon_{\ell-1} + h\mathcal{C}_\ell\Delta_\ell \quad , \qquad (**)}$$

wobei $\Upsilon_\ell = \begin{pmatrix}\mu\ell \\ \vdots \\ \mu_{\ell-k+1}\end{pmatrix} \in \mathbb{R}^{km}, \quad \Delta_\ell = \begin{pmatrix}\tau_\ell - \delta_\ell \\ \vdots \\ \tau_{\ell-k} - \delta_{\ell-k}\end{pmatrix} \in \mathbb{R}^{(k+1)m}$.

Für die folgenden Abschätzungen ist es wesentlich, eine durch eine Vektornorm $|\cdot|_*$ induzierte Matrixnorm $\|\cdot\|_*$ zu finden, so dass $\|\mathfrak{A}\|_* \le 1$ gilt. Eine solche Vektornorm kann man angeben, wenn die Eigenwerte der Matrix $\mathfrak{A}$ betragsmäßig kleiner gleich 1 sind und diejenigen Eigenwerte mit Betrag 1 einfach sind. Die Eigenwerte der Frobeniusmatrix $\mathfrak{A}$ sind aber gerade die Nullstellen des charakteristischen Polynoms

$$\rho(\lambda) := \alpha_0\lambda^k + \alpha_1\lambda^{k-1} + \cdots + \alpha_k$$

der BDF. Von linearen Mehrschrittverfahren für gewöhnliche Differentialgleichungen sagt man, sie erfüllen die **Dahlquist'sche Wurzelbedingung** falls die Nullstellen ihrer charakteristischen Polynome die obige Eigenschaft haben. Dabei ist für konsistente lineare Mehrschrittverfahren 1 immer eine Nullstelle (die Konsistenznullstelle) ihres charakteristischen Polynoms, da wegen der Konsistenz gilt $\rho(1) = \sum_{j=0}^{k} \alpha_j = 0$. Von der BDF ist bekannt , dass sie für $k \le 6$ das Dahlquistsche Wurzelkriterium erfüllt.
Sei also $|\cdot|_*$ eine geeignete Vektornorm mit $\|\mathfrak{A}\|_* = 1$. Dann können wir ausgehend von $(**)$ abschätzen:

$$\begin{aligned}
|\Upsilon_\ell|_* &\le |\mathfrak{A}\Upsilon_{\ell-1}|_* + h|\mathcal{B}_\ell\Upsilon_{\ell-1}|_* + h|\mathcal{C}_\ell\Delta_\ell|_* \\
&\le |\Upsilon_{\ell-1}|_* + hc_1|\Upsilon_{\ell-1}|_* + hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_* \\
&= (1+hc_1)|\Upsilon_{\ell-1}|_* + hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_* \\
&\le (1+hc_1)[(1+hc_1)|\Upsilon_{\ell-2}|_* + hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_*] + hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_* \\
&= (1+hc_1)^2|\Upsilon_{\ell-2}|_* + [(1+hc_1)+1]hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_* \\
&\ \vdots \\
&\le (1+hc_1)^{\ell-k+1}|\Upsilon_{k-1}|_* + \sum_{j=0}^{\ell-k}(1+hc_1)^j hc_2 \max_{i=0,\dots,n} |\tau_i - \delta_i|_*
\end{aligned}$$

Für die geometrische Reihe erhaltn wir

$$\begin{aligned}
hc_2\sum_{j=0}^{\ell-k}(1+hc_1)^j &= hc_2\frac{(1+hc_1)^{\ell-k+1}-1}{(1+hc_1)-1} \\
&= \frac{hc_2}{hc_1}((1+hc_1)^{\ell-k+1}-1) \le \frac{c_2}{c_1}(1+hc_1)^{\ell-k+1}
\end{aligned}$$

und der Term $(1+hc_1)^{\ell-k+1}$ kann durch einen Exponentialausdruck abgeschätzt werden:

$$(1+hc_1)^{\ell-k+1} \le (1+hc_1)^n \le (e^{hc_1})^n = e^{c_1nh} = e^{c_1(T-t_0)}$$

Insgesamt ergibt sich

$$|\Upsilon_\ell|_* \le e^{c_1(T-t_0)}\{|\Upsilon_{k-1}|_* + \frac{c_2}{c_1}\max_{i=0,\dots,n}|\tau_i - \delta_i|_*\} \quad .$$

Ausgehend hiervon kommt man zunächst zu einer Stabilitätsungleichung in den P-Komponenten $\mu_\ell = P_\ell\varepsilon_\ell$, indem man nun noch zum Maximum über alle $\ell = k,\dots,n$ übergeht, die Anfangsstörungen $\delta_\ell,\ \ell = 0,\dots,k-1$ durch die Störungen der Startwerte abschätzt und zu der ursprünglichen Norm $|\cdot|$ zurückgeht.
Mit Hilfe der Darstellung $(*)$ der Fehler folgt schließlich die behauptete Stabilitätsabschätzung.

$\square$

Bemerkung:

1.) bei expliziten Verfahren für explizite gewöhnliche Differentialgleichungen kann die Stabilitätskonstante bis auf Normumrechnungsfaktoren als $e^{L(T-t_0)}$ gewählt werden, wenn L eine Lipschitz-Konstante für die rechte Seite der Differentialgleichung ist. Im obigen Beweis verwendeten wir $e^{c_1(T-t_0)}$ für die Abschätzung der P-Komponenten μ_ℓ der Fehler, wobei c_1 eine Schranke für $\mathcal{B}_\ell$ war. Falls P konstant ist , ist dabei $\mathcal{B}_{\ell j} = \frac{\alpha_j}{\alpha_0^2} H_\ell^{-1} P_\ell G_{1\ell}^{-1} B_\ell$, wobei H_ℓ eine Störung der Einheitsmatrix und $P_\ell G_{1\ell}^{-1} B_\ell$ die Koeffizientenmatrix der inhärenten regulären Differentialgleichung ist. Hinzu kommt aus der Darstellung $(*)$ eine Schranke für die Norm des kanonischen Projektor $P_S(\cdot)$.

2.) Beim Rechnen mit variabler Schrittweite entstehen in der Rekursion für Υ_ℓ variable Matrizen $\mathfrak{A}_\ell$.Um einen Stabilitätsbeweis führen zu können , sind dann Bedingungen nötig, die garantieren ,dass eine Norm des Produkts $\prod_{\ell=k}^{n} \mathfrak{A}_\ell$ kleiner gleich 1 ist. Dieses Problem tritt schon bei der numerischen Integration von expliziten gewöhnlichen Differentialgleichungen auf.

4.2 Implizite Runge-Kutta-Verfahren

Die ersten Runge-Kutta-Verfahren zur näherungsweisen Integration expliziter gewöhnlicher Differentialgleichungen wurden in Arbeiten von Runge(1895), Kutta(1901) und Heun(1900) benutzt. Für Anfangswertaufgaben der Form

$$x'(t) = g(x(t), t), \quad x(t_0) = x_0$$

basieren sie auf der Idee, das Integral in

$$x(t_\ell) - x(t_{\ell-1}) = \int_{t_{\ell-1}}^{t_\ell} x'(t)dt$$

durch eine Quadraturformel

$$\int_{t_{\ell-1}}^{t_\ell} x'(t)dt \approx (t_\ell - t_{\ell-1}) \sum_{i=1}^{s} b_i x'(t_{\ell,i})$$

zu approximieren (äußere Quadratur). Die auf der rechten Seite enthaltenen Ableitungswerte $x'(t_{\ell,i}) = g(x(t_{\ell,i}), t_{\ell,i})$ verlangen die noch unbekannten Werte $x(t_{\ell,i})$, die wiederum durch Quadraturformeln für

$$x(t_{\ell,i}) - x(t_{\ell-1}) = \int_{t_{\ell-1}}^{t_{\ell,i}} x'(t)dt \approx (t_\ell - t_{\ell-1}) \sum_{j=1}^{s} \alpha_{ij} x'(t_{\ell,j})$$

angenähert werden (innere Quadratur).
Mit $h_\ell := t_\ell - t_{\ell-1}$ und $t_{\ell,i} := t_{\ell-1} + c_i h_\ell$ ergibt sich daraus die Verfahrensvorschrift für ein allgemeines s-stufiges Runge-Kutta-Verfahren:

$$\begin{aligned} x_\ell &= x_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i X'_{\ell,i} \\ X'_{\ell,i} &= g(\underbrace{x_{\ell-1} + h_\ell \sum_{j=1}^{s} \alpha_{ij} X'_{\ell,j}}_{=:X_{\ell,i}}, t_{\ell-1} + c_i h_\ell) \quad , \quad i = 1 \dots s. \end{aligned}$$

Hierbei stellen die Werte $X'_{\ell,i}$ Approximationen an $x'(t_{\ell,i})$ und $X_{\ell,i}$ Approximationen an $x(t_{\ell,i})$ dar. Manchmal findet man die Verfahren auch äquivalent in der Form

$$\begin{aligned} x_\ell &= x_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i g(X_{\ell,i}, t_{\ell,i}) \\ X_{\ell,i} &= x_{\ell-1} + h_\ell \sum_{j=1}^{s} \alpha_{ij} g(X_{\ell,j}, t_{\ell,j}) \quad , \quad i = 1 \dots s \end{aligned}$$

dargestellt. Das Verfahren ist durch die Koeffizienten

$$b = (b_1, \dots, b_s)^T \ , \ c = (c_1, \dots, c_s)^T \ , \ \mathcal{A} = (\alpha_{ij})_{i,j=1}^{s}$$

vollständig bestimmt und wird gewöhnlich durch ein von Butcher (1964) eingeführtes Schema, das nach ihm benannte Butcher-Tableau, beschrieben: $\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$.
Sind in der Koeffizientenmatrix $\mathcal{A}$ nur Werte unterhalb der Diagonalen von 0 verschieden, so können die Approximationen für die verschiedenen Stufen sukzessive durch s nacheinanderfolgende Aufrufe der rechten Seite g berechnet werden. Das "klassische" Verfahren

von Runge und Kutta ist ein solches explizites Verfahren mit $s = 4$ und den Koeffizienten

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \quad .$$

Sind auch Werte auf oder oberhalb der Diagonalen von $\mathcal{A}$ von Null verschieden, so sind die Approximationen der verschiedenen Stufen durch die Verfahrensvorschrift nur implizit definiert. Es ist zur Bestimmung jedes x_ℓ ein im allgemeinen nichtlineares Gleichungssystem der Dimension sm in den Unbekannten $X'_{\ell,1}, \ldots, X'_{\ell,s}$ (bzw. $X_{\ell,1}, \ldots, X_{\ell,s}$) zu lösen.
Zur Integration steifer Differentialgleichungen sind geeignete implizite Verfahren trotz dieses Aufwandes effektiver, da explizite Verfahren für derartige Aufgaben “irrsinnig“ kleine Schrittweiten verlangen würden.

Bei der Integration von ADGln können die Ableitungswerte nicht explizit durch eine rechte Seite der Differentialgleichung berechnet werden. Als geeigneter Verfahrensansatz (der so erstmals von Linda Petzold (1984) verwendet wurde) haben sich implizite Runge-Kutta (IRK) -Verfahren in folgender Form erwiesen:

$$x_\ell = x_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i X'_{\ell,i} \tag{4.3}$$

$$f(X'_{\ell,i}, x_{\ell-1} + h_\ell \sum_{j=1}^{s} \alpha_{ij} X'_{\ell,j}, t_{\ell-1} + c_i h_\ell) = 0 \quad , \quad i = 1 \ldots s. \tag{4.4}$$

Es ist in jedem Integrationsschritt ein sm-dimensionales, im allgemeinen nichtlineares Gleichungssystem in $X'_{\ell,1}, \ldots, X'_{\ell,s}$ zu lösen. Die Jacobimatrix hat dabei die Form

$$\begin{pmatrix} A_{\ell,1} & & \\ & \ddots & \\ & & A_{\ell,s} \end{pmatrix} + h \begin{pmatrix} \alpha_{11} B_{\ell,1} & \ldots & \alpha_{1s} B_{\ell,1} \\ \vdots & & \vdots \\ \alpha_{s1} B_{\ell,s} & \ldots & \alpha_{ss} B_{\ell,s} \end{pmatrix} \quad ,$$

wobei

$$A_{\ell,i} = f'_y(X'_{\ell,i}, x_{\ell-1} + h_\ell \sum_{j=1}^{s} \alpha_{ij} X'_{\ell,j}, t_{\ell-1} + c_i h_\ell) \quad , \quad B_{\ell,i} = f'_x(\ldots) \quad .$$

Für den Spezialfall $f(y, x, t) = y - g(x, t)$ ergeben sich die bekannten IRK-Verfahren für explizite gewöhnliche Differentialgleichungen.

Wir fragen nun nach der **Durchführbarkeit** und nach **Konsistenz- und Stabilitätseigenschaften** des IRK-Verfahrens (4.3),(4.4). Als Hilfsmittel für diese theoretischen Untersuchungen verwenden wir wieder eine Entkopplung der Verfahrensgleichung in die verschiedenen Lösungsanteile. Dazu bezeichne:

$$u_\ell := P_\ell x_\ell \quad , \quad P_\ell := P(t_\ell) \quad , \quad P_{\ell,i} := P(t_{\ell,i}) \quad ,$$

$$U_{\ell,i} := P_{\ell,i} X_{\ell,i} \quad , \quad U'_{\ell,i} := P_{\ell,i} X'_{\ell,i} \quad , \quad W_{\ell,i} := P_{\ell,i} X'_{\ell,i} + Q_{\ell,i} X_{\ell,i} \quad .$$

Mit diesen Bezeichnungen gilt

$$X_{\ell,i} = U_{\ell,i} + Q_{\ell,i}W_{\ell,i} \quad , \quad P_{\ell,i}X'_{\ell,i} = P_{\ell,i}W_{\ell,i}$$

und (4.4) bedeutet

$$f(W_{\ell,i}, U_{\ell,i} + Q_{\ell,i}W_{\ell,i}, t_{\ell,i}) = 0 \quad , \quad i = 1,\ldots,s \quad .$$

Diese Gleichung ist im index 1-Fall auflösbar zu

$$W_{\ell,i} = \omega(U_{\ell,i}, t_{\ell,i}) \quad , \quad i = 1,\ldots,s \quad ,$$

woraus die Beziehung

$$U'_{\ell,i} = P_{\ell,i}X'_{\ell,i} = P_{\ell,i}W_{\ell,i} = P_{\ell,i}\omega(U_{\ell,i}, t_{\ell,i}) \quad , i = 1,\ldots,s$$

folgt. Falls P konstant ist (dazu muss der Nullraum N konstant sein), folgt aus (4.3) nun eine analoge Beziehung für die P-Komponenten $u_\ell := Px(t_\ell)$:

$$u_\ell = Px_\ell = P(x_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i X'_{\ell,i}) = u_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i U'_{\ell,i} \quad .$$

Zusammen mit

$$U'_{\ell,i} \;=\; P\omega(u_{\ell-1} + h_\ell \sum_{j=1}^{s} \alpha_{ij} U'_{\ell,j} \;,\; t_{\ell,i}) \quad , i = 1,\ldots,s$$

haben wir so den Formelsatz erhalten, der sich auch bei Anwendung des IRK auf die (nur theoretisch verfügbare) inhärente reguläre DGL ergeben hätte. Falls der Nullraum der führenden Jacobimatrix konstant ist, sind also Entkopplung und Diskretisierung in diesem Sinne auch bei IRK vertauschbar und es übertragen sich die Verfahrenseigenschaften des benutzten IRKs für explizite gewöhnliche DGln auf die P-Anteile der Lösungen und ihrer RK-Approximationen.

Falls dieser Nullraum nicht konstant ist, gilt dies im allgemeinen nicht mehr. Es lassen sich dennoch Aussagen zu Konsistenz, Stabilität und Konvergenz treffen, jedoch verlangen die Beweise relativ hohen technischen Aufwand, weshalb wir an dieser Stelle darauf verzichten wollen.

Bisher haben wir jedoch auch im Fall konstanten Nullraumes N nur Aussagen für die P-Komponenten $u_\ell, u_{\ell,i}$ der Lösungsapproximationen hergeleitet. Diese gelten sogar für explizite RKV und speziell für das explizite Eulerverfahren (das auch ein RKV mit dem Butcher-Tableau $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \;\hat{=}\; x_\ell = x_{\ell-1} + h_\ell X'_{\ell,1} \quad , \quad f(X'_{\ell,1}, x_{\ell-1}, t_{\ell-1}) = 0$ ist).
Es bleibt zu untersuchen, wie sich die Q-Komponenten verhalten. Die erste Frage, die wir stellen, ist, ob diese überhaupt durch die Verfahrensgleichung bestimmt sind (beim expliziten Eulerverfahren war das ja nicht der Fall). Als zweites fragen wir, ob die Approximationen x_ℓ so bestimmt werden, dass sie den Nebenbedingungen genügen, d.h. ob gilt $x_\ell \in \mathcal{M}_0(t_\ell)$.
Wir beschränken uns dabei wieder auf den Fall konstanten Nullraumes. Sei also der Projektor Q auf den Nullraum N konstant. Dann folgt aus (4.3) :

$$Qx_\ell = Qx_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i Q X'_{\ell,i} \quad .$$

Um die $QX'_{\ell,i}$ zu bestimmen, nutzen wir den Zusammenhang

$$QX_{\ell,i} = Qx_{\ell-1} + h_\ell \sum_{i=1}^{s} \alpha_{ij} QX'_{\ell,j} \quad i = 1, \dots, s \quad .$$

Dies betrachten wir als lineares Gleichungssystem in den Variablen $QX'_{\ell,1}, \dots, QX'_{\ell,s}$:

$$\sum_{i=1}^{s} \alpha_{ij} QX'_{\ell,j} = \frac{1}{h_\ell}(QX_{\ell,i} - Qx_{\ell-1}) = \frac{1}{h_\ell}(Q\omega(U_{\ell,j}, t_{\ell,j}) - Qx_{\ell-1}) \;,\; i = 1, \dots, s\,.$$

Diese Beziehung lässt sich nur nach $QX'_{\ell,1}, \dots, QX'_{\ell,s}$ auflösen, wenn die Koeffizientenmatrix $\mathcal{A}$ des RKV regulär ist. Dies setzen wir im weiteren voraus. (Damit scheiden im folgenden alle expliziten RKV aus.)

Sei also $\mathcal{A}$ regulär und $\mathcal{A}^{-1} = (\hat{\alpha}_{ij})_{i,j=1}^{s}$. Dann ergibt sich

$$\begin{aligned} QX'_{\ell,i} &= \sum_{j=1}^{s} \hat{\alpha}_{ij} \frac{1}{h_\ell}(QX_{\ell,j} - x_{\ell-1}) \\ &= \sum_{j=1}^{s} \hat{\alpha}_{ij} \frac{1}{h_\ell}(Q\omega(U_{\ell,j}, t_{\ell,j}) - Qx_{\ell-1}) \quad , \quad i = 1, \dots, s\,. \end{aligned}$$

und

$$\begin{aligned} Qx_\ell &= Qx_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i QX'_{\ell,i} \\ &= Qx_{\ell-1} + h_\ell \sum_{i=1}^{s} b_i \sum_{j=1}^{s} \hat{\alpha}_{ij} \frac{1}{h_\ell}(Q\omega(U_{\ell,j}, t_{\ell,j}) - Qx_{\ell-1}) \\ &= \underbrace{(1 - \sum_{i,j=1}^{s} b_i \hat{\alpha}_{ij})}_{=:\rho} Qx_{\ell-1} + \sum_{i,j=1}^{s} b_i \hat{\alpha}_{ij} Q\omega(U_{\ell,j}, t_{\ell,j}) \end{aligned}$$

Dies soll $Qx_*(t_\ell) = Q\omega(u_*(t_\ell), t_\ell)$ approximieren.
Im Spezialfall linearer homogener ADGln mit konstanten Koeffizienten gilt für die Lösung $x_*(t) = P_S x_*(t)$ und damit $Q_S x_*(t) = 0$. Verwenden wir diesen speziellen Projektor $P := P_S$ auch für die Entkopplung und setzen $u(t) := P_S x(t)$, so folgt $Q_S \omega(u,t) = -\underbrace{Q_S G_{1S}^{-1} B}_{=Q_S} u = 0$ und die obige Betrachtung für Qx_ℓ führt zu

$$Q_S x_\ell = \rho \cdot Q_S x_{\ell-1} \;+\; 0 \;=\; \rho \cdot Q_S x_{\ell-1} = \rho^\ell \cdot Q_S x_0 \quad .$$

Nun gilt in diesem Fall für einen konsistenten Startwert x_0 , dass $Q_S x_0 = 0$, und damit folgt bei exakter Rechnung auch für die Iterierten $Q_S x_\ell = 0$. Kleine Fehler in $Q_S x_0$ wirken sich jedoch, falls $\rho \neq 0$ ist, auf alle Iterierten aus. Falls $|\rho| > 1$ werden sie verstärkt, falls $|\rho| < 1$ werden sie gedämpft. Für $|\rho| > 1$ sind die Verfahren exponentiell instabil (bedingt durch den Faktor ρ^ℓ in der Rekursion für die Q-Komponenten) und damit unbrauchbar. Für $|\rho| < 1$ kann wieder eine Stabilitätsungleichung bewiesen werden. Am günstigsten in dieser Hinsicht sind Verfahren, bei denen $\rho = 0$ gilt. Wann ist diese Eigenschaft erfüllt? Es war

$$\rho \;:=\; 1 - \sum_{i=1}^{s} b_i \sum_{j=1}^{s} \hat{\alpha}_{ij} \;=\; 1 - b^T \mathcal{A}^{-1} \mathbb{1} \quad , \quad \mathbb{1} \;=\; \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^s \quad .$$

Falls nun die Koeffizienten des IRK so gewählt sind, dass der Vektor b^T mit der letzten Zeile von $\mathcal{A}$ übereinstimmt, d. h.

$$b^T = e_s^T \mathcal{A} = (\alpha_{s1}, \ldots, \alpha_{ss}) \quad ,$$

so folgt

$$\rho = 1 - b^T \mathcal{A}^{-1} 1\!\!1 = 1 - e_s^T \mathcal{A}\mathcal{A}^{-1} 1\!\!1 = 1 - e_s^T 1\!\!1 = 1 - 1 = 0 \quad .$$

Verfahren mit derartigen Koeffizienten haben weiter die wichtige Eigenschaft, dass die Approximation x_ℓ für $x(t_\ell)$ mit der Approximation der letzten inneren Stufe $X_{\ell,s}$ für $x(t_{\ell,s})$ übereinstimmt. Aus (4.4) folgt, dass die Approximationen für die inneren Stufen die Nebenbedingungen der ADGl je an den entsprechenden Stützstellen $t_{\ell,i}$ erfüllen, d.h. $X_{\ell,i} \in \mathcal{M}_0(t_{\ell,i})\ , i = 1, \ldots, s$.
Nun ist es noch sinnvoll zu verlangen, dass $c_s = 1$, d.h. $t_{\ell,s} = t_\ell$ gilt, damit folgt $x_\ell \in \mathcal{M}_0(t_\ell)$ oder, äquivalent, $Qx_\ell = \omega(u_\ell, t_\ell)$.
($c_s = 1$ folgt in der Regel schon aus den Konsistenzeigenschaften der Verfahren:
Damit die inneren Quadraturformeln wenigstens konstante rechte Seiten exakt integrieren, muss gelten $\sum_{j=1}^{s} \alpha_{ij} = c_i$, $i = 1, \ldots, s$, und damit die äußeren Quadraturformeln wenigstens konstante rechte Seiten exakt integrieren, muss gelten $\sum_{i=1}^{s} b_i = 1$.)

Definition: *Ein RKV* $\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$ *mit den Eigenschaften*

- $\mathcal{A}$ *ist regulär,*
- $b^T = e_s^T \mathcal{A} = (\alpha_{s1}, \ldots, \alpha_{ss})$, $c_s = 1$

heißt IRK(DAE).

Im Buch von Hairer/Wanner werden diese Verfahren "stiffly accurat" genannt. Wie die BDF-Verfahren, sind auch die IRK(DAE) gut geeignet, um steife Differentialgleichungen effektiv zu integrieren.

Satz 4.5 *Die ADGl (4.1) habe den Index 1 und der Nullraum N sei konstant. Dann ist ein IRK(DAE) für hinreichend kleine Schrittweiten durchführbar und Diskretisierung und Entkopplung sind vertauschbar. Die Aproximationen x_ℓ sind konsistent (liegen in $\mathcal{M}_0(t_\ell)$). Es gilt eine Stabilitätsungleichung und die Konsistenz- und Konvergenzordnungen des Verfahrens bei Anwendung auf explizite gewöhnliche DGln übertragen sich.*

Beweis: Wir haben oben hergeleitet, dass unter den gemachten Voraussetzungen die P-Komponenten $u_\ell := Px_\ell$ dem IRK angewendet auf die inhärente reguläre DGl genügen. Damit übertragen sich die bekannten Stabilitäts-, Konsistenz- und Konvergenzeigenschaften des IRK bei Anwendung auf explizite gewöhnliche DGln auf die P-Komponenten.
Weiter haben wir oben gesehen, dass sich bei IRK(DAE) die Q-Komponenten mit $Qx_\ell := \omega(u_\ell, t_\ell)$ als Funktion der P-Komponenten und der Stützstellen ergeben. Mit Stetigkeitsargumenten für die Funktion ω auf einem kompakten Gebiet in Umgebung der Lösung folgt dann die Behauptung. □

Beispiele für IRK(DAE) sind die **RadauIIA-Verfahren**. Die Koeffizienten für ein solches s-stufiges Verfahren sind so gewählt, dass $c_s = 1$ gilt, die äußere und die innere Quadratur für Polynome bis zum Grad $s - 1$ als rechte Seiten exakt sind und die Konsistenzordnung $p = 2s - 1$ erreicht wird. Die Verfahren sind A-Stabil und L-stabil, die Überführungsfunktion ist die $(s - 1, s)$-Padé-Approximation der Exponentialfunktion (eine rationale Funktion mit Zählergrad $s - 1$ und Nennergrad s).